

May 2020

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

Stefaan Van Liefferinge

Columbia University, sv143@columbia.edu

Gabriel Rodriguez

Columbia University, gsr2101@columbia.edu

Lisa Peck

Columbia University, emp2201@columbia.edu

Tim Trombley

Columbia University, trt2115@columbia.edu

Kate Burch

Greater Portland Landmarks, ktebrch@gmail.com

Lauren Arnett

Columbia University, lba2138@columbia.edu

Karen Lin

Columbia University, kl2985@columbia.edu

Follow this and additional works at: <http://online.vraweb.org/vrab>

Recommended Citation

Van Liefferinge, Stefaan, et al. "Reverse Engineering the Image Library: a case study on the feasibility of using deep learning to identify significance in a 35mm slide collection." *VRA Bulletin*: Vol 46, Iss.2, Article 5, 2019. Available at: <https://online.vraweb.org/index.php/vrab/article/view/181>.

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

Abstract

The Columbia University Department of Art History and Archaeology holds approximately 400,000 35mm slides, but like other institutions without a master catalog, the collection is tremendously time-consuming to sort, leaving resources to languish in storage. To help resolve this, the Media Center for Art History at Columbia University used deep learning and optical character recognition software to detect original photographic images in the 35mm slide collection. Both technologies served to classify images as copywork or an original photo. This project aimed to apply transferable techniques that will enable other collections to partially automate the process of cataloging and identifying significant images to create an open source, scalable framework for archival discovery across humanities fields. This paper seeks to describe the methods and challenges and make clear the processes investigated.

This article has undergone a double-blind peer review process.

Keywords

35mm slides, collection access, technology, deep learning, neural net, computer vision, halftone, artificial intelligence (AI), automation, OCR, optical character recognition

Author Bio & Acknowledgements

The Media Center for Art History develops and supports fieldwork and research projects by documenting, presenting, and interpreting works of art, architecture, and cultural heritage sites. The Center's goal is to advance the digital humanities, explore digital technologies, and preserve and develop its visual collection. As part of the Department of Art History and Archaeology, the Media Center's specialized personnel and facilities serve the closely related fields of archaeology, art history, and architectural history. The staff is composed of the Director, Stefaan Van Liefferinge, Digital Curator Gabriel Rodriguez, Assistant Curator Lisa Peck, and Educational Technologist Tim Trombley. Media Center staff has expertise in art history, archaeology, computer science, and architectural history.

Acknowledgements:

Lauren Arnett (Columbia College '19) and Karen Lin (Columbia College '21) were student programmers. Manual cataloging and classification were carried out by Dominique Groffman (Columbia College '19) and Sophie Fox (Barnard College '19).

This project was generously supported by a Sparks! Ignition Grant from the Institute of Museum and Library Services.

Introduction

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 consists of approximately 400,000 slides collected from the 1940s through the early 2000s. The collection includes original photographic fieldwork by Columbia University faculty and Ph.D. students, commercial slide sets purchased from vendors and museums, slides donated by faculty, and copy stand photography, wherein faculty requested images from books and magazines to be reproduced as slides for use in teaching.

The Media Center for Art History (MCAH) is digitizing its slide collection for ease of use, allowing for remote access to the collection and increased findability. All 35mm slides are currently organized by subject and geographic region. No master catalog exists; the order of the slides themselves is the de facto catalog, disallowing any simple methods of retrieving significant materials from the collection.

Many university departments, libraries, and museums have faced similar problems with their slide collections. Of 112 slide libraries surveyed in 2014 by the Visual Resources Association Slide and Transitional Media Task Force, all possessed slide collections numbering from 80,000-550,000 slides.¹ As it would take years of labor to manually sort, catalog, and identify important images from these collections, most institutions surveyed 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.² In a personal conversation with Media Center staff, one Visual Resources Librarian described “agitation for space” by their university. Another Art Librarian explained via email that their 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, wrote on her methodology and experience weeding out the slide collection at the university to the bare minimum. 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.³

Abby Smith prophesied in 1999, “Digital will not and cannot replace analog. ... The real challenge is how to make those analog materials more accessible using the powerful tool of digital technology.”⁴ This project was designed to take on this challenge by investigating automated and algorithmic processes for identifying significant images. We experimented with techniques to lessen manual processing time for quickly identifying noteworthy resources in large collections, with the

¹ Slide and Transitional Media Task Force, “Tell Us Where Your Slides Are! Summary of Survey conducted in Fall 2014,” Visual Resources Association, 2015, <http://vraweb.org/survey-summary-from-slide-and-transitional-media-task-force/>.

² Ibid.

³ Karen A. Bouchard, “Now, Slides, Sail Thou Forth to Seek and Find! Facilitating a Slide and Photograph Diaspora,” *VRA Bulletin* 41, no. 2 (2015), <https://online.vraweb.org/vrab/vol41/iss2/10/>.

⁴ Abby Smith, *Why Digitize?* (Washington D.C.: Council on Library & Information Resources, 1999), 12-13.

additional goal of allowing slide libraries to establish individualized data-based retention criteria in cases where paring down a collection may be necessary.

Deep Learning and Algorithmic Image Analysis in Digital Collections

As digital collections become ubiquitous, image processing and deep learning techniques are increasingly being applied to the humanities. Deep learning is a type of Artificial Intelligence that uses computational models based on the human brain to allow computers to “learn” to detect features instead of manually programming solutions. Large sets of labeled data can be used to train computers to perform tasks such as facial recognition and image classification. North Carolina State University is currently exploring computer vision and algorithmic processing in the humanities in their Illustrated Newspaper Analytics project, which began in 2016. 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.⁵ Currently, many projects in digital humanities involve deep learning for image analysis and classification.⁶ These technologies 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.

In experimentation with deep learning, the Media Center seeks not to completely automate the digitization and weeding process but to use new tools to enhance the retention criteria and findability of analog source materials. While the creation of digital surrogates is especially important now for access reasons, the digitized image will never replace the original source material and is not intended to. This project attempts to provide an automated means to scour collections for retainable material in light of increasing pressure to deaccession.

Brief Project Summary

The Media Center for Art History at Columbia University was awarded a Sparks! Ignition Grant from The Institute of Museum and Library Services (IMLS) for the project “Reverse Engineering the Image Library: the feasibility of using deep learning to identify significance in a 35mm slide collection.” For this project, we identified the following procedures:

- Configuration of power processing workstation
- Manual copy cataloging of 7,815 slides for control data
- Development and testing of optical character recognition on slide labels

⁵ North Carolina State University, "Image Analytics Processes," Nineteenth-Century Newspaper Analytics, accessed December 18, 2018, <https://ncna.dh.chass.ncsu.edu/imageanalytics/techniques.php>.

⁶ University of Nebraska-Lincoln, Image Analysis for Archival Discovery (Aida), accessed December 18, 2018, <http://projectaida.org/>; The Frick Collection, "Photoarchive," accessed December 18, 2018, <https://www.frick.org/research/photoarchive>; PHAROS: The International Consortium of Photo Archives, accessed December 18, 2018, <http://pharosartresearch.org/>; John Resig, "Using Computer Vision to Increase the Research Potential of Photo Archives," *Journal of Digital Humanities* 3, no. 2 (Summer 2014) <http://journalofdigitalhumanities.org/3-2/using-computer-vision-to-increase-the-research-potential-of-photo-archives-by-john-resig/>; "The Art API - Artificial Intelligence for Art Recognition," ArtPI, The Art API, accessed March 1, 2019, <https://www.artpi.co/>; Computer Vision group, Heidelberg University, "Visual Search Tool," Digital Humanities, last modified 2018, <https://sabinelang254.wixsite.com/digital-humanities/visualesearchtool>.

- Development and testing of halftone detection on slide images
- Analysis of results

Below are details of each of these procedures. The overall project was preceded by the digitization of a sample set of slides from the collection.

Preliminary Work

A set of exactly 7,815 slides was digitized by work-study student assistants in the fall of 2017. These slides were pulled from 59 drawers representing varied sections of the slide collection; at least one drawer came from each major geographic distinction within the collection's organizational scheme and several varieties of media were represented. These images were digitized on a photostand using a Nikon D810 camera and a 60mm macro lens with a light table below. Each digital image was saved as a .tiff file, at approximately 4850x4850 pixels in dimension. Work-study student assistants cropped the images in small batches to ensure that no slide information was excluded from the final image.

Workstation Configuration

In order to process the materials created as part of the project and to provide an environment to develop scripting workflows, the Media Center constructed a high-powered workstation featuring a powerful multi-core processor, large amounts of RAM, and fast local storage as well as a dedicated graphics card.⁷ The graphics card was used in order to leverage GPU-targeted workloads common to many machine learning and imaging packages. To provide flexibility with open source software packages, the workstation uses Ubuntu Linux as its operating system, but many similar packages would be available on other operating systems. The workflows described in this paper could be accomplished with less capable hardware, but at the cost of longer processing time.

Manual Production of a Reference Set

Copy cataloging by Art History undergraduate student assistants generated a dataset of all textual information on each slide in the reference set. Cataloging found on slide labels was parsed to provide an accurate manual record of the text on each label.

During the manual cataloging process, variations in the spelling or grammar in both the artist/creator and title fields on the slide labels required the addition of two fields to the metadata scheme, one for the work agent's name as standardized in the Getty Union List of Artist Names, and one for the commonly known title of the work in question.⁸ Handwritten elements sometimes appeared on slide labels where corrections or fully handwritten labels were present on some of the oldest slides. A corresponding checkbox was added to indicate whether text was handwritten. An updated cataloging format was implemented to include these changes, as can be seen in Figure 1.

⁷ CPU: Intel Core i7 7700K; GPU: Nvidia 1080ti; SSD: 1TB Samsung 970 EVO; HDD: 4TB WD Black; RAM: 32GB DDR4.

⁸ Getty Research Institute, "Getty Union List of Artist Names," The J. Paul Getty Trust, last modified March 7, 2017, <http://www.getty.edu/research/tools/vocabularies/ulan/index.html>.

Figure 1: FileMaker Pro cataloging fields utilized by student catalogers. Top: Initial fields. Bottom: Final fields.

In addition to parsing, student catalogers noted whether or not they considered the transparency to have come from a print source based on the slide’s label. Student catalogers were asked to type “yes” in the “image_PrintSource” field only if the label text referenced a printed source, and leave the field blank if the label text did not provide this information.

By the end of the project, copy cataloging resulted in data records for each slide in the sample set. Positional information of specific metadata was not recorded due to the large disparity of the position of the text on the slides.

Examples of different typefaces and specific abbreviations and keywords typical of original transparencies were recorded in a separate document to assist student developers (Figure 2). Also recorded were physical attributes of slides that may impact digital signal processing, including noise such as dust, scratches, and other marks, observed on both transparencies and labels.

Automated Extraction of Slide Label Information

Optical character recognition (OCR) software analyzes the structure of images to determine

patterns and extract printed text. Applying this technology to slides facilitates batch extract of textual data from labels.

At the start of the project, technology was used from the open source package Tesseract, a

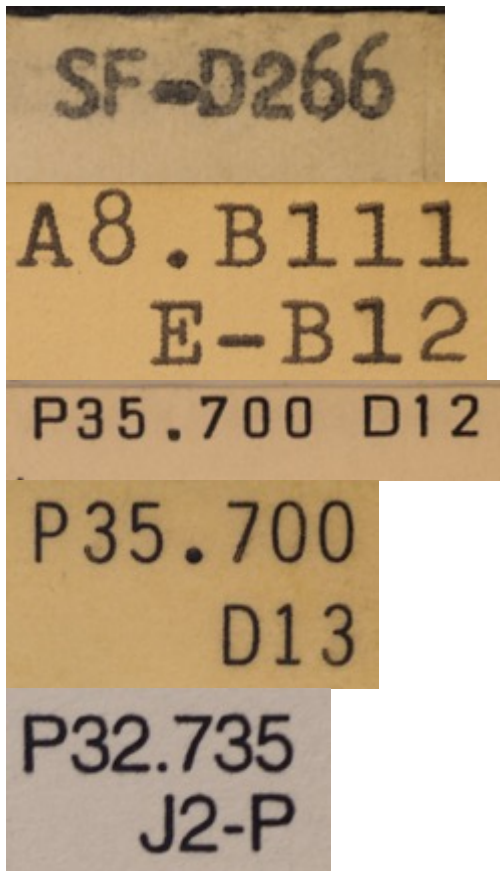


Figure 2: Five different typewriters from slide labels.

popular OCR software currently sponsored by Google.⁹ Tesseract’s documentation specifies certain pre-processing necessary for optimal results, including binarization to minimize variations in lighting, noise removal to increase clarity, and border removal to further increase text readability.¹⁰ Slide labels generally include a large red dot that serves to indicate the slide’s orientation, used to avoid upside-down projection in the classroom. To prevent interference with the automated OCR process, all digital images were preprocessed using a filter to remove red in the images. However, with the pre-processing recommended by Tesseract, results proved uneven. This was likely due to the wide range of typewriters and typefaces used in the slide collection, as seen in Figure 2. Tesseract seemed more accurate in tests using modern typefaces.

An attempt to resolve these weak results by following Dropbox’s OCR pipeline was explored with the goal of retraining Tesseract to more accurately recognize the older typefaces found in the slide library.¹¹

The effort required manually cropping letters, creating mock words using different combinations of these letters and feeding them into a neural net.¹²

Unfortunately, the irregularity of hand typed words from older typefaces still caused poor results. In the end, Tesseract proved too cumbersome to configure in order to obtain the needed level of accuracy. Additionally, the manual training, cleaning, and pre-processing necessary for a workable result would not be feasible in large-scale applications.

⁹ Tesseract-ocr, "Tesseract Open Source OCR Engine (main repository)," GitHub, last modified March 29, 2020, <https://github.com/tesseract-ocr/tesseract>.

¹⁰ Ibid.

¹¹ Brad Neuberg, "Creating a Modern OCR Pipeline Using Computer Vision and Deep Learning," *Dropbox.Tech*, April 12, 2017, <https://blogs.dropbox.com/tech/2017/04/creating-a-modern-ocr-pipeline-using-computer-vision-and-deep-learning/>.

¹² TensorFlow was used. See below for more about TensorFlow.

Original slide - manual transcription	Tesseract with Photoshop binarization	Tesseract with thresholding
 <p>Rodin, Auguste P32.765D Hugo, Victor (1884) Chicago: Art Institute</p> <p>22.8x18 drypt P/Univ Michigan 86/7 set CP #089</p> <p>COLUMBIA U ART HIST. SLIDE COLL.</p>	 <p>Rodin, Auguste P32.765D Hugo, Victor (1884) Chicago: Art Institute</p> <p>22.8x18 drypt P/Univ Michigan 86/7 set CP #089</p> <p>COLUMBIA U ART HIST. SLIDE COLL.</p>	 <p>Rodin, Auguste P32.765D Hugo, Victor (1884) Chicago: Art Institute</p> <p>22.8x18 drypt P/Univ Michigan 86/7 set CP #089</p>
<p>Rodin, Auguste P32.765D Hugo, Victor (1884) Chicago: Art Institute</p> <p>22.8x18 drypt P/Univ Michigan 86/7 set CP #089 COLUMBIA U ART HIST. SLIDE COLL.</p>	<p>ficciiriAug'uste P32.765</p> <p>Hu 0. Victor [1 34] Chicago: Art Institute 22.BXIE drgpt ' P/Univ Mic igan 85/7 set CF #069</p>	<p>ludifi.Augu\$ta PaZInsi' Hugo. Victor an)</p> <p>l éhicago: Art Institute igan 86/7</p> <p>x" Q</p>
<p>Tesseract with Photoshop editing - filters - added blurring for smoothing of noise, increased level of contrast</p>	<p>Tesseract with Photoshop editing - ideal image denoising - added blurring for smoothing of noise, increased level of contrast, hand-cleaned image</p>	<p>Google Cloud Vision API - image cropped out, no preprocessing of the label</p>
 <p>Rodin, Auguste P32.765D Hugo, Victor (1884) Chicago: Art Institute</p> <p>22.8x18 drypt P/Univ Michigan 86/7 set CP #089</p>	 <p>Rodin, Auguste P32.765D Hugo, Victor (1884) Chicago: Art Institute</p> <p>22.8x18 drypt P/Univ Michigan 86/7 set CP #089</p>	 <p>Rodin, Auguste P32.765D Hugo, Victor (1884) Chicago: Art Institute</p> <p>22.8x18 drypt P/Univ Michigan 86/7 set CP #089</p> <p>COLUMBIA U ART HIST. SLIDE COLL.</p>
<p>odln.Augusto "2.7655' :Huaoh Victor</p> <p>éhicngo: Art Institute gan 88/7</p> <p>K?' 9</p>	<p>Rodin.Augu5te P32.7650</p> <p>Hu 0. Victor [1 34] Chicago: Art Institute 22.BXIE dr p</p> <p>.. (349'; lUnzv M1: lgan 85/7 set CF #069</p>	<p>Rodin, Auguste P32.765D Hugo, Victor (1884) Chicago: Art Institute 22.8x18 drypt P/Univ Michigan 86/7 set CP #089 COLUMBIA U ART HIST. SLIDE COLL.</p>

Figure 3: Comparison of different OCR inputs and outputs with the original slide. Below each method and process description is the image input and OCR text output.

The second solution was to test Google Cloud Vision API.¹³ Cloud Vision features OCR with the ability to read multiple languages and handwriting. The Cloud Vision API is open to developers but is not open source. Despite questions of long-term availability, with the Cloud Vision API implemented, text output proved far more accurate even without pre-processing. In Figure 3, the input, output, and pre-processing of Tesseract and Cloud Vision can be compared. For a team with limited technical resources, using Cloud Vision provides results with few spelling errors or unnecessary special characters. In some cases, its OCR proved *too accurate*, outputting both label text and text visible within the transparency.

Classification Based on Slide Label Information

A script was written to utilize OCR output, creating a list of slides containing each keyword, sorted by keyword and by slide record ID. Quality control was conducted using the manually entered cataloging information for each corresponding set of slides. For further analysis and classification, a limited list of keywords derived from label text was used that would definitively indicate the source of the image, such as “plate” or “pg.” In Figure 4, an example of OCR output and keyword extraction can be compared.

Detection of the Presence of Halftone

Many slides in slide libraries are “copywork”; i.e., the slide transparency was photographed from a book, and is thus a low-quality reproduction of an extant printed image. Most of these slides can be identified by the presence of halftone, a translation of photographic tonalities into a continuous series of dots, and a visual artifact of mass production printing processes.¹⁴ If an image contains halftone, it is a reproduction from printed material.

Before a computer can be trained to make a distinction between an image with halftone and an image without halftone, the presence of this printing artifact must be made visually explicit through digital image processing. Digital image processing programs mathematically manipulate images to produce an output of an enhanced or filtered image. Applying a Discrete Fourier Transformation (DFT) produces a pattern that strongly correlates with the presence of halftone in a transparency, examples of which are visible in Figure 5. Using the open source image processing package OpenCV, a DFT was applied to the digitized slides.¹⁵

¹³ Google Cloud, "Detect text in images," Google AI & Machine Learning Products, last modified April 15, 2020, <https://cloud.google.com/vision/docs/ocr>.

¹⁴ Dusan C. Stulik and Art Kaplan, "Halftone," in *The Atlas of Analytical Signatures of Photographic Processes* (Los Angeles: Getty Conservation Institute, 2013), 5, http://hdl.handle.net/10020/gci_pubs/atlas_analytical.

¹⁵ OpenCV, accessed December 19, 2018, <https://opencv.org/>; Yun-Fu Liu, Jing-Ming Guo, and Jiann-Der Lee, "Halftone Image Classification Using LMS Algorithm and Naive Bayes," *IEEE Transactions on Image Processing* 20, no. 10 (2011), doi:10.1109/tip.2011.2136354.



Copywork slide	Original slide
	
OCR output	OCR output
<p>TISCHBEIN, J.H.W. IDYLL; 20. Girls' Heads in Red Sky; 1817-20 P33.860</p> <p>Oldenburg: Landesmuseum; o/p; 27x33cm Schulze, ed., Goethe und die Kunst (1994), p.372, no.250 AH_0203.21% UNIVERSITY DEPT. OF ART HISTORY</p>	<p>BAALBEK TEMPLE COMPLEX A8.B111 B</p> <p>K.LUNDE FOTO</p>
Keywords extracted - indicate slide is copywork	Keywords extracted - indicate slide is original
<p>ed. no. p.</p>	<p>foto</p>

Figure 4: Comparison of OCR output and keyword extraction for a copywork slide and an original slide.

Classification Based on Halftone

An image classifier was used with the open source deep learning framework TensorFlow to recognize the DFT patterns.¹⁶ In order to train the classifier, training sets were created from the slides that were preliminarily digitized by work-study students. Slides were batch cropped to isolate transparencies, removing the outer part of the slide with the label. Due to their large file size, these images needed to be downsized before testing for optimal results.

First tests led to inaccurate results, with the program outputting results indicating that the images were entirely all halftone or all non-halftone on training sets we knew to be diverse. Experimentation with different thresholding levels on the DFT seemed to have a positive effect. However, tests were still showing an improbably high percentage of halftone in folders, and concerns were raised about the diversity of training images as the majority presented halftone. To counterbalance this uniformity, images derived from the Media Center's original fieldwork photography were added to the training sets to positive effect.

¹⁶ TensorFlow, accessed December 19, 2018, <https://www.tensorflow.org/>.



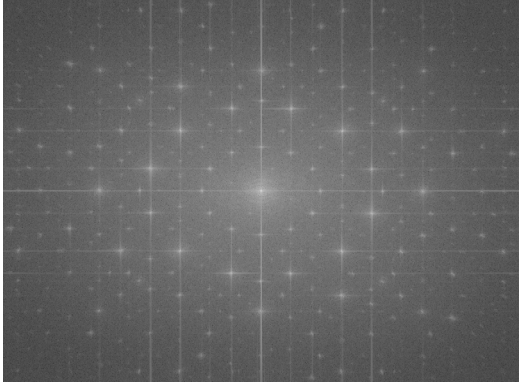
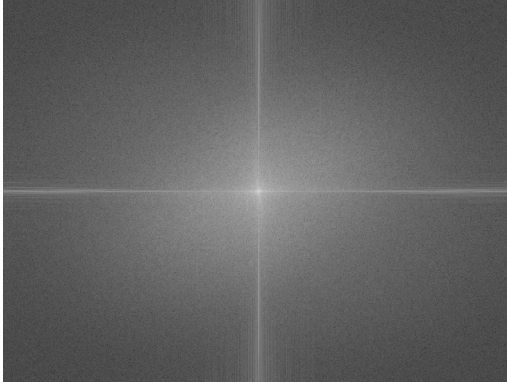
Detail of copywork transparency <i>- halftone visible</i>	Detail of original transparency <i>- no halftone visible</i>
	
DFT <i>- sparkles indicate halftone</i>	DFT <i>- lack of sparkles indicate original image</i>
	
TensorFlow output	TensorFlow output
Halftone; probability 0.999348	Nonhalftone; probability 0.528805

Figure 5: Comparison of DFT output for a transparency containing halftone and an original transparency. Visible sparkles in the digitally processed output indicate to the image classifier TensorFlow whether halftone is present.

Results

The project workflow needed to include manual identification of copywork slides to create a rigorously analyzed control group. To verify and quantify TensorFlow results, a careful manual analysis of slide transparencies was necessary. Visually searching for the presence of halftone yielded tangible results to measure the success of the image classifier. Student catalogers were assigned to manually analyze the original transparencies to look for the visual presence of halftone patterns. This data provided a reference set of information for comparison against the results of automation.

In order to quantify the OCR keyword search results, data from the “image_PrintSource” field was extracted from copy cataloging. This field was where student catalogers noted whether they considered the transparency to have come from a print source, based on the slide’s label text. For testing, a limited keyword search was implemented to target slides with a print source by searching for terms such as “pg.” or “fig.” Copy cataloging data was used for comparison against the results of the automated keyword search.

This comparison of manual and automated classification produced true positive, true negative, false positive, and false negative results. Each slide was assigned one of these markers, both for OCR testing and halftone detection. The definition for each marker can be seen in Figure 6, while Figure 7 shows the percentage of each marker in the entire testing set of 7,815 slides.

		Automated Analysis	
		Slide marked as copywork or as halftone	Slide not marked as copywork or halftone
Manual Analysis	Slide marked as copywork or as halftone	True Positive	False Negative
	Slide not marked as copywork or halftone	False Positive	True Negative

Figure 6: The comparison for each marker used. OCR keyword search returned a copywork or non-copywork result, while the image classifier returned a halftone or non-halftone result.

	True Positive	False Positive	True Negative	False Negative
Halftone Detection	48.2%	29.8%	12.4%	9.6%
OCR Keyword Search	41.2%	8.5%	27.9%	22.3%

Figure 7: Comparison of results from both tests.¹⁷

Automated classification takes significantly less time than manual analysis. To complete manual halftone analysis of all 7,815 slides, the effort took student assistants over 65 hours. Automated halftone classification of the same sample set can be completed in a single 3-hour session.¹⁸

Testing also quantifies how accurate the programs are overall, by calculating the sensitivity and specificity of each test. Sensitivity quantifies the percentage of actual positive results correctly identified by the test, while specificity quantifies the percentage of actual negatives correctly

¹⁷ Values rounded to 1 decimal.

¹⁸ Data from Media Center for Art History internal time tracking.

identified. Sensitivity and specificity are inherently linked, as both figures are vital to the health of the diagnostic test. Higher percentages are preferable (Figure 8).¹⁹

	Sensitivity	Specificity
Halftone Detection	83.4%	29.3%
OCR Keyword Search	64.9%	76.6%

Figure 8: Comparison of sensitivity and specificity from both tests.

Automated halftone classification is able to correctly identify 83.4% of all slide transparencies with halftone, while OCR keyword search is able to correctly identify 76.6% of all slide transparencies that do not originate from a printed source. Given this level of accuracy and the significant difference in speed between human and machine-assisted classification, it is clear that automated image classification is an effective way to reduce manual processing time and more quickly identify important images.

Issues and Inconsistencies

In some cases, transparencies with images representing mosaics or detail views of texture in paintings generated a false positive for the presence of halftone. Patterns visible on the slide would be mistaken for halftone, creating a false positive, as in the transparency shown in Figure 9. Likewise, large areas of tape masking, creating a makeshift crop of the transparency, may interfere with halftone detection and cause a false negative result.

An additional source of false negatives may be due to printing processes that do not employ the use of halftone. Some folders had a high quantity of black and white diagrams - images that we knew to be copywork slides, but that did not show visible halftone. These images therefore appeared to the classifier to be original images but were tagged as having a print source by student catalogers.

When working with 35mm slides, human error may be inevitable. Though the OCR and keyword search have a high level of accuracy, slide labels may have typos, illegible handwritten labels, or unusual terminology. These outliers will be missed by the keyword search, pointing to the need for a comprehensive and evolving list of terms. Student catalogers can help by noting discrepancies on slides. However exhaustive the keyword list may be, there is little to be done for slides with insufficient or inaccurate information, though we found these to be few and far between.

While cataloging the text on slide labels is often self-evident, it is possible that minor inconsistencies in data exist due to a difference in student employees and their individual interpretation of the labels. In an extended version of this project dependent on student assistants, turnover may be unavoidable and slight inconsistencies expected.

Development of Future Workflow

¹⁹ "Sensitivity and Specificity," Wikipedia, the Free Encyclopedia, last modified January 19, 2019, https://en.wikipedia.org/wiki/Sensitivity_and_specificity.

In looking towards the eventual digitizing and cataloging of the Department of Art History and Archaeology's entire 35mm slide collection, which totals over 400,000 slides, combining both measures of detection to utilize their individual strengths will streamline and reduce the time necessary to carry out the labor-intensive sorting process. Based on the sample set, about 60% of the collection can be expected to be made up of copywork slides. The remaining 40% of the collection, about 160,000 slides, is expected to be non-copywork slides made from original department photos or purchased from slide vendors. The following workflow is considered for the next stage of the project.

Slide digitization by the Media Center will continue using a photostand, DSLR, and light table as in the initial stages of this project. By first applying OCR to all digitized images, a large number of slides indicating a print source are effectively and automatically set aside by implementing a limited keyword search to target images with keywords such as "pg." or "fig." Next, the remaining slides are run through the halftone classifier. Images identified as non-halftone are most likely to be original photos taken and donated by a member of the Columbia Art History Department or photographs bought by the department from a slide vendor.

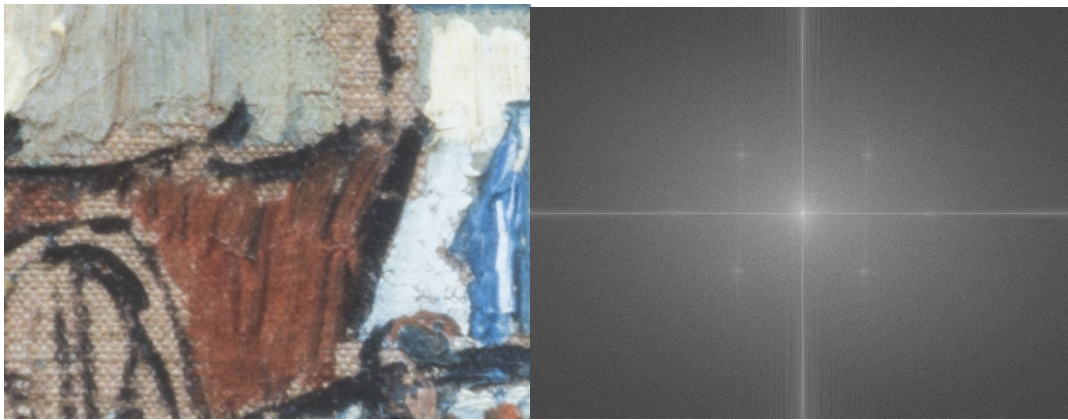


Figure 9: Left: detail of transparency without halftone. Canvas in original painting visible in a detail view of a slide. Right: DFT output for the slide. Note the presence of sparkles indicative of halftone, despite that there is no halftone present in the transparency.

From these results, the non-halftone slides are run through a secondary keyword search with a more expansive list, searching for known slide vendor names, such as "Saskia," "Alinari," and "Sandak." This second keyword search further pares down the results in favor of faculty and student contributions to the collection. Remaining images identified by the program as most likely to be original will be prioritized for manual verification and cataloging by MCAH staff, and added to the Media Center Image Database (MCID).²⁰ Throughout each process, the status of each image is logged in a CSV to allow for analysis and system improvements.

²⁰ Media Center for Art History, Columbia University, Media Center Image Database, accessed January 25, 2019, <https://mcid.mcah.columbia.edu/>.

With this workflow in place, the Media Center hopes to quickly and accurately sift through the entirety of the slide collection, seeking to first prioritize original images that may be made available to the public through MCID.

The composition of all image libraries is different, and these tools were built to allow a certain amount of customization, enabling libraries and other photograph collections to create individualized tools. The Media Center developed this workflow seeking original photography with our own 35mm slide library in mind. However, these tools can easily be adapted to fit many different queries. The Git repository and implementation documentation have been made available online.²¹

Conclusion

When used together as detailed above, both OCR and halftone detection processes complement each other to work efficiently in reducing the amount of human sorting necessary. In a workflow akin to the Media Center's, where everything will eventually be cataloged, the ability to effectively automate the process of sorting and prioritization is invaluable.

This project sought to employ the use of open source software in order to maximize its flexibility and reproducibility. All technologies used in the final iteration of this project, aside from the Google Cloud Vision API, are open source. The Media Center found the expense of the Cloud Vision API to be worth the vast improvement in accuracy and ease of use when compared to Tesseract (Figure 3).

This experiment was designed as a way for the Media Center to easily focus its efforts and to efficiently sort through a largely unknown collection. In some slide libraries, methodologies and prioritization for sorting and digitizing are defined solely through an individual's knowledge of the slide collection.²² The Media Center sought to reduce historical reliance on institutional memory, and instead define methodologies for qualitative slide analysis. This refined filtering mechanism allows slide libraries, when faced with culling or digitizing a slide collection, to effectively review the entire collection, making decisions easier and more logical. Automating the sorting process avoids the risk of dismissing large portions of a collection by memory alone.

²¹ Media Center for Art History, Columbia University, "Reverse Engineering the Image Library," accessed April 16, 2020, <https://learn.columbia.edu/imls>.

²² Karen A. Bouchard, "Now, Slides, Sail Thou Forth to Seek and Find."

Bibliography

- ArtPI. The Art API. Accessed March 1, 2019. <https://www.artpi.co/>.
- Bouchard, Karen A. "Now, Slides, Sail Thou Forth to Seek and Find:' Facilitating a Slide and Photograph Diaspora." *VR*A* Bulletin* 41, no. 2 (2015).
<https://online.vraweb.org/vrab/vol41/iss2/10/>.
- Computer Vision group, Heidelberg University. "Visual Search Tool." Digital Humanities. Last modified 2018. <https://sabelang254.wixsite.com/digital-humanities/visualesearchtool>.
- The Frick Collection. "Photoarchive." Accessed December 18, 2018.
<https://www.frick.org/research/photoarchive>.
- Getty Research Institute. "Getty Union List of Artist Names." The J. Paul Getty Trust. Last modified March 7, 2017.
<http://www.getty.edu/research/tools/vocabularies/ulan/index.html>.
- Google Cloud. "Detect text in images." Google AI & Machine Learning Products. Last modified April 15, 2020. <https://cloud.google.com/vision/docs/ocr>.
- 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, no. 10 (2011), 2837-2847. doi:10.1109/tip.2011.2136354.
- Media Center for Art History, Columbia University. Media Center Image Database. Accessed January 25, 2019. <https://mcid.mcah.columbia.edu/>.
- Media Center for Art History, Columbia University. "Reverse Engineering the Image Library." Accessed April 16, 2020. <https://learn.columbia.edu/imls>.
- Neuberg, Brad. "Creating a Modern OCR Pipeline Using Computer Vision and Deep Learning." *Dropbox.Tech* (blog). April 12, 2017.
<https://blogs.dropbox.com/tech/2017/04/creating-a-modern-ocr-pipeline-using-computer-vision-and-deep-learning/>.
- North Carolina State University. "Image Analytics Processes." Nineteenth-Century Newspaper Analytics. Accessed December 18, 2018.
<https://ncna.dh.chass.ncsu.edu/imageanalytics/techniques.php>.
- OpenCV. Accessed December 19, 2018. <https://opencv.org/>.

PHAROS: The International Consortium of Photo Archives. Accessed December 18, 2018. <http://pharosartresearch.org/>.

Resig, John. "Using Computer Vision to Increase the Research Potential of Photo Archives." *Journal of Digital Humanities* 3, no. 2 (Summer 2014). <http://journalofdigitalhumanities.org/3-2/using-computer-vision-to-increase-the-research-potential-of-photo-archives-by-john-resig/>.

"Sensitivity and Specificity." Wikipedia, the Free Encyclopedia. Last modified January 19, 2019. https://en.wikipedia.org/wiki/Sensitivity_and_specificity.

Slide and Transitional Media Task Force. "'Tell Us Where Your Slides Are!' Summary of Survey conducted in Fall 2014." Visual Resources Association, 2015. <http://vraweb.org/survey-summary-from-slide-and-transitional-media-task-force/>.

Smith, Abby. *Why Digitize?*. Washington D.C.: Council on Library & Information Resources, 1999.

Stulik, Dusan C., and Art Kaplan. "Halftone." In *The Atlas of Analytical Signatures of Photographic Processes*. Los Angeles: Getty Conservation Institute, 2013. http://hdl.handle.net/10020/gci_pubs/atlas_analytical.

TensorFlow. Accessed December 19, 2018. <https://www.tensorflow.org/>.

tesseract-ocr. "Tesseract Open Source OCR Engine (main repository)." GitHub. Last modified March 29, 2020. <https://github.com/tesseract-ocr/tesseract>.

University of Nebraska-Lincoln. Image Analysis for Archival Discovery (Aida). Accessed December 18, 2018. <http://projectaida.org/>.

"VRA CORE - a Data Standard for the Description of Works of Visual Culture." Library of Congress. Last modified February 15, 2018. <https://www.loc.gov/standards/vracore/>.

zdenop. "Qt-box-editor." GitHub. Last modified July 14, 2019. <https://github.com/zdenop/qt-box-editor>.