Reverse Engineering the Image Library: the feasibility of using deep learning to identify significance in a 35mm slide collection

The obsolescence of the 35mm slide library has been a prominent issue in visual resource collections for over a decade. As online image collections have grown, institutions have witnessed a precipitous drop in the use of their analog collections, and increased pressure to repurpose the spaces in which these collections are held. Many institutions have pursued a combination of deaccessioning and digitizing these analog resources. Because the majority of slide libraries exist without any master catalog, it is tremendously time-consuming to sort, catalog, and identify significant slides in legacy slide collections. Thus, these vast image libraries are largely left unused, and retention and digitization policies for slide libraries are devised without systematic surveying, based on personal experience rather than analyzed data.

During this project, the Media Center for Art History at Columbia University will apply established techniques in computer science to investigate methods to extract data from digitized 35mm slides for the purpose of archival discovery. Using a representative sample of slides from the collection of the Department of Art History and Archaeology at Columbia University, we will test image processing, deep learning, and optical character recognition software to assess the viability of automatic cataloging and image processing for 35mm slides. During this year-long project, we will conduct experiments to read and catalog the text on 35mm slide labels, and investigate the potential of using image classifiers to identify which slides are original photographs and which are copied from books. We hope to discover techniques that will allow for less time-consuming and more cost-effective processing of legacy slide collections, enabling libraries to partially automate the process of cataloging and identifying significant images. At the conclusion of the project, we will publish a white paper that describes our methods and findings, as well as a website that describes how to replicate our techniques. We hope to create an open-source, scalable framework for automated processes of data analysis and classification for archival discovery across humanities fields.

The 35mm slide library, once the cornerstone of teaching art history at all levels, now languishes in storage, with valuable image resources trapped inside its drawers. Through innovative applications of computer science and image processing, this project seeks to make these legacy collections accessible again, with the potential to increase available teaching resources, make significant images available for study and free-use publication, and to preserve the institutional memory and pedagogical history retained in these collections.

Reverse Engineering the Image Library: the feasibility of using deep learning to identify significance in a 35mm slide collection

Statement of National Need

In the last decade, the 35mm slide library has fallen into disuse at universities, museums, and libraries everywhere. The growing prominence of the online image search coincided with the decision of Kodak to stop producing and servicing slide projectors, leading to the abandonment of the 35mm slide collection for the ease of digital image resources. The slide collection of the Columbia University Department of Art History and Archaeology typifies 35mm slide libraries at institutions worldwide. It consists of approximately 400,000 slides collected by faculty and staff from the 1940s through the early 2000s. Each slide is composed of a photograph contained in a slide mount and labeled with information regarding the origin of the image. The collection includes original photographic fieldwork by Columbia faculty and PhD students, slide sets purchased from vendors and museums, slides donated by faculty, and copy stand photography, wherein faculty requested images from books to be reproduced as slides for use in teaching. All of these types of slides are integrated and organized by subject and geographic region. (the order of the slides themselves, organized by subject, is the de facto catalog). However, no master catalog exists, disallowing any simple methods of retrieving significant materials from the collection. Faculty and students now almost exclusively source their images from online image searches or by scanning from books. Any significant or unique images in the slide collection have to be searched for individually, handling each slide in every drawer that corresponds to the subject, a process both time-consuming and cumbersome.

For over 10 years, many university departments, libraries, and museums have been facing similar problems with their slide collections. Of 25 slide libraries surveyed, all possess slide collections numbering from 125,000 to 500,000 slides. Because it would take years of labor at high cost to manually sort, catalog, and identify important images from these collections, most have scanned at most a few thousand of their slides upon faculty request, then either left the collections unused, moved them to inaccessible storage facilities, or begun the process of deaccessioning them. In a personal conversation, one university Visual Resources Librarian described “agitation for space” by the university, and cited the concern that the time-consuming nature of identifying important slides would make it impossible to fully process the collection in a timely nature. Another university Art Librarian explained to us via email that the university digitized some of the slides earlier in the transition to digital media, but added, “if we were beginning the project now… we might make different decisions.” Karen Bouchard, the Scholarly Resources Librarian for Art and Architecture and Curator of the Instructional Image Collection at Brown University, has written an article on her methodology and experience weeding out the slide collection at the university to the bare minimum (VRA Bulletin, Vol. 41 [2014], Iss. 2, Art. 10). The methodology for slide retention she describes is based on her personal experience with the collection rather than a data-driven survey of the images and their metadata. She also describes the slow pace of student workers and the time crunch at the end of the entirely manual process that left her with no choice but to deaccession a large portion of the collection without going through it. Our project proposes to address this common problem by investigating automated and algorithmic processes for cataloging slides and identifying significant images. We plan to experiment with techniques that will lessen manual processing time, allowing libraries to

established data-based retention criteria and quickly identify significant image resources in large collections.

As digital collections grow in ubiquity, image processing and deep learning techniques are increasingly being applied to the humanities. Deep learning is a type of Artificial Intelligence that uses mathematical models based on the human brain to allow computers to "learn" to detect features instead of manually programming solutions. Using software packages such as Tensorflow, Google's open source deep learning framework, large sets of sorted data can be used to train computers to perform tasks such as facial recognition and image classification. The use of deep learning technology to extract information from digitized materials was explored in the University of Nebraska-Lincoln's Image Analysis for Archival Discovery (Aida) project (2013–2016) (http://aida.unl.edu/). The Aida team developed a deep learning classifier to identify poetic content in digitized historic newspapers based on visual signals. North Carolina State University is also exploring computer vision and algorithmic processing in the humanities in their Illustrated Newspaper Analytics project that began in 2016 and continues today (https://ncna.dh.chass.ncsu.edu/imageanalytics/techniques.php). This project applies image processing techniques to extract figures, match images, detect faces, and detect halftone in 19th century newspapers, and trains image classifiers to identify these features.

Algorithmic image analysis in art historical research has been explored by John Resig in several ongoing projects, including the Frick Photoarchive (http://www.frick.org/research/photoarchive) and Pharos: the International Consortium of Photo Archives (http://pharosartresearch.org/). Using deep learning for image similarity analysis, these databases automatically match visually similar images and correlate image metadata. Resig writes, "The potential for computer vision and image analysis to change how photographs and images are managed in archives, libraries, and museums is absolutely staggering. Tasks that previously were insurmountable (such as merging two million-photograph archives) are now in the realm of possibility. The implications of this technology are still being explored and are likely going to completely change photo archives as they currently exist." (2013). All of these projects demonstrate the potential for applying computer science in the digital image archive. These technologies will allow for advances in image retrieval, identifying similar or identical images across multiple archives, correcting errors in cataloging and metadata, and linking diverse digital archives across institutions to provide more comprehensive resources for teaching and learning.

Project Design

During this project, the Media Center for Art History at Columbia University (MCAH) will apply established techniques in computer science to assess the viability of new methods of automated processing and data extraction in a collection of 35mm slides. We have already digitized a sample of approximately 8,000 randomly selected slides from Columbia University's collection to use in experimental trials. Using open-source, industry standard software for image processing, deep learning, and optical character recognition, we aim to produce a scalable framework that addresses two major issues facing slide libraries: producing a searchable catalog, and identifying copy work.

To conduct these experiments, we will begin with building a workstation for parallel processing (a computer designed to handle simultaneous workloads.) We will seek input on metadata needs and standards from other visual resource librarians at peer institutions. With the advice we glean, we will build a database to be used for cataloging the metadata and physical attributes of the slides. All 8,000 slides in our sample set will be manually cataloged by a student worker trained by Media Center staff who possess expertise in visual resource cataloging and metadata standards. This manual catalog will be used as a control to compare with the automatically generated metadata. In addition to metadata regarding the artwork pictured on the slide, we will also manually catalog physical attributes of the slide that may impact digital signal processing, including noise such as dust, scratches, and other marks on the images and labels.

Because the vast majority of 35mm slide libraries exist without a written catalog, research and retrieval is difficult and time-consuming, and the prospect of manually cataloging a large collection of slides is prohibitively labor intensive. Using optical character recognition technology in the open-source package Tesseract, we will conduct trials to read slide labels and automatically input slide cataloging information. Tesseract is a popular open source optical character recognition currently sponsored by Google. The software analyzes the structure of images to determine patterns and extract printed text. Quality control will be conducted using the manually entered cataloging information for the corresponding set of slides, and the results will be used to refine and improve the software's performance.

Many slides in slide libraries are "copy work"; i.e., the slide image was photographed from a book, and is thus just a low-quality reproduction of an extant printed image. These copy work slides are generally less useful as better quality reproductions could be made by scanning the books. They also present copyright concerns. Separating out the copy work slides aids in identifying original photography and significant slides in the collection.
Copy work slides can be identified by the presence of halftone. Halftone is a translation of photographic tonalities into a continuous series of dots, and is a visual artifact of mass production mechanical or digital printing processes (Stulick 2013). If an image contains halftone, it is a reproduction from printed material and is not an original photograph. Using the open-source image processing package OpenCV, a discrete Fourier transformation (as described in Liu et. al.) can be applied to digital images. OpenCV and other digital image processing programs mathematically manipulate images to produce an output of an enhanced or filtered image. The discrete Fourier transformation produces a pattern that can be used to conclusively prove if an image was produced with halftone. We will apply this transformation to all the digitized slides. Quality control will be performed on this process by comparing the results from the transformed images with manual assessments of whether an image has halftone patterning.

Rather than individually identifying the presence of halftone by examining all transformed images, we will train an image classifier using the open-source deep learning framework Tensorflow to recognize the patterns that indicate halftone in transformed images. By inputting a smaller number of images sorted into halftone and non-halftone pattern categories, Tensorflow is used to yield an algorithm with the ability to identify future images as halftone or non-halftone. A successful application of this technique allows automated detection of whether a slide image is a more valuable and unique original photograph, or if it is an image copied from a mass-produced book. The results from this experiment will be analyzed using statistical

measures of binary classification that compare true positive, true negative, false positive, and false negative results. These results allow for the Tensorflow image classifier to be refined and "retrained" to produce more accurate identifications. For evidence of the validity of these methods, see results of preliminary tests on pages 5-6.

At the completion of these experiments, we will produce a comprehensive white paper, which we will make publicly available on the Media Center's website. This white paper will include our data and results, as well as a detailed manual of our methods and how to replicate them. Our white paper and website will provide an open-source, user-friendly, scalable framework for repeating these image analysis and automated cataloging techniques in other collections.

The project will be primarily completed by Media Center staff: Gabriel Rodriguez, Digital Curator; Kate Burch, Assistant Curator; and Tim Trombley, Educational Technologist. We will also hire two student employees to assist with the project: a Computer Science student assistant with basic programming skills and an interest in deep learning, and an Art History undergraduate student cataloger. The Media Center staff is supported during the academic year by three undergraduate work-study students. These students will also be available to assist with the transcription cataloging work for the project. The Media Center's director, Stefaan Van Liefferinge, will manage the project, and will also train and supervise the CS assistant who will run tests with the deep learning software. Van Liefferinge is well-equipped to bring this project to a successful conclusion. As described in his curriculum vitae, Van Liefferinge holds a graduate degree in Computer Science and held the position of Project Manager for Software Engineering for a company specializing in data extraction and document classification using deep learning. He is also well-qualified to supervise the CS assistant. Van Liefferinge previously served as the Principal Investigator of a NEH supported project in Artificial Intelligence in which CS students participated, and has served on Master's thesis committees of CS graduate students.

The Media Center uses the open-source project management web application Redmine to track all projects and work completed by staff; we will use this web application to track progress, keep notes, and ensure deadlines are met for this project. We will also keep extensive notes in Redmine to assist in preparing a thorough explanation of all of our methods in the white paper and website that we prepare at the end of the project.

Applying computer science techniques to our project includes some risks, which our work plan seeks to mitigate. The largest risk, as with any software-based product, is the effectiveness and longevity of the software used. Early tests of the software we plan to use have been successful. To mitigate the risk of obsolescence, we will use all software as a generic method of conducting our experiments, allowing for future adaptability. The senior project management experience of the PI will help in this regard.

Another risk is the interference of noise. As with any digitized image, noise created by condition issues on the slides, including dust, marking, and scratches, can interfere with the digital signal processing that will occur in our experiments. We will classify the types of noise and quantify and analyze the effects it has on our results. Additionally, we may draw upon the

expertise of Columbia University Computer Science and Data Science departments if needed, through the Data Science Institute's Campus Connection program, which matches qualified graduate students with on-campus research projects that could benefit from their skills.

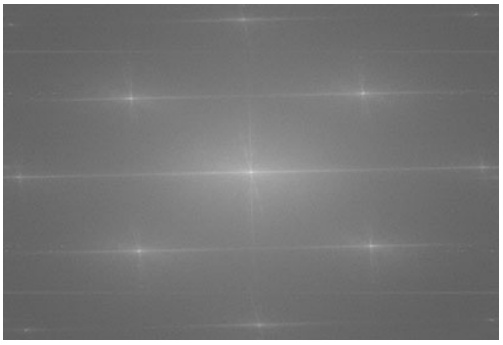
The effectiveness of our sample set is another risk that could affect results. We have decided on a random sampling method to choose 2% of Columbia University's slide collection (approx. 8,000 slides) for use in this project. However, the random sample may not accurately represent the variations issues we may face when applying these methods to an entire collection. Finally, as with any digital project, we must ensure that our methods are user-friendly enough to be applied across a wide variety of institutions with variable resources. We will mitigate this risk with thorough and in-depth documentation of all steps of the project on the public website we will create, and by providing our contact information for consultation. All the software we use will be open-source, and we will aim to use software packages in popular programming languages that include extensive documentation. These choices will also ensure the sustainability of the methods we develop.

Results of Preliminary Tests performed at the Media Center

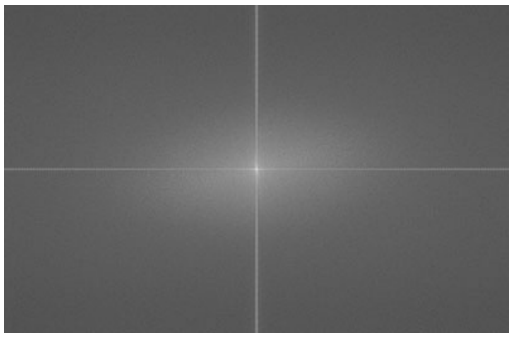
Slide A (detail to show presence of halftone)

Slide B (detail to show absence of halftone)

After applying a Discrete Fourier transformation (DFT), the slides produce recognizable patterns.



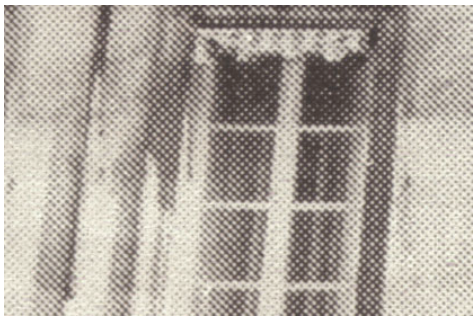
Slide A, after DFT



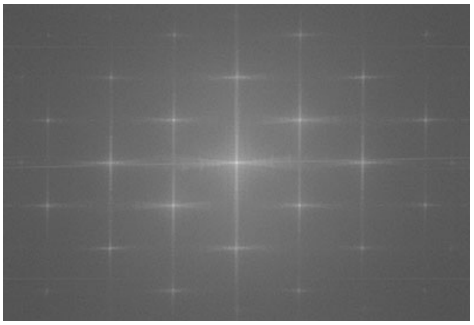
Slide B, after DFT

Using transformed slides such as these to train an image classifier in Tensorflow, the program is able to detect the presence of halftone in newly inputted slides, outputting a result with a degree of confidence.

After performing a Discrete Fourier transformation on Slide C, the image is inputted into Tensorflow, and the program correctly identifies it as a slide with halftone.



Slide C (detail to show presence of halftone)



Slide C, after DFT

Tensorflow output:
halftone (1): 0.993796
nonhalftone (0): 0.00620454

Tensorflow identifies Slide C as a slide with halftone with a 99.3796% degree of confidence.

Work Plan (see Schedule of Completion)
November:
-Build workstation for parallel processing (Trombley)
-Travel to peer institutions to research metadata needs and standards in image libraries, ensuring that the cataloging we produce is accordance with best practices in visual resources (Rodriguez)
-Analyze slides to determine metadata fields, such as artwork title, artist, location, date, etc. (Burch)
-Following travel and slide analysis, a metadata entry scheme will be developed and implemented (Rodriguez, Burch, CS student assistant)
-Following the completion of the metadata entry scheme, the undergraduate Art History student will begin copy cataloging all metadata from the first batch of approximately 2,000 slides (AH student cataloger)
- cataloguing database created (Trombley)

December:
-Copy cataloging continues (AH student cataloger)
-First image processing tests of scanned slides begin (Van Liefferinge, CS student assistant)

January:
-Copy cataloging of first batch will be completed (AH student cataloger)
-Using image processing results, image types for deep learning will be determined (Rodriguez, Burch)

February:
-Copy cataloging of second batch (metadata from 3,000 slides) will begin (AH student cataloger)
-First deep learning test, using slides from first cataloged batch of slides (Van Liefferinge, CS student assistant)
-First optical character recognition test, using slides from first cataloged batch (Van Liefferinge, Burch, CS student assistant)

March:
-Copy cataloging of second batch continues (AH student cataloger)
-Quality control assessments of deep learning and OCR results (Rodriguez, Burch)
-Statistical analysis of deep learning results (Rodriguez, Burch, CS student assistant)

April:
-Copy cataloging of second batch completed (AH student cataloger)
-Refine OCR based on results, and start batch-processing remaining scanned slides (Van Liefferinge, Burch, CS student assistant)

May:
-Copy cataloging of third batch (metadata from 3,000 slides) begins (AH student cataloger)
-Retrain and refine deep learning classifiers based on results (Van Liefferinge, CS student assistant)

June:
-Copy cataloging of third batch continues (AH student cataloger)
-Second deep learning test with refined classifiers (Burch, Van Liefferinge)

July:
-Copy cataloging of third batch completed (AH student cataloger)
-Quality control assessments and statistical analysis of second deep learning test (Rodriguez, Burch)
-Large batch OCR completed

August:
-Further refinement and final deep learning tests (Burch, Van Liefferinge)
-Final OCR refinement and processing of any remaining slides (Burch)
-Begin analysis of all results (Van Liefferinge, Rodriguez, Trombley, Burch)
-Begin draft of white paper (Van Liefferinge, Rodriguez, Burch)

September:
-Final OCR processing completed
-Quality control assessments and statistical analysis of final deep learning test (Rodriguez, Burch)
-Continue analysis and white paper (Van Liefferinge, Rodriguez, Trombley, Burch)

October:
-Complete analysis of all results (Van Liefferinge, Rodriguez, Trombley, Burch)
-Finalize white paper (Van Liefferinge, Rodriguez, Burch)
-Publish on MCAH website, make all results available on GitHub, Columbia's Academic Commons (https://academiccommons.columbia.edu/) as well as on academia.edu (Trombley)

National Impact

This project will produce a scalable framework to automatically catalog and identify visual information from 35mm slides, assessing methodologies for automated data extraction and classification and providing accessible tools for other institutions to apply to 35mm slide libraries and other digitization projects. Automating cataloging and image identification processes will significantly reduce the cost and time required to make legacy collections usable again, giving institutions the potential to increase image resources and broaden access to them. This will have the additional effect of improving collection management. Once a slide collection has been cataloged for future retrieval, with important images isolated, the collection can be moved to storage or partially deaccessioned, creating valuable space for libraries and university departments.

The basic cataloging created automatically can potentially be expanded on in the case of significant slides. The ease of archival discovery in 35mm slide collections enabled by this project will also allow institutions to find images that can be released into the public domain for research and publication, at great benefit to scholars nationwide. At Columbia, any photographs taken by faculty for the department may be made public, and other institutions can release images to the public domain based on their policies. These images can also be disseminated by their institutions through image databases such as Artstor, so the work done in this project has the potential to greatly increase valuable image resources accessible worldwide. The Media Center has worked closely with Artstor to share images from our collections in the past, and using the methodologies developed in this project, would potentially uncover many important art historical images to be shared on the Media Center's own image database, as well as with Artstor and other online image databases.

The slide library is an important archive of institutional and pedagogical history in the field of Art History. The ability to catalog and process these collections will provide valuable information for research on the historiography of the field, and will allow for cross-disciplinary exchange of image resources previously isolated by old and even obsolete cataloging distinctions. The automatic production of an RDF-based catalog will also allow for slide collections to be integrated with collaborative digital resources. The open source standards and linked data created by this framework will allow maximal use as a collaborative tool for scholars for integration with projects at the forefront of Semantic Web.

Our hope is to foster similar experiments for the multitude of large 35mm legacy collections. The methods explored by this project will be executed with user-friendly adaptability in mind. The website that accompanies our white paper will provide detailed step-by-step instructions on the technologies used, and all software will be open-source. The Media Center

will also provide advice and consultation for institutions who implement our methods in the future.

In addition to the potential to impact 35mm slide libraries, we hope our work will inspire the increased use of technology in digitized legacy resources across the humanities. All of the techniques explored in this project can be applied to any large digital database to make digital resources usable beyond simple facsimile representation. Optical character recognition, image processing, and deep learning technology can enable the efficient creation of metadata and automatic identification of content, and presents the potential to produce cross-disciplinary connections, in-depth data analysis and visualization, and new ways to create linkages between collections. We will produce a framework for methods that are reusable at low- or no-cost for institutions nationwide to increase their technological capabilities and bring their legacy collections into the digital age.

Bibliography

Blanke, Tobias, Michael Bryant, and Mark Hedges. "Open source optical character recognition for historical research." *Journal of Documentation* 68, no. 5 (2012): 659-683.

Burger, Wilhelm, et al. *Principles of digital image processing*. London: Springer, 2009.

Chen, Ching-Jung. "Analog to Digital: Conversion of the Image Libraries at the City College of New York." *Art Documentation: Journal of the Art Libraries Society of North America* 28.1 (2009): 36-39.

Chen, Tung-Shou, Jeanne Chen, and Yu-Mei Pan. "A new detection method of halftone images based on crisscross checking technique." *Multimedia Software Engineering, 2003. Proceedings. Fifth International Symposium on*. IEEE, 2003.

Liu, Yun-Fu, Jing-Ming Guo, and Jiann-Der Lee. "Halftone image classification using LMS algorithm and naive Bayes." *IEEE Transactions on Image Processing* 20.10 (2011): 2837-2847.

Lorang, Elizabeth et al. "Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections." *D-Lib Magazine* 21.7/8 (2015). Web.

Lundstrom, Jens, Verikas, Antanas. "Detecting halftone dots for offset print quality assessment using soft computing." International Conference on Fuzzy Systems, 2010, p. 1.

Manovich, Lev ---. "Media Visualization: Visual Techniques for Exploring Large Media Collections." *Media Studies Futures*. Ed. Kelly Gates. Blackwell, 2012. Web. http://softwarestudies.com/cultural_analytics/Manovich.Media_Visualization.web.2012.v2.doc

Nowviskie, Bethany. "Select Resources for Image-Based Humanities Computing." *Computers and the Humanities* 36.1 (2002): 109–131. Web.

Opar, Barbara Ann. "Discard to Retention: A specialized evaluation and digitization project for architecture slides at Syracuse University." *VRA Bulletin* 39.3 (2013): 5.

P. C. Chang and C. S. Yu, "Neural net classification and LMS reconstruction to halftone images," in *Proc. SPIE*, 1998, vol. 3309, no. 2, pp. 592–602.

Parsons, Sarah. "What Lies Beyond the Slide Library?: Facing the Digital Future of Art History." *RACAR: revue d'art canadienne/Canadian Art Review* (2005): 114-125.

Radke, Richard J., et al. "Image change detection algorithms: a systematic survey." *IEEE transactions on image processing* 14.3 (2005): 294-307.

Resig, John. "Using computer vision to increase the research potential of photo archives." *Journal of Digital Humanities* 3 (2013): 3-2. https://johnresig.com/research/computer-vision-photo-archives/

Stulik, Dusan C. *The Atlas of Analytical Signatures of Photographic Processes: Halftone*. Los Angeles: Getty Conservation Institute, 2013. Web. http://www.getty.edu/conservation/publications_resources/pdf_publications/pdf/atlas_halftone.pdf

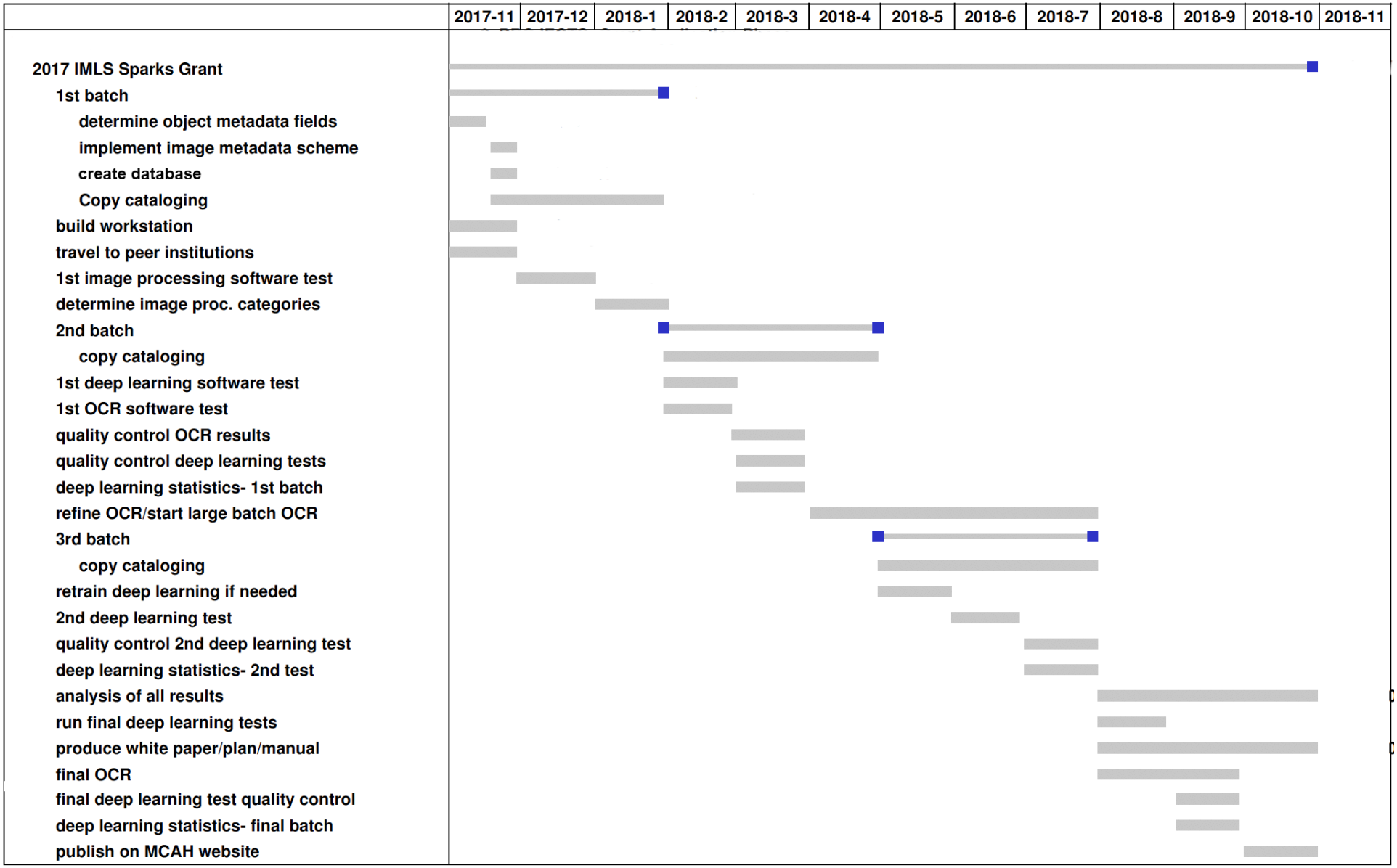
Tai, Chun-Jung, Robert Ulichney, and Jan P. Allebach. "Effects on Fourier Peaks Used for Periodic Pattern Detection." *Electronic Imaging* 2016.13 (2016): 1-8.

Terras, Melissa. "Image Processing in the Digital Humanities." *Digital Humanities in Practice* (2012): 71-90.

van der Maaten, Laurens, and Robert G. Erdmann. "Automatic thread-level canvas analysis: A machine-learning approach to analyzing the canvas of paintings." *IEEE Signal Processing Magazine* 32.4 (2015): 38-45.

Columbia University in the City of New York

Schedule of Completion



DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions
You must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

PART I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

This grant application is intended to fund experimentation to find a technique to automatically determine significant assets in a 35mm slide collection. The product will be a written resource detailing the most successful process and the best practices for the use of other visual resource collections. The results of our experimentation will be published without restrictions in a white paper that we will post publicly on our website learn.columbia.edu, Columbia Academic Commons and on Academia.edu. Additionally, we will also provide the test metadata and results from our experimentation in a CSV file. Both the metadata and the content of our white paper will be made available free of restrictions under the Creative Commons Zero Public Domain Dedication. This license gives the user the right to use our results for any purpose without having to give attribution. However, the Media Center will request that users actively acknowledge and give attribution whenever possible.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The Media Center will exert no ownership rights over the results of our experimentation for non-commercial use. We will notify users of the relevant terms and conditions as stated by the CC0 license as described in section A.1.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

We will create no products of this kind.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

Our research will be primarily accumulated in CSV spreadsheets as well as summarized in a white paper which will be published on our website. The results of the experimentation will include metadata for 8000 slides, both manually created by student workers and automatically created by OCR and sorted using deep learning software.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

We will use a workstation built for the purpose of expediting parallel processing workloads. The software utilized in this project will be OpenCV, TensorFlow, Tesseract OCR software, Microsoft Excel and Filemaker Pro. The images that will be analyzed have been captured in TIFF format using a Nikon D800 DSLR with 36 MP resolution.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

We will use digitized slide images in TIFF format at approx. 5000 x 5000 pixel resolution to conduct experiments. These images will be computer analyzed to determine whether or not they are derived from books or are original photography. The results of this analysis will be provided in a CSV file. The metadata created by this experimental process will also be made available in XML format in addition to a master CSV.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

The results of an automated workflow as determined through experimentation in this project will need to be periodically tested against the results of a manual workflow. Batches of slides will be cataloged in sets of 2000, 3000, and 3000 - each in 3 months cycles - during the grant period. Metadata for these batches will be manually entered by student workers, including specific data on the criteria that we have predetermined for our deep learning process (for examples, halftone vs. non-halftone images) in addition to the descriptive metadata on the slide label (which in the automated workflow is determined by OCR). This manual dataset will be compared to the results of the deep learning and OCR results for quality control.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

The products of the work will be stored on a storage server with redundant backups in MCAH server space. Regular maintenance is performed on the storage server to ensure data integrity and near-constant uptime for seamless network access by MCAH staff and collaborators. The server room is a

temperature controlled environment with a dedicated HVAC unit and an uninterruptible power supply to ensure stable conditions. The digital surrogates will also be duplicated in off-site cloud storage by a third-party to safeguard materials in the case that there is physical damage to local hardware and infrastructure.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

Transcription cataloging will be completed using Filemaker and stored in .csv files. The descriptive metadata will use VRA Core as a metadata scheme. Getty Vocabularies (ULAN, AAT, TGN, CONA) will be utilized as authorities with their associated reference IDs. Technical and preservation metadata concerning slide condition, slide type, extraneous markings and halftone vs. non-halftone, will receive their own custom fields.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

The textual cataloging resulting from Optical Character Recognition and manual efforts by MCAH staff and collaborators will be recorded and stored in comma separated value (CSV) files, the most accessible and stable format for tabular data. The correlation between the image files and the cataloging records will be established by a unique Record ID which will also serve as the title of each image file.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

All materials produced by the project for public consumption will be made available through the MCAH's website which will serve as a hub for all written materials and any open-source code related to the workflow. The published results of our experimentation, in the form of a white paper, will also be available on Columbia's Academic Commons (https://academiccommons.columbia.edu/), and on Academia.edu. Metadata in master .csv files will be uploaded to GitHub, a version control repository and internet hosting service that offers collaboration features. The metadata and a corresponding GitHub wiki will be available for public use and download.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

In addition to being published on GitHub, Academic Commons, and Academia.edu, the project website and all digital content will be publicly available on the Media Center's website. MCAH delivers digital assets using Drupal, which would be suitable for the hosting and integration of the deliverables into MCAH's existing web platforms. The project resources will be accessible via all standard web browsers in all of these places.

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources,

or assets your organization has created.

Media Center Site: http://learn.columbia.edu/
Art Atlas: https://mcid.mcah.columbia.edu/art-atlas
Media Center Image Database: https://mcid.mcah.columbia.edu/
GitHub: https://github.com/MCAH

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

N/A

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

N/A

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

N/A

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

N/A

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

N/A

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

N/A

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

N/A

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

N/A

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

N/A

C.3 Identify where you will deposit the source code for the software you intend to develop:

N/A

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

The results of our project will not be a dataset put rather a published resource describing our methodology. However, our experimentation will utilize descriptive and technical metadata from our sample set of 8000 slides. This data will be derived both from the slide labels themselves using transcription cataloging and using the automated processes described in the grant proposal (OCR and deep learning for image analysis). The digitized slides will be cataloged in a batch of 2000 for the first three months of the grant period, and then in batches of 3000 for each the remaining two intervals of three months.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

N/A

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

N/A

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

N/A

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

The methodology used to collect the data, as stated in section A.1, is the subject of our investigation. Technical and descriptive metadata on our sample set of 8000 slides will be collected both by our student cataloger as well as by the automated means of using Tesseract OCR software as well as Tensorflow and OpenCV to analyze the image for certain technical criteria. The data will be generated in flat text, CSV files readable by any text or spreadsheet editor. There are no technical requirements or dependencies necessary. Data will be stored in CSVs.

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

N/A

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

All of the findings for this project will be presented in an extended version of the white paper which the MCAH will host publicly on its website: http://learn.columbia.edu. The accompanying dataset will also be posted in addition to any documentation concerning the the digital capture setup. Any useful code or code snippets will be publicly accessible on the Media Center's website and GitHub. All software packages used in this project are open source, documented, and freely downloadable.

A.8 Identify where you will deposit the dataset(s):

see C.2, D.1, and A.6

Name of repository: Media Center for Art History

URL: learn.columbia.edu
GitHub: Https://github.com/MCAH

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

The data management plan will be reviewed ad hoc when needed, for example, when changes are made to the Media Center's server configuration or website. The data will remain available for the foreseeable future.