New Techniques for the Digitization of Art Historical Photographic Archives - the Case of the Cini Foundation in Venice

Benoit Seguin, Lisandra Costiner, Isabella di Lenardo, Frédéric Kaplan; DHLAB, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Abstract

Numerous libraries and museums hold large art historical photographic collections, numbering millions of images. Because of their non-standard format, these collections pose special challenges for digitization. This paper address these difficulties by proposing new techniques developed for the digitization of the photographic archive of the Cini Foundation. This included the creation of a custom-built circular, rotating scanner. The resulting digital images were then automatically indexed, while artificial intelligence techniques were employed in information extraction. Combined, these tools vastly sped processes which were traditionally undertaken manually, paving the way for new ways of exploring the collections.

Introduction

Museums and libraries hold large photographic archives to aid in art historical study, record changes in attribution and conditions of paintings, as well as document fragile architectural and monumental artwork. The Frick Museum in New York alone holds 1.2 million such images, while together, the fourteen largest photographic libraries house an estimated 31 million images [1]. In recent years, numerous initiatives have been launched to digitize this heritage, with the aim of ensuring preservation and fostering access [2]. Despite such efforts, problems persist in finding suitable methods for digitizing extensive collections of non-standard material both speedily and accurately. At the same time, there is a question of how to present the newly digitized information in a manner that would both aid current research and foster new scholarship.

This paper focuses on the case study of the digitization of the photographic archive of the Cini Foundation in Venice. It presents new technologies developed for the digitization, indexation and search of such images, presenting one approach of solving this complex problem. This research is part of the Replica project – a joint collaboration between the Cini Foundation and the DHLAB of the École Polytechnique Fédérale de Lausanne (EPFL).

Located on the island of San Giorgio Maggiore in Venice, the Cini Foundation houses roughly one million images, an especially rich repository of Venetian artistic and visual culture. The collection is divided into two sections: (1) a library comprising photographs mounted on cardboard cards with accompanying descriptive information, arranged thematically, and (2) the archives of a number of notable art historians. The first part of the archive has a standard format, namely of images mounted on cards of the same size, while the second varies widely. It can comprise of documents of different shapes, sizes, and include texts as well as images, single or double-sided.

This paper will focus solely on the digitization of the library and its cardboard cards, comprising roughly 330,000 documents.

The techniques developed in the process, and the challenges surpassed, however, can be generalized to the entire collection. They center on: (1) how to speedily and accurately digitize atypical materials found in a photographic archive, and (2) how to produce new digital methods for the automatic processing of this data - in particular, the extraction of textual and visual information from the raw scans.

New Digitization Tools

Given that the archival material came in a variety of sizes and contained visual as well as textual information on both sides of the page, a scanner was sought which could handle such constraints. Speed was equally important in the processing of the large collection. Another requirement was that the digitized file recorded the document's precise archival location. Finally, a design was sought that would minimize the operator isolation which occurs when undertaking the prolonged repetitive task of digitization over the course of months, ensuring both a more enjoyable work experience and heightened productivity.

To tackle these challenges, new scanning technologies were explored. Industry standards such as flatbed scanners, or the conveyor-belt scanner in use by the Smithsonian Museum [3], were unable to handle the complexity of the Cini material and its double-sided nature. Instead, this photographic material demanded a specialized tool, leading to the design of a custom-made scanner by Factum Arte (Madrid, Spain) in consultation with the DHLAB of EPFL.

The scanner was devised as a table with a circular, rotating top (diameter of 2 m) which comprised four image plates (Figure 1). Document sizes of up 594 x 420 mm, or A2 format, could be accommodated. The rotating top was controlled by a precision motor with variable speed enabling uninterrupted digitization of 1 image every 4 seconds. It was operated by a team of two people, one of whom placed the images on one of the glass plates, at the same time as digitization occurred on a second plate, and as a second operator removed the scanned images from a third plate. A sensor system would calculate the position and detect when a document was placed on the glass surface. Cameras mounted above and below the table simultaneously captured the recto and verso of each document placed on the glass plates. Flash units were designed and engineered by Factum Arte to provide the lowest level of light for the achievement of a high quality image while minimizing glare. Finally, the hardware consisted of two cameras connected to two controllers, which in turn led to a server.

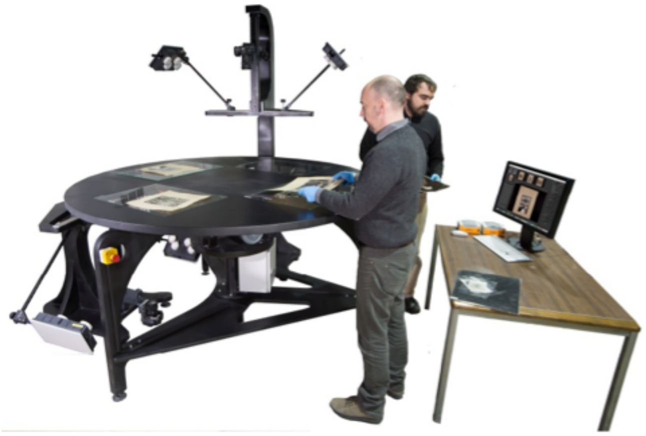

***Figure 1**. The Replica 360 scanner setup (Photograph: Factum Arte, 2016 [4])*

Two people operated the machine ensuring that work could be mutually checked and problems tackled collectively. The team aspect also fostered social interaction, minimizing operator isolation, enhancing work experience during the sustained repetitive task, and ensuring steady productivity.

In preparation for the scanning, bar-codes were created to record the archival position of each document, and were affixed to the verso of each cardboard card. To speed up the workflow, an xml file was also created. This automatically named files as they were being digitized with the title of the collection and the document's position within this, ensuring that scanning progressed continuously without the need to stop and name files individually. For each document in the archive, four files were created: one for the digitization of the recto side, one for the verso, and one corresponding md5 signature file for each of these to verify authenticity and integrity. Accuracy and integrity of files could likewise be checked by comparing the automatically generated name with the bar-code.

In the case of the Cini Foundation, the design of the scanner allowed for the digitization of 1500 images per day, with the digitization of the entire photographic archive of 330,000 images completed in roughly 18 months.

Image Processing Pipeline

The scanner output resulted in high-resolution raw images (of 400 ppi and 5424 x 3616 pixels). As the scanning of the material took place on-site at the Cini Foundation in Venice, and the data was processed by the DHLAB of EPFL, Switzerland, transferring the data posed a challenge in itself. The 50 terabytes of RAW files, were converted to JPEG images of 90% quality. Only the scans of the document rectos were transmitted, reducing the size to 2.6 terabytes.

Figure 2 illustrates a typical scan of the recto side of a document. Each of the Cini scans contained the reproduced artwork pasted on a cardboard card. Metadata information about the artwork is typed at the top of the card. The scan likewise includes a black border surrounding the image, as well as a color control bar. To process the image, the relevant information had to be extracted. This included, cropping the art reproduction from the scan. Secondly, the structured metadata, written at the top of each cardboard, had to be extracted from the image and rendered into text, to enable a future textual search of this information. Because thousands of scans awaited processing, devising an automatic approach to complete these tasks was essential.

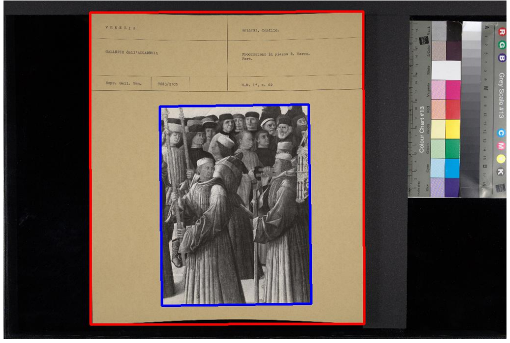

***Figure 2**. Scanned image. The red and blue rectangles denote the areas to be extracted – respectively, the cardboard card, and the art reproduction.*

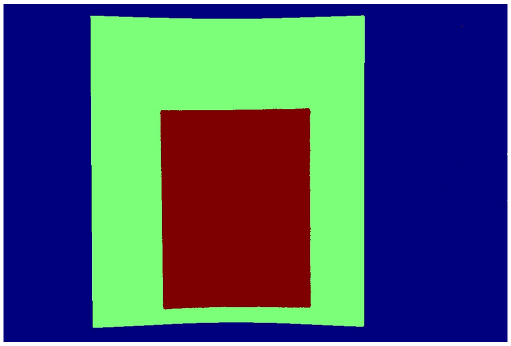

***Figure 3**. Predicted mask from the corresponding scanned image in Figure 1. The annotated training mask is extremely similar to this.*

Image segmentation

Segmenting the different areas of interest proved complicated. Within each scanned image, the cardboard could appear in different positions, and orientations, at times being rotated up to 90°. Aging and humidity also affected some documents, deforming the cardboard support so that it no longer appeared rectangular. Difficulties were compounded by inconsistent scanning practices in the early days of digitization, which likewise produced non-standard layouts with the color control bar at times overlapping the area of the cardboard. Upon these cardboard supports, the art reproductions also varied in position, shape and orientation. At times, these images filled the entire area of the cardboard obscuring the metadata. Their colors and textures likewise differed. Some monochrome photographs even resembled in color their cardboard backing. This variability posed challenges for automatic extraction, as there was no consistent digital marker across images.

Given this variety, standard techniques for image segmentation that rely on only one visual cue were inadequate. Color and texture were not sufficiently discriminant. The detection of lines which were predominantly straight and the creation of a rectangle out of these, was the technique that appeared most promising, but it too failed in yielding perfect accuracy.

The approach which produced the best results was based upon artificial intelligence, namely deep learning and convolutional neural networks (CNN). More specifically, we used a specific form of CNN called a fully-convolutional neural network (FCN). These networks took as input an image and returned as output an image of the same size where each pixel contained a number of class probabilities. In the past, these techniques proved effective for pixel-wise predictive tasks, and they currently underlie the best performing systems for semantic and instance segmentation [5]. FCNs have recently been introduced successfully to tackle problems of document processing, especially historical document processing. The model used in the course of this project is based upon a U-net architecture [6] where most parameters derive from a pre-trained version of the Resnet-50 architecture [7]. Details of the architecture of this system and how it was trained rest outside the scope of this paper and are described elsewhere [8].

In short, to process the Cini documents, a FCN was trained to predict for each pixel whether it belonged to the background, cardboard backing, or the art reproduction. Training the system required a corpus of correctly annotated images. These were produced by taking a number of scanned images and drawing upon them with an image editing software. Here, the background was painted one solid color, the cardboard area another color, and the art reproduction area with a third color (as in Figure 3). Through this process, a user annotated 120 training images in the course of 2 hours.

The trained FCN generated an image along with the probability of each pixel of being part of the background, the art reproduction, or the cardboard. The highest probability was taken, yielding a prediction image, where each pixel was assigned to one of the three classes (Figure 3).

From this prediction image, the smallest rectangle that would enclose all the pixels of a given class was extracted. Additional cleaning, based on morphological operations, further removed small artifacts, or outlying pixels, that arose in the course of the prediction. Additional logic was used to improve the results in specific cases. For instance, at times, the FCN confused black non-textured areas of the art reproduction with the background of the scan. Imposing the rule that the art reproduction area has to be located within the space of the extracted cardboard corrected this issue.

Reading the metadata

The next step in image processing centered on extracting the textual metadata that was printed at the top of each cardboard card. Fortunately, the large majority of this metadata was typed and not handwritten, enabling optical character recognition. For this task, standard optical character recognition (OCR) libraries were tested, including the open-source Tesseract and the commercial GoogleVision [9]. The latter performed much better than the former, so GoogleVision was used in this study.

An OCR algorithm provided bounding boxes around each word and its respective transcription. These words were then clustered together into small paragraphs so that each paragraph accounted for a single metadata entry.

Next, each paragraph had to be assigned the correct metadata label (author, description, city, etc.). Here too the metadata layout of the Cini cardboard cards proved inconsistent. Although the same categories were typically relegated to roughly the same areas of the cardboard, the spaces allocated to these fields differed in height. Furthermore, the lines denoting the separation between fields were not always straight because of the bending of the cardboard, rendering this structure unreliable for establishing divisions.

To account for these inconsistencies, the assignment of metadata labels relied on the position of the extracted paragraphs. Given a layout model (which is a collection of pairs <position, label>), we found the best label for each paragraph by looking at the distances between the paragraph's centroid and its position within the layout. Multiple layout models (encoding the variability of the headers) were also tested, picking the one that best matched the extracted paragraphs (based on the sum of the distances between the paragraphs' centroids and their respective assigned position in the layout).

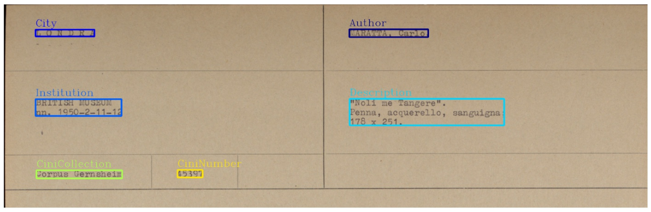

***Figure 4**. Close-up of a simple metadata header, with the detected text and the corresponding metadata label assigned plotted in color on top.*

Evaluation

To test the efficiently of the processing pipeline and its chosen tools, the quality of the results was evaluated. In particular, we evaluated: (1) the extraction of the visual elements, (2) the detection of the text, (3) the label assignment, and (4) the OCR quality.

For the extraction of the cardboard cards and the art reproductions, we annotated 120 scans that were used for training the FCN (100 training + 20 validation), and then tested the performance on 150 additional examples. The metric used to evaluate the bounding rectangle and its correspondence to the ground-truth was the intersection-over-union (IoU) of these areas. As the results show (Table 1), near perfect extraction accuracy was achieved.

In terms of text detection and the evaluation of the OCR quality, we manually examined 412 random cardboard cards. For this step, the biggest problem was created by the archival practice of manually correcting the typed metadata with handwritten annotations. Unsurprisingly, in these cases, the OCR rarely transcribed the handwritten text. Two types of error were predominant in this case: ones where the cardboard omitted information in one category (i.e. where a particular field was left blank), or where information was partially missing. For instance, the OCR might miss the second line of the author field which contained additional information, or omit the handwritten annotation which was a precision and not a full correction. Results are shown in Table 2.

For the label assignment, we evaluated 337 cardboard cards where all the text elements were properly detected, corresponding to 2028 metadata fields. The only issues recorded were 5 cases of fields being merged together and 3 cases of one field being separated into two.

For the OCR quality, we manually corrected the transcription of the author and description fields (which are the most important for our project), in a set of 150 cardboard cards. The results are displayed in Table 3, and show a good accuracy for the OCR algorithm. Most of the fields are perfectly transcribed with at most a 1-character error, which makes the transcriptions good-enough for undertaking 'fuzzy searches', and allows further automatic correction of these errors with an artist name dictionary, such as the Getty's Union List of Artist Names [10]. Still, some challenges remain, including the detection and reading of handwritten annotations, and the flagging of ambiguous cases that require manual review.

	Cardboard mean-IoU	*Photograph mean-IoU*	*Photograph IoU>95%*
-Base predictions	99.3%	98.2%	96.7%
-Cleaned predictions	99.3%	98.2%	97.3%
-Additional logic	99.3%	99.0%	100.0%

***Table 1**. Results of the quality of the image extraction across 150 randomly chosen scans. The second row correspond to simple post-processing cleaning operations of the predictions (morphological operations). The last row corresponds to adding the additional logic of the photograph being inside the cardboard.*

Metadata Field	*Completely missing*	*Partially missing*	***Total***
Author	1.7%	1.7%	**3.4%**
Description	0.5%	2.4%	**2.9%**
Institution	0.2%	1.0%	**1.2%**
City	1.7%	0.7%	**2.4%**

***Table 2**. Text detection error-rate for some fields, evaluated on 412 random scans.*

Metadata Field	*Perfect*	*<=1 error*	*Perfect* *(normalized)*	*<=1 error* *(normalized)*
Author	77.3%	96.62%	83.8%	97.3%
Description	77.3%	93.92%	85.8%	96.6%

***Table 3**. Quality of the OCR transcription for 150 cardboard cards where text is detected. The average character-error-rate is 2.0% and 1.4% for the author field and the description field respectively. The "normalized" column indicates that we do not consider lowercase/uppercase errors, nor blank spaces/punctuation.*

Conclusion

This study details the digitization of the photographic archive of the Cini Foundation introducing techniques that address the challenges posed by complex archival materials. It proposes new tools for the rapid digitization of photographic archives and details digital techniques for image segmentation and information extraction. As cultural institutions have been undertaking much of this work manually, these techniques are poised to vastly speed the process of digitizing art historical collections, while at the same time facilitating access and assisting research.

The final results of the image extraction and metadata reading techniques outlined here display a high accuracy which renders this extracted information immediately usable. In fact, these images and their metadata are already being employed as the basis for a new search engine, Replica. This tool enables the search of similar images, or of details within images, further showcasing the usefulness of new digital tools in advancing art historical research [11] [12].

Acknowledgments

We would like to thank the Cini Foundation, in particular the Secretary General, Pasquale Gagliardi, and all of the staff and collaborators who oversaw and participated in the digitization process. We are also grateful to Factum Arte, and in particular to Adam Lowe and his team for the development and creation of the scanner. This research forms part of the Venice Time Machine project, generously supported by the Fondation Lombard Odier.

References

[1] T. Loos, "'Photo Archives Are Sleeping Beauties.' Pharos Is Their Prince." *New York Times*, 14 Mar. 2017, https://www.nytimes.com/2017/03/14/arts/design/art-history-digital-archive-museums-pharos.html.

[2] Pharos: The International Consortium of Photo Archives, http://pharosartresearch.org/.

[3] S. Overly, "The Smithsonian turned to conveyor belts, cameras to digitize its many artifacts." *The Washington Post,* 25 Jan. 2015.

[4] Factum Arte, "Press Release: Replica 360 Recto/Verso Scanner," 2016, http://www.factum-arte.com/pag/757/Replica-360-Recto-Verso-Scanner.

[5] K. He, et al, "Mask R-CNN", in ICCV, Venice, Italy, 2017.

[6] O. Ronneberger, et al, "U-Net: Convolutional Networks for Biomedical Image Segmentation", in MICCAI, 2015.

[7] K. He, et al, "Deep Residual Learning for Image Recognition", in CVPR, 2016.

[8] B. Seguin, S. Ares Oliveira, et al, "A generic segmentation approach for baseline detection, document layout analysis and object extraction in historical documents", 2018.

[9] GoogleVision, https://cloud.google.com/vision.

[10] B., Seguin, et al. "Extracting and Aligning Artist Names in Digitized Art Historical Archives", in Digital Humanities Conference, Mexico City, Mexico, 2018.

[11] B. Seguin, et al. "Visual Link Retrieval in a Database of Paintings", in *ECCV Workshops*, pp. 753–767 DOI: 10.1007/978-3-319-46604-0 52, 2016.

[12] I. Di Lenardo, et al. "Visual Patterns Discovery in Large Databases of Paintings", in Digital Humanities Conference, Krakow, Poland, 2016.

Author Biography

Benoit Seguin is a PhD Candidate at the Digital Humanities Laboratory at the École Polytechnique Fédérale de Lausanne. His research focuses on employing modern computer vision techniques on large visual databases of art, leading him to develop new algorithms for the digitization and exploration of these photographic collections. He received a Diplôme d'Ingénieur from École Polytechnique Paristech and a Master of Science in computer science from the École Polytechnique Fédérale de Lausanne.

Lisandra S. Costiner is a Post-Doctoral Fellow in the Digital Humanities Laboratory of the École Polytechnique Fédérale de Lausanne. Her doctoral dissertation (University of Oxford, 2017) focuses on Italian medieval and Renaissance visual and devotional culture. Before turning to art history, she pursued a Bachelor's degree in visual and environmental studies at Harvard University.

Isabella di Lenardo, holds a doctorate in Theories and Art History, and is a Researcher in Digital Humanities at the École Polytechnique Fédérale de Lausanne. Her interests center on the production and circulation of artistic knowledge in the XVIth-XVIIIth centuries. She collaborates with numerous institutions in the fields of digitization and preservation of Cultural Heritage.

Frédéric Kaplan holds the Digital Humanities Chair at École Polytechnique Fédérale de Lausanne and directs the Digital Humanities Lab, leading projects that combine archive digitization, information modelling and museographic design. With his team, he is developing the Venice Time Machine, a project to model the evolution and history of Venice over a thousand years. Frédéric holds a PhD in Artificial Intelligence from University Paris VI, and previously worked as researcher at Sony Computer Science Laboratory.