NEH Application Cover Sheet (HAA-256249)
Digital Humanities Advancement Grants

PROJECT DIRECTOR

Mr. Golan Levin
Director, Frank Ratchye STUDIO for Creative I
5000 Forbes Avenue
Pittsburgh, PA 15213-3815
USA

E-mail: golan@andrew.cmu.edu
Phone: 917 520 7456
Fax:

Field of expertise: Interdisciplinary Studies, Other

INSTITUTION

Carnegie Mellon University
Pittsburgh, PA 15213-3815

APPLICATION INFORMATION

Title: *Supporting Cultural Heritage Research in Historic Photography Archives with Machine Learning and Computer Vision*

Grant period: From 2017-09-01 to 2019-02-28

Project field(s): Interdisciplinary Studies, General; African American History

Description of project: We address the challenges faced in the research and annotation of large digital image archives by creating prototype software tools that use machine learning and computer vision. Specifically, we are developing software tools to aid research into the Carnegie Museum of Art's publicly available Teenie Harris Archive, a major photography collection documenting 20th century African American life in Pittsburgh. Our goal is to create open-source software that uses state-of-the-art techniques to help identify and annotate visually distinctive features across this large (80,000 item) set of digitized photographs, to improve and expedite the Museum's archiving and cataloging process. Through compatibility with International Image Interoperability Framework (IIIF) standards, our project will furthermore provide free tools and reproducible, computer-vision based workflows that other museums, libraries and archives can use to help organize their own digital collections.

BUDGET

Outright Request	71,458.00	**Cost Sharing**	0.00
Matching Request	0.00	**Total Budget**	71,458.00
Total NEH	71,458.00		

GRANT ADMINISTRATOR

Mr. Robert Kearns
5000 Forbes Avenue
Pittsburgh, PA 15213-3815
USA

E-mail: osp-preaward@andrew.cmu.edu
Phone: 412-268-5837
Fax: 412-268-6279

Table of Contents

List of Participants **1**

Abstract **2**

Narrative **3**
Enhancing the humanities 3
Environmental scan 5
History of the project 5
Work plan 6
Staff 7

Data Management Plan **8**

Project Budget **9**

Biographies **11**

Letters of Commitment and Support **12**

Appendices **13**

List of participants

Levin, Golan, Project Director
Director, The Frank-Ratchye STUDIO for Creative Inquiry, Carnegie Mellon University
Associate Professor of Electronic Arts, School of Art, Carnegie Mellon University
Associate Professor (by Courtesy) of Computer Science, Carnegie Mellon University

Newbury, David, Co-Director
Principal, Workergnome Studios

Hughes, Tom, Project Manager
Associate Director, The Frank-Ratchye STUDIO for Creative Inquiry, Carnegie Mellon University

Zelevansky, Lynn, Institutional Partner
The Henry J. Heinz II Director, Carnegie Museum of Art

Abstract

Supporting Cultural Heritage Research in Historic Photography Archives with Machine Learning and Computer Vision

Abstract

We address the challenges faced in the research and annotation of large digital image archives by creating prototype software tools that use machine learning and computer vision. Specifically, we are developing software tools to aid research into the Carnegie Museum of Art's publicly available Teenie Harris Archive, a major photography collection documenting 20th century African American life in Pittsburgh. Our goal is to create open-source software that uses state-of-the-art techniques to help identify and annotate visually distinctive features across this large (80,000 item) set of digitized photographs, to improve and expedite the Museum's archiving and cataloging process. Through compatibility with International Image Interoperability Framework (IIIF) standards, our project will furthermore provide free tools and reproducible, computer-vision based workflows that other museums, libraries and archives can use to help organize their own digital collections.

Narrative

Enhancing the humanities

The Frank-Ratchye STUDIO for Creative Inquiry requests an NEH Digital Humanities Level II Grant to develop a set of prototype software tools that use advanced machine learning and computer vision techniques to identify, annotate, and organize visually distinctive features in a large, culturally significant photography archive. Our subject is the Teenie Harris photography archive of the Carnegie Museum of Art (CMoA), a forty-year visual record of African-American life in Pittsburgh comprised of more than 80,000 digitized negatives and photographs. By developing software tools compatible with the International Image Interoperability Framework (IIIF) -- a collection of open standards for accessing and presenting digital images -- we seek to assist the CMoA in their research into the Teenie Harris archive, while contributing new workflows to the larger open-source software community of museums, libraries and other institutions seeking technologically innovative methods for researching large, digitized image collections.

The idea for our project emerged from conversations with our colleagues at the CMoA, as they discussed the research challenges they faced in their work with the Teenie Harris Archive (THA). Charles "Teenie" Harris was a photographer for *The Pittsburgh Courier,* one of the most influential black newspapers of the 20th century. In a career spanning more than four decades, Harris captured the events and everyday experience of African American life. In 2001, the CMoA accessioned his more than 80,000 photographs and negatives, creating the Teenie Harris Archive. Much of the rich history of Pittsburgh's African American community, from the 1930's through the 1970's, is recorded in this extensive photographic collection. Unfortunately, although this archive is a unique and historically significant visual record, information about the specific people, places and events in his images are often unknown, as very few of the negative and photographs contain titles, labels, or even approximate dates.

With the help of previous support from the NEH, archivists at CMoA have digitized the Teenie Harris Archive, and have begun the laborious process of captioning, tagging, and adding other metadata to the photographs. (It is important to note that this work is performed in close coordination with representatives from Pittsburgh's African American community, with guidance and oversight by an advisory committee comprised of Harris family members, academic specialists, and community leaders who have insisted on the African American community's ownership of the history represented in Harris' images.) Much of this annotation work has been conducted through interviews collected directly from Teenie Harris' contemporaries, and, wherever possible, with the original community members documented in his photographs. Through a combination of outreach, exhibitions, interviews, and other public programs, some 2,000 images, or about 2% of the archive, have been positively identified with the help the Pittsburgh community. Unfortunately, many of the persons who can best contribute to the annotation of the archive are now advanced in age. Time is therefore of the essence in gathering their first-person accounts and descriptions. Better tools are needed, and soon, to help cross-reference positively identified people, places, and events across 80,000 different photographs.

Our proposal is to develop open-source software to help address the pressing need for more efficient and effective methods of organizing, searching and cross-referencing large photography collections. To this end, we intend to incorporate recently-released code libraries for image analysis into easy-to-use interfaces that can help archivists and other domain experts to identify, sort, and tag visually distinctive subjects. Our tools will be compliant with the International Image Interoperability Framework (IIIF) standards, significantly expanding the community of researchers and institutions who can build upon our efforts.
The Teenie Harris archive represents the ideal dataset for our project. This catalog of images spans four decades and traces out the lives of a community of individuals -- including children, everyday neighbors, musicians, athletes, and community leaders -- in a consistent geographic area, the city of Pittsburgh. This

Narrative

consistency offers numerous instances of the same faces, places, etcetera, creating a perfect testbed for software prototypes using state-of-the-art algorithms for image feature recognition.

The first component of our research will be to create machine-learning-based tools for computational image analysis and annotation. An example of this in practice would be to identify specific individual faces that appear throughout different photographs in the archive, potentially even identifying instances of that same person across different decades -- an incredibly laborious process to conduct by hand. (This technology can already be found in the for-profit sector, most notably in the face-identification algorithms employed by Facebook, Google, and Apple.) Such algorithms can make it possible for a photographic subject singled out during a first-person interview -- whether a face, vehicle, building, or other visually distinctive pattern -- to be identified and rigorously cross-referenced throughout the entire Teenie Harris archive. We imagine a system in which potential matches are automatically presented to interview subjects for their consideration, to radically accelerate the process of annotation and data collection.

While expediting user-generated queries in community interviews are one planned outcome for our proposed research, our team will also explore methodologies for a more automated form of tagging system. In this workflow, a computer running our software tools would go through the otherwise laborious process of analyzing each image in the archive to find instances of (for example) vehicles. The final confirmation of whether or not the program has accurately identified a car would carried out by a human user for the final confirmation. For example, our software could query tens of thousands of photographs, and return 100 instances of images it 'thinks' contains a car, with some confidence. Human users -- whether archivists, museum visitors, or crowdsourced information workers -- can now examine just 100 images as opposed to thousands, and help confirm or reject the software's finding. This feedback creates a smarter program over time, utilizing multiple tiers of expertise to better train the system. Our team has seen recent success with such methods; for example, with our recent "Terrapattern" project, we used machine learning (in the form of convolutional neural networks) and computer vision to help everyday people discover visually distinctive patterns in satellite imagery.

The second major objective of our research is to align the tools we build with a methodology and workflow that is useful and relevant to the cultural heritage community. In addition to releasing our prototype tools (as well as accompanying source code, documentation, and workflows) on our laboratory's GitHub site, we will also build our tools to comply with the IIIF set of metadata standards. The choice to use IIIF lies in both the high quality open-source APIs that are already available for image management, annotation and presentation, as well as the strong existing support for this standard across cultural heritage institutions.

IIIF, or the International Image Interoperability Framework, is an existing Linked Open Data standard for image interoperability. It defines metadata standards for dealing with high-resolution images, providing a consistent API for accessing both images, the metadata that surrounds them, and how to present and associate images together. It is being used at CMoA, as well as at the Internet Archive, the Bibliothèque nationale de France, the Vatican Library, and other major museums, archives, and national libraries around the world. By employing this emerging digital standard to host image metadata, it allows image resources to be easily shared, incorporated, and recontextualized without loss of authority or human intervention.

By developing tools that directly integrate with existing IIIF implementations, our techniques will be immediately usable by any other institution that implements the IIIF framework. This will significantly increase the opportunities for collaboration and reuse across the cultural heritage sector. IIIF provides the framework for how to access images as well as a standardized format for integrating our data output into standardized annotations. Our proposed software fills the key step in the middle of locating, identifying, and

Narrative

generating that data. We have prototyped this workflow with the palette extractions tool found at http://palette.davidnewbury.com (see appendix).

Through the release of these open-source tools, it is our intention to provide a clear path for the usage of machine learning techniques to enhance visual research in the humanities.

Environmental Scan
Over the past three years, there has been rapid progress in machine learning algorithms, especially those using neural networks for image understanding. Moreover, the open-sourcing of these algorithms has led to a renaissance in experimentation. The application of machine learning techniques for annotating and organizing institutional image archives, however, is a field still in its infancy. To date, there are only a few publicly accessible examples.

The best known of these may be a recent collaboration between the Google Arts & Culture research group and French new-media artist, Cyril Diagne; this project presents a map (https://artsexperiments.withgoogle.com/#/art_map) that organizes nearly a half-million artworks, from more than 600 cultural institutions, according to their visual similarity. In a related vein, the British Library recently collaborated with German new-media artist, Mario Klingemann; using the library's 1-million-image Flickr collection, Klingemann used a hybrid approach mixing automatic classification with manual confirmation to identify and tag tens of thousands of the images – ranging from maps and portraits to flora and fauna. Among other discoveries, Klingemann has uncovered previously unknown examples of image plagiarism.

The Cooper Hewitt museum has been working with image augmentation using computational analysis tools, and their collections website provides many interesting examples of using images to augment collection metadata. Their software libraries are open source, and, in fact, our prototype palette extraction system was generated using their library (https://github.com/cooperhewitt/plumbing-palette-server). However, while their algorithms have been released as open source, implementing them require significant computer skills and are not integrated into any existing ecosystem of tooling.

History of the project
CMoA first released the public dataset of the Teenie Harris archive on their GitHub site in the spring of 2016. During this time period, Project Co-Director David Newbury and CMoA Curator Lulu Lippincott brought the dataset to Project Director Golan Levin's course in information visualization at Carnegie Mellon University. CMU Students, working with Levin and Newbury, built experimental computer vision tools as part of their coursework. The positive initial outcomes from that pilot study inspired the more in-depth development of this project.

Work plan
Phase I: Initial Image Dataset Anaylsis and Preparation (September - December 2017)

The primary goal of this initial phase is to upload the nearly 80,000 digitized images from the THA, along with the accompanying metadata, into an IIIF Image server. This will require provisioning a server, generating IIIF Image API and Presentation API Manifests for each image, and verifying that these images are accessible and usable by our framework. Additionally, we will develop a system that will allow our generated annotations to be appended to these manifests.

Narrative

Phase II: Prototype Tool Workflow Development(January - May 2018)

In this phase, we will take the existing machine learning annotation tools developed in our prior prototyping and modify them to integrate with the IIIF server from Phase I. We will also develop a standardized way of accessing these tools, and we will verify the ability of our tools and workflow to work across the complete image set.

The goal of this phase is to prove the effectiveness of our proposed workflow and develop and document consistent system of patterns for applying, describing, and integrating the machine learning tools.

Phase III: Development of Prototype Tools (May - August 2018)

In this phase, we will prototype and develop a collection of new machine learning tools that use the workflow and patterns developed in Phase II. These tools will extend our prototyping work and create new annotation systems and integrations. During this phase, we will also engage both CMOA and our Machine Learning consultant to advise and verify that our techniques are both computationally correct and applicable and useful to the cultural heritage sector.

Phase IV: Documentation and Project Dissemination (September 2018 - February 2019)

In this final phase of our proposed project, we will complete the documentation of the tools and develop a small website that presents examples and provides documentation to institutions that are interested in applying these tools to their own images.

We will also provide all of the annotations developed over the course of this project to CMoA for potential permanent integration into their collection metadata.

We will also present both a public lecture and a physical interface showcasing our research at The Frank-Ratchye STUDIO for Creative Inquiry. This lecture will be recorded and made available online to aid in outreach to the cultural heritage community.

Throughout this work plan, we will be in contact with the IIIF community to assure that the tools and workflow are well-integrated into existing and future community best practices. Newbury is a current participant in both the general and Museums IIIF working groups, and will demonstrate the tools and techniques to that community.

In keeping with the STUDIO's commitment to open source development, our prototype software tools, associated project source code, workflows and use cases will be posted on the STUDIO's GitHub site. This will be released under the MIT open source media license. Full documentation of project outcomes will be posted on the STUDIO's website, www.studioforcreativeinquiry.org, and the core project team will produce the NEH white paper. Photo and video documentation of the project will be posted publicly on the STUDIO's Flickr and Vimeo accounts. During the final phase of the project, we will submit the results of our research and key findings for presentation at either the Museum Computer Network conference or the IIIF Conference.

Narrative

Staff

Golan Levin and David Newbury will serve as Co-Directors on the project, in charge of direction of the project and as the primary developers of the prototype software tools. In addition to project direction, David Newbury will serve as Lead Software Developer, overseeing the technical direction of the project. Tom Hughes will serve as Project Manager and will oversee logistics and coordination between project partners and consultants as well as budget oversight. Carnegie Mellon University students will be engaged as assistant researchers, working with Levin and Newbury in the development of the software elements of the project. In addition to this core development team, a machine learning consultant to advise and verify that our techniques are both computationally correct and applicable. This consultant may come from the Machine Learning Department at Carnegie Mellon University or an external partner based on project needs. CMoA staff (Curators, Archivists) will be engaged as consultants to advise on the applicability and usefulness of our project to the cultural heritage sector. For testing crowdsourcing methodologies, the STUDIO will engage users via the Amazon Mechanical Turk online service.

Final product and dissemination

Please see Phase IV under Work Plan

Appendices

The website links listed below are work samples by the STUDIO team as well as related websites mentioned in our proposal:

The Teenie Harris Archive
http://teenie.cmoa.org/

The International Image Interoperability Framework
http://iiif.io/

Work Sample I: IIIF Palette Service
http://palette.davidnewbury.com/

This is an example of an IIIF Service that encodes color palettes into the IIIF Image API. It also provides a demonstration of a viewer for it, as well as a functional transform generator for adding this service to existing info.json files.

Work Sample II: Terrapattern.com
http://www.terrapattern.com/
Terrapattern is an online, visual search tool for satellite imagery. The project provides journalists, citizen scientists, and other researchers with the ability to quickly scan large geographical regions for specific visual features.

Work Sample III: Student Investigations with the Teenie Harris Archive
http://zariahoward.github.io
Preliminary set of investigations by a CMU student Zaria Howard examining the Teenie Harris Archive using machine learning and computer vision techniques

Appendices

Data Management Plan

This project will primarily result in three forms of data: computer source code, machine learning models, and documentation.

In keeping with the STUDIO's commitment to open source development, our prototype software tools, models, associated project source code, workflows, documentation, and use cases will be posted on the STUDIO's GitHub site. This will be released under the MIT open source media license or CC BY 4.0 Creative Commons license as applicable to allow for the largest possible application. These will remain available on GitHub for a period of not less than five years, and we anticipate them remaining there for the foreseeable future. We will also host a mirror of the source code, documentation, and models on the STUDIO's website to protect against the event that GitHub is no longer available for hosting.

Our primary data set, the Teenie Harris Archive, consists of images and metadata that are publically available through the Carnegie Museum of Art. CMOA is the custodian of these data sets and will continue to maintain the canonical copy of this data. Throughout the project, our use of the Teenie Harris Archive will generate annotations in the Open Annotation (http://www.openannotation.org/spec/core/) standard that augment that collection. We have agreed to provide these annotations at the completion of the project to the Carnegie Museum of Art for potential integration into their permanent collection. Additionally, through the use of our tools, these annotations can be regenerated as needed.

(It is possible that IIIF will replace Open Annotation with W3C Web Annotations (https://www.w3.org/TR/annotation-model/) before this project can be completed. In that case, our annotations would use the Web Annotation Model.