Making large art historical photo archives searchable

THÈSE N^{O} 8857 (2018)

PRÉSENTÉE LE 23 NOVEMBRE 2018
AU COLLÈGE DES HUMANITÉS
LABORATOIRE D'HUMANITÉS DIGITALES
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Benoît Laurent Auguste SEGUIN

acceptée sur proposition du jury:

Prof. D. Gatica-Perez, président du jury
Prof. F. Kaplan, Dr I. di Lenardo, directeurs de thèse
Prof. O. Chum, rapporteur
Prof. B. Sherman, rapporteur
Prof. S. Süsstrunk, rapporteuse

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

To my family…

Remerciements

Avant tout, je voudrais remercier mes superviseurs de thèse, Frédéric Kaplan et Isabella di Lenardo, pour m'avoir permis de prendre part à ce très beau projet, et pour m'avoir soutenu avec bienveillance tout au long de ce travail de doctorat.

Si ces années ont été un plaisir de travail et de recherche, c'est aussi grâce aux personnes que j'ai eu la chance de côtoyer au sein du laboratoire des Humanités Digitales. Une pensée chaleureuse donc à Maud, Sofia, Orlin, Matteo, Giovanni, Vincent, Cyril, Nils, Dario, Bastien, Yannick et notre incomparable secrétaire Alicia.

Au cours de l'avancement de ce projet, j'ai pu aussi collaborer avec des gens aussi divers que passionants. Le résultat final n'aurait pu être ce qu'il est sans le concours de Carlotta, Lia, ainsi que le duo Andrea et Remko de la Cini.

Finalement, cette aventure n'aurait pu avoir eu lieu sans le soutien de la Fondation Cini, qui nous a permis de travailler sur sa fantastique collection photographique. Un grand merci à tous les opérateurs qui ont passé de longues et nombreuses heures à scanner les quelques 300'000 documents qui ont servi de base à ce travail.

Sur une note plus personnelle, je voudrais remercier tous mes amis, anciens ou nouveaux, musicaux ou scientifiques, Lausannois ou plus lointains, qui m'ont aidé à me construire et à façonner qui je suis aujourd'hui.

Enfin, je voudrais dédier ce travail à mes parents et à ma famille, qui m'ont soutenu à tout moment de ma vie. Je suis conscient chaque jour un peu plus de la chance que j'ai eu d'avoir grandi dans un tel environement, merci pout tout...

Lausanne, 19 Octobre 2018 B. S.

Abstract

In recent years, museums, archives and other cultural institutions have initiated important programs to digitize their collections. Millions of artefacts (paintings, engravings, drawings, ancient photographs) are now represented in digital photographic format. Furthermore, through progress in standardization, a growing portion of these images are now available online, in an easily accessible manner. This thesis studies how such large-scale art history collection can be made *searchable* using new deep learning approaches for processing and comparing images. It takes as a case study the processing of the photo archive of the Foundation Giorgio Cini, where more than 300'000 images have been digitized. We demonstrate how a generic processing pipeline can reliably extract the visual and textual content of scanned images, opening up ways to efficiently digitize large photo-collections. Then, by leveraging an annotated graph of visual connections, a metric is learnt that allows clustering and searching through artwork reproductions independently of their medium, effectively solving a difficult problem of cross-domain image search. Finally, the thesis studies how a complex Web Interface allows users to perform different searches based on this metric. We also evaluate the process by which users can annotate elements of interest during their navigation to be added to the database, allowing the system to be trained further and give better results. By documenting a complete approach on how to go from a physical photo-archive to a state-of-the-art navigation system, this thesis paves the way for a global search engine across the world's photo archives.

Keywords: digitization ; large image collections ; visual search ; deep learning ; computer vision ; archives ; art history.

Résumé

Au cours des dernières années, musées, archives et autres institutions culturelles ont lancé des programmes importants de numérisation de leurs collections. Des millions d'objets (peintures, gravures, dessins, photographies anciennes) sont désormais disponible sous forme de photographies numériques. De plus, une partie croissante de ces images est désormais disponible en ligne, de manière facilement accessible. Cette thèse étudie comment ces grandes collections d'histoire de l'art peuvent être rendue *cherchable* en utilisant les nouvelles approches d'apprentissage profond pour le traitement et la comparaison d'images. Il prend comme étude de cas le traitement des archives photographiques de la Fondation Giorgio Cini, où plus de 300 000 images ont été numérisées. Nous montrons comment un pipeline de traitement générique peut extraire de manière fiable le contenu visuel et textuel des images numérisées, ouvrant ainsi la voie à la numérisation efficace des grandes collections d'archives photographiques Ensuite, en exploitant un graphe des connexions visuelles, une métrique visuelle est apprise permettant de regrouper et de rechercher des reproductions d'œuvres d'art indépendamment de leur support, résolvant ainsi efficacement un problème difficile de recherche d'images interdomaine. Enfin, la thèse étudie comment une interface Web complexe permet aux utilisateurs d'effectuer différentes recherches en fonction de cette métrique visuelle, et le processus par lequel les utilisateurs peuvent annoter des éléments d'intérêt lors de leur navigation pour les ajouter à la base de données, améliorant ainsi la qualité des résultats. En documentant une approche complète sur la manière de passer d'une photo-archive physique à un système de navigation à la pointe de la technologie, cette thèse ouvre la voie à un moteur de recherche global dans les archives de photos du monde entier.

Mots-clés : numérisation ; large collection d'images ; recherche visuelle ; apprentissage profond ; vision par ordinateur ; archives ; histoire de l'art.

Contents

Remerciements **v**

Abstract (English/Français) **vii**

List of figures **xiii**

List of tables **xvii**

Introduction **1**

1 Background **9**
- 1.1 Image Features . . . 9
  - 1.1.1 Hand-crafted features . . . 10
  - 1.1.2 Deep Learning Approaches . . . 11
- 1.2 Computer Vision applied to Art Images . . . 14
- 1.3 Computer Vision for Image Retrieval . . . 15

2 From the Physical Archive to the Digital Archive **19**
- 2.1 Scanning infrastructure, logistics . . . 20
  - 2.1.1 Challenges . . . 20
  - 2.1.2 Scanner design and characteristics . . . 21
  - 2.1.3 Digitization process . . . 22
- 2.2 Pipeline for the processing of the photo-collection . . . 23
  - 2.2.1 Raw Conversion . . . 24
  - 2.2.2 Cardboard/Photograph extraction . . . 25
  - 2.2.3 Reading the metadata . . . 29
- 2.3 Aligning artist names with a knowledge database . . . 37
  - 2.3.1 Knowledge databases for artists . . . 38
  - 2.3.2 Challenges . . . 40
  - 2.3.3 Matching approach . . . 42
  - 2.3.4 Results . . . 44
- 2.4 Genericity of the proposed pipeline . . . 46

2.4.1 dhSegment: a generic approach for document segmentation 48
2.4.2 The state of handwritten text recognition 53
2.4.3 Extracting the metadata elements? 55

3 Organizing the Visual Information **57**
3.1 Goals 58
3.1.1 The Original Question : Tracking Patterns in Paintings 58
3.1.2 The Practical Question : the Problem of the Duplicates 58
3.2 Formalization of the Problem 61
3.2.1 Challenges 61
3.2.2 Formalization 62
3.2.3 Morphographs 68
3.3 Approach 72
3.3.1 Outline 72
3.3.2 Elements of the Framework 74
3.4 Automatic Classification of the Physical Connections 76
3.4.1 Approach 77
3.4.2 Results 82
3.5 Learning a Visual Metric 87
3.5.1 Problem Statement 88
3.5.2 Approach 90
3.5.3 Implementation Details 97
3.5.4 Results 98

4 The Replica Search Engine **103**
4.1 Goals 103
4.2 Indexation of Ressources 105
4.2.1 International Image Interoperability Framework (IIIF) 105
4.2.2 Our Approach 108
4.3 Searching Capabilities of the System 110
4.3.1 Metadata query 110
4.3.2 Single and multiple images queries 111
4.3.3 Image region query 112
4.3.4 Other capabilities 113
4.4 Displaying Results 114
4.4.1 Interface 114
4.4.2 Visual link discovery experiment 118

Conclusion **126**

A “Discoveries” in the collection **133**

Bibliography **147**

List of Figures

1 Example of restoration process documented in a photo-archive 2
2 The propagation of the Sleeping Venus motif 3
3 The pattern of the “kneeling lady” in the works of the BASSANO family 3
4 Pipeline of the project. 7

1.1 Convolutional layer and Alexnet 11
1.2 U-net architecture 14

2.1 General pipeline of the digitization 20
2.2 Doubled-sided example 21
2.3 Scanner design 23
2.4 Scanner delivery 24
2.5 Photo-collection of the Cini Foundation 25
2.6 Documents digitized per day 26
2.7 Example of scan 27
2.8 Processing pipeline of the documents. 28
2.9 Network architecture 29
2.10 Cardboard extraction 29
2.11 Prediction failure 30
2.12 Cardboard extraction examples 31
2.13 OCR process 32
2.14 Organization of the metadata elements. 33
2.15 Handwritten correction types 35
2.16 Missed printed text 35
2.17 Knowledge database examples 39
2.18 Alignment process 42
2.19 Distribution of images per artists 46
2.20 Cumulative distribution of images per artists 46
2.21 Temporal and spatial distribution of the collection 47
2.22 dhSegment framework 49
2.23 Historical Layout Analysis example 50

2.24 Ornament extraction 51
2.25 Baseline detection examples 53
2.26 Handwritten transcription examples 54

3.1 Examples of interesting connections 59
3.2 Example of pattern propagation 60
3.3 Physical link example 64
3.4 Partial order agreement 65
3.5 Example of consistency 67
3.6 Examples of morphographs 70
3.7 Visual knowledge diagram 73
3.8 Proposed framework for acquiring the morphograph 74
3.9 Photographs with conflicting metadata 76
3.10 Examples of physical connections 78
3.11 Examples of non-physical connections 79
3.12 Matching distances for local descriptors 80
3.13 Highlighting the differences of images 81
3.14 Duplicate detection example 83
3.15 Precision-Recall curves for the binary classification of physical links. 84
3.16 Cartoon example 88
3.17 Limits of feaure points 90
3.18 Network architecture for transforming an input image to the corresponding visual descriptor. 92
3.19 Hinge loss 93
3.20 Triplet learning 94
3.21 Limitation of the triangular inequality 96
3.22 Spatial filtering of cnn matches 97
3.23 Example of retrieval queries 101

4.1 IIIF Image API 106
4.2 The Cini as IIIF 107
4.3 IIIF to Replica conversion 108
4.4 Replica architecture 109
4.5 Complex image search 112
4.6 Region image search example 113
4.7 Replica interface 115
4.8 Visualization modes 116
4.9 Map visualization 118
4.10 Map visualization with/without the duplicate images 119
4.11 Experiment annotations example 122

4.12 Found connections relative to their distance in the visualization 124
4.13 Pipeline of the project, completed 129
4.14 Leonardeschi variations of the Nativity 130

List of Tables

2.1 Cardboard-Photograph extraction results 30
2.2 Performance on handwritten elements 36
2.3 Performance on typewritten elements 36
2.4 OCR performance 36
2.5 Coverage obtained by the matching process. 45
2.6 Most common artists 47
2.7 Layout Analysis for Manuscripts results 50
2.8 Ornament extraction results 51
2.9 Baseline detection results 52

3.1 Number of physical connections extracted from the collection. 85
3.2 Attribution conflicts in the Cini corpus. 86
3.3 Retrieval performance 100
3.4 Retrieval with reranking performance 100

4.1 Survey question 123
4.2 Number of correctly connected images 123
4.3 Percentage of success of finding informative connections 125

Introduction

> Immature poets imitate; mature poets steal;
> bad poets deface what they take,
> and good poets make it into something better.
> — T. S. Elliot

The Importance of Photo Collections

Historically, the need to physically see objects of study has always forced scholars to travel through the world. In order to avoid the requirement of traveling around the globe, working with reproductions has almost always been a part of the way scholars have managed to compare artworks with each other. These reproductions, which were originally in the form of drawings or engravings, provided only a limited coverage of the artistic production, and consequently limited the extend of possible studies by scholars.

The development of photography has dramatically changed Art History as a discipline, as it has largely improved the quality of these reproductions, as well as their availability. This evolution is also contemporary of the development of the art market, prompting the need for precise attribution and estimates. The rise of the *connoisseurship*[1] was an effect of this evolution, with the figure of the *connoisseur* leaning heavily on his personal collection of photographs to give his expertise.

Early pioneers of the use of photographic technology emerged, including illustrious names such as Aby Warburg, Bernard Berenson and Helen Clay Frick[2]. Their collections, which they acquired during their lives, have contributed in creating the most important photo archive institutions in the world.

Today, these photo-archives scattered throughout the world represent a vast library of knowledge about many works of art, with information that is difficult to obtain by other means. For example, they contain information about the history of some of these artworks, such

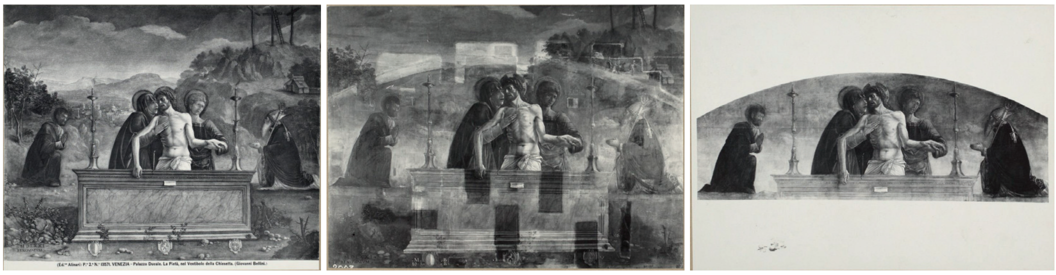


Figure 1: Restoration process of the *Pieta* by Giovanni Bellini (or Gentile Bellini), in the Palazzo Duccale (Venice). The original painting was enlarged with a landscape in 1571 by Paolo Farinati. From left to right respectively, the artwork before, during, and after the restoration of 1948 bringing back the original simpler composition. (Images 80C_224, 80C_227, and 80C_233 from the Cini photo-archive.)

as the multiple restoration processes a painting might have gone through, as can be seen in Figure 1. Furthermore, they document a much larger range of objects than conventional online image databases, since not only masterpieces but also the complete spectrum of the Art production is represented, including minor works. As these photo-collections were acquired during a long time period, they also record many artworks not publicly available. For instance, paintings which are in private collections, but which went through the art market at one point of their lives, could have been photographed at the moment of the auction. Despite the recent efforts in digitization by museums around the world, these initiatives usually focus on very high quality acquisitions of the important works in their collections, thus not reaching the incredible coverage these photo-collections offer. Finally, these photo-collections are also a testament of the practices of these Art Historians, giving precious insights about their research and artistic period interests, thus allowing insight into the personalities of these scholars.

One task that benefits from having a large coverage of the artistic production is the ability to re-establish genealogies of motifs, filiation of models, and groups of compositions, i.e the history of “forms”. By being able to explore these visual collections, one can indeed put in relation preceding and subsequent works, effectively documenting the artistic contexts, and the dynamics of production. For instance, we have represented in Figure 2 the original prototype of the composition of the sleeping Venus by Giorgione, one (of the many) variations done by Titian, and almost 300 years later, how this composition had an impact on the painting of Edouard Manet. A different form of motif reappearance is displayed in Figure 3, where the reappearance of a very similar character is a manifestation of the design process.

As the number of images available grows, so does the quality of the possible investigations. While for *connoisseurs*, it allows a more detailed expertise; it also permits the establishment of more plausible connections between potentially distant artists (geographically and/or historically), through the transmission of artistic novelties. Since the scale of the collection one

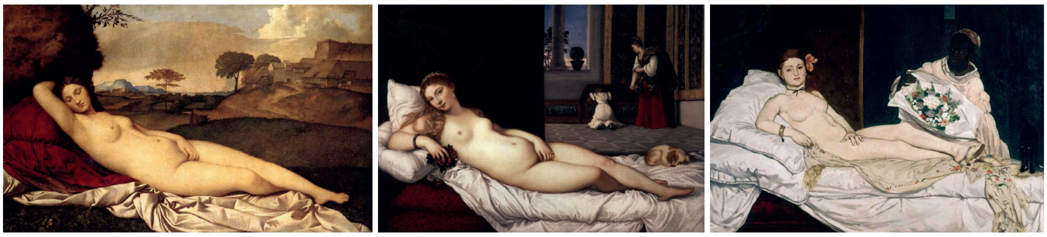


Figure 2: From left to right: *Sleeping Venus* by Giorgione (c.1510) (landscape often attributed to Titian) ; *Venus of Urbino* by Titian (1538) ; *Olympia* by Edouard Manet (1863). This is a well-documented case of a transmission and reuse of a composition.

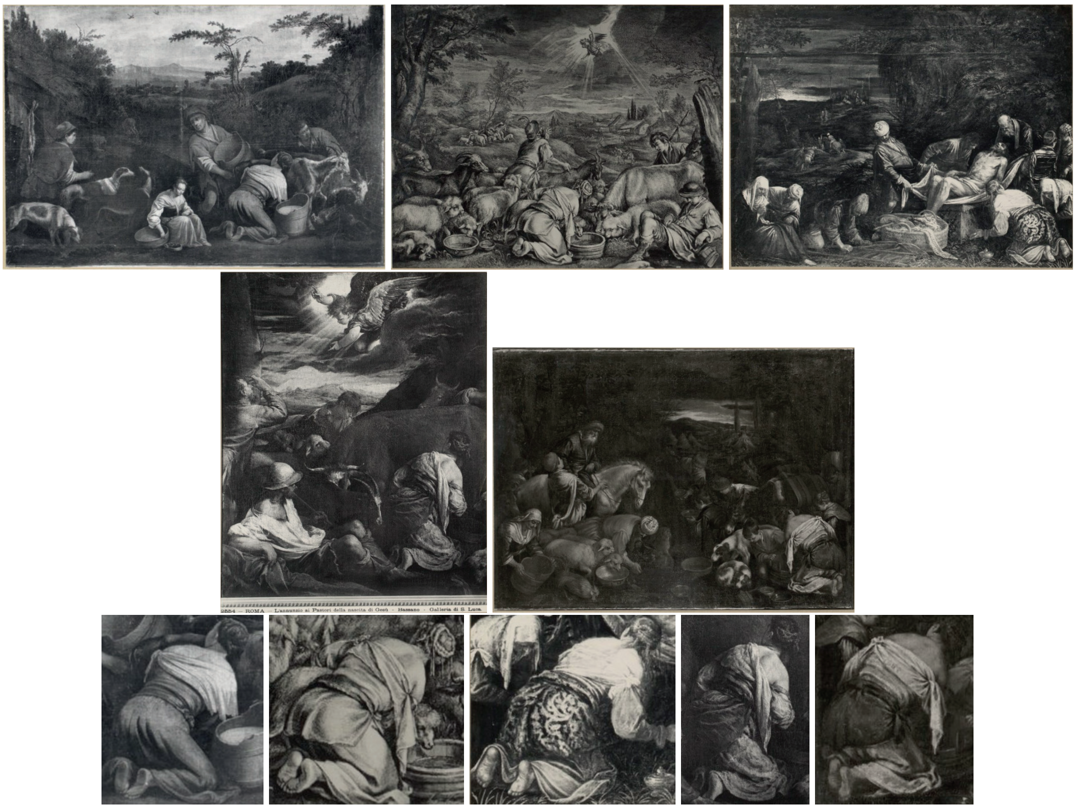


Figure 3: Multiple artworks coming from the Bassano Family workshops. Notice how, despite the varying compositions, the pattern of the kneeling woman at the front is very consistent, giving interesting insights in the design process at these workshops. (Images respectively 158C_626, 45C_170, 162B_569, 110C_10, 158B_321 from the Cini photo-archive.)

works with is relevant, the fantastic coverage of photo archives offers incredible opportunities, assuming one is able to effectively use the amount of data they represent.

The Current State of Photo Collections

However the current state of photo collections does not allow effective use of the information they represent.

Firstly, most of the collections are not yet in digital format yet. Indeed, the sheer number of individual documents to be acquired is large, and represents a huge amount of work, that institutions are not always capable of doing. Additionally, most of the current digitization initiatives are limited by the manual work of converting the metadata to its digital representation. Indeed, documents in photo archives are often complex semi-structured documents, with a mix of visual and textual information. The notes attached to these documents are often free text that does not translate well to structured information, resulting in a costly conversion process. Finally, as the documents are still in a physical form, accessing them requires physically moving to the location of the archive, which hinders the research process.

Secondly, another challenge for the exploitation of this data is *fragmentation*. This fragmentation is two-fold. First, even if they were all digitized, each institution having its own photo collection, the information would be scattered in different data silos. Since the organization of each archive usually corresponds to local historical practice, inter-operation between these archives can be difficult. However, without the conjunction of these separate archives, one will always have only a partial view of the complete information, as each collection often has a specific coverage, be it spatially or temporally. The second form of fragmentation is inside the archives themselves. As these archives were often created as the aggregate of the photo collections of multiple individuals, there can be wide variations in the way they are organized. For instance, the Cini photo archive is made of no less than 29 individual separate collections.

Finally, even in the case of digitized images, the searching capabilities of these collections are always based on textual metadata only. First, this assumes that each image was actually correctly transformed in a set of tags and/or textual description, which can be a very costly process, as precise standardized textual descriptions (like Iconclass [3]) requires specialized annotators. Such an expensive operation cannot be manually carried through complete collections, often leaving us with sparse and incomplete metadata. Also, a textual transcription of an image is always a projection of the original information, limiting the possibilities for searching. In the end, the elements we are looking for are visual artworks, and it is natural to want to query these archives according to their visual information rather than their textual information.

Machine Vision to the Rescue

Extracting the information (both visual and textual), and visually index these millions of documents is a difficult task. Given the quantity of what needs to be processed, a purely manual approach would be impractical, hence the need to use machine vision to our advantage.

In recent years, the advancements in computer vision have been staggering. Since the reappearance of neural networks in the field of machine learning in 2012, we have seen incredible improvements in almost every domain of the field. Very difficult tasks, like object classification, are now almost considered a solved problem, if enough training data is available. The fact that these new methods of deep learning have very good prediction performances, can be used very efficiently once trained, and adapt much better to unseen training data, have opened new areas in the applications of computer vision in the real world.

This makes us believe that we have reached the tipping point where technology is now mature enough to have these photo archives reach their proper potential. The difficult task of automatically processing these large collections while adapting to their specificities is now possible with the help of deep learning. Similarly, the advances in learning visual similarity can potentially allow us to build meaningful search capabilities on top of these digitized collections.

Challenges

The large scale digitization of photo collections poses multiple problems. Apart from the physical constraints to efficiently convert heterogeneous documents to a digital format, being able to extract the relevant parts (textual, visual, etc.) with a method that could adapt to the variety of formats available is not something which has been investigated previously. Given the potential diversity among documents of the archive, this is a crucial point in order to allow a generic processing pipeline to be applied to more than one situation.

Additionally, since this sort of data has not been made available before, previous works on large collections of photographs of artworks are very limited. Since research is often driven by which datasets are available, when it comes to art images research has focused on style classification and object detection. As such, the task of image search in this domain is under-evaluated.

Visual search in art images presents challenges because of the large textural variations across images. Indeed, the same pattern can appear as a sketch, a print, or in a painting with a variety of style. Additionally, the quality of images in photo archives can be another challenge. Indeed, they are often old black-and-white photographs, sometimes in below average lighting conditions. Finally, the matching element might only occupy part of the image, making it more difficult to retrieve.

Aim of this thesis

The aim of this thesis is to develop methods and techniques to (i) make the largest number of photo-collections accessible, and to (ii) navigate relevantly in this large dataset of images.

When it comes to the first question, the goal is to propose building blocks for the efficient digitization of these photo archives. Indeed, as the International Consortium of Photo-Archives[4] claims 27 millions images among 14 institutions, with only a limited amount digitized, reusable methodologies are key. But accessibility of the images is not only digitizing them but also publishing them online so that anyone can interact with the different collections in a standardized way.

The second question is about navigating such large iconographic collections. A focus of this work will be to provide a way to search *visually* in a database of artworks, which includes acquiring a corresponding dataset for training/validation, as well as proposing a well-performing solution for it. But as photo archives are not just simple lists of images, we also aim at providing specific interfaces for exploring the complexity of these collections.

Structure of the thesis

In this thesis, we answer both these questions by describing how we built a complex search engine on top of the photo-collection of the Cini Foundation in Venice, one of the world's leading institutions for the study of Art History. We will describe the complete pipeline that allowed us to go from the physical documents in the archive to a capable navigation interface, tailored for the needs of Art Historians.

After a chapter dedicated to documenting related works, the thesis is roughly divided with respect to the successive steps of the pipeline displayed on Figure 4.

The first chapter will describe how we efficiently digitize and extract the information for more than 330'000 documents from the photo-collection of the Cini. Additionally, we will discuss the potential generalizability of our approach to other photo archives.

The second chapter focuses on how to organize and index these large collections of images. We introduce a formal model allowing us to represent the complexity of the relationships between these images, and leverage it to learn a visual similarity metric. We also solve the problem of duplicate photographs, automatically finding hundreds of artworks displayed by different photographs with conflicting metadata.

Finally, the last chapter deals with the navigation system created, discussing both its infrastructure and interface, which allows navigating these large collections of documents.

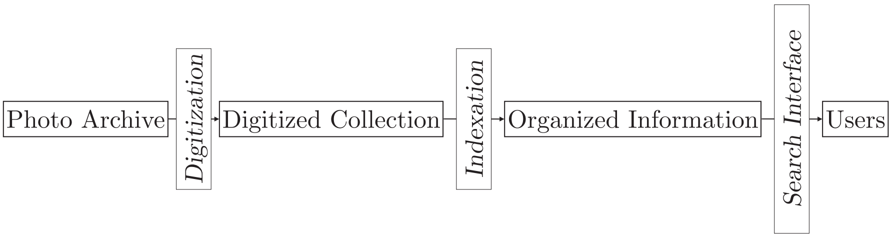


Figure 4: Pipeline of the project.

1 Background

In this chapter, we review a range of literature related to the issues we are trying to tackle, serving as a suitable background for the research we conducted. Firstly, we will review the most common feature representations used to encode images for computer vision tasks, and how they have evolved, from hand-crafted techniques to pre-trained neural networks. Secondly, we cover the previous works done in applying computer vision to Art images, and its relation to domain adaptation. Finally, we will look at the work previously done in the case of image retrieval and its relation with metric learning.

1.1 Image Features

The goal of the field of Computer Vision is to give high-level understanding of images and videos to computers. Corresponding tasks includes object localization/classification (given an image, what are the object represented?, and where are they?), face recognition (given the photograph of a person, identify him/her from a database of persons), or handwritten recognition. In practice, the input visual signal is almost always acquired via a digital camera in the form of an array of pixels, each of them encoded as a (Red,Green,Blue) triplet, creating a tensor of values of size $H \times W \times 3$ (where H and W are respectively the height and the width of the image). However, for the high-level tasks Computer Vision is dealing with, such a representation is not a very good input to be able to output a decision.

Indeed, even a small translation or a change of illumination can have a drastic change on the raw pixel values of an image. Other possible transformations include a change of viewpoint, scale, occlusion of objects, etc. In order to be able to handle these variations in a better way, an image is often converted to a *feature representation*, a vector of much lower dimension than the number of pixels represented, but with a good expressive power of the visual information present in the image.

These feature representations are then used in-place of the images in order, for example, to learn an object classifier, or to retrieve similar images. Because of their importance, we present some of the most prevalent image features. First, we go over the handcrafted features, i.e computed from manually engineered equations, then we will present the features coming from deep networks, which have been recently dominating the field.

1.1.1 Handcrafted features

Most of the popular handcrafted features are based on encoding the orientations of the gradient, as it naturally brings some invariance to color and illumination. The GIST descriptor for instance [5], is based on dividing an input image on a regular grid, computing an orientation histogram for each cell of this grid, and concatenating all the values in a single vector, representing the global image.

A similar approach was taken for the HoG descriptor [6], where the image is represented by a grid of the gradient histograms, however they show that a local normalization of the histogram cells was producing better performance. Using this representation, they could learn a discriminative template (through a linear classifier) and apply it in a sliding-window manner to perform pedestrian detection on images. This approach was then extended in [7] where in addition to the main filter used for matching, part-filters are used as well, increasing the performance for object detection.

Despite its performance, a large drawback of the aforementioned descriptors are that they are very sensitive to scale and rotation. For instance, in the case of object detection with HoG descriptors, the comparison of the search template has to be done exhaustively with many resized version of the input images to match the approximate scale of the object, which is an expensive process.

An extremely important way of analyzing images are based on feature points and their associated descriptors. With the SIFT algorithm [8], the authors proposed a complete pipeline to (i) detect stable points of interests with a corresponding scale and orientation, thus making them transformation invariant, and (ii) compute a local descriptor for the detected points. A lot of additional research was done in improving the different steps of the pipeline, for instance detecting affine regions instead of circular ones [9], or noticing that simply taking the square root of the descriptor actually improved performance [10]. Alternative feature points detectors and descriptors have been proposed over the year, one of the most popular being SURF [11], an approximated and faster version of SIFT.

Initially proposed for object matching between images, local descriptors also became popular for many other tasks. However, as the output is a variable set of local descriptors, it does not produce a practical fixed-length representation. The "Bag of Visual Words" approach

[12] consists of encoding the image with an histogram of visual words. First, a dictionary is learned from a large number of local descriptors by computing a set of D representative visual words through K-means. Then given an image, local descriptors are computed, and each descriptor is then assigned to the closest visual word. This list of visual words are aggregated in an histogram, effectively transforming the input image to a D dimensional vector.

1.1.2 Deep Learning Approaches

Deep learning has, in the recent years, become dominant in many areas of machine learning, and computer vision is not an exception. The most important systems for deep learning and images are the Convolutional Neural Networks (CNN).

CNN are based on convolutional layers. A convolutional layer can be seen as parameterized function that takes as input an image of size H, W with C channels, which can be represented as a tensor of size $H \times W \times C$. The parameters of a convolutional layer are mainly concentrated in its C' convolutional filters. Each filter, of size $F \times F \times C$, is used to compute a scalar response over the input tensor, at every position of the input image-tensor. A non-linear operation (called an activation function) is applied to each of these scalar responses, creating a feature-map. The concatenation of these feature maps (one per filter), is of size $H \times W \times C'$ and is the output of the convolutional layer.

Because the input and the output formats of a convolutional layer are of the same type (3-dimensional tensors), they can be chained together[1]. This allows creating a complex parameterized function that, for instance, takes an input image and outputs a probability distribution. Such a construction is a Convolutional Neural Network.

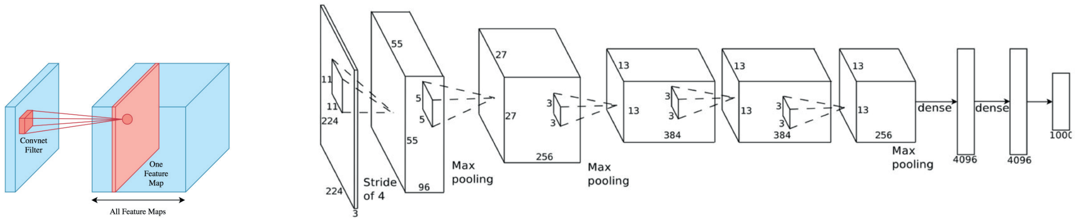


Figure 1.1: Left: a single convolutional layer.
Right: the Alexnet architecture[2]. The input is a 224 × 224 RGB image, while the output is a 1000d vector representing the classification score for each class.

As these networks can have millions of parameters through their convolutional filters, they

---

[1] For the sake of completeness, other operations which works on tensors are usually needed as well, like the pooling operation which reduces the spatial dimension by aggregating values (going from $H \times W \times C$ to $\frac{H}{2} \times W \times C$ for instance). However, the core of the CNNs lies in their convolutional layers.

[2] Both images taken from https://brilliant.org/wiki/convolutional-neural-network/.

need to be trained to perform a given task. Fortunately, as every building block of the function is differentiable, the parameters can be optimized through gradient descent. More precisely, one can show an image to the network, compare the network's output with the desired value, and slightly update the parameters of the network so that its output is closer to the target. This training process is performed iteratively on a large number of annotated samples, progressively improving the performance of the model.

The use of CNN is actually relatively old, as it was used with great effectiveness for the task of handwritten recognition as early as 1989 [13, 14]. However, it was in 2012 when the first "deep" (i.e. many stacked convolutional layers) CNN [15] appeared that the expressive power of CNN architectures came to light. This network, dubbed "AlexNet" in honor of its creator, was using a stack of 8 layers (see Figure 1.1), and shattered the competition at a difficult object-classification task[16]. Given an input image, the network could output the correct object label in its top-5 predictions (out of 1'000 classes) with an error-rate of 15.3%, the closest competitor being at 26.2%.

This performance showcase had a dramatic impact on almost all fields of computer vision, where CNN have become the prevalent method. However, this important shift was made possible by three main evolutions: (i) the increasing size of the datasets, (ii) large computing power becoming available, especially using graphic processing units (GPU), and (iii) better optimization algorithms. Indeed, as these networks are huge functions with a large number of parameters[3], optimizing them required both a lot of data, and a lot of computation, which were not available in the 90's. Similarly, this milestone was built on various contributions that came in the years prior, allowing for faster and more stable training. For instance, using a simpler Rectified Linear Unit (ReLU)[18, 19] instead of the more traditional $tanh$ as the activation function was instrumental in the CNN revolution, as well as the dropout technique[20], which allowed to avoid co-adaptation of the learned features.

Since the first deep CNN, the community has proposed various improvements, both boosting the performance and understanding of these networks. For instance, in [21], the authors showed that a simple very-deep architecture (16 or 19 layers) built only on 3 × 3 convolutional filters could bring the performance on the object-classification task to 11.1% top-5 error. Some of the most important algorithmic contributions of the recent years are probably the batch-normalization [22] (a way to stabilize the distribution of inner activation values, hence improving training), and residual connections [23, 24]. By using identity connections bypassing the convolutional layers, the first residual networks (going up to 152 layers[23]) brought the error rate to 6.8%. Both the very deep 16-layers network and the 50-layers residual network have become very popular architecture to be used in various areas of computer vision.

---

[3] Improvements in the CNN architectures of the following years made the models drastically smaller (so much that they can now easily be embedded in phones [17]), but it was a major drawback for the first generation of deep CNN.

Fully-Convolutional Neural Networks

Most of the architectures presented above are in the context of image classification, where the output is a fixed-size vector representing the probabilities for each label. But by reorganizing the building blocks of neural networks, different architectures for different needs can be created. One example is the Fully-Convolutional Neural Networks (FCN) which, given an input image, can predict separated attributes for each pixel. These FCN usually only use convolutional layers (hence their names), and downsampling/upsampling layers. Traditionally, they are formed of two parts: an *encoder* and a *decoder*. The encoder part is very similar to a standard object-recognition-oriented architecture, progressively reducing spatial resolution while increasing the number of channels to get a higher-level representation of the corresponding part of the image. The decoder, on the opposite, use the feature maps outputted by the encoder to upsample the signal back to the original resolution. An early example of such architecture is [25] where the authors train an architecture end-to-end for semantic segmentation of objects.

Initially proposed for the task of cell segmentation in microscope images, the U-Net architecture [26] (Figure 1.2) has become the standard for pixelwise segmentation tasks. The fundamental characteristic is that the encoder and decoder are symmetric, resulting in a u-shaped organization. At each resolution level, the upsampled feature map of the decoder is merged with the corresponding feature map of the encoder through a concatenation. This permits the iterative increase of resolution to be more precise as intermediate representation are sequentially integrated.

These types of neural networks are interesting to us, as they are used more and more in the task of document analysis. For instance, [27] and [28] both uses FCN for separating the page of a scanned document, while [29] tackles the issue of segmenting the elements of old manuscripts.

CNN Features

When one computes the output of a CNN given an input image, the successive feature maps are computed as intermediate values for the final result. If one considers these intermediate outputs by themselves, a CNN implicitly computes a list of hierarchical feature representation, useful for computing its final decision. In that sense, by training a CNN, we also train image features in the form of these feature maps.

Multiple works have investigated the power of these *CNN features* extracted from large deep networks trained on the ImageNet dataset [16]. In [30], they show how these features can be used as a strong base for other object classification tasks. In [31], they additionally show that they perform surprisingly well across domains (i.e. training a classifier on a product catalog, and then applying it on camera images), and for fine-grained classification (bird species), both

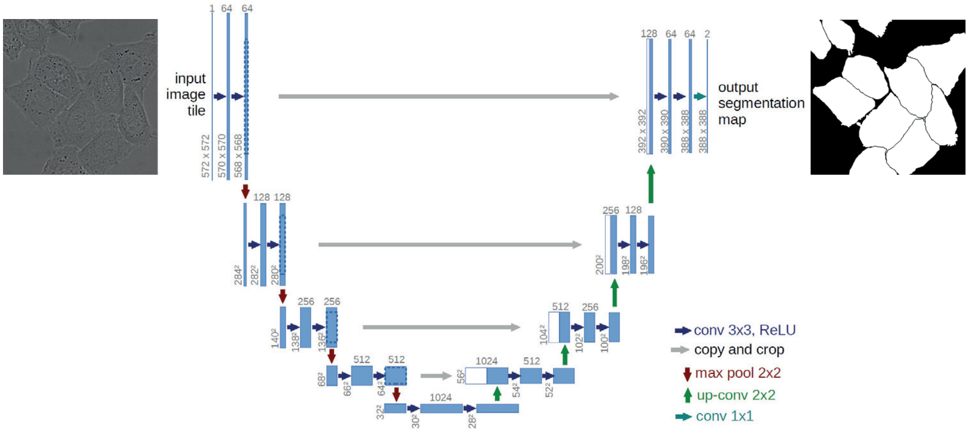


Figure 1.2: U-Net architecture. Encoder and decoder part on the left and right respectively. Skip-connections are displayed in grey. Image taken from [26].

these tasks not being directly related to the original problem the CNN was trained on.

A most exhaustive analysis is performed in [32]. In this work, they analyze a wide range of computer vision tasks, and evaluate the difference in performances of different pre-trained networks, as well as the differences coming by which level of intermediate representation is used as input features.

1.2 Computer Vision applied to Art Images

As far as analysis of paintings is concerned, a lot of the precursor work actually comes from the Image Processing community. By using powerful signal processing techniques (such as wavelet decomposition), researchers have tried to use statistics of the visual signal as a signature of the painter in order to identify forgeries. We do not plan to be exhaustive on this topic and we just mention the joint efforts of multiple labs on VAN GOGH paintings [33], as well as the smaller scale analysis of some of PIETER BRUEGEL drawings [34].

Computer vision applications to art images are often attempts at direct translations of standard machine vision tasks to the art domain. Image classification for instance, has seen some interest by using the large dataset of modern art from Wikiart [35]. Because each artwork is assigned to a “style”, a “genre” and an artist, multiple works have tried to evaluate classifiers for these classification tasks. In [36], they use pre-trained CNN features to perform these distinctions, while in [37] they combine them with a mix of handcrafted features. By combining these labels, the authors of [38] learn a general metric space between artworks, which they use [39] to detect “influential” paintings (i.e. dissimilar to artworks created before it, but with similar ones painted after).

Some institutions have released some datasets for researchers as well. The Bibliothèque Nationale de France, for instance, was interested in the task of metadata annotation [40] and how to use semi-supervised learning to add tags to their collection in a more efficient manner. The RKD challenge [41], is about predicting the multiple characteristics of the online collection of the Rijksmuseum. It is probably the most consistent dataset for automatic author attribution, used by [42] for instance, which does an exhaustive analysis of CNN classification on it.

The task of object classification/detection has received a certain interest as well, mainly as it is an interesting case of *domain adaptation*. The goal is often to see if a classifier/detector trained on natural images can generalize to art images. For instance, in [43], researchers have interested themselves in detecting faces in cubist art. Similarly, by using the images from ArtUK[44], an object classification dataset following the same classes as the PASCAL dataset [45] was created [46]. Subsequent works on this dataset have investigated the performance for object classification by using discriminative regions [47], and CNN features [48, 49]. Overall, most of these works show that Bag-of-Words approaches perform poorly on art images, and that CNN features have often the best transferability i.e. they perform well on a domain they were not trained on. HoG and especially its part based version (DPM) can also outperforms vanilla CNN approaches in certain tasks, for instance in cubist art [43]. As such, adaptability of a trained model to the art domain is sometimes used as a diagnosis of its performance, for instance when introducing a new CNN architecture for object detection [50].

However, as far as visual search in artworks is concerned, there is only a limited amount of prior-work. In [51], they use HoG descriptors combined with exemplar-SVM learning [52] to find out the salient characteristics of a query image, before performing an exhaustive comparison on their search database. While getting good results, the process is extremely slow and computationally intensive for a single query. Another related work is [53], where the authors tackle the difficult task of matching a painting to a 3D model. In a similar fashion, they have a computationally intensive process where they render multiple views of a 3D model to identify discriminative patches, which are then matched to the paintings, allowing to identify an approximate point of view from which the object is seen in the painting.

1.3 Computer Vision for Image Retrieval

For the task of image retrieval, or image search, the amount of prior work is very substantial and we are not planning to be exhaustive.

The most common methods are variations on the original bag of local descriptors mentioned previously[12]. Each image of the database is converted to the list of visual words detected in it. Given a query image, one can extract its local descriptors, and then use an inverted index to quickly find images sharing the same visual words. The number of matches finally gives a score that is used to rank the images of the database with respect of the input query. While the original pipeline is still very similar, multiple improvements have been proposed. For instance, in [54], an additional binary code is added to each visual word, improving the matching of the visual words. Reranking the images retrieved with the bag-of-words based on their spatial consistency has been introduced in [55]. It relies on fitting a geometric transformation between two images based on the local descriptors matches, hence filtering the inconsistent matches and getting a more accurate score. Finally, another strategy is to use query expansion (first introduced in the visual domain in [56]), where by using the information in the top-retrieved images, one can improve the recall of the initial query.

However, the evaluation of these image retrieval techniques are almost always done on the same image retrieval datasets, which are either focused on buildings (Oxford5k[55], Paris6k[57]), places (Holidays[54]) or objects (UKB[58], INSTRE[59]). In all cases however, the same physical element is present both in the query and the images to be retrieved. Matching the same object is a task that local descriptors excel at, but when there is domain variability like we do for art images, their performance drop considerably (see previous section). However, there is no proper dataset for this evaluation, the closest being sketch retrieval (retrieving photographs from a sketch), but where the query domain (sketches) is not the same as the search domain (photographs).

Deep learning methods have been catching up with the performance of handcrafted systems. They are mostly based on using a deep CNN to compute a global descriptor for every image. Given a query image, its descriptor is extracted and then compared with respect to every element of the database via its euclidean distance. Initial approaches leveraged state-of-the-art CNN pretrained on ImageNet, extracting the result of its first fully connected layer [60]. It was then observed that aggregating convolutional feature maps could be more efficient[61], for instance with the R-MAC descriptor [62] which does local average of the feature maps before a global aggregation.

Apart from the feature extraction architecture, another direction for improving deep features for image retrieval is to fine-tune the underlying network so it is more relevant to the task at hand. In [63], they acquire a classification dataset made of landmarks to retrain the last layer of the network, then use the CNN features from this fine-tuned model, showing improvements in retrieval tasks. Instead of training a classification network to extract its intermediate representation, one can also directly learn the descriptor function in a metric learning fashion. A complete survey of metric learning can be found at [64], but in the case of supervised deep learning, two main approaches can be distinguished. The first [65] consists of using as constraints pairs of positive images (whose descriptors should have a small distance between each other), and pairs of negative images (whose descriptors should not be close). The second [66] is based on constraints in the form of triplets of images A, P, N where the distance between A and P should be smaller than the distance between A and N. In the cases of landmarks, these constraints can be obtained by using a large number of photographs and spatially verifying them with local descriptors [67] and [68]. Despite performing very well, these fine-tuned networks are again tailored for the specific task of landmark recognition, which does not correspond to our task at hand.

2 From the Physical Archive to the Digital Archive

In this chapter, we will tackle the problem of efficiently digitizing large photo-collections. Indeed, almost every institution dedicated to the study of Art History has a varying number of photographs gathered over the years by individuals, often Art Historians themselves. Before the recent development of the Internet, such collections were a primordial source for accessing images of artwork without having to physically go to their locations, and as such were of primary importance for any serious research institution.

The amount of work that went into acquiring these large photo-collections over the year is gigantic, and represents huge volumes of under-used data. In the prospect of trying to work with the biggest corpus possible of images, digitizing these collections seems a natural step compared to going back to digitizing the objects themselves. If the quality of the obtained images will not on par with direct acquisition (as we are digitizing old black-and-white photographs), the fact that the information is already gathered in a single physical location allows to scale much more easily.

These collections can have very varying formats and structures that are mainly dependent of the institution's archival system. They are usually a mix of semi-structured documents including photographs, typewritten text and handwritten notes. As such they offer multiple challenges for efficiently (and hopefully almost automatically) digitize them. An operation which, given the scale involved (millions of elements), has to be made efficient.

We will use the case study of the photo-collection of the Cini Foundation in Venice. A world leading institution for the study of Art History, it hosts the biggest photo-archive related to Venetian art, with a million of images. Its large collection is divided among multiple *fondi*, often the former personal photo-collection of scholars, which were gathered over the years by the foundation. Out of the one million photographs, around a third of them (330'000) have been glued on standardized large pieces of cardboard, the rest being in varied formats and

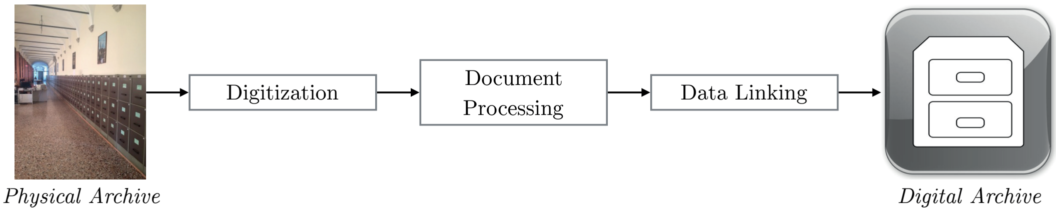


Figure 2.1: General pipeline of the process described in this chapter.

organization.

We will present the pipeline we developed in order to go from an archive of physical photographs to an organized digital collection. The pipeline can be divided into three main parts (see Figure 2.1) which correspond to the first three sections of this chapter. We will first present the digitization effort on the standardized part of the collection, made possible by a scanner which was specifically designed by a third party for this use-case. Then we will discuss how the information of the scanned documents, both textual and visual, is automatically extracted, before showing how a semi-automatic approach allowed us to align the extracted artist names with respect to a knowledge database. Finally, we will discuss the generalization of the proposed approach to other photo-archives, with the remaining challenges.

2.1 Scanning infrastructure, logistics

As a disclaimer, it must be noted that this section describes work that is not a direct contribution of the writer of this thesis. Indeed, our colleagues of Factum Arte in Madrid did the design of the scanner, while the management of the digitization process was the result of the work done directly at the Cini Foundation. Nevertheless, a description of the challenges and of the digitization work is included here for completeness.

2.1.1 Challenges

As stated in the introduction, the Cini photo-collection comes from the aggregation of different sub-collections (*fondi*) that were gathered by separate individuals for decades. Because of the very nature of the collection, the scanning infrastructure for this project should answer to three fundamental characteristics.

The first challenge is the *heterogeneity* of the different parts of the collections. As a collection of sub-collections, there is no standard size or shape for the photograph throughout the archive. The biggest might be as large as A4 pages with the smallest slightly larger than a stamp. Also,

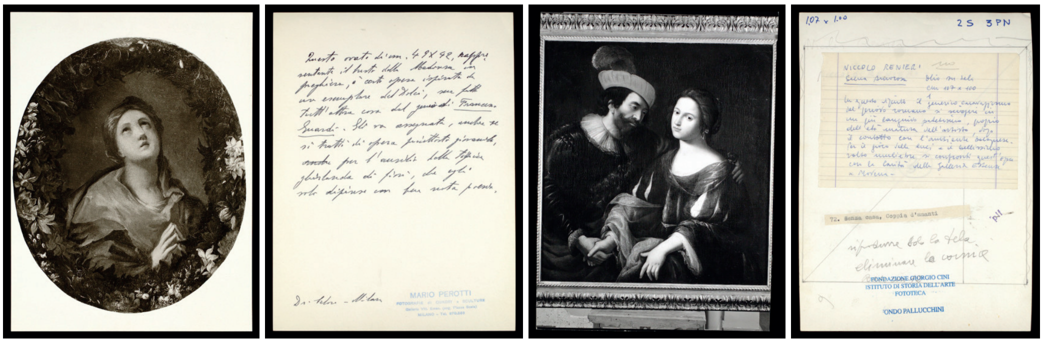


Figure 2.2: Two examples of photographs having notes on the back.

some photos are glued on large flat cardboards, while others may be irregularly shaped and grouped in anonymous envelopes.

The second challenge comes from the *double-sided* nature of a part of the collection. Since many of these photographs were part of the personal collections of famous 20-th century art historians, they actually were working tools for their studies, and it was common practice for scholars to add handwritten notes on the back of their photographs. These notes are extremely interesting as it reflects the thoughts of the scholar about the depicted object. For instance, there might be information about the location of the object, the actual physical size, but also an attribution proposed by the historian. Some examples can be seen on Figure 2.2.

The third challenge concerns the *throughput* necessary for such an enterprise. If one desires to digitize 1’000’000 elements, the process should be sufficiently efficient for the project to take a reasonable amount of time. For instance, for a system capable of digitizing 1’000 documents per day, 5 days a week, the process would still take more than 4 years.

Unfortunately, it seems clear that these different requirements are not really compatible, as it is always hard for any given system to be *fast* and *flexible* at the same time. Industry standards such as flatbed scanners would not allow the necessary throughput nor handle the double-sided nature of the collection. If some institutions have turned to a novel conveyor-belt scanner which allows a fast digitization process, they still can not digitize in a recto-verso manner.

2.1.2 Scanner design and characteristics

The task of designing the scanning system was given in September 2015 to Factum Arte [69], a Madrid based company specialized in all forms of digitization, which has an history of collaboration with the Cini Foundation.

The proposed scanner was unveiled at the Foundation in early March 2016. It is following a very unique design that can be seen on Figure 2.3. It was devised as a table with a circular, rotating top (diameter of 2 m) which comprised four image plates. 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 gathering the acquired data. The cameras were actually standard off-the-shelf high quality cameras, allowing a 400ppi resolution of the documents (5424 x 3616 pixels).

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.

The design of the scanner and of all the components have actually been open-sourced by Factum Arte, and the system underwent a European level certification. This potentially allows other institutions to replicate the device for their own needs.

The data output of the scanner corresponds to two high-resolution raw files (Canon .cr2) representing 60MB each, so 120MB per digitized document. Given the speed of digitization (1 image every 4 second), this accounts to a rate of 30MB/s of data produced on average.

2.1.3 Digitization process

The digitization process has focused on the“easy” part of the photo-collection first, i.e. the photographs glued on large standardized pieces of cardboard, also referred as *schedoni*. These documents are grouped thematically in large drawers occupying two corridors of the Foundation (see Figure 2.5). There are 512 drawers, each containing on average 600 documents.

The process started in July 2016 and finished in August 2017, so roughly 13 months later. During that time, 337’000 documents were digitized; the number of pages digitized per day can be seen on Figure 2.6. The most obvious pattern is the 5 workdays per week during the whole process, but also a relatively large intra-day variability based on the amount of manpower available. Additionally, one can notice the significant dip during the autumn-winter period

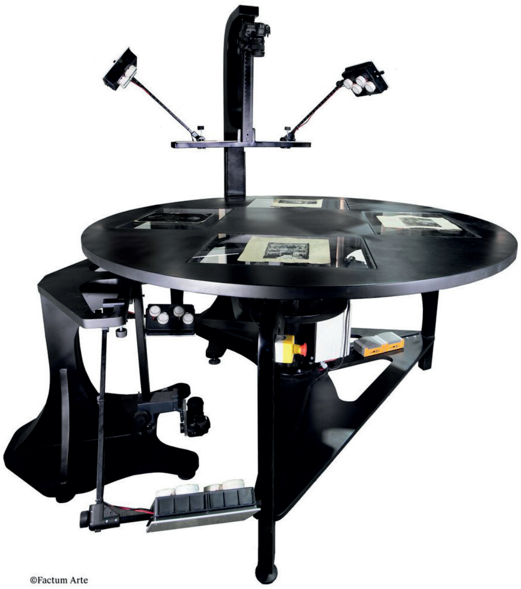


Figure 2.3: The designed scanner. Notice the two cameras for simultaneously digitize the two side of the document, and their respective sets of flashes.

of 2017 where significant changes were experimented/made with the scanning process. For instance, adding a Plexiglas separator between the color balance markers to avoid overlaps between them and the scanned object, or using a transparent Plexiglas weight to flatten objects that lost their planar shape due to age and/or humidity. Overall, after the initial variations, operators were able to process 1’500 documents per day on average.

As previously stated, the scanner creates two raw files for the recto and the verso of each scanned document respectively. In the end, the digitized 337’000 documents brings the total amount of data up to roughly 50TB. The scanned documents were continuously uploaded from Venice to Lausanne[1] as the digitization was undergoing.

2.2 Pipeline for the processing of the photo-collection

In this section, we will focus on an automatic pipeline for extracting the information in the scans. Indeed, given the amount of documents crossing the Alps everyday, an automatic process was a necessity. Fortunately, since the first-phase of digitization focused on the

---

[1] The internet connection of the Foundation actually had to be upgraded to support the transfer, which allowed to reach a transfer rate of 160GB/day. Eventually though, the first stage of the processing (compressing the raw files to jpeg format) was done on site directly, drastically reducing the amount of data needed to be transferred. Indeed, if one only needs a compressed version of the recto scan, the 337’000 documents represent “only” 2.6TB.

Figure 2.4: Delivery of the scanner at the Cini Foundation in March 2016.

standardized pieces of cardboard (the easiest part of the photo-collection to process), most of the information present is typewritten and structured already, which permits an automated process.

An example of a scan can be seen on Figure2.7. There are three relevant parts we are interested to extract:

- the area of the scan containing the cardboard, in order to remove the unnecessary parts of the image (color markers mainly), and have a cropped image of the document.
- the area of the photograph, in order to extract it and input it in a database.
- the metadata information at the top of the cardboard, which describes attribution, description of the image, location and institution hosting the corresponding artwork, etc.

A diagram of the pipeline can be seen on Figure 2.8, the following subsections will mirror the different blocks represented.

2.2.1 Raw Conversion

The very first step of the pipeline consists on a simple conversion task. Indeed, if for archival purposes, the complete raw information has to be kept; for most practical tasks, a standard

Figure 2.5: One of the two walls of drawers containing the standardized pieces of cardboard of the photo-collection. We can see half of the 177 columns, each column containing 3 drawers.

(and compressed) image format is most suitable. Thanks to the presence of the color markers on the scan, a standard white-balance process can be performed to project the color information to RGB. The obtained 3-channels image is then compressed to a JPEG format (90% quality), effectively reducing the size of the file by a factor of 15.

2.2.2 Cardboard/Photograph extraction

Cropping the cardboard and the photograph from the scans is actually a challenging task. Within each 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 area. Their colors and textures likewise differed with some photographs having a color and texture very close to the cardboard itself.

Because of this variability, a handcrafted technique involving a single type of visual clue (color, texture, edges or shapes) is not discriminative enough to perform the separation of the different layout elements correctly. In order to design a flexible framework that could

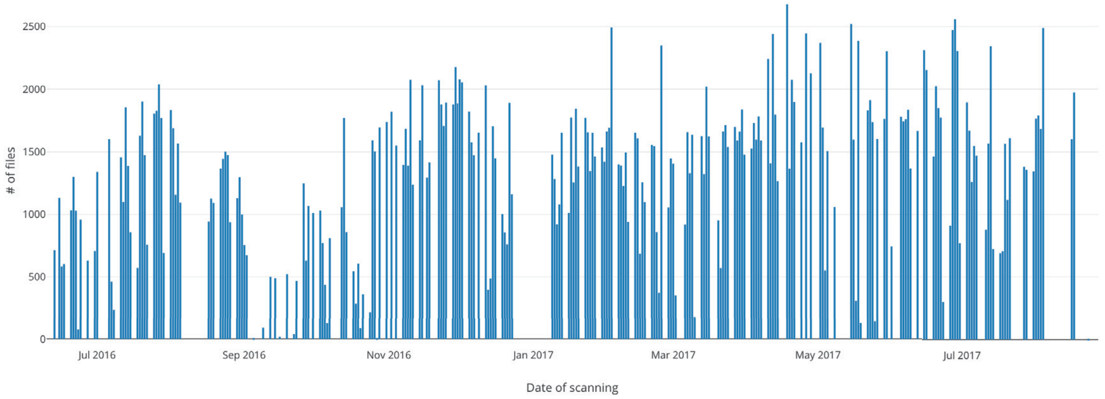


Figure 2.6: Number of documents digitized per day.

potentially be used for other use-cases, we used a trainable segmentation system based on a CNN architecture. More precisely, given an input scan, our model should be able to predict for each pixel which class (background, cardboard, or photograph) it is part of.

Network architecture

A very popular network architecture for the task of segmentation in images is the U-net[70] (shown in the previous chapter, see Figure 1.2). However, because we wanted to lower the need for training samples to the minimum, our goal was to make our architecture as data-efficient as possible. One first standard trick is using on-the-fly data augmentation, i.e. optimizing the network with rotated/zoomed/flipped versions of the training images, acting as additional training samples. Another strategy is to leverage networks which were pre-trained on another task, assuming that the intermediate representations they learned are generic enough so that they can be used to initialize a model. A standard architecture for this is the Resnet50[23] network pre-trained for image classification on ImageNet. We use it to replace the encoder part of the U-Net network. Additionally, we simplify the decoder part of the network, (i) by using bilinear upsampling instead of learned transposed convolutions, and (ii) by keeping only one convolution layer per level. The corresponding architecture is displayed on Figure 2.9.

Thanks to these modifications, the final architecture may have more total parameters (32.8M instead of 23.6M in our U-Net implementations) but since most of them are part of the pre-trained encoder, only 7.79M have to be fully-trained[2]. Also, because of the bottleneck blocks of the Resnet architecture (reducing the number of channels before the 3×3 convolution), our

---

[2] we do not count here the 1.57M parameters coming from the dimensionality reduction blocks. Indeed, they are initialized as random projections, which is a valid way of reducing dimensionality, one can see them as a part of fine-tuning the pre-trained network as well.

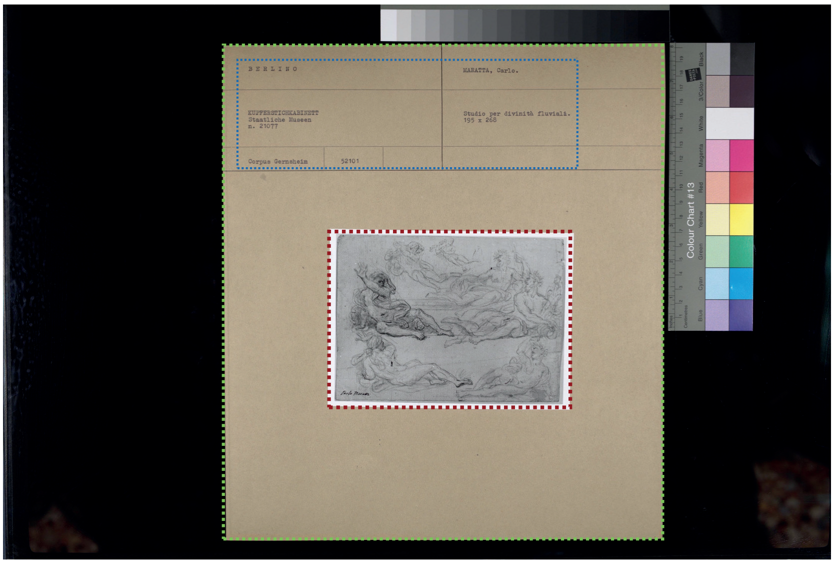


Figure 2.7: Example of an image produced by the scanner. The cardboard, photograph and metadata area are outlined in green, red and blue respectively.

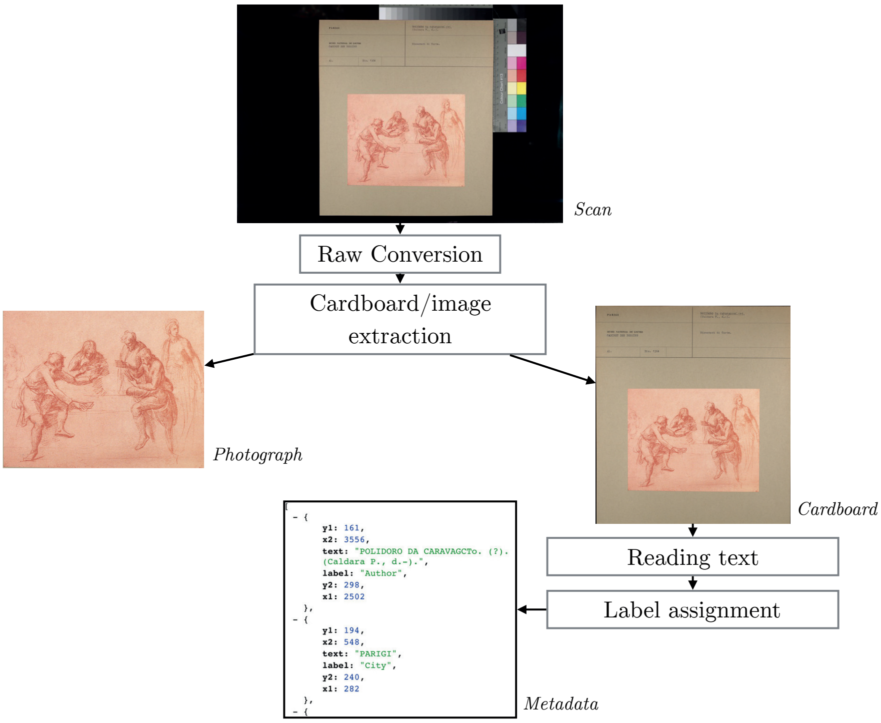
model needs less operations than the standard U-net allowing faster training and inference times (by around 40%).

Training and application

For the acquisition of training/testing data, random images were selected from the acquired scans and manually annotated at the pixelwise level. Using standard image-editing software (Photoshop), one can efficiently just draw the objects of interest with different colors and thus create accurate segmentation masks relatively quickly. Using this process, around 50 scans can be annotated per hour by a single person.

The network is trained to output the probability that each pixel belongs to one of the three classes: background, cardboard or photo. The loss is a standard cross-entropy for each pixel averaged for the whole image. Training is done with ADAM[70], exponential learning rate decay starting at 5.10^{-5}, with batch-renormalization[71] for all layers, during 40 epochs. Random rotations and zooms are applied on the fly during training, and training is done with patches of 400x400 pixels randomly extracted from the training images. Original images are resized to 1M pixels. The training process only takes 20 minutes with a modern GPU of 2016.

In order to turn the predictions into regions, we simply clean the predicted masks for the non-background classes with morphological operations and extract the smallest enclosing rectangle around them. In some cases (see Figure 2.11), parts of the photograph can be



```
[
 - {
       y1: 161,
       x2: 3556,
       text: "POLIDORO DA CARAVAGCTo. (?).
       (Caldara P., d.-).",
       label: "Author",
       y2: 298,
       x1: 2502
   },
 - {
       y1: 194,
       x2: 548,
       text: "PARIGI",
       label: "City",
       y2: 240,
       x1: 282
   },
 - {
```

Figure 2.8: Processing pipeline of the documents.

confused with the background, but that can be mitigated by using the additional layout constraint that there can be no background pixel inside the cardboard.[3]

Evaluation

In total, 270 scans were annotated, 100 for training 20 for validation and 150 for testing. The evaluation metric used is the Intersection-over-Union (IoU) between the extracted rectangle and the minimal rectangle containing the region of interest. Results can be seen in Table 2.1, and show extremely good accuracy.

In particular, we can notice a difference between the vanilla U-net and our pre-trained encoder version, especially for high threshold values. This is a manifestation of the added generalization (especially around borders between segmentations) that a pre-trained architecture gives.

---

[3] Note that a cardboard area predictor alone would still not be enough as there is not always a margin around the photograph that is part of the cardboard.

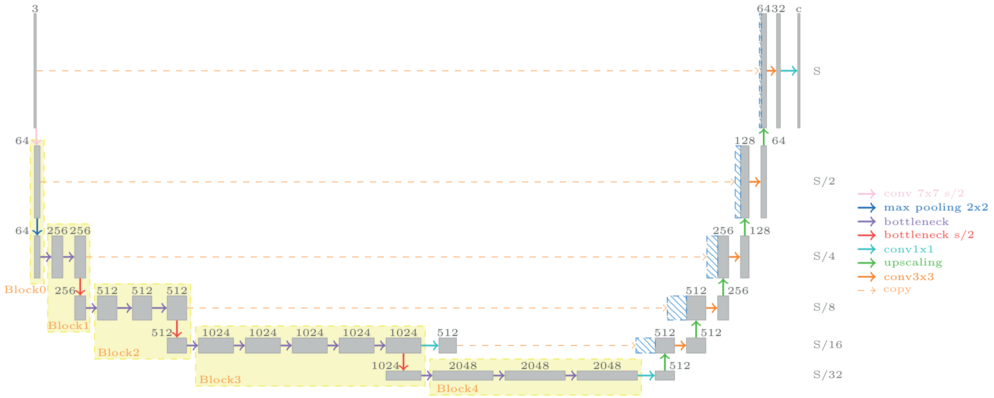


Figure 2.9: The used network architecture, using a pre-trained Resnet50 (yellow) and a simplified decoder. An additional detail is the use of two 1x1 convolutions (light blue) in order to reduce the dimensionality of the higher layers before up-sampling.

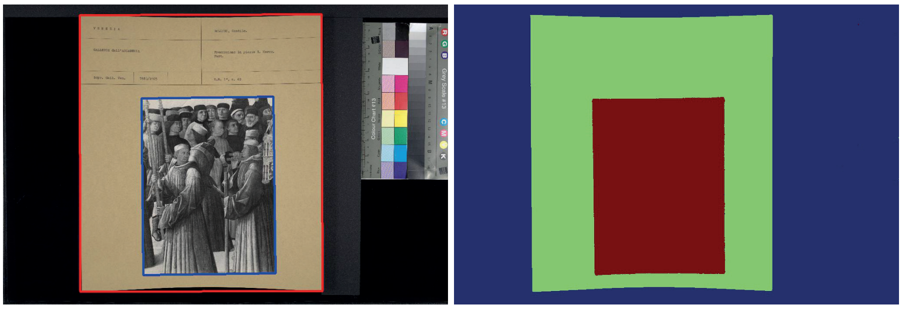


Figure 2.10: Example of prediction map obtained from the network (right) and the extracted areas overlaid on the original scan (left).

2.2.3 Reading the metadata

After having separated the different visual elements of the documents, the next step is about extracting and reading the textual metadata that was printed at the top of each cardboard document. This metadata is organized in a tabular layout where each cell contains a small paragraph of information about the represented object in the photograph.

Multiple variations in the representation can happen, which can make a given extraction algorithm fails. The position of the text might be slightly not standard, with letters going over the line borders of the table, and big photographs partly covering the text. The tabular organization of the metadata also has multiple possible formats (smaller height, slightly different cells layout).

Another factor of variability is the presence of handwritten notes on the documents added by scholars over the years. These usually provide important information such as an attribution

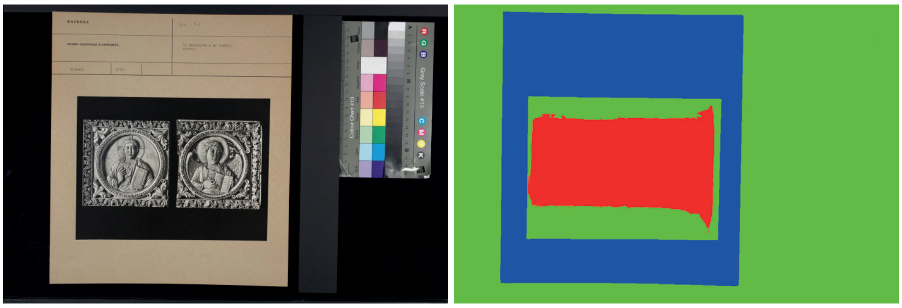


Figure 2.11: Example of case where part of the image is labelled as background,.

Table 2.1: Results of the extraction process. mIoU stands for the mean of the Intersection-over-Union of all elements, R@0.85 represents the Recall achieved for a minimum threshold of IoU$\geq$ 0.85.

Method	Cardboard	Photo			
	mIoU	mIoU	R@0.85	R@0.95	R@0.98
Predictions-only	0.992	0.982	0.980	0.967	0.900
+ layout constraint	0.992	0.988	1.000	0.993	0.947
U-net (+ constraint)	0.991	0.973	0.980	0.940	0.560

correction, or the size of the artwork. However, as we will show later in the evaluation part (page 34), they are only present on a small percentage of the documents and we decided to ignore their detection in this study.[4]

Reading and separating the text

Since we are only facing typewritten text, we will rely on optical character recognition (OCR) software as it has achieved a mature state. The open source library Tesseract and the commercial Google Vision OCR were tested.

The commercial software proved much more effective for two reasons. First, the recognition was much more accurate than with its open-source counterpart, especially if given hints about the encountered languages in the documents (in our case, we set it to Italian, French, English and German). Second, it was also able to detect the typewritten textual part automatically, which had to be specifically implemented in order to use Tesseract.

The output of an OCR algorithm is a list of extracted bounding boxes around each word and their respective transcription, as can be seen on the top of Figure 2.13. The minimum distance between every two pairs of rectangles is computed and used to cluster the words together.

---

[4] The general problem of detecting and reading handwritten text will be discussed further in the generalization section of this chapter both in Section 2.4.1 and 2.4.2.

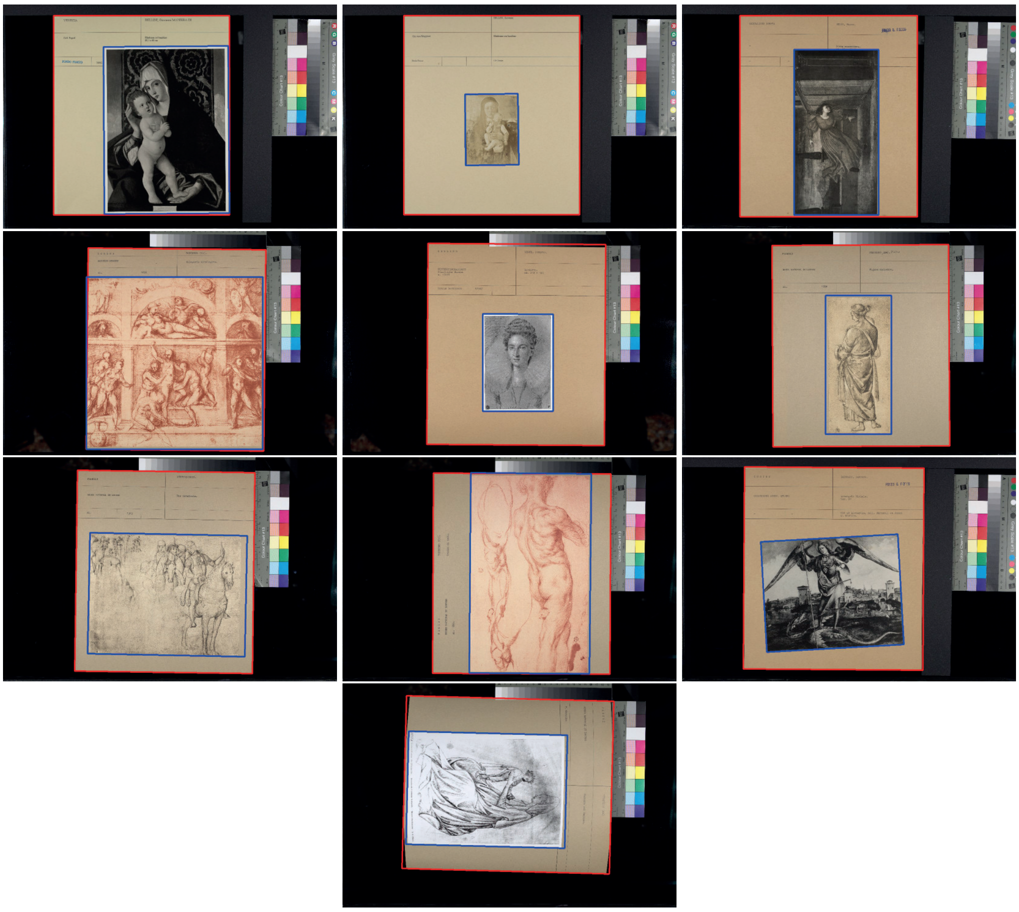


Figure 2.12: Example of extraction of the layout elements.

More precisely two words closer than a certain threshold τ have to be part of the same cluster[5]. The clustered words are then ordered according to their positions and their transcriptions combined to reform the text of the paragraph, line breaks included, as can be seen on the second part of Figure 2.13.

Label assignment

In order to obtain structured information, each extracted paragraph has to be assigned to its corresponding metadata label. In practice, the structure and positions of the metadata elements are relatively consistent and follow the corresponding schema shown on Figure 2.14. They respectively represent:

---

[5] This is actually a special case of the DBSCAN[72] clustering algorithm, with $min\_samples\_cluster = 1$.

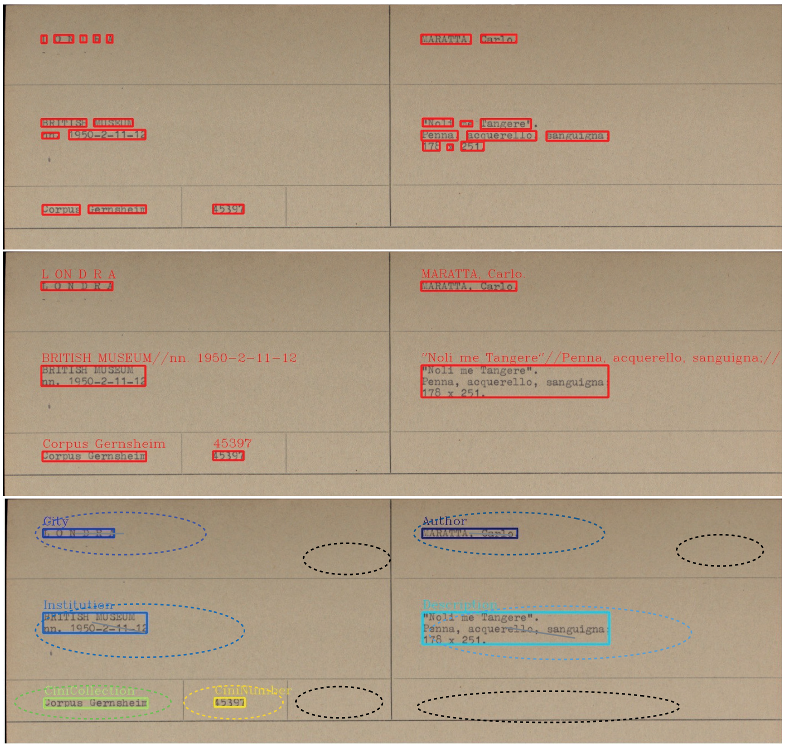


Figure 2.13: Words are detected, then clustered in paragraphs before being assigned to their corresponding label with the layout model.

- Location: the location information with the *City* where the photographed artwork is located, sometimes the corresponding *Country* will be written in the same cell as well. If the location is unknown it is usually written here.

- *Author*: the name of the artist(s) if known. If not properly attributed, it can be a generic statement like "PITTORE TEDESCO sec XVII".

- *Fondo' Stamp*: the original photo-collection (*Fondo*) this photograph comes from, printed with a stamp.

- *Institution*: the name of the museum/church/gallery which hosts the artwork. There can be additional information about the identifier number in the institution archival system, or about the precise location inside the church, etc.

- *Description*: usually a short sentence describing the object and/or its representation (*Madonna col bambino*, *Venere e Adonis*). There is also often some information about the material, and the size of the photographed object.

- *Photographer*: the name of the photographer who took the picture. If a *fondo* was only acquired by one person, the corresponding stamp might be used here instead.

- *Identifier*: the number assigned to this photograph in this *fondo*, if available.

- *Date*: the date when the photograph was taken, if available.

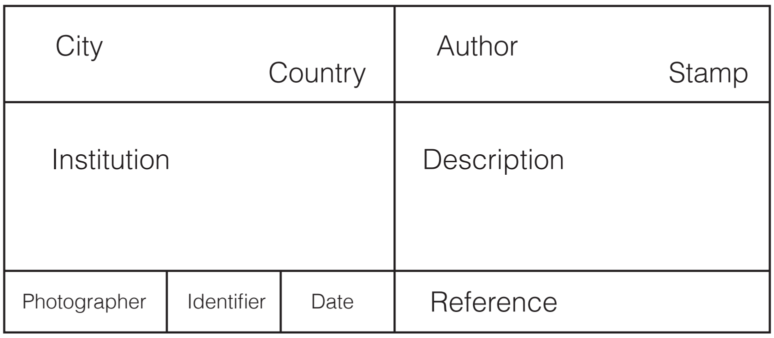


Figure 2.14: Organization of the metadata elements.

Since there are a couple of possible layouts, with different heights, we specified a set of *layout models*. Each *layout model* l simply consists of a list of layout elements $l_i = (p_i, \Delta_i)$ where $p_i = (p_i(x), p_i(y))$ encodes the defined position of the element and $\Delta_i = (\Delta_i(x), \Delta_i(y))$ represents its spread. We can then define the distance between a text-paragraph (represented by his centroid t_j) and a layout element l_i as:

$$d(l_i, t_j) = \sqrt{\left(\frac{p_i(x) - t_j(x)}{\Delta_j(x)}\right)^2 + \left(\frac{p_i(y) - t_j(y)}{\Delta_j(y)}\right)^2}$$

Given a layout model l and the centroids of the extracted paragraph, we want to minimize the sum of distances between the text blocks and their assigned layout elements (see last row of Figure 2.13). This is a standard assignment problem that can be solved in polynomial time. The optimization is done for each layout model, selecting the solution minimizing the assignment score.

Evaluation

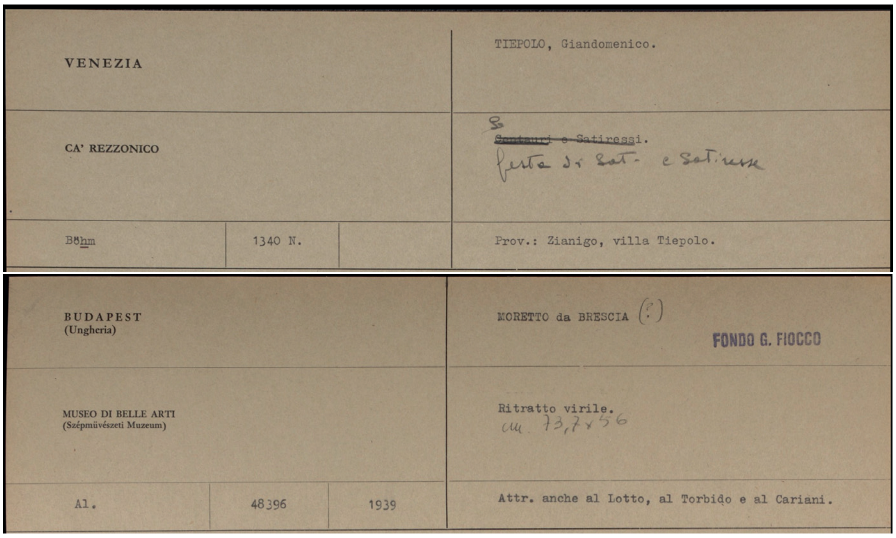
In order to evaluate the quality of the metadata parsing, we manually went through 412 scans randomly selected from the digitized photo-collection. Out of these, 31 ($7.5\% \pm 2.5\%$) had at least one handwritten correction on them. As expected, the OCR algorithm does not usually capture them, and if it does the transcription is usually unusable. Some of these handwritten elements are complete or major correction of the corresponding field, while others might be small correction or additional information. The recapitulating numbers can be seen on Table 2.2. Unsurprisingly, the most common manually modified element is the description field, where the size, material or additional description is added. Also, references are often added, as well as additional information about the institution (often the corresponding inventory number of the museum). Finally, modifications on the author field often include additional insight about an unknown attribution (spatial and/or temporal localization, for instance “SEC XVII" or “PITTORE TEDESCO").

We also recorded the number of elements of typewritten text not detected on Table 2.3. The most missed elements are the ones usually represented with one word or with an abbreviation. The photographer field which can be as short as “Al.” (standing for the Alinari brothers, famous Italian photographers) is especially hurt (4.9% missed).

However, even if we combine the total errors caused by handwritten elements *and* by mistakes of the OCR detection, we see that most fields (including the Author, and Description fields) have an **detection error-rate of around 3%.**

As far as the label assignment of each text paragraph is concerned, we manually corrected 337 scans where all the textual elements are properly extracted. This accounts to 2’028 unique metadata entries. When compared to our extraction, we counted 8 errors for the clustering step: 5 cases of two fields being merged together, corresponding mainly to very long descriptions bleeding into the reference boxes, and 3 cases of a given metadata entry being split into two parts. However, after the clustering step, no paragraph was assigned to a wrong metadata label.

Finally, we also evaluated the OCR transcription quality. For this task, a simple web interface was designed showing the top of the extracted cardboard and allowing to quickly transcribe the Author and Description fields. 148 documents where the text is properly detected were manually transcribed, and the resulting annotations compared with the automatic OCR transcriptions. Full results can be seen on Table 2.4, and show the high-accuracy of the OCR algorithm, with a character-error-rate of 2.0%. Indeed, almost **97% of elements are transcribed with at most a one-character error**.



VENEZIA

CA' REZZONICO

Böhm
1340 N.

TIEPOLO, Giandomenico.

S
~~Centauri e Satiressi~~.
festa di sat. e satiresse

Prov.: Zianigo, villa Tiepolo.

BUDAPEST
(Ungheria)

MUSEO DI BELLE ARTI
(Szépmüvészeti Muzeum)

Al.
48396
1939

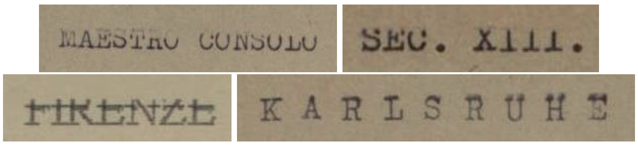
MORETTO da BRESCIA (?)

FONDO G. FIOCCO

Ritratto virile.
cm. 73,7 x 56

Attr. anche al Lotto, al Torbido e al Cariani.

Figure 2.15: Examples of handwritten corrections for the description field: complete correction (top), and additional information (bottom).



MAESTRO CONSOLO
SEC. XIII.
FIRENZE
KARLSRUHE

Figure 2.16: Examples of missed printed text, it is usually associated with a bad printing process, but not always (bottom right).

In the end, the quality of the metadata extraction, while not perfect, is very decent. There are two possible ways to view this level of quality: for the sort of large scale analysis, text-searches, or data mining we are interested in, this is an acceptable level ; but for Art Historians, knowing that there is a non-negligible probability (3%) that the converted digital version does not

---

[6] As we have a number of situation with 0 appearance (which would create a confidence interval of 0%) we takes the best practice coming from [73] and lower-bound it by $3/n \approx 0.73\%$.

Table 2.2: Proportion of **handwritten elements missed** in 412 cardboard, with 95% confidence intervals[6], per field.

Field	Complete/major correction	Additional information	Total
Author	0.5% ± 0.7%	1.0% ± 0.9%	1.5% ± 1.2%
Description	0.5% ± 0.7%	2.4% ± 1.5%	2.9% ± 1.6%
City	0.2% ± 0.7%	0.0% ± 0.7%	0.2% ± 0.7%
Institution	0.2% ± 0.7%	1.0% ± 0.9%	1.2% ± 1.1%
Photographer	0.2% ± 0.7%	0.0% ± 0.7%	0.2% ± 0.7%
Number	0.2% ± 0.7%	0.0% ± 0.7%	0.2% ± 0.7%
Date	0.0% ± 0.7%	0.0% ± 0.7%	0.0% ± 0.7%
Reference	1.7% ± 1.2%	0.7% ± 0.8%	2.4% ± 1.5%

Table 2.3: Proportion of **printed elements missed** in 412 cardboard, with 95% confidence intervals, per field.

Field	Completely missed	Partly missed	Total
Author	1.0% ± 0.9%	0.7% ± 0.8%	1.7% ± 1.2%
Description	0.0% ± 0.7%	0.0% ± 0.7%	0.0% ± 0.7%
City	1.5% ± 1.2%	0.7% ± 0.8%	2.2% ± 1.4%
Fondo	0.7% ± 0.8%	0.0% ± 0.7%	0.7% ± 0.8%
Institution	0.0% ± 0.7%	0.0% ± 0.7%	0.0% ± 0.7%
Photographer	4.9% ± 2.1%	0.0% ± 0.7%	4.9% ± 2.1%
Number	1.0% ± 0.9%	0.5% ± 0.7%	1.5% ± 1.2%
Date	0.0% ± 0.7%	0.0% ± 0.7%	0.0% ± 0.7%
Reference	0.0% ± 0.7%	0.0% ± 0.7%	0.0% ± 0.7%

Table 2.4: Evaluation of the transcription errors made by the OCR system on 150 elements. We display the number of perfect transcription (*0 error*), the number of transcriptions which contains at most one character wrong, and the average character-error-rate (*CER*). Also, we compute the same metrics by ignoring punctuation or blank-spaces errors.

Field	Standard			Normalized		
	0 error	≤ 1 error	CER	0 error	≤ 1 error	CER
Author	77.3%	96.62%	2.04%	83.8%	97.3%	1.50%
Description	77.3%	93.92%	1.35%	85.8%	96.6%	0.93%

represent faithfully the original document is not acceptable. As such, being able to always go back to the primary source is a concern that will be addressed in the last chapter of the thesis.

Application of the pipeline

The described pipeline was applied on the 330’000 scanned documents (2.6TB), corresponding to the complete standardized part of the photo archive of the Cini Foundation. Processing the collection took almost a week on a 48-cores machine with two GPUs, mainly because of the high resolution of the original files involved. Indeed, the most costly operations were decoding the large JPEG files, cropping and saving them, whereas the processing part (OCR, segmentation) was done at a lower resolution.

At the end of the processing, we managed to convert each original high resolution scans to three elements:

- an image containing the primary source with *all* the relevant information, including possible handwritten notes or corrections of valuable historical interest.
- an image only containing the photograph.
- a dictionary of key-value pairs representing the read metadata entry of the document.

2.3 Aligning artist names with a knowledge database

The third and last step of our physical-to-digital pipeline is to link the information with a knowledge database. Indeed, even with the successful parsing process of the previous step, the information we extracted is still limited. More precisely, all what we transcribed is raw text, words which by themselves have no significant meaning and do not tell us much about what was digitized. For instance, which artists are represented in the collection and how many photographs about their work is there? What if a user only wants to look at the images from Italian painters? or from the 16th century? These are questions we cannot answer with what we extracted.

To get additional context, the Author field is the most interesting to look at. Indeed, even if it is just the name of an artist, for an Art Historian this very name automatically gets associated with the corresponding time period, as well as the city where this artist was active, already giving a rough context about the creation of the artwork. However, in order to be able to harvest that knowledge, we need a form of external database indexing artists. Being able to match the extracted artist names with the corresponding entry in the database is then a problem of *record-linkage*.

In the following section, we will first cover the useful open databases for artist names available online, and then expose the specific challenges we face in trying to align the extracted text of the Cini collection with a well-formatted online database. A semi-supervised method is then proposed and evaluated.

2.3.1 Knowledge databases for artists

In the landscape of Linked Open Data, there are a couple of research institutions which have organized their internal information about artists and made it available on the Internet. We will focus here on the biggest one available, coming from the Getty Institute in Los Angeles, and on a different approach taken by Wikidata.

Union List of Artist Names

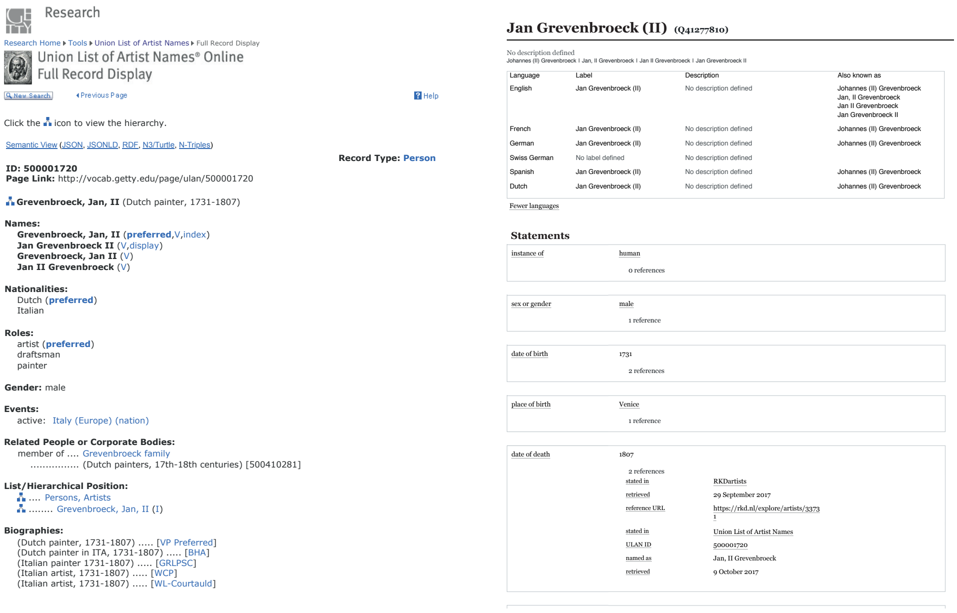
Originally created in 1984 for the needs of merging and coordinating controlled vocabulary resources inside the J. Paul Getty Trust's projects. The Union List of Artist Names (ULAN) evolved over the years through the supervised contributions of numerous dedicated editors. If it was available as a hard copy at a time, it is now possible to completely query the content of the database.

Constantly evolving thanks to a custom built editorial system allowing staff to edit information, it contains information about more than 192'000 artists.

Wikidata

Wikidata is, according to its own Wikipedia page, "a collaboratively edited knowledge base hosted by the Wikimedia Foundation. It is intended to provide a common source of data which can be used by Wikimedia projects such as Wikipedia, and by anyone else, under a public domain license." It was created much more recently in October 2012, and has not even officially reach maturity yet.

It is based on a document-oriented database, where each item (representing a topic) contains a list of statements. Statements are key-value pairs representing facts, and are validated by external references in a very wikipedian fashion. As such, for the domain of Art History knowledge, Wikidata aims to work as an aggregate of the multiple data sources available from institutions.



Research
Research Home ▸ Tools ▸ Union List of Artist Names ▸ Full Record Display
Union List of Artist Names® Online
Full Record Display
New Search
◂Previous Page
Help

Click the icon to view the hierarchy.

Semantic View (JSON, JSONLD, RDF, N3/Turtle, N-Triples)

Record Type: Person

ID: 500001720
Page Link: http://vocab.getty.edu/page/ulan/500001720

Grevenbroeck, Jan, II (Dutch painter, 1731-1807)

Names:
Grevenbroeck, Jan, II (**preferred**,V,index)
Jan Grevenbroeck II (V,display)
Grevenbroeck, Jan II (V)
Jan II Grevenbroeck (V)

Nationalities:
Dutch (**preferred**)
Italian

Roles:
artist (**preferred**)
draftsman
painter

Gender: male

Events:
active: Italy (Europe) (nation)

Related People or Corporate Bodies:
member of .... Grevenbroeck family
.... (Dutch painters, 17th-18th centuries) [500410281]

List/Hierarchical Position:
.... Persons, Artists
.... Grevenbroeck, Jan, II (I)

Biographies:
(Dutch painter, 1731-1807) ..... [VP Preferred]
(Dutch painter in ITA, 1731-1807) ..... [BHA]
(Italian painter 1731-1807) ..... [GRLPSC]
(Italian artist, 1731-1807) ..... [WCP]
(Italian artist, 1731-1807) ..... [WL-Courtauld]

Jan Grevenbroeck (II) (Q41277810)

No description defined
Johannes (II) Grevenbroeck | Jan, II Grevenbroeck | Jan II Grevenbroeck | Jan Grevenbroeck II

Language	Label	Description	Also known as
English	Jan Grevenbroeck (II)	No description defined	Johannes (II) Grevenbroeck Jan, II Grevenbroeck Jan II Grevenbroeck Jan Grevenbroeck II
French	Jan Grevenbroeck (II)	No description defined	Johannes (II) Grevenbroeck
German	Jan Grevenbroeck (II)	No description defined	Johannes (II) Grevenbroeck
Swiss German	No label defined	No description defined	
Spanish	Jan Grevenbroeck (II)	No description defined	Johannes (II) Grevenbroeck
Dutch	Jan Grevenbroeck (II)	No description defined	Johannes (II) Grevenbroeck

Fewer languages

Statements

instance of — human
0 references

sex or gender — male
1 reference

date of birth — 1731
2 references

place of birth — Venice
1 reference

date of death — 1807
2 references

stated in	RKDartists
retrieved	29 September 2017
reference URL	https://rkd.nl/explore/artists/3373 1
stated in	Union List of Artist Names
ULAN ID	500001720
named as	Jan, II Grevenbroeck
retrieved	9 October 2017

Figure 2.17: Partial web views of the corresponding pages of the same artist (Grevenbroeck Jan) on the ULAN (left) and Wikidata (right). Organized information can be seen on both of them, including possible naming variations. One can notice on Wikidata the expanded statement about the date of death which shows from which other knowledge database it is basing its information on.

Comparison

Both knowledge systems can be queried with SPARQL queries which makes relatively easy the possibility to scrape the necessary information we need[7]. We extracted all the artists born before 1900 on both systems (though only engravers, painters, sculptors, and architects for Wikidata, as it has a much broader coverage with musicians, poets, etc.). This process gave us 127'959 entities for ULAN and 75'734 for Wikidata.

Eventually, we found the additional coverage provided by the surplus of entities from ULAN to be crucial for getting better results. However, given the novelty of the Wikidata system, and the fact that it is gathering information from more than one institution (including ULAN), we expect it to grow quickly and be a better choice in the future[8].

2.3.2 Challenges

When one needs to align names with a knowledge database, typos and OCR errors are a common problem. But working with an old Italian photo-archive, we found that there were three additional challenges that we needed to overcome.

Names variation

The first one is about the fact that the same artist might be recorded under different representations, which might not be in the knowledge database. Multiple reasons can be the cause of it:

- For instance, one artist might sometimes be referred with his pseudonym, or via his full name (i.e CANALETTO against GIOVANNI ANTONIA CANAL).
- Regional variations are also present as well, and are a distinct characteristic of the Italian language. The first name GIOVANNI BATTISTA, for example, might be also be recorded as GIANBATTISTA, GIAMBATTISTA, GIOVAMBATTISTA, GIAN BATTISTA, etc.
- Finally some names might be totally not standard, coming for instance from a translation of foreign name, the French painter JACQUES CALLOT being often mentioned as GIACOMO CALLOT.

---

[7] However, the ontologies and organization of the information is quite different, hence the requests involved are completely different.

[8] A quick comparison at the time of writing shows that the number of entities on Wikidata has already increased to 173'862, which is more than twice the number we were getting at the time of these experiments!

Implicit knowledge

Conversely, sometimes the naming might actually not be precise enough. This is related with the pragmatics of the annotation process. Understanding that if one archivist writes LEONARDO on a file, he or she is referring to LEONARDO DA VINCI implies modeling a series of implicit assumptions which are changing based on the local cataloging practices. As such, a simple algorithm trying to directly find a correspondence in the database of artists will not be able to“guess" what is by far the most likely outcome, not knowing which Leonardo is LEONARDO.

This can happen when multiple people have been referred to in the same way at different points of time, and it is especially the case in our task at hand because of the families of painters which existed in Italy. For instance, TIZIANO VECELLIO could technically refer to the well known TIZIANO, or his relative TIZIANELLO, but no one would doubt that without further precision, it corresponds to the much more famous older one.

Sometimes, it is also due to the choices made for this very collection. For instance, there are three DAVID TENIERS part of the same painting family[9]. Despite the second one being the most famous and having produced the most, he is referred as DAVID TENIERS (IL GIOVANE) in the collection with the simpler naming reserved for his father[10].

Compositional structure

The last challenge is linked with the practice of archivists to describe particular unknown authors using specific syntactic processes in order to refer to workshop productions, copies and unattributed works related to a more well-known artist. Some examples include TIZIANO (BOTTEGA DI-), MICHELANGELO (COPIA DA), or LEONARDO (SCUOLA DI-). In the case of the Cini collection, this is a relatively consistent process in the form of “<Artist-name> (<modifier>)”.

Such modifiers do not only give a connection to an identified person but also qualify a relationship between the unknown author of the artwork, and the mentioned artist. For instance, the *modifier* used describes how strongly an artist was involved in the creation process of a painting, or whether the artwork is a copy of a Master’s work. In some cases, it also allows to argue whether or not the artwork was produced in the same location and at the same time the mentioned artist was alive (SOTTO LA DIREZIONE DI compared to *imitatore di*).

Finally, the *modifier* can also qualify the confidence of the attribution (DAVID TENIERS, DAVID

---

[9] More precisely the oldest (1582-1649) is the father of the second (1610-1690), himself father of the third (1638-1685).

[10] Alternatively, the spelling DAVID TENIERS I and DAVID TENIERS II might also be found, with a very inconsistent DAVID TENIERS (IL TERZO) for the rare works of the grandson.

TENIERS (ATTR), DAVID TENIERS (?)).

2.3.3 Matching approach

In order to deal with all the challenges presented above, we designed the following two stages approach (see Figure 2.18):

- First a global dictionary linking a name with the corresponding entity is created from the scraped data and additional collection-specific knowledge.
- This dictionary is used to perform a first-pass of matching on all the input names which needs to be matched.
- The names which have been matched during the first-pass act as a basis of candidates used to generate a second dictionary using a pre-defined grammar encoding the possible modifiers used in this collection.
- A second and final matching pass is performed between the second dictionary and the remaining non-matched elements.

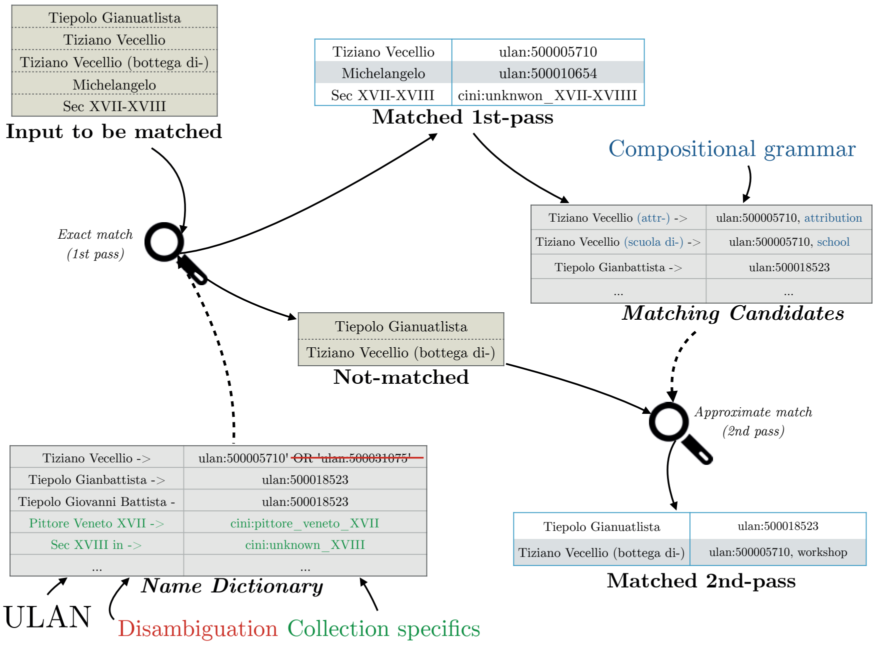

Tiepolo Gianuatlista
Tiziano Vecellio
Tiziano Vecellio (bottega di-)
Michelangelo
Sec XVII-XVIII
Input to be matched
Tiziano Vecellio | ulan:500005710
Michelangelo | ulan:500010654
Sec XVII-XVIII | cini:unknwon_XVII-XVIIII
Matched 1st-pass
Compositional grammar
Exact match (1st pass)
Tiziano Vecellio (attr-) -> | ulan:500005710, attribution
Tiziano Vecellio (scuola di-) -> | ulan:500005710, school
Tiepolo Gianbattista -> | ulan:500018523
... | ...
Matching Candidates
Tiepolo Gianuatlista
Tiziano Vecellio (bottega di-)
Not-matched
Approximate match (2nd pass)
Tiziano Vecellio -> | ulan:500005710' ~~OR 'ulan:500031075'~~
Tiepolo Gianbattista -> | ulan:500018523
Tiepolo Giovanni Battista - | ulan:500018523
Pittore Veneto XVII -> | cini:pittore_veneto_XVII
Sec XVIII in -> | cini:unknown_XVIII
... | ...
Name Dictionary
ULAN
Disambiguation
Collection specifics
Tiepolo Gianuatlista | ulan:500018523
Tiziano Vecellio (bottega di-) | ulan:500005710, workshop
Matched 2nd-pass

Figure 2.18: Schema of the alignment process.

First dictionary construction and match

The scraped data from the knowledge database gave us a list of possible names for a given entity. We augment the list of names with *generative substitutions* (for instance replace THE YOUNGER by IL GIOVANE, if GIROLAMO is present, add the version with GEROLAMO). Also, additional names are added which are often very specific to the Italian language. For example, JAN GREVENBROECK is recorded as JAN GREVEMBROCH in the Cini. These actions aim at improving the coverage of the data we have, and tackle the issue of the multiple naming.

Related to this coverage problem is the set of unknown attributions, but that still contains information. For instance PITTORE VENETO XVII or SEC XVIII IN are providing spatial and temporal information. Because the date range is relatively standard (SEC VI, SEC XVI-XVII, SEC XIV M[11]), one can generate them all, and combine them with a list of *generic unknowns* which can be added to the dictionary as 2'089 special Cini-specific unknown entities.

In total, we get a list of 398'832 different possible names. However, at the end of building this dictionary, the "implicit knowledge" problem appears as the dictionary not being an injective function: some names correspond to a list of possible entities. To resolve some of them we build a list of *disambiguations* that override, for a given name, the entity it should be matched with.

The exact match step between the input elements and this built dictionary is done by normalizing each text string on both side (removing punctuation and special characters, as well as bringing every character lowercase) before an exact comparison. If no match is found with the normalized exact comparison, we fall back to a bag-of-words comparison, which allows for instance to match LORENZO DI CREDI with DI CREDI LORENZO. For any match found, we add the original field with the corresponding matched entity to a *matching dictionary*.

Second dictionary construction and match

The second round of matching is based on using the elements that were successfully during the first phase. The assumption is to use these found matches as an indication of the way persons are referred in the collection, and to propagate these names recognition. A list of possible *modifiers* which can characterize the attribution are used in conjunction with the *matching dictionary* to generate 288'269 candidates for the second pass matching.

Again for the matching process, both the input element and the matching candidate are normalized before comparison. However, in this stage we actually use approximate matching

---

[11] For the non-initiated, like I was, SEC XIV IN, SEC XIV M, SEC XIV EX represents respectively the beginning, middle and end of the 14th century.

in order to correct the OCR errors. We compute the Levheinstein distance[12] between an input element and a candidate, a small ratio between the distance and the length of the compared strings is considered a match. As we have 105’121 unique elements left to be matched, the number of pairs $(input\_element, candidate)$ to be compared is greater than 30 billions. While this represents a large computational effort, by using a very efficient implementation of the Levheinstein distance and by spreading the work on 48-cores, the process can be done in less than 2 hours.

Flexibility of the approach

To recapitulate, during the process, five configurable inputs are used:

- the generative substitutions (THE YOUNGER <-> IL GIOVANE)
- the additional namings (JAN GREVEMBROCH -> ULAN:500001720)
- the possible base unknowns (PITTORE VENETO, MOSAICISTA)
- the disambiguations (LEONARDO -> ULAN:500010879)
- the modifiers (BOTTEGA DI, IMITATORE DI)

All these parameters are simple spreadsheets that anybody can modify based on the missed elements to improve the quality of the results. For instance, one can look at the most common ambiguous situations and just add the corresponding resolving elements. This allows a form of active cleaning of the data, where we display the most common elements which were not successfully matched so that an expert can modify the configurable lists, thus improving the coverage of the matching process.

2.3.4 Results

Coverage

The coverage results can be seen on Table 2.5. As we can see, almost 50’000 elements (14.6%) actually have an empty Author field, most of them not being artworks but photographs of remote Italian villages or aerial photographs of Venice. We see the usefulness of the two-passes system which allows to bump the coverage, matching 35’000 additional elements. Eventually,

---

[12] The Levenshtein distance is a string metric for measuring the difference between two sequences of characters. It is representing the minimum number of single-character edits (insertions, deletions or substitutions) required to change one string into the other. For instance, LEONARDO and LAONARDOO are separated by a Levenshtein distance of 2 (1 substitution and 1 insertion).

Table 2.5: Coverage obtained by the matching process.

	Number of elements	Relative to non-empty
Total	330’078	-
Total non-empty	282’606	100%
Matched 1st pass	173’571	61.6%
Ambiguous non-resolved	3’882	1.4%
Matched 2nd pass	208’510	73.8%

the total coverage obtained, with 73.8% of the non-empty elements being matched with an actual entity, is relatively satisfactory considering the difficulty of the task at hand.

Missed matches

Out of the 74’000 elements which are not properly matched. We can distinguish two categories.

The first category consists in Author names which may have been matched if the algorithm were to be improved (e.g. in terms of author name variation or possible compositional structure). It is predominant in the elements where the process was not successful. Apart from large OCR errors, the most typical unmatched strings correspond to collective works in which several authors are named. For instance, the string BASSANO JACOPO E FRANCESCO (father and son) corresponds to 134 records. Given the way we tackled the matching problem, we do not allow mentions of multiple people at the same time, which is definitely a current limitation, but could be solved in the future.

The second category of elements which were not matched with ULAN, are in fact not a product of misalignment but represent people, often minor artists, who are not recorded in the knowledge database. In the present study, a number of artists who do not feature in ULAN were uncovered in the Cini archive. These include, AUGUSTO CARATTI, a minor artist from nineteenth-century Padua, who is represented by 65 works in the Cini collection, and NATALE MELCHIORI an early eighteenth-century painter from Castelfranco, Veneto, represented by 39 works, who also has a street named after him in Treviso. Another artist who does not feature in the ULAN database but nevertheless has a significant presence in the Cini archive with 110 drawing, is ANTONIO CONTESTABILE, an eighteenth-century draftsman from Piacenza.

Global views

A proper analysis of the photo-collection is outside the scope of this project. However, with 74% of the assignments of the images done, we are in a unique position to look for the first time in a quantitative way at the global content of the collection.

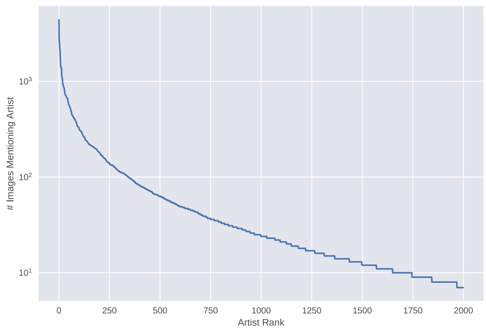


Figure 2.19: Distribution of number of images assigned for each artist.

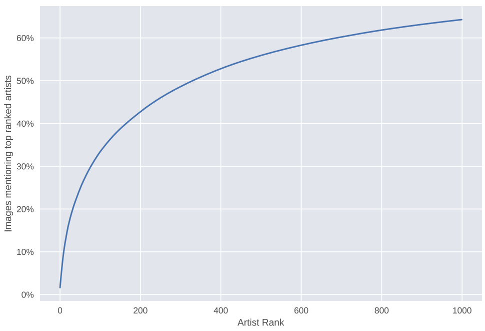


Figure 2.20: Proportion of images assigned with respect to the most common artists. The 200 most represented artists represent 43% of the collection.

First, we can look at the distribution of images per artist (Figure 2.19). One can notice that the decay of number of images assigned to an artist is surprisingly steep: 18 artists with at least 1'000 images, 346 with more than 100, and 1'746 creators with 10 images assigned to them. This shows a very uneven representation of the artists (as can be seen on Figure 2.20), with the 200 most common persons representing 43% of the collection. In these top artists, we unsurprisingly encounter the most famous Venetian painters (see Table 2.6).

Another information we can get from the knowledge database is the spatial and temporal context, based on the dates of birth/death, and the citizenship of the creators. On Figure 2.21 , we can see that a large majority of the mentioned artists are Italian of the 16th and 17th century. This is coherent with the rationale behind the creation of the collection of gathering photographs about Venetian art.

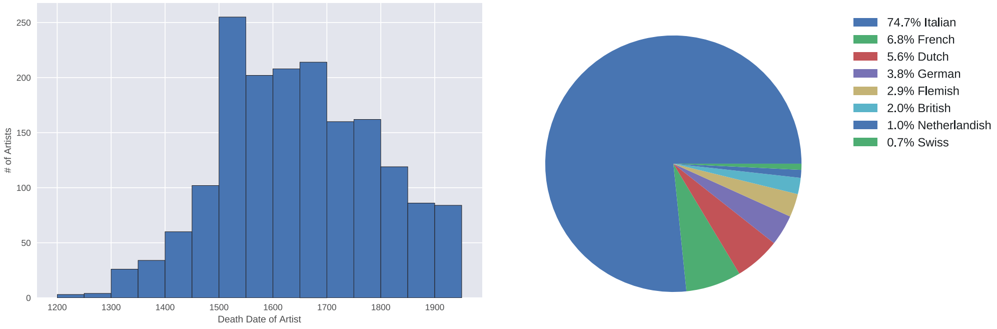
2.4 Genericity of the proposed pipeline

In the previous sections, we presented our solution for the automatic processing of the digitized part of the photo collection. However, this only corresponded to the standardized part of the archive, which are photographs glued on structured pieces of cardboard. A similar standardized format can be found in other institutions (and in that case the application of our process is almost direct), but a majority of photographs in these archives are in a much more heterogeneous state. In this section, we discuss the potential genericity of our approach and the likelihood of success of the different steps of the pipeline.

We do not consider here the challenges related to the scanning process, as the scanning infrastructure is already very capable of coping with the complexity of the documents of the

Artist	Number images
Tiepolo, Giovanni Battista	4419
Guardi, Francesco	2848
Tintoretto, Jacopo	2576
Buonarroti, Michelangelo	2517
Veronese, Paolo	2184
Titian	2174
Raphael	1788
Palladio, Andrea	1781
Giotto	1446
Bellini, Giovanni	1443
Palma, Jacopo, il giovane	1393
Canaletto	1383
Tiepolo, Giovanni Domenico	1376
Carracci, Annibale	1361
Piranesi, Giovanni Battista	1137
Grevenbroeck, Jan, II	1113
Carpaccio, Vittore	1079
Magnasco, Alessandro	1067
Canova, Antonio	991
Bartolommeo, Fra	938

Table 2.6: Most common artists in the digitized collection.



(a) Temporal distribution

(b) Spatial distribution

Figure 2.21: Temporal and spatial distribution of the 1'746 artists mentionned by at least 10 documents.

archive. When it comes to the rest of the processing pipeline however, we believe it can be reduced to the following three key problems:

- image analysis of the document, which includes detecting and extracting the visual and textual components of the scan.
- being able to read the textual parts. While for typewritten text, it is mostly solved, most of the information on the documents actually is handwritten.
- extracting and organizing the read textual information.

2.4.1 dhSegment: a generic approach for document segmentation

As we described previously, we used a segmentation-based technique leveraging a pre-trained CNN in order to separate the different visual components (photograph/cardboard) of the scan image. Despite proving very successful, one can wonder if such an approach generalizes to other document processing tasks gracefully, or if our solution was only tailored to our original problem. Indeed, while the variability of document processing tasks is large, research works often focus on a single problem. As such, we believe a generic solution for historical document processing is valuable for the research community.

Our original approach was based on a two-stages process: (i) predicting local-characteristics at a pixelwise level, and then (ii) performing some simple post-processing operations on the predicted probabilities. The goal of this subsection is to generalize this approach so that it can apply to other types of document processing tasks, creating the dhSegment framework.

The same training procedure, resizing strategy, and network architecture is used in order to generate the class probabilities for each pixel. However, because each problem asks for a slightly different output, we allow ourselves some simple operations to be performed on the generated probability map:

- thresholding operations: in order to convert the probability maps to binary maps. Can be done class-wise if multiple labels are predicted per pixel.
- morphological operations: used for simple cleaning of the predictions.
- connected components analysis: separating a binary image to its separate components is useful for distinguishing different objects, and maybe filtering some of them based on their size.
- shape vectorization: a vectorization step is needed in order to transform the detected region into a set of coordinates. To do so, the blobs in the binary image are extracted

as polygonal shapes. In fact, the polygons are usually bounding boxes represented by four corner points, which may be the minimum rectangle enclosing the object or quadrilaterals. The detected shape can also be a line and in this case, the vectorization consists in a path reduction.

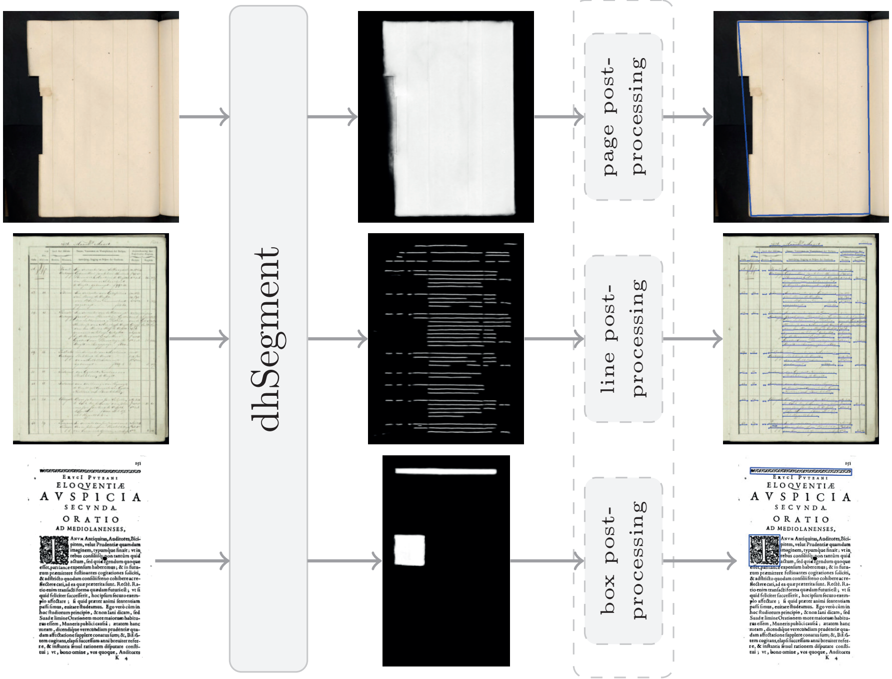


Figure 2.22: The document processing pipeline resulting from the generalization of the method employed for the processing of the scans. The documents (left) are processed through a generic neural architecture that is trained for the specific task. The predicted probabilities (middle) are then converted to the desired output (mask, regions, polygons, etc.)

We evaluated this simple and generic strategy on three different scenarios which we present below. For each experiment, five independent training were done and every evaluation metric includes its computed variance.

Layout analysis for Medieval Manuscripts

Document Layout Analysis refers to the task of segmenting a given document into semantically meaningful regions. In the experiment, we use the DIVA-HisDB dataset[74] and perform the task formulated in [75]. The dataset is composed of three manuscripts with 30 training, 10 evaluation and 10 testing images for each manuscript. In this task, the layout analysis focuses on assigning each pixel a label among the following classes : text regions, decorations, comments and background, with the possibility of multi-class labels (e.g. a pixel can be part of the main-text-body but at the same time be part of a decoration).

The system is trained to directly predict for each pixel the classes it is part of. A threshold of 0.5 is used, and small separated components are removed. The results we get are competitive with the other contestants of the competition.

Table 2.7: Results for the ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts [75] - Task-1 (IoU)

Method	CB55	CSG18	CSG863	Overall
System-1 (KFUPM)	.7150	.6469	.5988	.6535
System-6 (IAIS)	.7178	.7496	.7546	.7407
System-4.2 (MindGarage-2)	.9366	.8837	.8670	.8958
System-2 (BYU)	.9639	.8772	.8642	.9018
System-3 (Demokritos)	.9675	.9069	.8936	.9227
dhSegment	.974±.001	.928±.002	.905±.007	.936±.004
System-4.1 (MindGarage-1)	.9864	.9357	.8963	.9395
System-5 (NLPR)	.9835	.9365	.9271	.9490

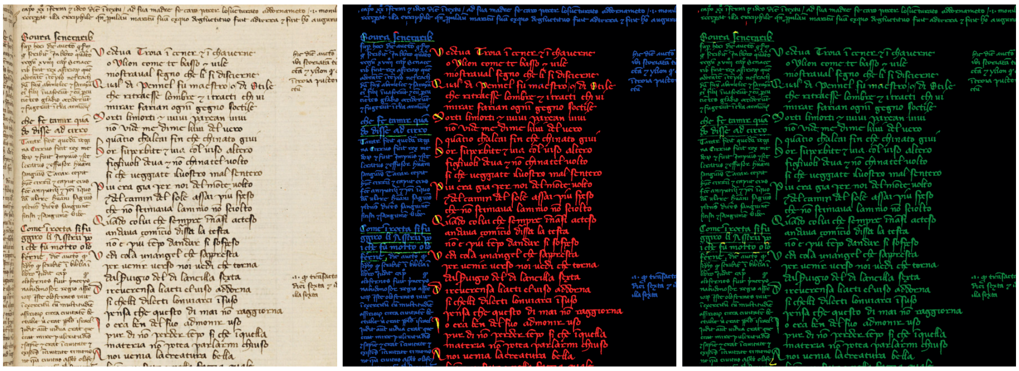
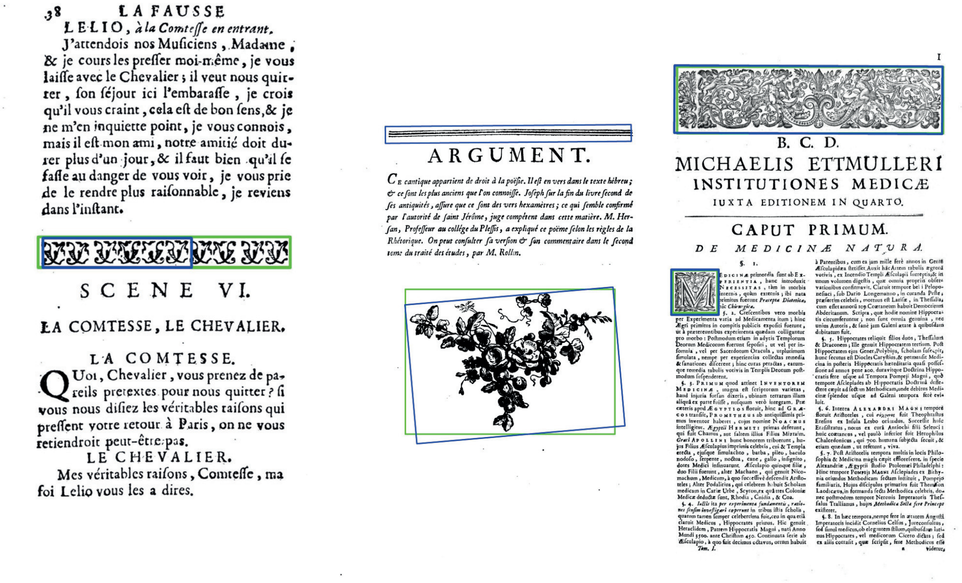


Figure 2.23: Example of layout analysis on the DIVA-HisDB test set. On the left the original manuscript image, in the middle the classes pixel-wise labeled by the dhSegment and on the right the comparison with the ground-truth (refer to the evaluation tool[76] for the signification of colors, green means perfect prediction)

Ornaments extraction

The second task is about extracting ornaments in printed books of the digitized collection of the Bibliothèque Cantonale Universitaire (BCU) of the University of Lausanne. Ornaments are of interest to book historians in order to track the productions of printers, or identify copies of the same edition.

Bounding box annotations for 912 pages (612 containing at least one ornaments) were acquired as part of a Master Thesis [77], where the student was using a combination of region proposal techniques coupled with a CNN classifier to filter the false positives. We train a network to predict for each pixel if it is part of a bounding box annotation of an ornament or not. Then by simply thresholding the predicted probabilities, we get a binary mask where each connected component is converted to a rectangle region. This proved much more effective than the original approach of [77] where a complex combination of region proposal techniques were coupled with a CNN classifier to filter the false positives.



38 LA FAUSSE

LELIO, *à la Comteſſe en entrant.*

J'attendois nos Muſiciens, Madame, & je cours les preſſer moi-même, je vous laiſſe avec le Chevalier; il veut nous quitter, ſon ſéjour ici l'embaraſſe, je crois qu'il vous craint, cela eſt de bon ſens, & je ne m'en inquiette point, je vous connois, mais il eſt mon ami, notre amitié doit durer plus d'un jour, & il faut bien qu'il ſe faſſe au danger de vous voir, je vous prie de le rendre plus raiſonnable, je reviens dans l'inſtant.

SCENE VI.

LA COMTESSE, LE CHEVALIER.

LA COMTESSE.

QUoi, Chevalier, vous prenez de pareils pretextes pour nous quitter? ſi vous nous diſiez les véritables raiſons qui preſſent votre retour à Paris, on ne vous retiendroit peut-être pas.

LE CHEVALIER.

Mes véritables raiſons, Comteſſe, ma foi Lelio vous les a dires.

ARGUMENT.

CE cantique appartient de droit à la poëſie. Il eſt en vers dans le texte hébreu; & ce ſont les plus anciens que l'on connoiſſe. Joſeph ſur la fin du livre ſecond de ſes antiquités, aſſure que ce ſont des vers hexamètres; ce qui ſemble confirmé par l'autorité de ſaint Jérôme, juge compétent dans cette matière. M. Herſan, Profeſſeur au collége du Pleſſis, a expliqué ce poëme ſelon les règles de la Rhétorique. On peut conſulter ſa verſion & ſon commentaire dans le ſecond tome du traité des études, par M. Rollin.

I

B. C. D.

MICHAELIS ETTMULLERI

INSTITUTIONES MEDICÆ

IUXTA EDITIONEM IN QUARTO.

CAPUT PRIMUM.

DE MEDICINÆ NATURA.

Figure 2.24: The left image illustrates the case of a partially detected ornament, the middle one shows the detection of an illustration but also a false positive detection of the banner and the right image is a correct example of multiple ornaments extraction.

Table 2.8: Ornaments detection task. Evaluation at different IoU thresholds on test set

Method	IoU threshold	F-val	P-val	R-val	mIoU
[77]-config1	0.5	0.560	0.800	0.430	-
[77]-config2	0.5	0.527	0.470	0.600	-
dhSegment	0.7	0.94±.023	0.96±.036	0.92±.013	0.87±.016
	0.8	0.87±.033	0.84±.049	0.91±.016	
	0.9	0.56±.054	0.42±.053	0.83±.036	

Textline extraction

A last and maybe more challenging example of the flexibility and efficiency of our network for historical document processing is the task of detecting the baselines of handwritten text, an important first step for handwritten recognition.

Here the network trains directly to predict which pixels are in a small (5pixels) radius of the annotated training baselines. The predicted probability map is then filtered with a Gaussian filter ($\sigma = 1.5$) before using hysteresis thresholding[13] ($p_{high} = 0.4$, $p_{low} = 0.2$). The obtained binary mask is then decomposed in connected components, and each component is converted to a polygonal line.

The READ-BAD dataset [78] was used, and we put our results in comparison with the competitors of the cBAD: ICDAR2017 Competition [79]. Despite the simplicity of our approach compared to the other domain-specific methods, we still equal the state-of-the-art on the Complex Track of the competition, and are competitive on the Simple Track.

Table 2.9: Results for the cBAD : ICDAR2017 Competition on baseline detection [79] (test set)

Method	Simple Track			Complex Track		
	P-val	R-val	F-val	P-val	R-val	F-val
LITIS	0.780	0.836	0.807	-	-	-
IRISA	0.883	0.877	0.880	0.692	0.772	0.730
UPVLC	0.937	0.855	0.894	0.833	0.606	0.702
BYU	0.878	0.907	0.892	0.773	0.820	0.796
DMRZ	**0.973**	**0.970**	**0.971**	**0.854**	0.863	**0.859**
dhSegment	0.88±.023	**0.97**±.003	0.92±.011	0.79±.021	**0.95**±.005	**0.86**±.011

Conclusion

Through these experiments, we proved that the same generic deep learning approach can be applied to a wide range of document processing task and perform competitively specialized state-of-the-art methods.

What these experiments also show, is that even if we do not have (yet) more complex digitized documents from photo archives to test, the quality of the results give us confidence that extracting complex visual and textual elements from the scanned images is a challenge we already have good solutions for.

---

[13] Applying thresholding with p_{low} then only keep connected components which contains at least a pixel $\geq p_{high}$

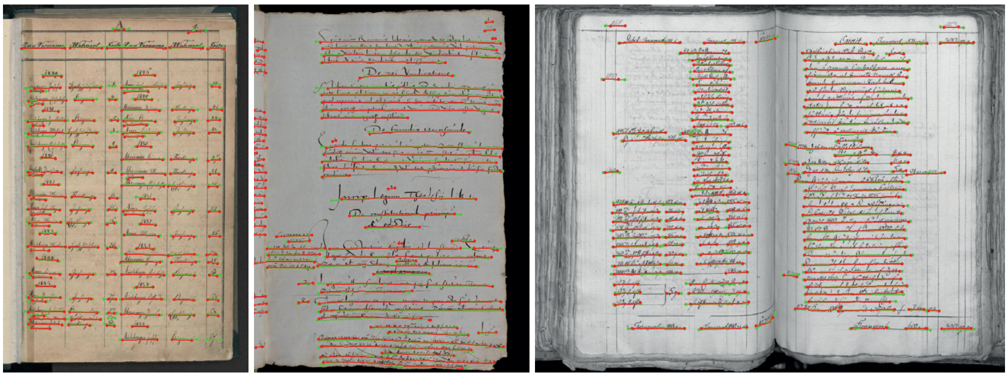


Figure 2.25: Examples of baseline extraction on the complex track of the cBAD dataset. The ground-truth and our predicted baselines are displayed in green and red respectively. Some limitations of the simple approach we propose can be seen here, for instance detecting text on the neighboring page (top right), or merging close text lines together (bottom and top left). These issues could be addressed with a more complex pipeline incorporating, for instance, page segmentation or by having the network predict additional features, but this goes beyond the scope of this experiment.

2.4.2 The state of handwritten text recognition

The second step for the generalization of our approach is about reading the extracted text areas. During the processing of the Cini, we showed that OCR software for recognizing typewritten was mature and performing well, but we ignored the problem of transcribing the handwritten information as it was only present on a small portion of the documents. However, for the rest of the Cini photo archive (the 700'000 remaining photographs), the text is most often in handwritten and not typewritten format, which leads us to wonder if we would be able to extract this information.

In a standard handwritten transcription pipeline, the first step is to identify and locate the lines of handwritten text, despite a possibly complex layout. This task was actually already evaluated as part of the third experiment of the previous subsection, where we showed the generality of our document segmentation pipeline. While imperfect, the performance is already good enough to be used in practice.

Given the detected lines in a document, the second step is to convert the corresponding areas to their corresponding textual form. While it is a difficult problem, it is another area where deep-learning methods have made tremendous improvements in the last years. For example, working with the difficult manuscripts of the Venetian State Archive (see Figure 2.26), researchers could show that an automatic transcription system could outperform good amateur transcribers [80], reaching a character-error-rate of 7.2%.

More generally, reports from the READ[14] project [81] indicate large improvements by using deep learning architectures, especially in the case where the training data contains examples coming from the same writer as the testing data.

Fortunately, when transcribing an archive, one usually annotates training examples directly on documents coming from the collection itself, which corresponds to the case where the same writing hands are present in the training and testing data. Moreover, for photo-collections, the number of different writers is most likely limited to the original owners of the photographs, some archivists, and professors who felt they had the credentials to add their own notes on the documents during consultation. Thus, given the proper annotation interfaces, historians could transcribe just a subset of documents, quickly covering all the writing styles of the collection. A trained system could then be able to transcribe the rest of the documents with a limited number of errors.

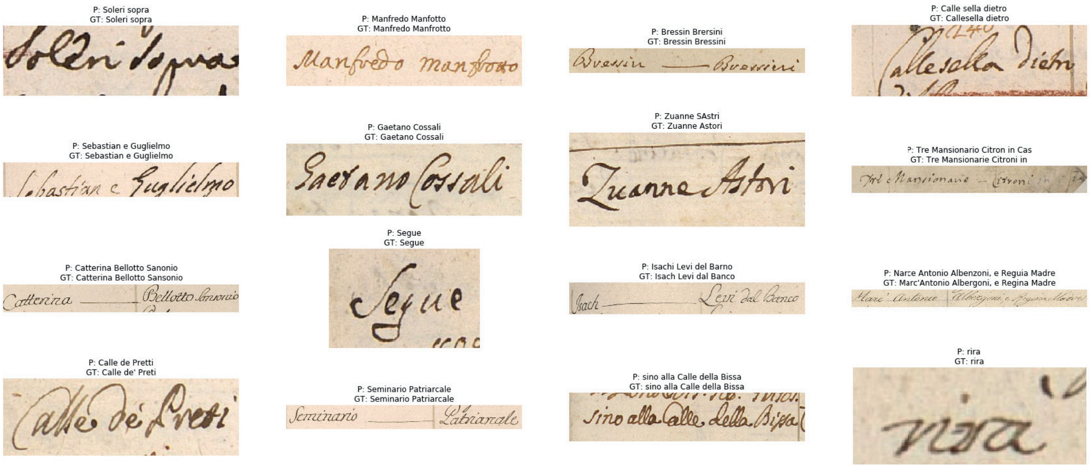


Figure 2.26: Examples of automatic transcriptions (P) compared with their groundtruth (GT) for Venetian manuscripts. Taken from [80].

There would be an obvious trade-off between the amount of annotation work and the quality of the automatic transcriptions obtained, while most likely never reaching perfect accuracy. However, even with a character-error-rate of 7%, a textual search system already allows to retrieve a significant portion of corresponding document, especially if a form of fuzzy-matching[15] is enabled.

---

[14] Standing for "Recognition and Enrichment of Archival Documents", the READ project is an European project which aims at “revolutionizing access to archival documents with the support of cutting-edge technology such as Handwritten Text Recognition (HTR)”.

[15] Modern text search engine are able to retrieve documents with limited OCR errors, for instance finding “Raffaalo” even if the query was “Raffaello”

2.4.3 Extracting the metadata elements?

Assuming textual elements were properly read, transforming the document to a list of formatted metadata (*Attribution*, *Location*, etc.) is more of an open-ended question. In our case, data was simple and formatted, as the position of the text was a clear indicator of which label corresponds to a given part of text.

Our approach could be used on other collections which share the same level of standardization. The Zeri Foundation (Bologna) for instance, uses similar storage, and the layout models can easily be adapted for every *fondo* that needs to be processed. However, in the case of imperfectly transcribed handwritten notes, extracting information is more difficult and requires higher-level textual understanding. This operation then seems a bit undecidable without knowing more about the specificity of each photo-collection.

However, we argue that properly organizing metadata is not necessary to start valorizing a digitized collection. Suppose we only have images and raw text extracted, users can still search the collection with a full-text search, allowing already navigating the digitized documents, using keywords like artist names. Also, visual search (which will be covered in the following parts of this work) allows finding relevant documents without even any transcription of their textual components. These search methods show that properly parsing the metadata is not a requirement to already unlock the knowledge of these photo-collections and make them accessible.

Conclusion

In this chapter, we considered the problem of efficiently digitizing a large photo-archive. Using the case study of the Cini Foundation, we presented an approach that was successful in transforming a physical archive to a structured digital collection. With the help of a specifically designed scanner, the first part of the collection of the Venetian institution (more than 330’000 photographs) was digitized in 13 months. Using a combination of deep learning and commercial OCR software, we showed that the visual and textual information of the scanned documents could be extracted with very good accuracy. Finally, we successfully connected a large portion of the attribution information of these photographs with a knowledge database, which allowed us to assign a spatial and geographical context to each document, effectively giving a first global view of the content of the archive.

In the last section, we considered how this pipeline could be potentially applied to other archives. We extensively demonstrated the genericity of our segmentation framework for various document-processing tasks, and argued for the possibility of being able to read the textual information, may it be typewritten or even handwritten. These observations show that the large-scale digitization of the photo archives is possible, potentially making the incredible information they contain available and searchable to the general public.

3 Organizing the Visual Information

As millions of photographs from archive get digitized, unprecedented opportunities for their use arise. Currently, most indexing and search systems in institutions are only based on metadata and tags, effectively making them blind, and ill suited for search questions like the transmission of patterns. In this chapter, we take the opposite approach of only looking at the visual information of these documents.

More precisely, assuming one is given a large collection of photographs coming from these photo archives, we interest ourselves in the task of organizing them purely based on their visual content, and leveraging modern techniques of computer vision to aid us in the process. This includes detecting photographs of the same object, and learning to search visually the collection of images.

We start by describing the type of organization we were interested in, resulting from an original Art History question combined with the specificity of dealing with images coming from photo-archives. Then we describe a mathematical formalization encoding the complexity of the visual organization into the form of graphs with *physical* and *visual* connections. In Section 3.3, we propose a global framework incorporating this formalization, machine vision, and the interface system Chapter 4.

The last two sections correspond to the applications of computer vision to our large collection. First, in Section 3.4, we interest ourselves in the task of clustering the photographs representing the exact same artwork. Then in Section 3.5, we show how we use annotated visual links in the collection, and deep learning, to train a powerful visual metric that enables visually searching the archive with improved accuracy.

3.1 Goals

3.1.1 The Original Question : Tracking Patterns in Paintings

The main goal of this research was to make photo archives searchable. The assumption being that if a tool allowed us to find similar images, it would foster Art Historical research and creates new knowledge.

However, visual similarity is an ill-defined concept. Depending on its research interest, one might consider artworks similar or not based on the color proximity, the style of the artworks, the shapes of the represented elements, or the semantic content. In this research, we focus on the tracking of “patterns” reappearing in the art production.

Examples of connections between artworks we are interested to unravel are shown on Figure 3.1. These visual connections between artworks are extremely important for Art Historians as they can give an insight about the creative process of the painters. For instance, the reuse of a specific motif might imply that the painter was exposed to the work of another artist. A larger scale example can be seen on Figure 3.2, where the same female pose is reused across a variety of different subject.

Because this problem is about the propagation of forms, and shapes, the corresponding search system will reflect this desired to be invariant to the style of the represented artwork or its medium (painting, print, drawing, etc.). As such, the visual similarity we are interested in is a specific one that does not represent the diversity of potential visual searches one could do on a corpus of images.

3.1.2 The Practical Question : the Problem of the Duplicates

Working with the data of a photo archive allows having access to a much larger corpus of images than conventional online databases. However, unlike online databases which have a nicely organized list of artworks, each represented by a single high resolution image, archives are more complicated? When gathering multiple photo-collections or data repositories, many images might correspond to the same object. This is actually very natural, as multiple people would have photographed famous artworks independently. Additionally, images of details of paintings (background, the face of the main character, a secondary character, etc.) are often recorded as a completely separate photograph as well. Finally, these duplicate photographs of a simple object might have a different description or a different attribution, depending on the original notes of the photographer.

This creates a practical problem in the exploration of images, especially with a visual search. When the results are 20 images extremely similar, it can be hard to distinguish how many

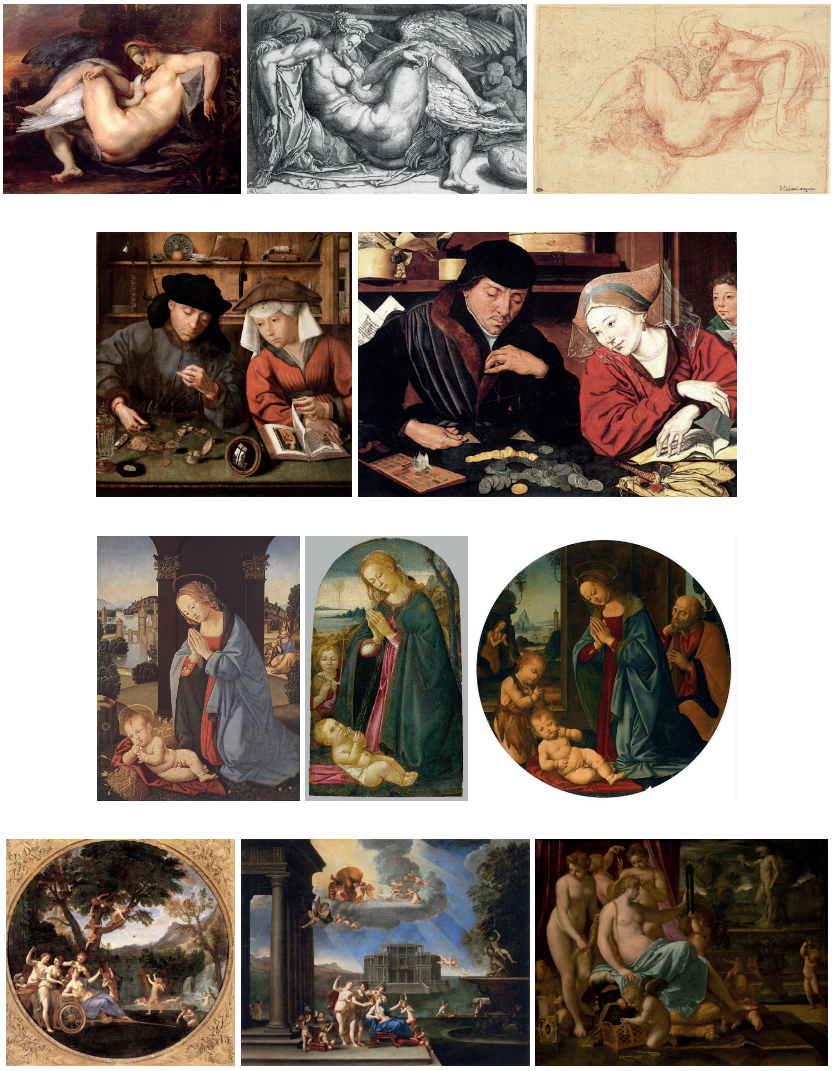


Figure 3.1: Examples of connections we are interested in being able to find. **First row** : *Leda and the swan* different mediums (RUBENS, Peter Paul : painting ; CORT, Cornelis : engraving ; BUONARROTI, Michelangelo : drawing) **Second row** : similar composition (MASSYS, Quentin *The Moneylender and his Wife* ; REYMERSWAELE, Marinus van *The Banker and His Wife*) **Tirth row** : *Adoration of the Child* different authors (DI CREDI, Lorenzo ; DEL SELLAIO, Jacopo ; DI CREDI, Tommaso) **Fourth row** : similar element in the *Toilet of Venus* (ALBANI, Francesco first two ; CARRACCI, Annibale)

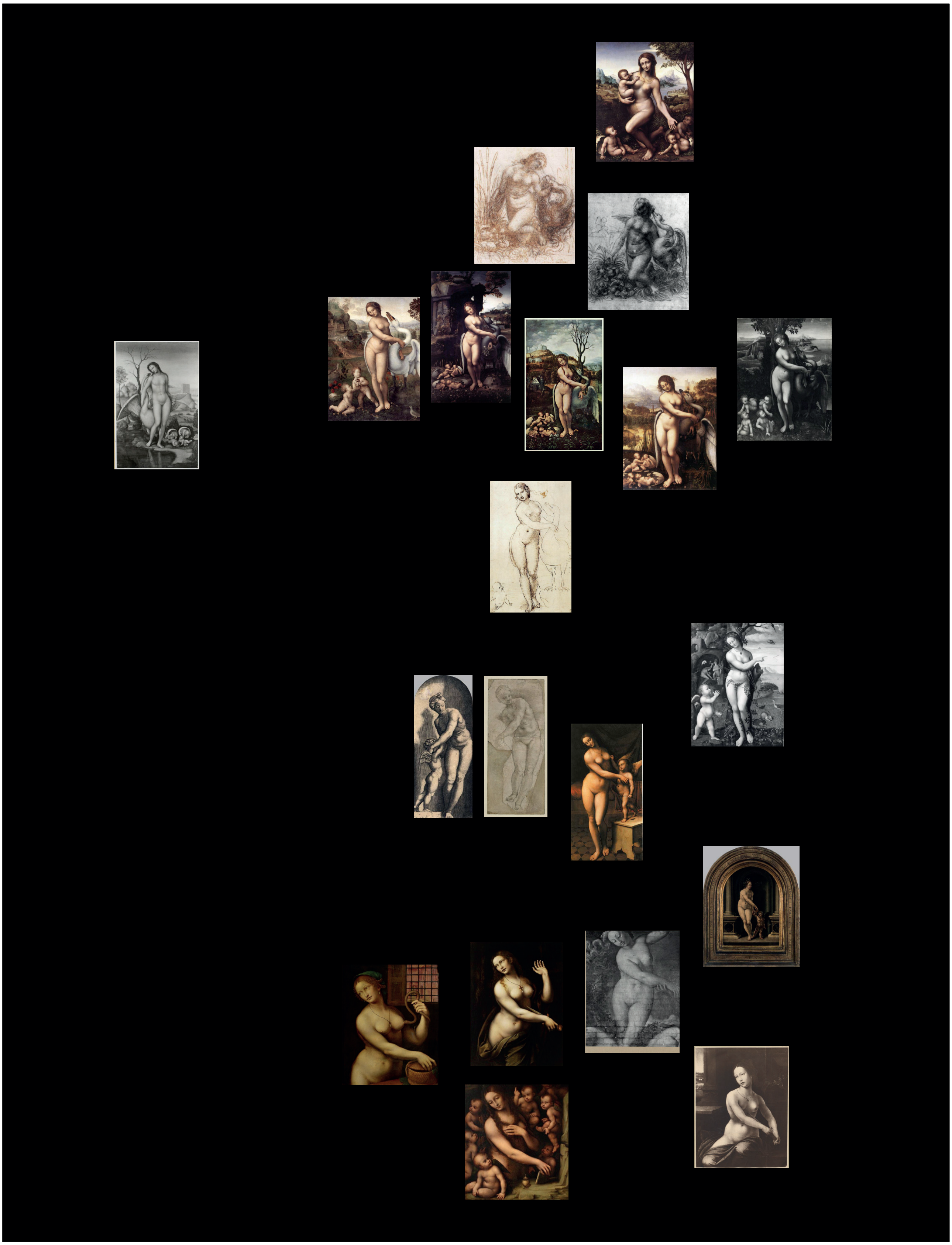


Figure 3.2: These works of art were created by nine different artists, but one can easily notice how some patterns and motifs were reused and modified. It is also interesting to note the central female character who represents Leda (and the swan) in the upper part in the visualization, then becomes Venus in the central set of images, to finally be reused as Lucretia, Cleopatra, Mary Magdalene, and even Eve in the bottom five images.

different objects there are actually, and how many are just photographs of the same artwork. Such over-representation can quickly clutter the search results, making it a painful experience to navigate a collection visually. Moreover, as we will see later, it is sometimes extremely tricky to distinguish an actual copy from a duplicate photograph.

On the other hand, being able to detect these duplicates is of great practical importance. For instance, many photographs are completely anonymous with no information of authorship, description or provenance. If one is able to match an anonymous image to a corresponding record in an existing database, proper metadata can be automatically imported, effectively describing these originally unknown photographs. Additionally, detecting duplicates images in a collection might allow to discover conflicts in the metadata (different attribution?, changed location?, etc.) creating fruitful debates.

3.2 Formalization of the Problem

The goal of this section is to propose a formalization that allows representing the complexity of both the tasks at hand. More precisely, given a collection of images, our end goal is to be able to represent information in the form of a graph, i.e. connections between the images.

3.2.1 Challenges

Ambiguity of Sameness

The task of detecting if two images are part of the same object implicitly implies a definition of what is the extent of the object itself. For instance, in the case of a triptych, if photographs of the side panels are taken, should they be considered representing the same object? Also, for certain famous paintings, multiple photographs of details are present. For example, in the case of the *Gioconda* of LEONARDO, one image focusing on the face, and one on the crossed hands. These two images would represent the same artwork, but since there is no overlap between them, it is impossible to know it just from their visual information.

Also, and this is a point that will be described further in the next section, the concept of sameness for artworks can be more fluid than one might think. For instance, what about two photographs depicting the same painting before and after a light restoration? a heavy restoration? Sometimes part of the paintings would be partially covered or expanded by other painters. In these situations, what relationship the objects (and hence the photographs representing them) have between them?

Finally, another difficult situation is the case of serial productions like engravings. Two prints might have been created from the same woodblock matrix, and then the photographs representing them can be considered a specific form of duplicates. Indeed, despite being physically separate objects, the objects share the same precise production matrix.

Formalizing Visual Influence?

As we stated above, our goal is to be able to help scholars track the propagation of patterns, motifs, and iconography in large collections of images.

First, there is a certain *diversity* of ways two paintings might appear “connected” with each other (Figure 3.1). It can be due to a global composition being reused, while local elements do not match. On the other hand, it might be due to the reuse of a similar element in both artworks, while changing the rest of the layout. Finally, we are also interested in finding copies across medium. In practice however, things are not so clear-cut as the visual relationship between two artworks is often a mix of global and local elements. As such, it is hard to define how artworks are related to each other, and to the best of our knowledge an ontology of describing visual connections does not exist.

Second, when we consider two artworks to be “related”, we are implicitly referring to the rest of the historical visual production. For instance, if we think about standard iconographies like the Baptism or the Crucifixion, most of them follow very standard representations, but we would not consider all of them connected with each other, as it is not really relevant. However, if two Crucifixion are following an extremely similar recipe, or if two non-standard compositions are sharing a moderate similarity, these are interesting pairs of images to consider. This shows the *salient* aspect of the connections we are looking for, as they arise from their local *relative* strength.

3.2.2 Formalization

This section presents an almost mathematical definition of how we represent the information in order to tackle the two problems we described.

Basis

The first thing we need to define is the working space, i.e. the basic elements we are dealing with. In our case, we made clear that we were not interested in metadata and only want to consider a large number of images. Also, we are interested in the relationships between these basic elements, which we call connections

Definition 1. *A* collection *is a set of images. If an image A is part of a collection $\mathscr{C}$, we note $A \in \mathscr{C}$.*

Definition 2. *Given a collection* $\mathscr{C}$*, a* connection *is a pair of images in* $\mathscr{C}$*. For* $A \in \mathscr{C}, B \in \mathscr{C}$*, we use the notation* $(A\text{–}B)$*. The set of all possible connections (i.e. the set of unordered pairs of images) induced by* $\mathscr{C}$ *is noted* $\mathfrak{C}(\mathscr{C})$*.*

Physical Connections

Since it is difficult to decide if two images are part of the same object, either because the extent of an object is unclear, and/or the images alone are not enough to decide if it is the case, we propose the following operationalization:

Definition 3 (Physical Connection)**.** *Given a collection* $\mathscr{C}$*, and* $A \in \mathscr{C}, B \in \mathscr{C}$*, the connection* $(A\text{–}B)$ *is said to be* physical *if it is possible to assert, from the visual information only, that they have an overlap covering the same physicality i.e. the same object or the output of the same serial production. In such a case, we use the notation* $A \cong B$*.*

According to this definition, we can resolve some of the edge cases we mentioned before. Given A and B images :

- if A is a close-up of B, then $A \cong B$.
- if A and B are non-overlapping close-ups of the same painting (for instance, hands and face of the *Gioconda*), then $A \ncong B$.
- if A and B are engravings coming from the same wood-block, then $A \cong B$.
- if A and B are engravings coming from different wood-blocks, then $A \ncong B$.

Additionally, a definition that will become useful later is the following:

Definition 4 (Physical Closure)**.** *Given a collection* $\mathscr{C}$*, the* physical closure *of an image* $A \in \mathscr{C}$ *is the set of images for which there is path of physical connection leading to A. More precisely:*

$$P_{\mathscr{C}}(A) = \left\{ B \in \mathscr{C} \middle| \exists C_i \in \mathscr{C}, A \cong C_0 \cong C_1 \cong \cdots \cong B \right\}$$

Partial Order for Connections

Defining the strength of a connection between two artworks is a difficult question. As we stated previously, the strength of a connection (i.e. the similarity between the images) is mostly a relative concept. If we try to reduce the cognitive process of a person comparing the relations

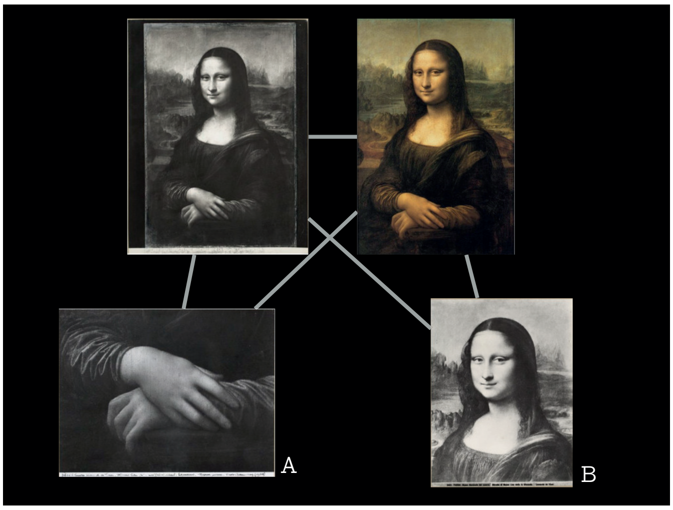


Figure 3.3: Example of physical connections. Note that even if A and B are not connected, they are in their respective physical closure $B \in P_{\mathscr{C}}(A)$.

between images to the minimum, we would get only comparison operations: "given three images A, B, C, one *might* be able to say that the connection $(A–B)$ is stronger than $(A–C)$".

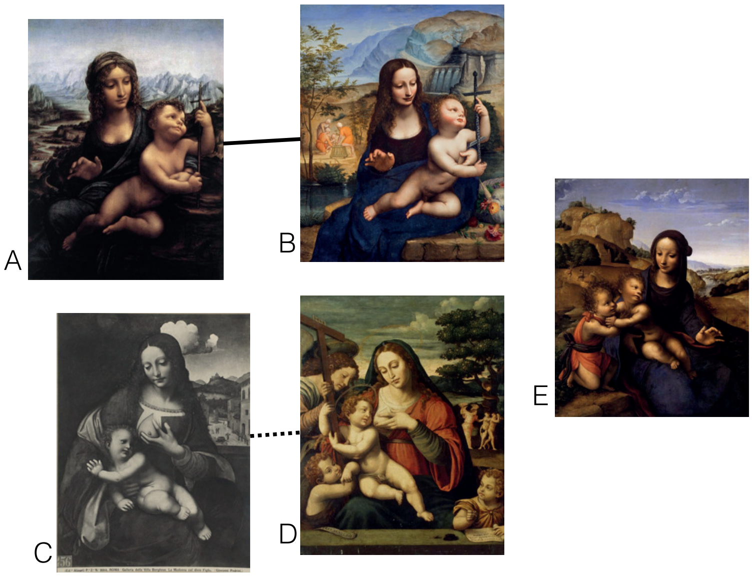
In short, for some pairs of connections, one can say that one is stronger than the other. Mathematically, this translates to a *strict partial order* over the set of connections $\mathbb{C}(\mathscr{C})$. It is important to note that depending on the person answering the question of comparing the similarity between images, the answers can differ. Shortly put, the judgment of each person can be encoded as a separate strict partial order.

As such, multiple partial orderings are possible. For a given strict partial order "<", we note that the strength of the connection c_1 is stronger than the strength of the connection c_2 according to "<" as $c_1 > c_2$. Similarly, if we have images A, B, C we would note that B is more related to A than C as $(A–B) > (A–C)$.

However, since there might be more than one opinion about the relative strength of connections between images, we want to define the set of information where people agree i.e. the clear positive answers:

Property 1 (Agreement of Partial Orders)**.** *Given a collection $\mathscr{C}$ and a non-empty set of strict partial orders $<_1, <_2, \cdots, <_n$, the* agreement *(or intersection) of them (defined such that $c_1 < c_2 \iff \forall i, c_1 <_i c_2$) is a strict partial order as well.*

Proof. A judgment is only a partial order, so if we call $<$ the agreement of $\{<_i\}_i$, we need to prove the three properties of a strict partial order:



c_1	c_2	$c_1 <_1 c_2$	$c_1 <_2 c_2$	$c_1 < c_2$ (with $<$ representing $<_1 \cap <_2$)
$(A–B)$	$(B–E)$	$>_1$	$>_2$	$>$
$(D–E)$	$(C–D)$	$<_1$	$<_2$	$<$
$(A–B)$	$(C–D)$	$>_1$	$?_2$	?
$(B–E)$	$(D–E)$	$>_1$	$<_2$	?

Figure 3.4: Five paintings by five different artists. They all have a striking similarity with each other, but if one were to try to order the strength of their relationships between each other, ambiguous choices and cases appear. If $(A–B)$ is clearly the strongest connection, and $(C–D)$ a strong one as well, the rest does not appear as obvious. In the table, some decisions of two examples of possible judgments are shown with their corresponding intersection.

- *reflexive* for any c, $c <_1 c$ is false (reflexivity of $<_1$) so $c < c$ is false as well.
- *transitive* suppose we have $a > b$ and $b > c$, it means $\forall i, a >_i b$ and $\forall i, b >_i b$. Using the transitivity of each $<_i$ we have $\forall i, a >_i c$, which is equivalent to $a > c$.
- *asymmetric* suppose a and b two connections, and $a < b$, we then have $a <_1 b$, which implies that $a >_1 b$ is not true (asymmetricity of $<_1$) so $a > b$ is false as well.

The agreement $<$ is then a strict partial order. □

What Property 1 means is that given a set of experts, if we only consider as true the statements where they all agree, it is equivalent as having a single person acting as a meta-expert. Thus, considering the specific judgment of a single expert, or the set of clear comparisons between images (i.e. where most people agree), is mathematically equivalent.

Local and global consistency

We showed that the complete information about our visual similarity is a strict partial order on the connections between images. However, such information is difficult to represent directly. Because a graph structure is practical to edit, understand and represent, we ideally would want to use it to encode the visual relations (influence, patterns) between images. In this subsection, we propose a formalization allowing a standard graph to represent some of this partial ordering information.

We consider a graph to be a set of connections. For instance, given a collection $\mathscr{C}$, the graph of physical connection is the set $\mathscr{G}_p = \left\{(A\text{–}B) \in \mathbb{C}(\mathscr{C}) \middle| A \cong B\right\}$.

Definition 5 (Local consistency, strict)**.** *Given a collection $\mathscr{C}$, and a strict partial order "$<$", a graph $\mathscr{G}$ is said to be* strictly locally consistent in $A \in \mathscr{C}$ *with respect to "$<$" and $\mathscr{C}' \subset \mathscr{C}$ iff*

$$\forall B \in \mathscr{C}, \forall C \in \mathscr{C}', \ \big((A\text{–}B) \in \mathscr{G} \ \textit{and} \ (A\text{–}C) \notin \mathscr{G}\big) \implies (A\text{–}B) > (A\text{–}C)$$

In standard terms, what this means is that if a graph is strictly locally consistent in A, out of all the connections originating from A, the strongest ones are all edges of the graph. This implies that the partial order allows a complete separation of all the connections implicating A into two sets, where all the elements of one set are stronger than the elements of the other.

In practice, because of the duplicate images, a strictly locally consistent graph would get cluttered quickly. Indeed, given a consistent graph $\mathscr{G}$, if $(A\text{–}B)$ is in $\mathscr{G}$, then for every photograph B' which is a duplicate of B, $(A\text{–}B')$ would have to be in $\mathscr{G}$ as well. In order to simplify the task of editing the graph, we use a relaxed version of local consistency that takes into account the fact that many images are depicting the same object. Instead of considering all the images not connected with A in the graph, we ignore the photographs which are physically connected with neighbors of A in the graph (i.e. the physical closure of the neighbors of A).

Definition 6 (Local consistency)**.** *Given a collection $\mathscr{C}$, and a strict partial order "$<$", a graph $\mathscr{G}$ is said to be* locally consistent in $A \in \mathscr{C}$ *with respect to "$<$", and $\mathscr{C}' \subset \mathscr{C}$ iff*

$$\forall B \in \mathscr{C}, \forall C \in \mathscr{C}', \ \Big((A\text{–}B) \in \mathscr{G} \ \textit{and} \ \big(\nexists A' \in P_{\mathscr{C}}(A), \nexists C' \in P_{\mathscr{C}}(C) \ \ (A'\text{–}C') \in \mathscr{G}\big)\Big) \implies (A\text{–}B) > (A\text{–}C)$$

The local definition can be extended to the complete graph:

Definition 7 (Global consistency)**.** *Given a collection $\mathscr{C}$, and a strict partial order "$<$", a graph is said to be* globally consistent *if for every $A \in \mathscr{C}$ it is locally consistent with respect to "$<$" and $\mathscr{C}$. We call the set of globally consistent graphs $\mathbb{GC}_{\mathscr{C}}$.*

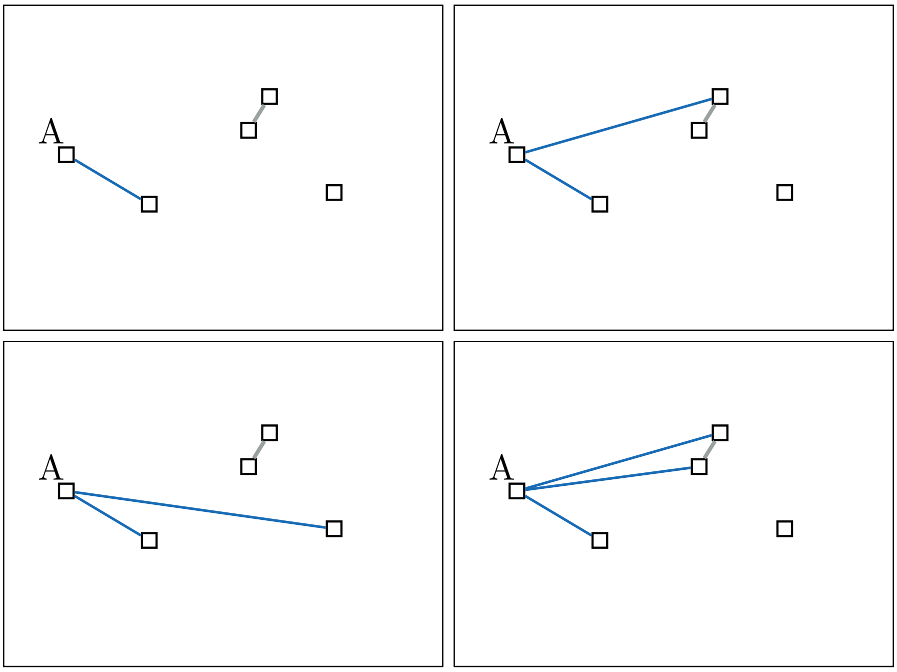


Figure 3.5: Examples of local consistency in A, assuming the spatial distance between the elements correspond to their similarity. Physical connections are in grey.
Top left: strictly locally consistent. Top right: locally consistent (not strictly)
Bottom left: not locally consistent. Bottom right: strictly locally consistent.

Simply speaking, a globally consistent graph is a graph that is locally consistent everywhere. This allows us to define a class of graphs where information about the underlying partial order is encoded, which is what we were looking for.

The space of globally consistent graphs (which is a subset of the set of graphs defined on $\mathscr{C}$) has some interesting properties.

Property 2. *Given a collection $\mathscr{C}$, and a strict partial order "$<$", the following statements are true:*

- *the empty graph $\mathscr{G} = \emptyset$ is globally consistent.*
- *the complete graph (everything connected with everything) $\mathscr{G} = \mathbb{C}(\mathscr{C})$ is globally consistent.*

Property 3. *The space of globally consistent graphs is stable with respect to the union operation. (i.e. if $\mathscr{G}_1 \in \mathbb{GC}_{\mathscr{C}}$ and $\mathscr{G}_2 \in \mathbb{GC}_{\mathscr{C}}$, then $\mathscr{G}_1 \cup \mathscr{G}_2 \in \mathbb{GC}_{\mathscr{C}}$).*[1]

Proof. For every image A, B and C in $\mathscr{C}$, suppose that proposition $(A–B) \in \mathscr{G}_1 \cup \mathscr{G}_2$ and that $\nexists A' \in P_{\mathscr{C}}(A), \nexists C' \in P_{\mathscr{C}}(C) \;\; (A'–C') \in \mathscr{G}_1 \cup \mathscr{G}_2$. Then to prove global consistency of $\mathscr{G}_1 \cup \mathscr{G}_2$, we have to show that $(A–B) > (A–C)$.

Since $(A–B) \in \mathscr{G}_1 \cup \mathscr{G}_2$, then it has to be at least in one the two graphs. Without loss of generality, we can assume $(A–B) \in \mathscr{G}_1$. Additionally, $\nexists A' \in P_{\mathscr{C}}(A), \nexists C' \in P_{\mathscr{C}}(C) \;\; (A'–C') \in \mathscr{G}_1 \cup \mathscr{G}_2$ implies that $\nexists A' \in P_{\mathscr{C}}(A), \nexists C' \in P_{\mathscr{C}}(C) \;\; (A'–C') \in \mathscr{G}_1$. From the global consistency of $\mathscr{G}_1$ we can say that $(A–B) > (A–C)$. □

An important point to remember is that there is not just one graph that satisfies the condition of global consistency, as highlighted by Property 2. There is a space of globally consistent graphs, and each graph in this space will encode a different subset of the information about the underlying partial order. For instance, both the degenerate cases of Property 2 are not encoding any information about the corresponding partial order.

3.2.3 Morphographs

Our goal is to be able to represent the relative strength of connections between artworks in a given collection $\mathscr{C}$. In the previous section, we have presented a formalization of a class

---

[1] If we want to be completely precise, we can actually go slightly further as $\mathbb{GC}_{\mathscr{C}}$ with the union operator forms a *commutative monoid*. This can be seen as a commutative group without the inverse axiom. This comes from the fact the union operator is stable, associative and commutative, and that there exists an identity element ($\emptyset$).

of graphs that have nice properties to encode the information of an underlying strict partial ordering for connections between objects. Here, we present how the previous characterization can be applied to our use-case.

The Choice of one Partial Order

In the previous formalization, we assume a unique strict partial order about the strength of relations between images "<". However, it is important to remember that *there is not a unique partial ordering*. According to which Art Historian you ask and his domain of interest, different opinions might be possible for the same question: "which pair of images look the most similar?".

As far as individuality is concerned, how do we tackle the possible variety of opinions? One solution is to take the consensus i.e. where people agree. We already argued (see Property 1 and Figure 3.4) that by taking the agreement of multiple people, we can keep a strict partial order, and mostly remove this variability.

However, by trying to satisfy everyone, the corresponding partial order is never true anymore as there will be always one entity disagreeing. Thus, there needs to be a form of guideline to stabilize the consensus process.

In our case, we define the partial order we consider according to the co-occurrence of patterns or composition in them. Because these graphs encode the transmission of shape and form across artworks, we call the class of globally consistent graphs with respect to that partial ordering *morphographs*.

While morphographs are a class of possible graphs on a collection of images, we will often refer to the graph we have been acquiring as *the* morphograph, despite not being a unique construction.

Advantages of a Graph Representation

The use of a graph to represent the sort of visual similarity we are interested in has multiple advantages. The first one, and perhaps the most important, is that it is relatively easy to work with, as it is a standard data structure for computers, and simultaneously easy to reason with for humans. It is also possible for users to easily edit it, modify it, and visualize it.

The fact that it is easy to visualize fosters discussion and sharing of the information. Also the created connections, which we tend to refer as *Visual Links* (or *Visual Connections*) between artworks, can directly be exported to the greater audience, effectively creating an object of knowledge in itself.

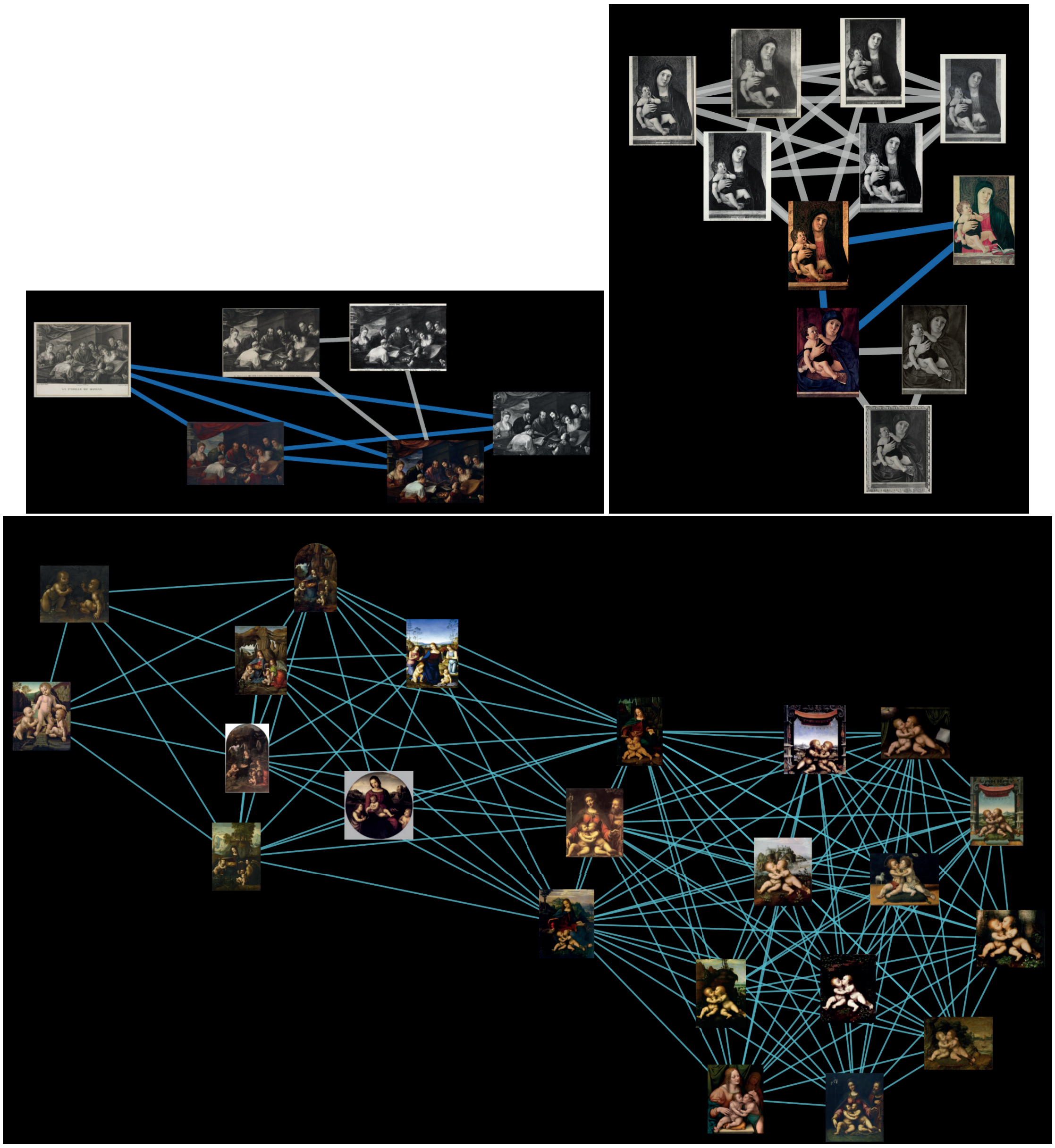


Figure 3.6: Examples of morphographs, physical connections are displayed in grey, visual connections in blue. Top left: four artworks based on an original composition by Leandro Bassano. Top right: three *Madonna and child* by Giovanni Bellini and the multiple corresponding duplicate photographs.
Bottom: visual links can represent complex pattern propagation, such as in this graph. The starting point is the series of two *Virgin of the Rocks* paintings based upon the original by Leonardo da Vinci (center left). On the center right, there are variations where the Virgin is replicated but the two infants in the foreground are altered. In both cases, the two infants became a composition of their own on the far right and the far left of the graph.

Finally, it is of practical interest because its construction can be done as an iterative process. As stated by Property 2, the empty graph is a valid morphograph, which can be used as a starting point. Then users can incrementally edit the graph by adding connections, which as long as they are careful about keeping local consistency for the elements they modify, keep the graph globally consistent. Also, such a process can be undertaken by multiple users at the same time, as long as they are not editing the exact same images at the same time. This enables potential crowd-sourcing through a common interface. Both of these properties (incremental and parallel construction) are highlighted by the fact that the union of two morphographs is a morphograph as well (Property 3).

The Mismatch of Visual Knowledge

However, the construction of our morphograph comes with one major difficulty: how can one user, modifying the graph, be sure that the consistency is kept? Indeed, the property of consistency is actually very strong: when a user connects A with B, he implicitly says that *every* other image $C \in \mathscr{C}$ (not physically connected with A or B) shares a weaker connection with A than B does.

The pitfall is that it is almost impossible for a user to know the complete collection $\mathscr{C}$ and to be sure that the consistency property is not violated when he edits the graph. In practice, one can rely on its knowledge of visual production and its own memory of previously seen images, but it is unlikely that will cover the complete collection.

We can formalize the concept of *Visual Knowledge* as what one knows about the human production in the visual arts[2]. This *visual knowledge* comes from past experiences, for instance places visited, objects seen in person or through photographs. Each Art Historian, depending on its field of study and level of expertise, will have a different knowledge of the visual production. We can note that it is a finite concept, as it is bounded by the theoretical knowledge of all created objects/artworks in History, and practically bounded by the knowledge of all objects/artworks which have survived up to this day (including in other forms: photographs, drawing, etc.). A collection of photographs, like the one we have, can also be seen as a form of visual knowledge in itself, as it encodes information about the represented artworks (see Figure 3.7).

Having said that, the issue we presented just above can then be seen as a mismatch between the visual knowledge of a user and the collection itself. We can rightfully expect an Art Historian, with a visual knowledge ω, to edit the graph so that it stays globally consistent with respect to $\omega \cap \mathscr{C}$ i.e. the images he knows of and which are part of the collection $\mathscr{C}$. But that means we

---

[2] It has to be precised that we still are in a framework of purely visual information, and only dealing with images. In that sense, we do not include metadata, or textual information in such a concept of visual knowledge.

cannot expect proper consistency with respect to the images which are part of the connections but not known by the user ($\mathscr{C} \setminus \omega$). This poses a fundamental challenge to the edition of such a morphograph given a large collection.

3.3 Approach

Despite being a satisfying formalization, the concept of morphographs poses challenges as to how properly construct them. In this section, we describe the global approach that was taken allowing us to solve these issues.

3.3.1 Outline

A Search System as an Extension of Knowledge

In the last section, we highlighted that acquiring such a morphograph was a challenge due to the impossibility of the users to know all the images which are in the collection.

In order to tackle this issue, e couple the graph annotation tool with a powerful visual search/-navigation system. Indeed, in practice, when one is trying to modify a part of the visual graph, only a tiny portion of the collection is relevant for this modification, may it be for evaluating the relative strength of two connections, or in order to find missing elements (that the user did not know of) which needs to be connected to keep the consistency property of the morphograph.

Such an approach can be seen as allowing the users to emulate the knowledge of the complete collection $\mathscr{C}$, while only reviewing a tiny portion of it. A search engine, in essence, is a suitable answer as it will (hopefully) bring only the relevant images in the collection for the current question at hand.

Framework

The resulting framework (Figure 3.8) is a combination of three components:

- first, the *graph of annotated connections* made by the users, i.e. the morphograph.
- second, *computer vision techniques* which, from the graph of annotated connections, learn how to help organizing the images in the collection.
- third, an *interface/search tool* that helps the users leverage the computer vision models, in order to navigate the collection and better annotate the morphograph.

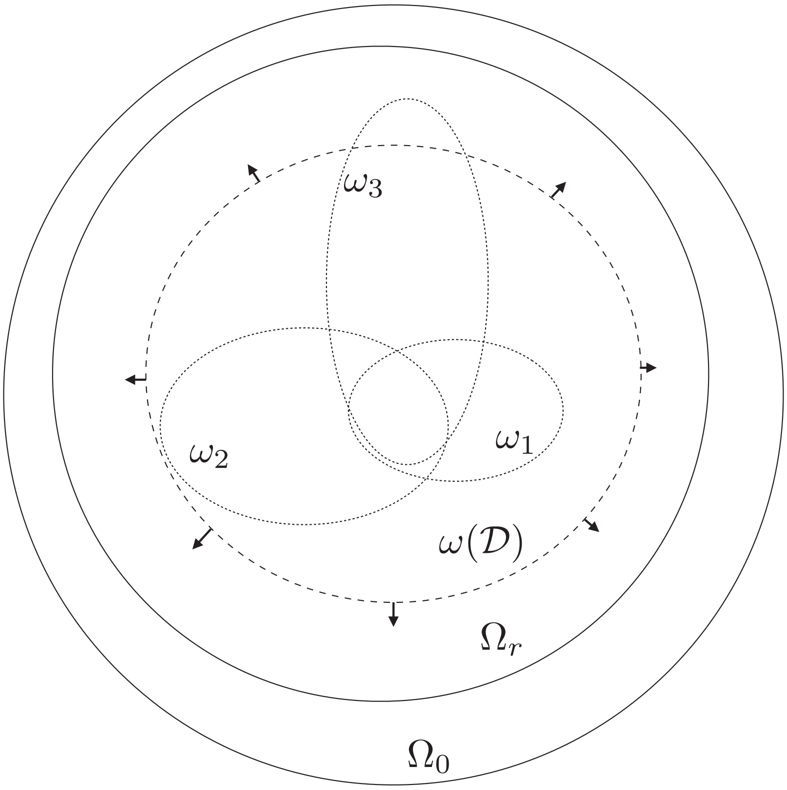


Figure 3.7: Ensemble diagram of different scales of visual knowledge. Ω_0 represents the visual knowledge induced by all the created objects in History, and Ω_r would be the knowledge coming from the surviving visual information of today. Here we call $\mathscr{D}$ the collection of all digitized images, and $\omega(\mathscr{D})$ the corresponding induced knowledge, expanding as digitization carries on. ω_i represents the visual knowledge of individuals with different levels and domains of expertise. If there is a common set that everybody is aware of, the direction and the scale of what is known by each person varies, and can sometimes go beyond all current digitized information (ω_3).

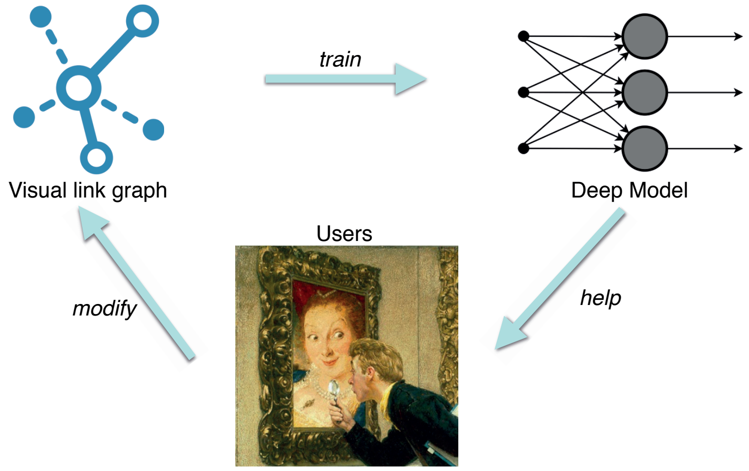


Figure 3.8: Schematic of the proposed framework. The users annotate the morphograph (graph of visual links), which is then used to train a visual similarity system based on a deep network (Section 3.5). This similarity function is leveraged by an interface/annotation system (Chapter 4), effectively helping the users to find more elements to be added in the graph.

Such an organization creates a feedback loop (as seen in Figure 3.8). Indeed, the more connections there are in the graph, the better the visual model will become (as the training data increases), which in turn increase the performance of the search engine, which allows the users to find even more connections, further improving the model, etc.

While the definition of the morphograph made up a significant portion of this chapter already, the computer vision techniques used for helping to organize the collection will occupy the rest of this very chapter. The search/annotation system, however, will be covered in the next chapter.

3.3.2 Elements of the Framework

Collection and Graph

Collection: the basis of our framework is of course the collection of images we are working with. In the rest of this thesis, we will focus on a corpus made of the digitized images coming from two main sources. The first part comes from the digitization of the Cini photo archive. This corresponds to the data which was automatically processed by the pipeline described in the previous chapter, and accounts for around 330'000 images. This is a corpus focused on Italian (and especially Venetian) art of the 15th to 17th century.

The second part of the corpus is made of the drawings, engravings and paintings of the Web

Gallery of Art [82], a website aggregating a large amount of copyright-free images of artworks. This corpus accounts for around 35'000 additional images. The use of this set of data is mainly for historical reasons, as the data from the Cini only became available during the course of the project. At the time, it was the best suited online dataset, as it focuses on the same period we are interested in (15th to 18th century), unlike Wikiart[35] for instance (where most artworks were created after 1800).

Acquired graph: we annotated two types of connections between images of the collection. The first type represents the physical connections corresponding to two images representing the same object, or the same physicality. The second type is the set of visual connections of the morphograph.

Since the beginning of the system in Spring 2016, a couple of Art Historians and myself have navigated, searched and annotated connections in the collection. For the first months, connections were mostly found by using information about pattern transmission from the handful of resources available discussing the topic. But as time went by, and data poured from the Cini, we relied more on exploring the collection directly. As of August 2018, more than 2'800 physical connections, and 6'300 visual connections are part of the annotated data. Although the search system is designed to minimize the errors in global consistency of the annotated graph, we do not expect it to be a perfectly annotated dataset either.

Computer Vision techniques

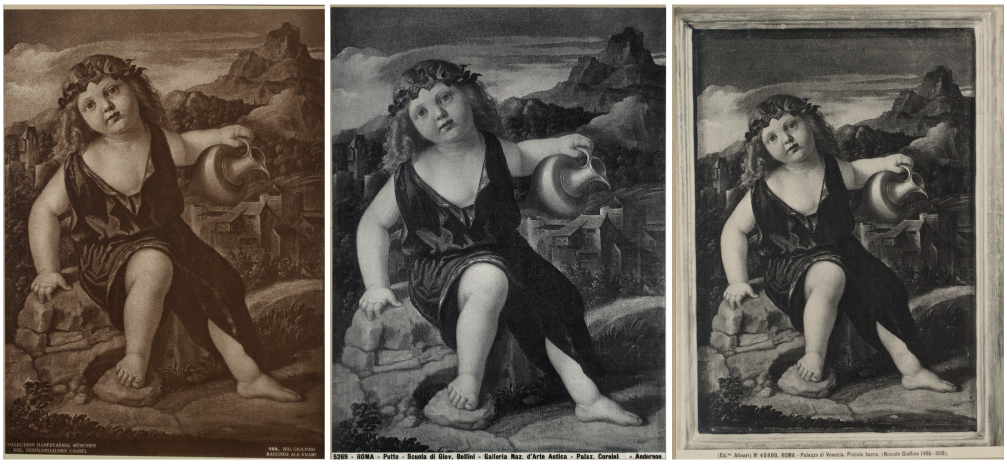
The backbone of the framework is the set of computer vision techniques that allow generalizing the annotations done on the graph of connections. We distinguish two fundamental tasks that our system solves.

The first task is about the automatic detection of the physical connections. By using the set of physical connections annotated, can we automatically propagate this information? and what are the corresponding difficulties? This will be covered in *Section 3.4*.

The second task is about generalizing the knowledge coming from the annotated morphograph. Since the graph is encoding some information of an underlying partial ordering, we will show how we can learn a visual similarity function that generalizes the comparison encoded by the morphograph to the rest of the collection. Our approach will be described in *Section 3.5*.

Search System

The last element of the framework is the searching system. It will be described in *Chapter 4*.


COLLECTION HANFSTAENGL MÜNCHEN
KGL. GEMÄLDEGALERIE CASSEL
229. NIC. GIOLFINO
BACCHUS ALS KNABE
5269 - ROMA - Putto - Scuola di Giov. Bellini - Galleria Naz. d'Arte Antica - Palaz. Corsini - Anderson
(Ed.ne Alinari) N° 40800. ROMA - Palazzo di Venezia. Piccolo bacco. (Niccolò Giolfino 1486 - 1518).

Position in the Cini	47A_538	110C_258	113B_38
Attribution	BELLINI, Giovanni	BELLINI, Giovanni (scuola di)	GIOLFINO, Niccolò
Title	Bacco	Putto	Piccolo bacco
Location	Gemäldegalerie, KASSEL	Galleria Nazionale d'Arte Antica, ROMA	Museo di Palazzo Venezia, ROMA

Figure 3.9: Three photographs representing the same object, but at completely different positions in the Cini collection. We can note how metadata here does not help at all to recover the fact that they represent the same object, as neither attribution, title nor location is consistent.

3.4 Automatic Classification of the Physical Connections

In this section, we will focus on the problem of automatically finding physical connections in our collection. As we outlined previously in Section 3.1.2, this task is of great practical importance for the organization of a collection of photographs. In the case of the Cini corpus for instance, the collection was created as an aggregate of personal collection acquired concurrently. As such, many artworks were photographed multiple times, at different dates, etc. Additionally, the creators of these individual collections might have given different attributions and/or description to the same artwork (see Figure 3.9). This made difficult for archivists to regroup these duplicate photographs together among the hundreds of thousands of images they have available.

We will first present the challenges of the task through different examples, then we will outline the methods used both to diagnose edge cases and automatically classify connections as physical. Finally, we will show the conclusions of applying this method to the Cini corpus.

Examples/Challenges

We defined physical connections as two images where one can assert they have the same *physicality*, meaning that the same object or two objects of a serial production (engraving) is/are represented on the images. In practice, such a definition includes the most obvious cases such as full duplicate photographs of the same artworks, two issues of an engraving made with the same woodblock, or a detail view of a large painting (first rows of Figure 3.10).

But there are other more complex cases. For instance, restoration works may have changed a significant portion of a painting. Back in the introduction (Figure 1), one can see the restoration process of a BELLINI fresco that was extended by later painters. Other surprising situations can also happen, for instance with photographs which were sometimes manually edited by modifying directly the positive or negative film, altering the acquired image. A striking example can be seen on the third row of Figure 3.10, where multiple photographers wanted a good crop of the character of Flora in the *Primavera* painting of BOTTICELLI, which is partly occluded by a neighboring character.

On the other hand, there are also images which are extremely similar but actually correspond to different objects. Copies by different artists are usually relatively distinguishable if the two images are next to each other. But in the cases of a series of painting made by the same artist or his workshop, visual comparison (looking at the images side-by-side) might not even be enough.

These two sets (positive and negative) of examples highlights some of the difficulties we face for this task. Local or even global characteristics can be changed because of a restoration/alteration process, and the point of view can be different. Additionally, the quality of these photographs and their illumination settings are often sub-optimal, effectively adding noise to the measurements, which does not help our cause. At the same time, if we look at workshop series, we have artworks which were made to be as similar as possible to each other. In a way, the best artists of History are actively trying to confuse us...

3.4.1 Approach

Detecting Spatial Coherence

While it is true that some artworks were made to be almost indistinguishable to each other, it is interesting to note that *they were made to be indistinguishable for the human eye.* It is well known that the human vision system does not distinguish well small local variations as we abstract a lot of the visual information. Most of the time, reproducing this invariance is a desired property for computer vision techniques, and CNN have been extremely well performing at it, enabling much better generalization on high-level tasks.

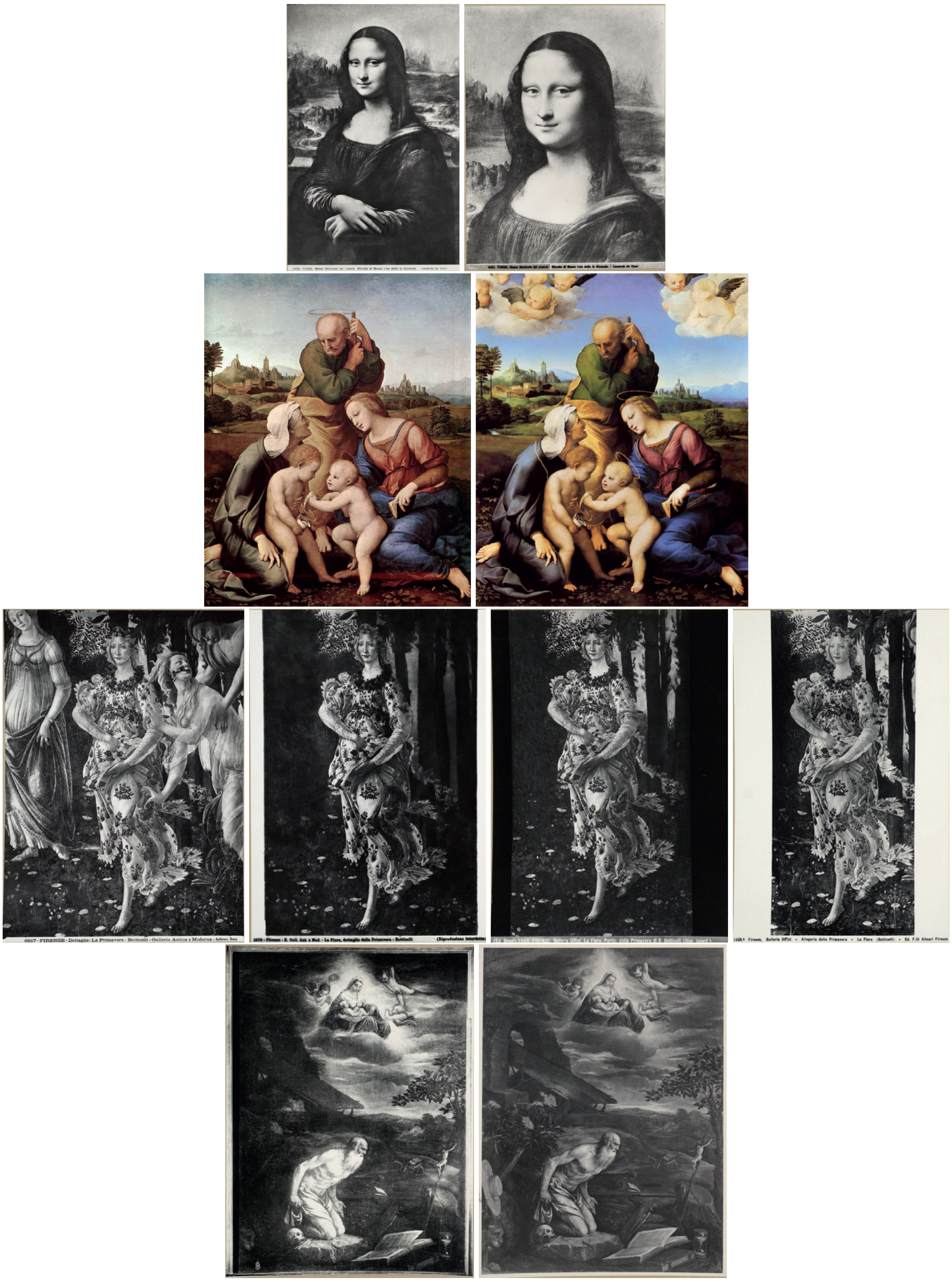


Figure 3.10: Examples of physical connections.
First row: *La Gioconda*, full-artwork and close-up.
Second row: *Madonna Canigiani* of RAFFAELLO, before and after the restoration of 1982.
Third row: Flora from the *Primavera* of BOTTICELLI, crop of the original and different versions in the Cini, modified by photographers. Notice how in the version on the right, extreme attention was given to reproduce the floral background after having removed the occluding character on the right.
Fourth row: *S. Gerolamo* by JACOPO BASSANO, different lighting and expositions make elements on the left of the picture completely disappear.

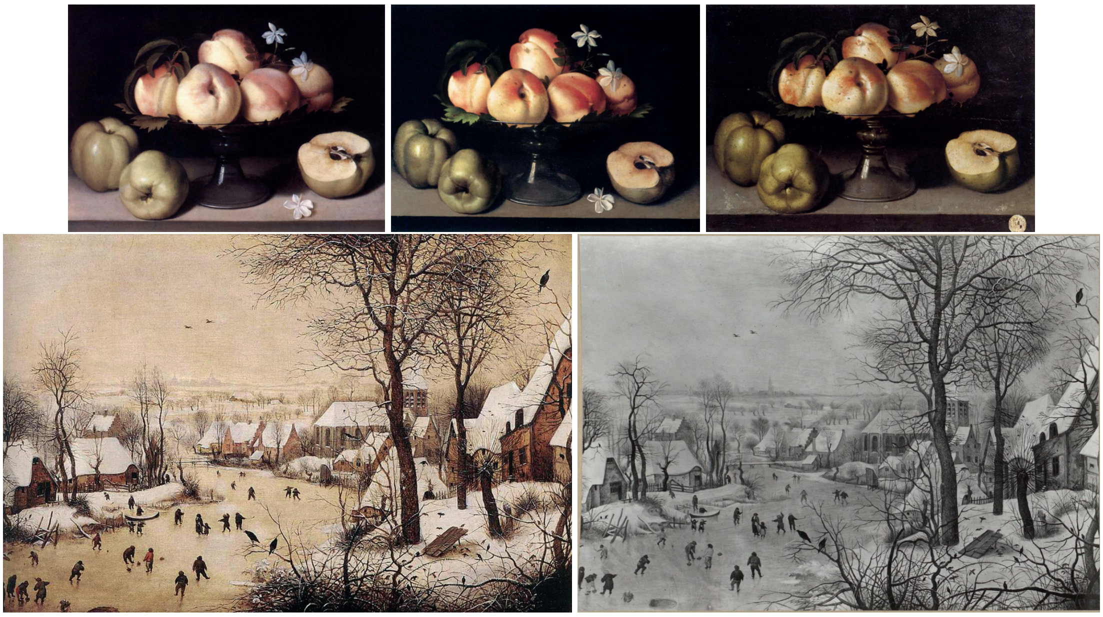


Figure 3.11: Examples of close pairs of images which are **not** physical connections.
Top: different still-lifes by Galizia, Fede
Bottom: *Winter* by Bruegel, Peter the Elder (left) and Bruegel, Peter the Younger (right)

However, computer vision techniques do not have to try to emulate the way we see. When it comes to precise measurements, checking perspective, and spatial consistency, computers are much better than we are. In this specific case, one can guess that geometric consistency will be important as we are trying to decide if they are the same objects are not. This tends to drive us to the techniques of feature point matching as they have a very good spatial sensitivity. Indeed, feature points precisely identify a stable point in the image, since we know the corresponding object is planar, we can check the spatial coherence of these matches between the images, which is a great indicator of checking same physical sameness.

Aiding Human Decision

Even for deciding the label of a connection given two images, a side-by-side comparison proved not to be reliant enough. In a way, deciding if two images are representing the same object can be compared to a game of “spot the differences”: a form of visual comparison. For most people[3], this include a not very efficient nor reliable eyeball search, going back and forth

---

[3] Apparently, some people with good enough visual parallax control, can directly go cross-eyed on both images set next to each other. That way, they basically use their depth-perception mechanism in the brain to automatically match the two images, and the differences stand up right away.

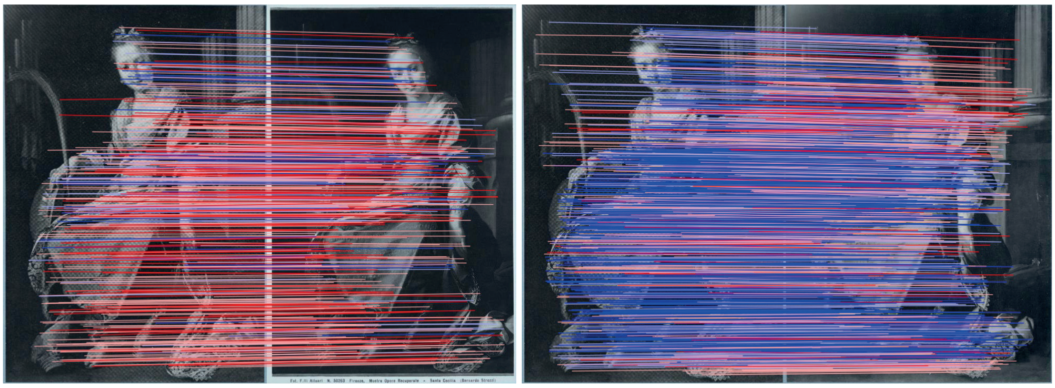


Figure 3.12: Matching local descriptors, color-coded based on the threshold during inlier filtering i.e. how spatially coherent they are. Blue matches are obtained with a low threshold, red with a higher one. On the left is a pair of different artworks, while on the right is a pair of photographs of the same object. When matching the same object, the number of matches is higher and more spatially coherent.

between both images.

Since our objects of interest are mainly planar objects: paintings, frescos, drawings, etc. We know that two views of the same object are related by an *homography* [83] (representing the change of viewpoint between the two photographs), which means that there exists $H \in \mathbb{R}^{3\times 3}$ such that for every physical point of the planar object, if its coordinates are (x, y) in the first image and (x', y') in the second, we have:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \text{ with } h_{33} = 1$$

Given the candidate pairs of points generated by a feature detector/descriptor, we have a list of potential matching positions in the two images, though with a substantial amount of false matches (outliers). As is common practice, outliers can be automatically removed with RANSAC [84] and we obtained the best transformation matrix H fitting the good matches (inliers) correspondences.

After H is evaluated, one can use it to back-project the second view of the object in the coordinate system of the first view, perfectly aligning them. In order to show the differences between the two aligned images, we found very effective to create a looping animation blending the two aligned images together. In such a representation, small differences that were almost invisible before pop right away, making it a precious tool for comparing these close variations.

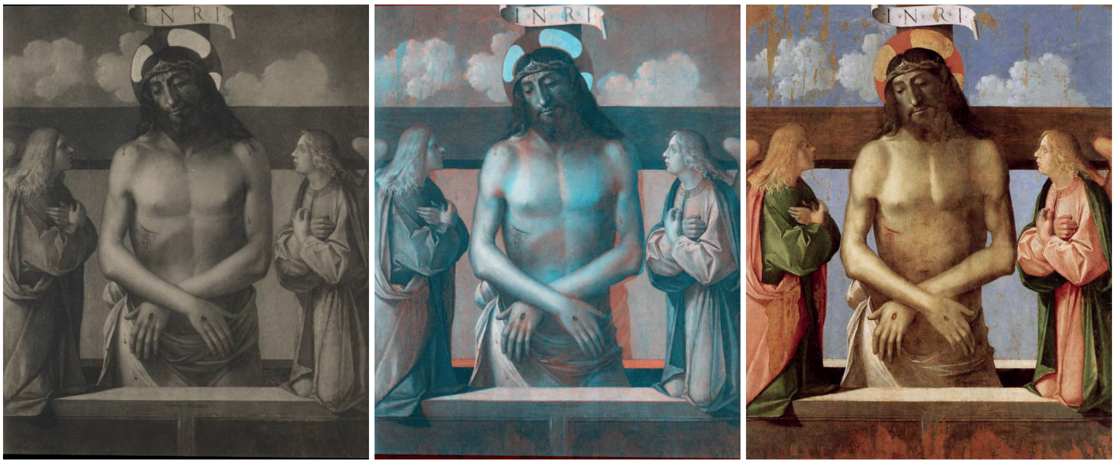


Figure 3.13: Example of how properly aligning images can show differences easily. The *Pietà* by BENEDETTO DIANA is represented before (left) and after (right) a restoration work. It is impossible to show the animated transition that we use to visualize the differences, so the best we can print is an overlay with well-separated colors (center). The modified area, especially at the right of the Christ, appears very clearly and highlights how the painting was modified during its lifetime.

Algorithm

As highlighted by the views generated in Figure 3.13, spatial coherence is a key factor to distinguish between very similar images and physical connections.

A similar problem was encountered in [85], where they try to detect whether illustrations in printed books were based on the same wood-block or not. But in their case, because it is a printed medium, they can binarize the images and perform a difference of the aligned binary images, an approach we cannot apply here.

The main idea we use here is that in the case of close copies, some feature points do match but are slightly off after the alignment process. A simple way to get this information, is to perform the inlier/outlier separation of feature points with separate thresholds, getting a different number of matching inliers depending on how “selective” we are. Our guess is that the good matches that are not perfectly aligned are a sign of a close copy.

Given two images I_1 and I_2, the set of features we use are:

- the number of detected feature points in each image n_1 and n_2.
- the number of candidate correspondences extracted n_G. Correspondences (p_1, p_2) are extracted such that for each feature point p_1 in I_1, the closest feature descriptor in I_2 is

associated with p_2 and respectively (the closest descriptor to p_2 in I_1 is p_1).

- the portion of each image s_1 and s_2 which is matched during this process. The matched area is computed as the bounding-box containing all the spatially coherent matches.
- the number of spatially coherent matches $\{n_s(\tau)\}_\tau$ for different values of the inlier threshold τ during the fitting of the homography. More precisely, given the n_G candidate pairs, RANSAC is used to robustly estimate the homography transformation H. The back-projection error is then computed for every pair to compute the actual number of inliers for the corresponding threshold τ.

Given d different values of τ the final feature vector encoding the pair of images is then of dimension $d+5$:

$$\begin{bmatrix} \min(n_1,n_2) & \max(n_1,n_2) & n_G & \min(s_1,s_2) & \max(s_1,s_2) & \frac{n_s(\tau_1)}{n_G} & \cdots & \frac{n_s(\tau_d)}{n_G} \end{bmatrix}$$

3.4.2 Results

Prediction Performance

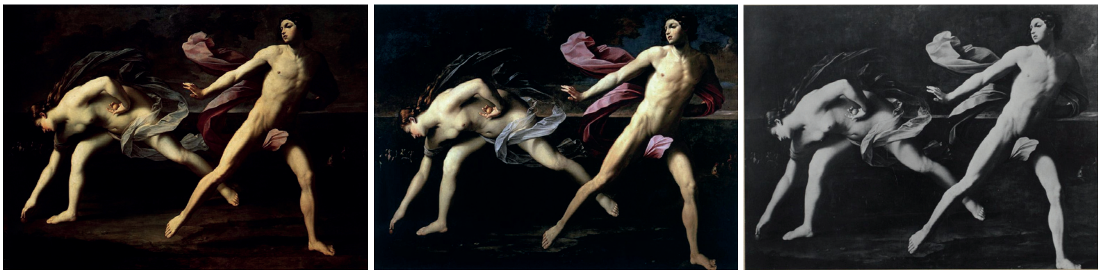
For training/testing data, we rely on the annotated graph we acquired. We have more than 8’000 connections annotated, including more than 3’000 physical connections. Many close copies appear in the morphograph of visual connections, acting as difficult negative cases.

As far as implementation is concerned, each image is resized so that its bigger dimension is 720px before computing the feature points. Feature points are extracted with the upright version of SURF [11], as it is slightly faster for the large comparison we will do next, and orientation detection is not important in our situation. The thresholds used are $\tau \in \{1.0, 2.0, 5.0, 10.0, 20.0\}$ (in pixels).

Different binary classification methods were tried but the most successful was a kernelized Support Vector Machine [86] using Radial Basis Function [87]. As is standard practice, each feature is whitened to zero mean and unit variance, and we optimize the parameters with 10-fold cross-validation (RBF: $\gamma = \frac{1}{d+5}$, and SVM: $C = 30.0$).

Results are show in Figure 3.15 for different values of precision. As we can see, performance increases when information about the area of the matched region, or the number of inliers correspondences for different thresholds are added to the feature vector. Combining both gives the best performance allowing making very few mistakes while catching most of the physical links.

Interestingly, we could estimate the number of errors we did during the annotation process. By



Pair of images	Probability physical connection
(Madrid)-(Naples)	3.8%
(Madrid)-(Cini)	3.1%
(Naples)-(Cini)	99.9%

Figure 3.14: Example of the learned model properly distinguishing close artworks. Two versions of *Atalanta and Hippomenes* by GUIDO RENI, *left*: the version in Madrid at the Prado Museum, *center*: the version in Naples at the National Museum of Capodimonte. On the *right* is a photograph from the Cini, despite the extreme similarity between the artworks, the model is able to very confidently identify that the photograph is indeed representing the version at the Naples museum.

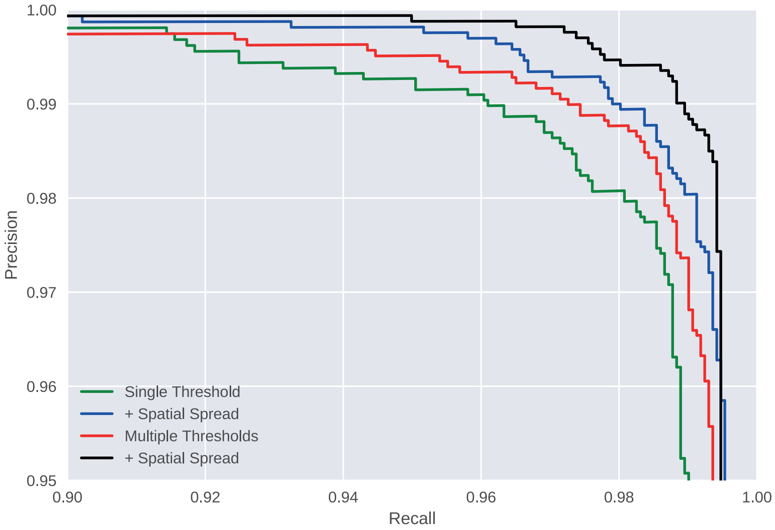
performing cross-validation on the complete dataset, we could look at the failure cases of the prediction model and easily check for errors without having to reparse the complete dataset. During the process, we found 49 connections wrongly flagged as physical (False Positive) and 21 wrongly flagged as non-physical (False Negative). Given the size of the total dataset, this can give an estimation of human performance with a precision of 98.5% for a recall of 99.3%, which is quite comparable with the performance of our learned model[4]. Much of these errors can be attributed to attention errors, but also to the unexpected closeness of some of these artworks, which made us sometimes flag a connection as duplicate photographs without double-checking provenance or computing the blended images overlay.

Application to the Cini Corpus

The learned model was applied to the complete set of images in the collection. However, because the number of possible pairs of images is too high to evaluate them exhaustively ($330'000 \times 330'000 \simeq 109$ Billions), we need to pre-select them. As a simple selection scheme, we compute the distance of CNN features (as they will be presented in the next section) between every pair of images, which can be done in a relatively short time with efficient GPU implementations like [88]. Eventually, the 2'000'000 pairs of images with the smallest distances between their CNN descriptors are selected as candidates for being physical connections.

---

[4] The performance given in the table of Figure 3.15 is based on the corrected dataset.



Precision		0.98	0.99	0.995
Recall	Single Threshold	0.981	0.961	0.925
	Spatial Spread (SP)	0.991	0.980	0.966
	Multiple Thresholds (MT)	0.987	0.973	0.954
	MT + SP	0.994	0.990	0.978

Figure 3.15: Precision-Recall curves for the binary classification of physical links.

Number of physical connections found	113’510
Estimation of recall (from training data)	90.6%
Cini + WGA	
Number of objects	39’916
Number of images involved	108’555
Cini only	
Number of objects	36’836
Number of images involved	100’549
Number of images with clear attribution	63’963 (63.6%)

Table 3.1: Number of physical connections extracted from the collection.

The results are presented in Table 3.1. More than 110’000 physical connections were automatically found, which by comparing with the annotated physical connections in the graph represent a recall of 90.6%. The missed connections are usually the ones where there is a large scale change (i.e. between images of the full artwork and close-ups of details), which are often not properly caught by our simple selection scheme based on global CNN descriptors.

Once we have connected images with each other, we can compare their corresponding metadata. In the case of the Cini, we showed in the previous chapter how we semi automatically parsed the attribution field of the documents, effectively linking them with an artist identifier in a knowledge database. We only keep the images with a single “clear” attribution, (meaning we discard entity linkage of the form “*scuola di*” or “*modi di*”). Then for every object with at least two attributions (coming from two different photographs), we check if the ULAN identifier is the same for all of them. Results are reported in Table 3.2: we **automatically found more than 1’200 objects with conflicting attributions, which corresponds to 5.78% of the objects with at least two clear attributions**. What is interesting is that a large portion of these conflicts are between photographs situated in completely different drawers of the collection, showing how it would have been extremely difficult to catch them without this machine vision based approach.

Looking at the most common attribution conflicts, it is unsurprising to see the most famous Venetian painters represented. Also, some of the most prominent conflicts are about painters of the same family (brothers, father and son). However, if we remove the top 6 cases, there is no combination of painters with more than 8 conflicting objects for the remaining 1’000 artworks. This indicates a long tail distribution with many conflicting configurations.

Cini attribution		
Objects with agreeing attribution	19’784	53.7%
Objects with conflicting attribution	1’217	3.3%
Estimation conflicting rate	5.79%	

Most common attribution conflicts		Number of artworks
Tiepolo, Giovanni Battista	Tiepolo, Giovanni Domenico	89
Tintoretto, Jacopo	Tintoretto, Domenico	17
Guardi, Antonio	Guardi, Francesco	17
Palma, Jacopo, il giovane	Tintoretto, Jacopo	14
Orcagna	Nardo di Cione	12
Titian	Campagnola, Domenico	10
Carracci, Annibale	Carracci, Agostino	8
Morlaiter, Michelangelo	Ceruti, Giacomo	8
Toschi, Paolo	Correggio	7
Peruzzi, Baldassare	Pinturicchio, Bernardino	7

Table 3.2: Attribution conflicts in the Cini corpus.

Limits and possible extensions

We presented how we could help users compare images with small differences, helping to distinguish physical connections in the data. We also presented how we could automatically detect these connections in a large collection, which allows automatically highlighting inconsistencies in the metadata which would be hard to detect otherwise.

One common situation in digital humanities is that after a proper definition of the concepts, examples in the data often challenge these very definitions. By looking at some of the failure cases of the predictive model, we found that connections between two different paintings that were probably created with the aid of a cartoon were often predicted as physical.

As a reminder, a cartoon is a full-scale preparatory drawing for a fresco, an oil painting or a tapestry. Using various techniques, the artist could transfer their design to the wall or the canvas as a starting point for the creation of the artwork. Since we considered prints made from the same woodblock as being physical connections because they are serial productions sharing the same “physicality”, the case of paintings made from the same cartoon is relatively similar. It is beyond the scope of this work to investigate this distinction further, but it highlights the difficulty to have binary definitions, as edge cases are bound to happen in such complex interdisciplinary fields.

However, we also believe the simple visualization of overlaying two artworks automatically aligned with feature points is of great practicality to investigate the usage of cartoons. In

Figure 3.16, we show the famous painting by RAFFAELLO, the *Madonna Bridgewater* (situated in Edinburgh) and its closest variation (out of at least 5 other paintings based on this composition) which comes from the Museo Nazionale of Naples. The photograph metadata states that this is a copy, or a painting from the school of Raffaello, but as we can see on Figure 3.16, the similarity is much too important to be a simple copy. Also, RAFFAELLO was known for his use of cartoon[5].

Additionally, in this case, the size of the artworks is mentioned on the documents: the Bridgewater version being measured as 81x56cm, while the version in Naples is reported as slightly bigger with 87x64cm. This allows us to compute the pixel/cm resolution of each of the photograph we have and the corresponding normalized homography matrix:

$$H_n = \begin{bmatrix} 1.011 & 0.04 & 1.081 \\ -0.034 & 0.995 & 5.988 \\ 0. & 0. & 1. \end{bmatrix}$$

The obtained matrix is almost a perfect rotation matrix[6] ($\det(H_n) = 1.008$), which indicate that despite the images being at different resolutions and the artworks being different size, the matching parts are physically *exactly the same size*, which greatly reinforces the hypothesis of the use of a cartoon.

A complete study of cartoon usages is beyond the scope of this work, however this small case study highlights the effectiveness of combining moderate quality photographs, smart visualizations, and computer vision techniques. In this very situation, we are even able to validate a hypothesis without having to physically access neither objects.

3.5 Learning a Visual Metric

In this section, our goal is to be able to generalize the information represented in the morphograph. Indeed, as we have described in Section 3.2, the annotated connections of the graph encodes some information about a partial ordering between the strength of the connections between images.

One can wonder if it is possible to learn to generalize the partial ordering encoded by the morphograph to the rest of the collection: given two connections which are not related to

---

[5] For instance, seven of the ten cartoons made by him for the creation of the tapestries of the Sistine chapel are to be seen in the Victoria and Albert Museum in London

[6] As a reminder, a transformation which only perform a rotation of an angle θ and a shift of Δ_x, Δ_y can be written $\begin{bmatrix} \cos\theta & -\sin\theta & \Delta_x \\ \sin\theta & \cos\theta & \Delta_y \\ 0. & 0. & 1. \end{bmatrix}$

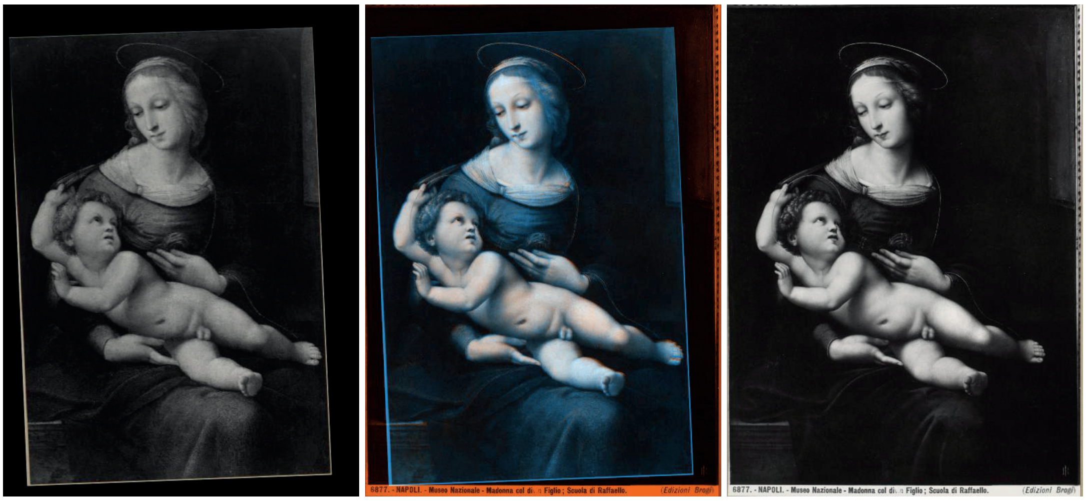

6877. - NAPOLI. - Museo Nazionale - Madonna col di[illegible] Figlio; Scuola di Raffaello. (Edizioni Brogi)
6877. - NAPOLI. - Museo Nazionale - Madonna col di[illegible] Figlio; Scuola di Raffaello. (Edizioni Brogi)

Figure 3.16: *Bridgewater Madonna* (left) aligned and compared with the version in Naples (right). Almost no differences can be noticed on the overlay, apart from the window on the right being shifted. Notice how even the step the Virgin is sitting on (bottom left) is perfectly aligned.

the acquired morphograph, are we able to predict if one is a stronger visual link than the other? Such an estimator would have great consequences in the way one could explore the collection. For instance, it would allow performing image search by finding the images with the strongest connection, or would allow to make a visualization of the image space where connected images would be located together.

3.5.1 Problem Statement

We suppose we are starting with a collection $\mathscr{C}$ and a corresponding morphograph $\mathscr{G}$ (globally consistent with respect to $\mathscr{C}$ and a partial order "$<_{\mathscr{C}}$"). Given its definition, such a graph implies a set of constraints on the connections of $\mathscr{C}$ of the form: $(A–B) <_{\mathscr{C}} (A–C)$ (where A, B, C are three images in $\mathscr{C}$).

In order to learn to generalize from this set of constraints, we want a similarity function which given two images computes a score estimating the strength of the connections between these two images. More precisely, we want to estimate a function $s : \mathscr{C} \times \mathscr{C} \to \mathbb{R}$ such that[7]:

$$\forall A, B, C \in \mathscr{C}^3, \big((A–B) <_{\mathscr{C}} (A–C) \implies s(A, B) <_{\mathbb{R}} s(A, C)\big)$$

---

[7] In order theory terms, this means s is a *monotone* function from the ordered set of connections $(\mathfrak{C}(\mathscr{C}), <_{\mathscr{C}})$ to the ordered set $(\mathbb{R}, <_{\mathbb{R}})$

By doing so, we effectively learn a total ordering of the connections, which coincides with the set of constraints we are given initially. However, as in any standard machine learning task, there is more than one possible function s given a set of training data so, what we are interested in is the ability of the learned function to generalize to non-annotated parts of the collection.

Evaluation procedure

In order to measure the generalization power of s, we first need some proper testing data, but in practice we only have one big annotated morphograph, and not separated training-testing dataset. Fortunately, another interesting property of a morphograph is that if one is to divide it into its connected components, each of these very component is a valid morphograph by itself. By grouping these components, we can break down the complete initial morphograph in independent sub-morphographs. This allows us to divide easily divide our total morphograph into training, validation and testing sub-morphographs.

Given a testing morphograph $\mathcal{G}$, we frame the evaluation as an information retrieval problem, as the goal of the framework is to retrieve related images. More precisely, given the definition of a morphograph, for each image A which is part of the graph, its neighbours (i.e. images connected with A in the graph) are the most relevant images with respect to A. This enables us to generate testing search queries where given an image in the test graph, a search system should be able retrieve the images connected with the query. The corresponding search system is made such that given an image will retrieve the elements in the collection with the higher similarity with the query. The standard evaluation metrics of information retrieval can then be used, for instance the percentage of elements retrieved in the top N result (Recall at N, that we denote as $R[N]$ later), or the mean Average Precision (mAP).

Challenges

Image similarity and image retrieval have been established domains of work for some years already, but the precise task we are facing is presenting some unique challenges that we are describing here.

First, the visual domain we are tackling is unusual. Almost all the image retrieval datasets are based on color photographs of real modern scenes. In our case, our images are often poor quality black-and-white photographs that were acquired in the beginning of the 20th century. More importantly, we want our similarity metric to be invariant to medium changes (engravings, paintings, drawings, etc.) or painting style. This is a form of cross-domain similarity, which can be encountered in the case of sketch-based image retrieval, where a system tries to retrieve images from the drawing of a user. However, unlike in the case of

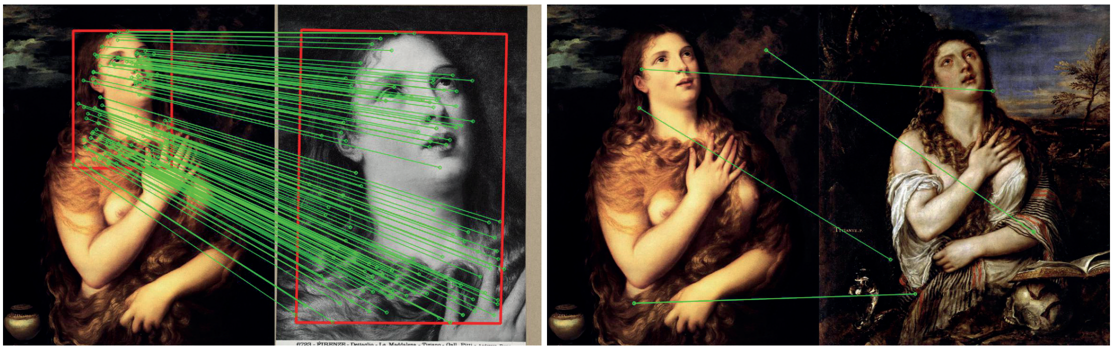


Figure 3.17: Example of the power and limits of handcrafted local descriptors approaches. They allow precise matching of the same image despite large cropping (left) but fail completely when the similarity is higher level (right) despite the two *Maddalena* having been made by the workshop of TIZIANO.

sketch-retrieval, we do not have two distinct domains as we are just given a large set of images coming from a multiple different visual domains.

Most image retrieval datasets consist on retrieving instances of the same object or building. As such, one of the key challenges is usually to handle the change of viewpoint and scale, a task at which feature point descriptors excel at (see Figure 3.17). In our situation, we are looking at the transmission of 2D semi rigid patterns, and will focus on CNN approaches to tackle the cross-domain invariance.

3.5.2 Approach

In this section, we will present the different approaches we used to estimate the similarity function s.

Pre-trained Convolutional Feature Vectors

Deep Convolutional Neural Networks are extremely powerful models for almost all vision tasks. More interestingly, when trained on very large corpus like ImageNet[16], it has been shown that the intermediate representations they learn constitutes a very good base representation of the visual information [31, 60], allowing them to be reused directly in other computer vision pipelines.

More specifically, applications of these pre-trained CNN to the problem of visual instance retrieval have been studied in [63, 89] on the classic *Oxford5k*, *Paris5k* and *Holidays* image retrieval benchmarks. In these works, an important point of discussion is which intermediate representation of these pre-trained network should be reused, and how to derive a fixed-length vector from it. More precisely, we need a function $f : \mathscr{C} \rightarrow \mathbb{R}^D$ that transforms an image to fixed-length vector.

Given an initial image A being a 3-channels RGB array of size $[h, w]$, it can be represented as a 3D tensor of dimensions $h \times w \times 3$. Using that tensor as input for a CNN, the intermediate representations are usually in the form of convolutional feature maps $F_{j,k,l}$, i.e. 3D tensors of dimensions $h' \times w' \times d$ where $h = s.h'$ and $w = s.w'$ [8]. In order to convert this feature map (whose first two dimensions depend in the size of the input image) to a fixed-length vector, an aggregation operation is necessary. One can consider the feature maps as the spatial aggregate (first two dimensions) of individual feature vectors (last dimension): a list of $h' * w'$ visual words, each of dimension d. From this observation, a common aggregation scheme is to aggregate these visual words together, losing their spatial position, like a soft bag-of-words. This would result in a single fixed-length representation of size d. Popular reduction functions are the sum, or the maximum pooling operation:

$$f_{mean}(A)[l] = \sum_{j,k} F_{j,k,l}$$

$$f_{max}(A)[l] = \max_{j,k} F_{j,k,l}$$

Both these equations can actually be generalized to the generalized p-mean. Indeed, with $p = 1$ or $p \rightarrow \inf$ for sum pooling or max pooling respectively, the previous equation are special cases of:

$$f_p(A)[l] = \Big(\sum_{j,k} F^p_{j,k,l}\Big)^{\frac{1}{p}}$$

After the aggregation, the descriptor is then normalized to unit-norm. The full pipeline of the function f can be seen on Figure 3.18. Based on this, a simple similarity function s between two images can be defined as the inner product of their respective visual descriptors:

$$s(A,B) = f(A).f(B)$$

---

[8] s being the stride of the network at the level. Often $s = 2^i$ with i representing the number of spatial reducing operations (2x2 pooling, or convolution with stride =2) the input went through before reach this intermediate representation

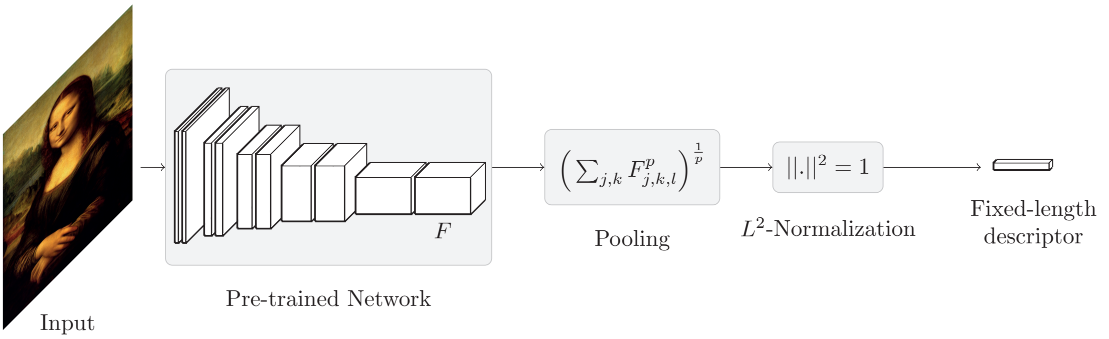


Figure 3.18: Network architecture for transforming an input image to the corresponding visual descriptor.

Fine-tuning the Network with Hard Triplets

Most pre-trained networks are trained on image classification. In the most common cases, object classification data (ImageNet[16]) is used as it is the biggest and most diverse standard dataset for image classification. As such, available networks pre-trained on this very dataset are plenty. However, while these networks have learned powerful general vision features, they are still best performing for the case of object classification which is not what we are interested in here.

Starting from a pre-trained network, we want to leverage a training morphograph to fine-tune the network parameters to improve the performance of our model. Remember that a morphograph induces many constraints of the form $(A_i–B_i) >_{\mathscr{C}} (A_i–C_i)$ because of that fact it is locally consistent everywhere (see Definition 6 on page 66). Because s should emulate the original partial ordering in the set of connections $\mathcal{C}(\mathscr{C})$, but in $\mathbb{R}$, the constraints should be valid for s as well: $s(A_i, B_i) > s(A_i, C_i)$, which translates to $\Delta_i = s(A_i, B_i) - s(A_i, C_i) > 0$.

In order to optimize for these sets of constraints $\left\{s(A_i, B_i) - s(A_i, C_i) > 0\right\}_i$, we need a differentiable loss function forcing these constraints to be satisfied. The most common choice for an ordering constraint is to apply the *Hinge Loss* : $l_m(\Delta_i) = \max(0, m - \Delta_i)$, which is a linear penalization of how violated the constraint is, but is equal to 0 when the constraint is validated (see Figure 3.19).

This loss function applied in this fashion for similarity learning is often referred to as the *Triplet Loss*[66], as each constraint relies on three input elements $\left(A_i, B_i, C_i\right)$. From this, we

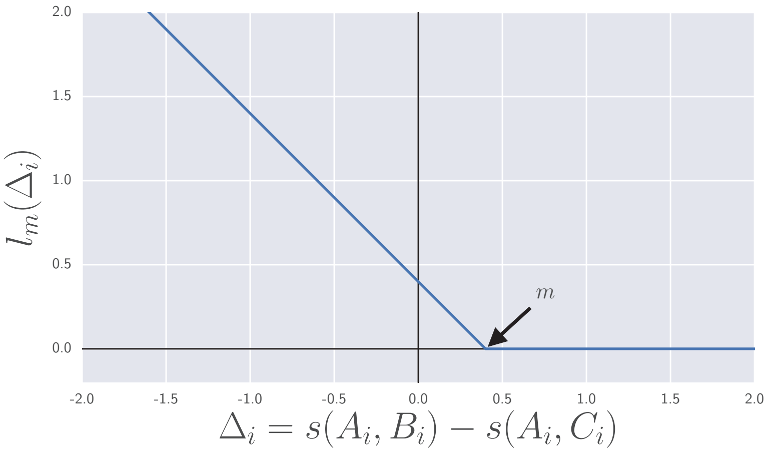


Figure 3.19: Plot of the hinge loss function. Notice that when the constraint is satisfied ($\Delta_i > m$ i.e. $s(A_i, B_i) > s(A_i, C_i) + m$), the loss is effectively equal to 0. The parameter m pushes the constraints to be valid with a certain margin, it can be very small or even equal to 0.

can derive the total loss $\mathscr{L}$ we try to minimize as the sum of the loss for each constraint:

$$\begin{aligned}\mathscr{L} &= \sum_i l_m(\Delta_i) \\ &= \sum_i \max\big(0, m - s(A_i, B_i) + s(A_i, C_i)\big)\end{aligned}$$

Our goal is to optimize the parameters of s in order to minimize $\mathscr{L}$. Assuming s is differentiable, we can optimize its parameters with gradient descent using batches of triplets of images (A_i, B_i, C_i). However, for a given morphograph the number of triplets to choose from can be extremely large, and most of them will have a zero loss (as the values of the similarity are already in the right order), hence not creating any update for the parameters of s.

More precisely, in our case, the triplets (A_i, B_i, C_i) we can generate from the morphograph are when A_i and B_i are connected in the graph, but A_i and C_i are not, even through physical connections (see page 66). Thus, the choice of A_i and B_i is relatively limited as it is bounded by the number of edges of the graph. On the other hand, given a choice of two connected images (A, B), almost any image of the collection can be added to complete a valid triplet (A, B, C).

Similar to [67] and [68], we mine hard examples during the optimization process to improve our visual similarity function s. More precisely we sample *hard triplets*: triplets where the corresponding constraint is violated using s, which are then informative for refining s. For this, we use a search index to find, for a given pair (A, B), the images C not connected with A but

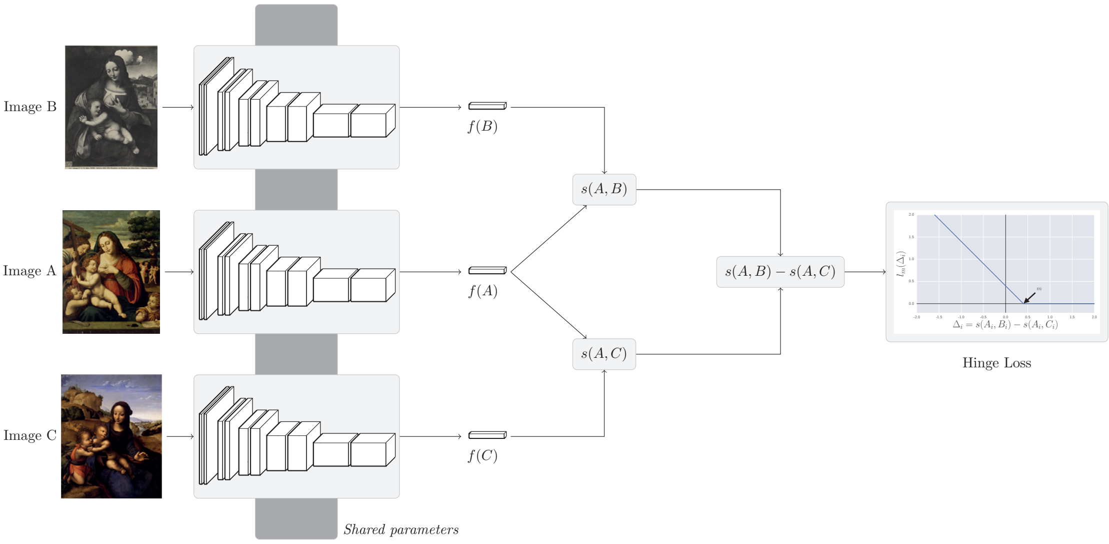


Figure 3.20: Schematic for triplet learning where the similarity function is defined based on fixed-length global descriptor (like in the previous sub-section). The three paths of the computation share the same network parameters, which get updated if necessary (i.e. if $s(A,B) \not> s(A,C)$).

with the highest $s(A,C)$. However, since the parameters of the similarity function s are updated during the training process, so does the set of hard triplets. Thus, we need to iteratively mine hard-triplets, and optimize the function s, as shown by the following pseudo-code:

```
Require: Initial similarity function $s$, collection $\mathscr{C}$, morphograph $\mathscr{G}$
for $n = 1..$Number of training epochs do
    $index \leftarrow$ BUILDSEARCHINDEX($\mathscr{C}, s$)
    $\left\{(A_i, B_i, C_i)\right\}_i \leftarrow$ SAMPLEHARDTRIPLETS($\mathscr{G}, index$)
    $s \leftarrow$ OPTIMIZESIMILARITYFUNCTION($s, \left\{(A_i, B_i, C_i)\right\}_i$)
end for
```

Spatial re-ranking

Transforming an image to a fixed-length descriptor is very practical. Comparison between images can be done with a single distance computation between vectors (or a correlation in our case). All the descriptors can be pre-computed in advance, and they can easily be stored to build a search index by leveraging standard libraries that perform efficient nearest-neighbor searches.

However, such a projection might be too much of a constraint to properly represent the complexity of the similarity function s. For instance, if for three images A, B, C, $s(A,B)$ and

$s(A, C)$ should be high, while B and C could be unrelated and should have a low similarity (see Figure 3.21). Because of the triangular inequality, this is a fundamental limitation of using a metric on an embedding space[9].

In order to tackle these limitations, and improve the quality of the similarity, we propose to leverage the spatial matching mechanism used on bag-of-words descriptors. Indeed, in the pooling operation, we aggregate the feature map $F_{i,j,k}$ (3D tensor of shape $h' \times w' \times d$ for an input image of shape $h \times w$), to form a fixed length descriptor, effectively projecting the two spatial dimensions (indexed by i and j here). However, we could also consider the feature map as a collection of $h' * w'$ regularly sampled high-level area descriptors of size d each. Each of this descriptor is also coarsely localized according to its i, j coordinates in the 3D tensor.

Based on this remark, one can use a similar approach as with SIFT-matching. Given two input images A and B, we compute the two corresponding feature maps, $F(A)$ and $F(B)$ respectively. Each of these feature maps represents a set of descriptors (as described above), correspondences between the descriptors are found based on a cross-check criteria[10].

These correspondences $\mathfrak{C} = \left\{\left((i, j), (i', j')\right)\right\}$ between coordinates (i, j) in $F(A)$ and (i', j') in $F(B)$ are then filtered according to the best found spatial transformation between them. Because the transformation is coarse, the coordinates unprecise and the number of descriptors low, we use a more simple transformation than a full homography:

$$\begin{bmatrix} j' \\ i' \\ 1 \end{bmatrix} = H_{s,\Delta_h,\Delta_w,\delta_m} \begin{bmatrix} j \\ i \\ 1 \end{bmatrix} = \begin{bmatrix} \delta_m s & 0 & \Delta_w \\ 0 & s & \Delta_h \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} j \\ i \\ 1 \end{bmatrix}$$

Here, the number of free parameters is only three: the scale s, and the spatial shifts Δ_h and Δ_w. Left-right mirroring is also allowed with the binary value δ_m as it is relatively common, for instance when a print is made based on a painting (in that case, the wood-block is carved with the proper orientation but the printing process will flip the representation). Also, it means only small rotations are allowed.

Using this scheme, we can then define another similarity measure between two images A and B as the maximum number of correspondences which are spatially consistent with respect to

---

[9] Note that even if we used the cosine similarity via the correlation of the vector embedding $f(A)$ and $f(B)$ in the computation of $s(A, B)$, it is very much linked with the euclidean metric in our case. Indeed, because $f(A)$ and $f(B)$ are normalized, $||f(A) - f(B)||^2 = 2(1 - f(A).f(B)) = 2(1 - s(A, B))$

[10] Given two sets of descriptors $\mathscr{D} = \{D_i\}$ and $\mathscr{D}' = \{D'_j\}$, we consider the pair i, j to be a match if D'_j is the closest descriptor to D_i in $\mathscr{D}'$, and respectively D_i is the closest descriptor to D'_j in $\mathscr{D}$

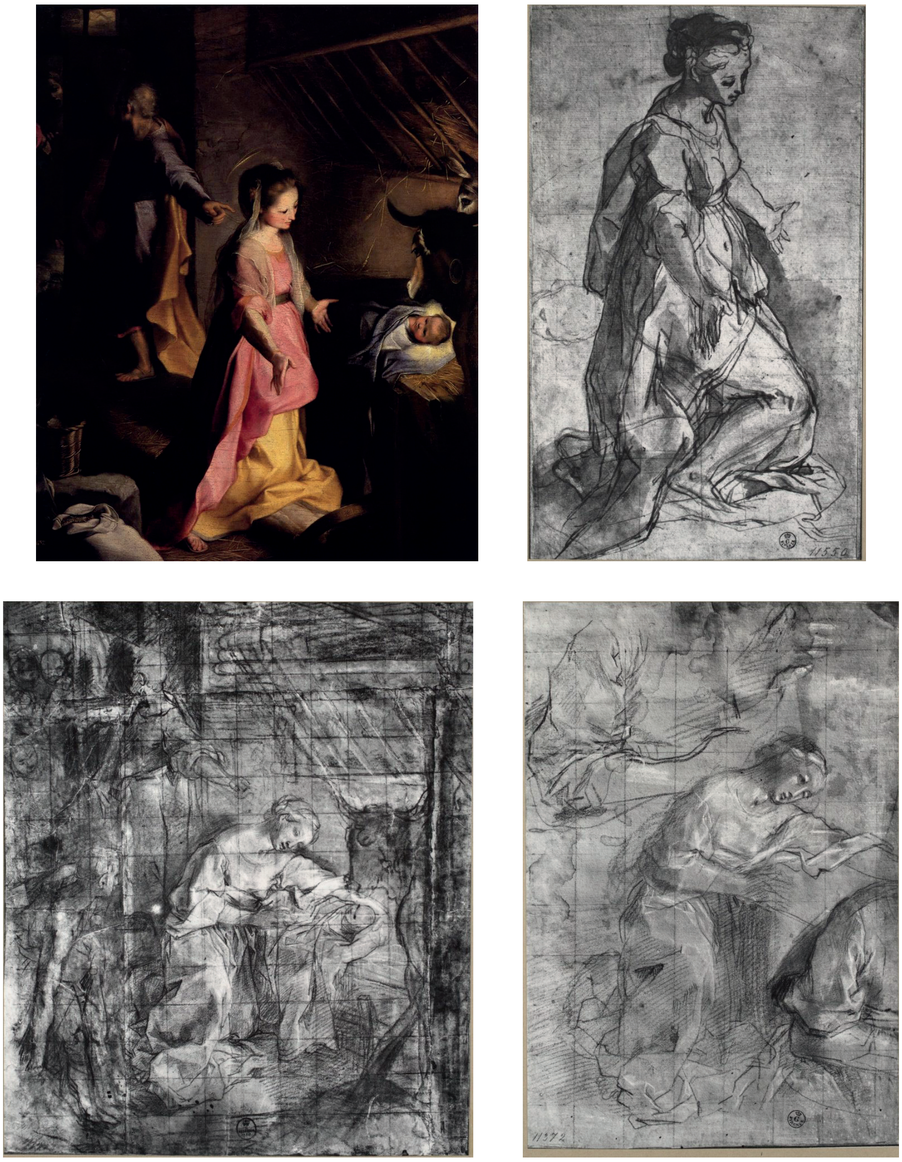


Figure 3.21: Multiple works by FEDERICO BAROCCI. On the top left, the painting of *The Nativity* from the Prado Museum in Madrid, is clearly related to two preparatory sketches (top right and bottom left), while the two studies have little in common. Finally, the drawing at the bottom right is also clearly related to the second study, but not to the original painting. Interestingly, the Cini description of the documents do not really help finding these connections, the one on the top right being referred to as *Studio per una figura femminile in atto di inginocchiarsi* and the bottom right as *La Samaritana (?)*.

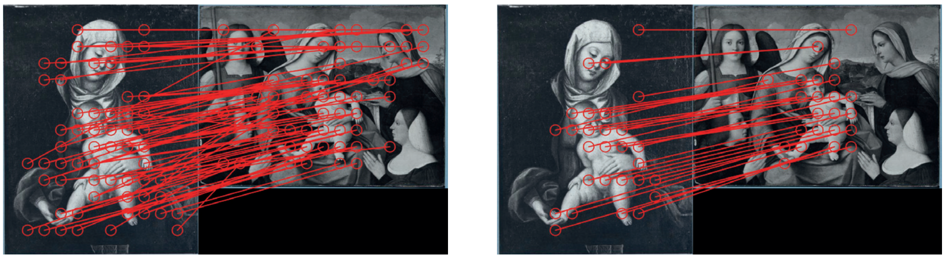


Figure 3.22: Example of using the CNN feature maps to match images. Left: candidate matches computed from the CNN descriptors (each cell of the feature map). Right: resulting matches based on spatial filtering.

a margin ϵ:

$$s(A,B) = \max_{s,\Delta_h,\Delta_w,\delta_m} \#\left\{ \big((i,j),(i',j')\big) \in \mathfrak{C} \mid \left\| \begin{bmatrix} j' \\ i' \\ 1 \end{bmatrix} - H_{s,\Delta_h,\Delta_w,\delta_m} \begin{bmatrix} j \\ i \\ 1 \end{bmatrix} \right\| < \epsilon \right\}$$

3.5.3 Implementation Details

The evaluation is done on the images extracted from the Web Gallery of Art, complete with all the images sharing at least a visual connection in our constructed morphograph. This gives a total of 41’595 images, with 3’146 images being part of at least one visual connection. The visual graph is divided into a training, validation, and testing subgraph (40%, 10% and 50% of the original graph respectively). From the testing subgraph, we generate one query for each connected node, with its connected elements as targets for the query[11]. This gives us a benchmark of 1’658 queries, on which we compute the standard metrics for image retrieval.

The networks used are the VGG16 [21] and Resnet50 [23] architectures, pre-trained on ImageNet [16]. For VGG16, we only use the convolutional parts i.e. up to *pool5*, leading to a feature map with a depth channel of dimension d=512, and a stride of 32. For Resnet50, we use the full network, giving a d=2048 feature map, with the same stride. We experimented with values of p equal to 1 (sum-pooling), 2, and inf (max-pooling), during the pooling operation before normalization.

The input images are resized so that their larger side is equal to a fixed max-size (320, 480 or 640 px). This allows us to batch the images during the fine-tuning process in a square of constant max-size. This greatly helps the computation as images have varying ratios. Training is performed with batches of 8 triplets at a time, using the ADAM optimizer[70] with an initial learning-rate of 10^{-5} (slowly decaying) and l2-weight decay of 2.10^{-5} for Resnet50 and 5.10^{-4}

---

[11] Additionally, images sharing a physical connection with the query are ignored during the evaluation process, in a similar way that the Oxford5k dataset [55] treats the “junk” images.

for VGG16. The hinge loss margin is set to 0.01. An important detail is that we do not update the batch-norm parameters during the fine-tuning process, as the statistics of the images we work with are drastically different compared to ImageNet images, leading to a drastic loss of performance.

As for the spatially consistent similarity, we cannot efficiently compute every pair in order to perform an image search. We use the standard re-ranking approach, fetching a fixed number of candidates (400) based on their fixed-length descriptors first, before computing the spatially consistent similarity. The feature maps of all the images are computed in advance with the VGG network, and retrieved on the fly as necessary to perform the re-ranking computation.

One might think that storing dozens (or hundreds) of thousands of pre-computed feature maps would be a problem storage wise. Indeed, for a max-size of 480, the average size of a VGG16 feature map is 330KB (assuming a floating point precision of 4 bytes). However, these feature maps are very sparse signals, as only 14.2% of the values are non-zeros (a side-effect of the ReLU activation function). In a similar but less drastic approach than [62], we leverage this sparsity for compression, by only storing the non-zero values, and the position offsets between them. Additionally, we quantize the non-zero values to a single byte, which is a common operation for compressing neural networks, especially for embedded devices. Finally, the obtained position offsets and quantized values are compressed with Zstandard[90]. The final size of the compressed feature map is on average 12KB, which is more than 27 times smaller than the reconstructed data, and is the same size of a 3'000 dimension floating point descriptor (with 4 bytes precision). This makes the computed index for the 41'595 images slightly bigger than 500MB.

Also, in order to speed up the re-ranking process, because the number of correspondences is relatively low, and the transformation coarse, we found that during the estimation of the spatial transformation H, it was faster to use a voting algorithm (similar to a Hough transform and [91]), compared to the RANSAC method more traditionally used in this situation.

The decompression and spatial-verification steps are done on the fly during the re-ranking process. It is computationally intensive but as it can be easily parallelized, we can perform a single search (re-ranking 400 elements) in 420ms on 12 threads.

3.5.4 Results

The mean average precision (mAP) and recall at different ranks are displayed on Table 3.3. From these results, it is apparent the performance of the descriptor coming out of the Resnet50 network is superior to the one produced by the VGG16 network. This is not surprising as it is a more modern architecture and it produces features with a higher dimension (2048 vs 512). Also, across both networks, the fine-tuning process drastically improves the performance, making the mAP jumps by almost 30% and 26% for the VGG and the Resnet networks respectively. Another point of similarity is the fact that both architectures perform best with the input image being resized with its maximum dimension equal to 480px. On the other hand, there is a discrepancy where the VGG network performs much better with a generalized pooling of $p = 2$ while for the Resnet architecture, $p = Inf$ (i.e. max-pooling) is more effective. Globally, the relative performance of the different configurations stays the same before and after the fine-tuning process.

The performance of the retrieval system after re-ranking the top 400 elements (both for the best fine-tuned system and the best pre-trained one), are displayed on Figure 3.4. We can see the large improvement in performance compared to just using the fixed length descriptors, going from 61.8% to 76.7% (mAP) and from 72.3% to 82.5% (recall-at-20), showing the importance of spatial consistency for the type of visual similarity we are looking for. While the best performing resolution is still 480px, the parameter p used for fine-tuning does not seem to have a clear effect on the performance. Qualitative results of search queries and the respective ranks of their connected elements are displayed on Figure 3.23.

Conclusion

In this chapter, we showed how computer vision can help with organizing images of a photo archive.

First, in order to have an editable data structure encoding the complexity of visual similarity and object sameness, we described the concepts of physical links and morphographs. As a graph structure, this formalization allows us an incremental construction of a knowledge database, which interacts nicely with our interface/search system and our machine vision algorithms.

When it comes to detecting photographs of the same artwork, we showed through various examples that the spatial coherence was the best indicator for predicting the “sameness” of two images. By training a classifier on handcrafted measures of this spatial consistency, we could detect more than 100’000 pairs of matching photographs in the Cini collection. Looking at the attribution of each photograph allowed us to automatically identify 1’217 artworks with conflicting attributions in the Venetian archive.

Then, we showed that one can fine-tune a visual metric based on an acquired morphograph. Leveraging an initial pre-trained CNN, important improvements in performance can be attained by enforcing the constraints in similarity encoded by the connections of the morphograph. Additionally, using a re-ranking step based on a more fine-grained indexing of the visual information improved further the retrieval performance of the system, showing the

Network	ft	Max-size	Pooling p	mAP	R[20]	R[50]	R[100]	R[200]	R[400]
VGG16	N	320	Inf	25.2	36.4	43.9	49.9	55.8	62.0
VGG16	N	480	1	25.2	36.1	43.2	48.4	55.4	62.3
VGG16	N	480	2	28.7	40.0	47.2	52.4	58.7	65.2
VGG16	N	480	Inf	25.4	36.5	42.9	48.3	54.9	61.5
VGG16	N	640	Inf	23.1	34.1	40.7	46.1	52.8	58.6
VGG16	Y	320	Inf	53.5	64.2	70.8	75.9	80.4	84.4
VGG16	Y	480	1	51.7	63.4	71.1	76.1	80.7	85.4
VGG16	Y	480	2	57.2	67.5	74.0	78.3	82.3	86.5
VGG16	Y	480	Inf	54.4	63.9	71.4	76.4	80.8	84.8
VGG16	Y	640	Inf	51.7	62.0	69.5	74.1	78.8	83.8
Resnet50	N	320	1	24.1	36.8	43.3	49.1	54.8	60.8
Resnet50	N	320	2	29.7	41.7	48.8	54.3	60.0	67.1
Resnet50	N	320	Inf	32.7	44.4	51.6	57.7	64.1	70.5
Resnet50	N	480	1	23.4	36.4	43.5	47.8	53.6	59.2
Resnet50	N	480	2	31.3	43.3	50.2	55.9	61.6	67.1
Resnet50	N	480	Inf	35.7	47.7	55.0	60.2	66.0	72.5
Resnet50	N	640	Inf	35.3	46.8	54.1	59.8	65.2	70.9
Resnet50	Y	320	1	46.6	59.8	68.1	73.7	79.4	84.7
Resnet50	Y	320	2	52.4	65.0	73.0	78.3	83.1	87.3
Resnet50	Y	320	Inf	54.9	66.3	73.2	78.4	83.5	88.4
Resnet50	Y	480	1	52.4	64.6	73.5	78.8	83.6	88.1
Resnet50	Y	480	2	57.7	69.0	76.1	81.1	86.0	89.5
Resnet50	Y	480	Inf	**61.8**	**72.3**	**78.8**	**82.9**	**87.3**	**90.2**
Resnet50	Y	640	Inf	59.4	69.4	76.9	81.3	84.7	88.6

Table 3.3: Summary of the retrieval performance for global descriptors.

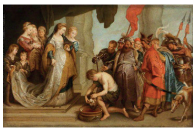
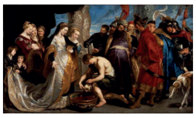
ft	Max-size	Pooling p	mAP	R[20]	R[50]	R[100]	R[200]	R[400]
N	320	N/A	58.6	65.4	68.3	69.9	71.3	72.5
N	480	N/A	63.3	67.5	70.0	71.2	71.8	72.5
N	640	N/A	61.2	66.2	68.8	70.3	71.7	72.5
Y	320	Inf	70.6	78.2	83.1	86.3	88.5	90.2
Y	480	1	**76.7**	**82.5**	**85.9**	**88.0**	**89.3**	90.2
Y	480	2	76.4	81.7	85.8	87.4	89.0	90.2
Y	480	Inf	76.6	82.2	85.9	87.7	89.1	90.2
Y	640	Inf	74.4	80.5	84.9	87.2	89.1	90.2

Table 3.4: Summary of the retrieval performance for global descriptors with the re-ranking process performed with the VGG feature maps on the top 400 candidates. “ft” indicates wether or not fine-tuning of the networks were done. “Max-size” corresponds to the maximum input size used, both for the fine-tuning of the network, and for the computation of the feature maps. “Pooling p” represents the parameter of the generalized mean used during training.

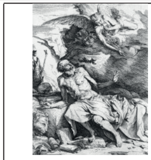
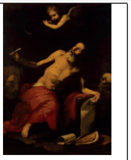
Pretrained	>400	>400	141
Fine-Tuned	177	38	1
Reranked	2	65	1

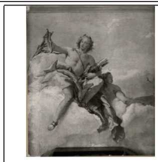
Pretrained	>400	3	>400	2	229
Fine-Tuned	2	13	4	3	1
Reranked	5	3	4	2	1

Pretrained	216	>400	>400	>400	>400
Fine-Tuned	9	1	10	166	29
Reranked	4	2	1	3	5

				
Pretrained	146	2	293	>400
Fine-Tuned	2	1	3	5
Reranked	2	1	4	3

Pretrained	81
Fine-Tuned	210
Reranked	1

	
Pretrained	>400
Fine-Tuned	>400
Reranked	>400

	
Pretrained	>400
Fine-Tuned	393
Reranked	1

Pretrained	>400
Fine-Tuned	39
Reranked	3

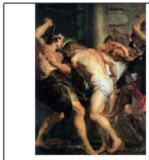
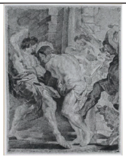
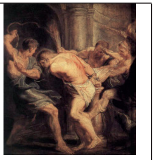
		
Pretrained	>400	102
Fine-Tuned	1	6
Reranked	1	2

Figure 3.23: Examples of retrieval queries. The query image is displayed on the top left, and for each target, its rank in the search results is displayed according to the corresponding search system. The best configurations for the respective subsets (Pre-trained, Fine-Tuned, and with Re-ranking) are used.

importance of spatial organization when doing cross-domain image similarity.

Both the detected physical connections and the learned visual similarity are fundamental to enhance the navigation and search experience of these large collections of images. Indeed, clustering photographs of the same object is fundamental in order to properly aggregate multiple overlapping collections coming from different institutions. Here we showed that this could be done even without looking at the metadata of the documents. Also, removing redundant images can facilitate the exploration of these large corpuses. On the other hand, the visual similarity function is the most fundamental component in our search system. Thus being able to continuously improve its performance as the morphograph expands is crucial to the effectiveness of our complete system.

4 The Replica Search Engine

In the previous chapter, we mentioned that the complete framework for our global approach relied on a search/annotation system (Section 3.3). It is the key tool allowing the users to interact with the large collection of images. As such, we needed to build an interface system that could be used by archivists to search and navigate their collections (at the Cini for instance), and at the same time works for us as an annotation tool to acquire our training data. In this chapter, we will present the design and implementation of the corresponding answer that is the Replica search engine.

First, we will go over the different features our system needed to satisfy. Then, we will show how we leverage IIIF to index distributed collections of image resources from potentially multiple institutions. Next, the multiple search modes of the interface will be covered. Finally, we demonstrate how we can leverage the similarity metric learned in the previous chapter in order to display search results in a meaningful way, fostering the discovery of additional connections in the morphograph.

4.1 Goals

Aggregate Collections

The most basic block of a search system is the data it is indexing. Indeed, the usefulness of a search system is only relative to the quantity of information it allows retrieving. In our case, we are talking about large collections of photographs with potential metadata attached to them.

Unlike most image search engines, we do not want to randomly scrap the web for every visual resource available, as the amount of relevant images (artworks) we would get that way would be minimal. Instead, what we want is to index the digitized collections of institutions which are hopefully published online. This means that our system needs to work as a indexer of a distributed set of collections, each being an independent data silo delivering metadata and visual content.

However, as a data indexer, another constraint is that we would prefer to not host the visual content of the other institutions ourselves. Even in the case of permissive licensing, it is preferable for the content to be clearly linked with and distributed by the institution itself.

Powerful Exploration Capabilities

Having a large collection of images and their respective metadata, searching capabilities are of course the fundamental feature of a search engine. Because it is a research platform, we want to be able to provide as many exploration capabilities of the indexed corpus as possible.

In terms of searching capabilities, each document is multimodal as it contains visual and textual information. Then, queries based on textual input or visual input should be allowed by the system. Additionally, the users could want to explore the collection based on both types of inputs at the same time for instance: “show me photographs similar to the one I selected, but attributed to Leonardo da Vinci”.

Also, aside from searching documents, representing the results is important as well. Given a large set of images to display, a single list might not be the most efficient way of representing the information to the user. In order to give a better “big picture” of the visual results, we might want to leverage more advanced visualization techniques.

Graph Edition

Finally, as presented in the global framework in the previous chapter, the interface system should be the tool for the users to annotate and modify the connections between the images of the collection. Indeed, being able to modify the morphograph is what fuels the improvements of the visual metric as described in the previous chapter.

As such, there should be efficient ways of adding/removing connections between the images of the indexed collections, while leveraging the navigation capabilities of the system. Additionally, because some actions might be done programmatically (the physical connections detection for instance), it might be necessary to track *who* (human or bot), and *when* the modifications were done.

4.2 Indexation of Resources

In this first section, we will deal with the gathering of the data. As described above, we are faced with a set of distributed collections of images to aggregate. At the same time we are trying to avoid delivering and collecting the images ourselves, as we are not owner of the data.

First, we will describe the IIIF standard, and explain how it is a suited solution for our task at hand. Then, we will present how we use IIIF as a basis for our system.

4.2.1 International Image Interoperability Framework (IIIF)

The International Image Interoperability Framework (IIIF for short) is (according to its website [92]) “a set of shared application programming interface (API) specifications for interoperable functionality in digital image repositories.” It comes from the realization that an important part of scholar (digitized) resources are visual data: may it be manuscript pages, newspapers, maps, scrolls, or any photographs. But at the same time, these image resources are “locked up in silos, with access restricted to bespoke, locally built applications.”

By providing a definition of a common language through their APIs, data providers and data consumers can separately develop long term and interoperable solutions. For instance, software for efficiently delivering high-resolution multi-scale images can be improved, enhancing the delivery of data, while a plethora of web applications leveraging these resources appear.

IIIF Image API and Presentation API

There are multiple APIs defined by IIIF, but here we will only cover the two relevant ones for this work, the *Image* and *Presentation* API.

Perhaps the most fundamental part of IIIF is how image files can be delivered from one provider to a consuming application. The Image API is the solution to this. Each image resource is represented by an URL which provides a standard way for an application to request different parts (or the totality) of the image at multiple resolutions. On the other hand, the image provider can also publish information about the resource in a standardized way, such as its license, its maximum resolution, the sizes that can be retrieved efficiently, etc. This offers an elegant solution to the problem of displaying thumbnails and potentially very high-resolution details of the same visual resource, from a single Unique Resource Identifier. Indeed, visualizers can request on the fly what they need, may it be a low-resolution full view of an artwork or a very detailed corner to examine individual brush-strokes.

The Presentation API, on the other hand, is about describing views using the image resources. Since an object can be representing the pages of a book, a single image for a painting, or the

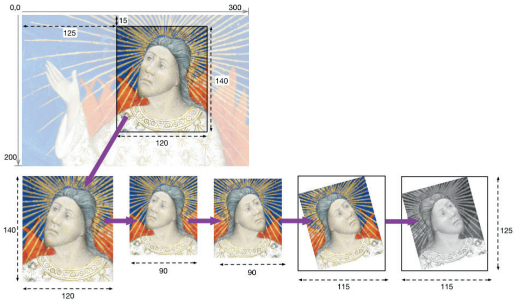


Figure 4.1: Example of a Image API request, where a portion of an image is selected, resized, rotated and converted to grayscale from a single image request.

multiple views for a statue, these multiple images need to be organized as a single entity. In order to attain this goal, in the Presentation API, an object is represented through a *manifest*.

A manifest contains two things: a *description of the object*, and *how it should be displayed*. About the former, simple information is given such as a short description of the object, the institution it comes from, the webpage describing it (usually on the institution website), simple metadata (as a list of (key, value) pairs, such as ("author" -> "Leonardo da Vinci")), etc. On the other hand, most of the rest of the manifest describes the visual content of the document: how many pages are there, in which order should they be displayed, which IIIF Image resources should be used, etc.

Additionally, these manifests (which can be referenced with the URL they can be accessed from), can be grouped into collections[1]. This allows institutions advertising the list of objects it is publishing online, all from a single URL end-point.

In the end, it seems the IIIF standard is a very good fit for our situation. The Presentation API permits a simple publication of the visual resources in an unified way across institutions, while keeping a reference to their own website with additional information. Also, the Image API allows for them to be the true deliverer of the images in a direct manner, even if their content is embedded in another application, hence keeping a proper citation of the resource.

Converting our data to IIIF standard

In our case, we needed to convert our collection of images to a IIIF repository. Our data comes from two main sources, the Web Gallery of Art (WGA) [82] and the photo-collection of the Cini foundation.

---

[1] Collections can also contain collection, which permits a more fine-grained organization than a pure flat list of manifests.

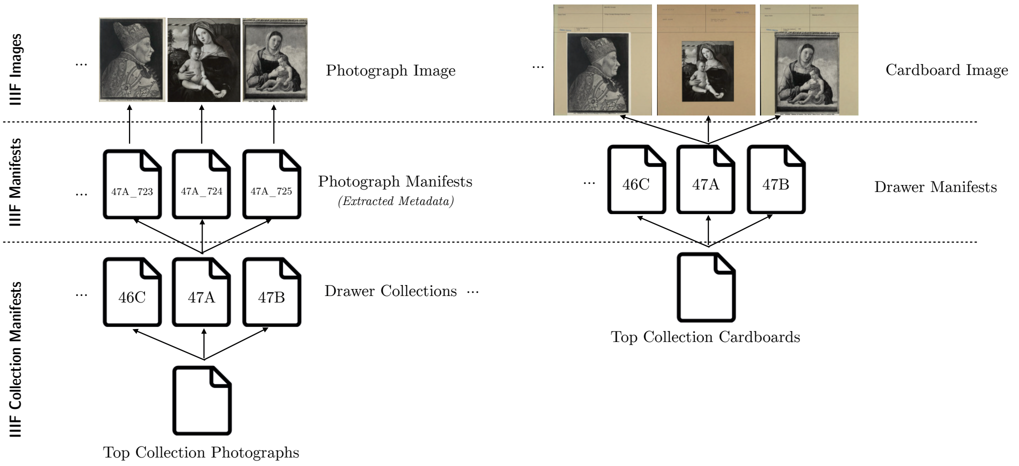


Figure 4.2: Organization of the data extracted from the Cini photo-archive (see Chapter 2).

In the case of the WGA, the complete metadata (with image links) can easily be downloaded in a tabular form from the website itself. Since they do not have a IIIF service, we downloaded everything and converted it to a single IIIF collection of manifests (one manifest containing only a single image), with images hosted on our own image server. However, the "homepage" of each object is the actual web page of the WGA, where there is often additional information about the artwork.

In the case of the Cini, the story is slightly different. For each document, we have successfully extracted an image of the document (cardboard), an image of the photograph and a structured metadata representation (see Chapter 2). As a basis, it is natural to form a manifest for each extracted photograph, with its corresponding metadata. However, something we highlighted was that the metadata extraction process, while well performing, is not perfect. As such, being always able to go back to the raw document is a necessity.

In order to handle this, we also host the extracted cardboards as IIIF images, and have one manifest per physical drawer collect the corresponding documents. Given such a manifest, any IIIF viewer allows scrolling through the original cardboards in the same order they are physically in the corresponding drawer of the photo-archive, which allows getting the feeling how the collection was originally organized. This also permits to have a URL to be set as "homepage" in the manifest of the extracted photograph, effectively linking the image to its primary source in the original archive. This is also important as related photographs might have been put next to each other during the archivistic process, a proximity which would not show when all the images of all collection are gathered in a flat indexing system.

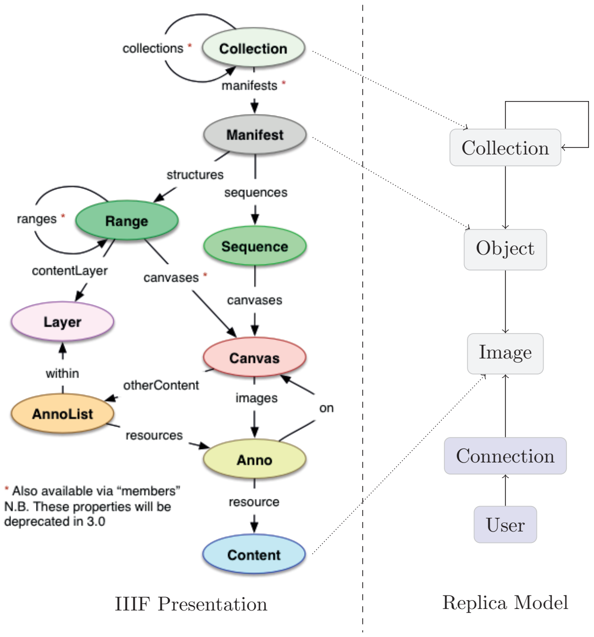


Figure 4.3: Conversion of the IIIF Presentation data model to our Replica representation. We only extract the relevant information i.e. the images displayed and the metadata of the object. In light blue, we have the elements specific to the Replica model, they correspond to the annotation system (annotated connections between the images, and annotators).

4.2.2 Our Approach

Because we have converted our collections to a generic IIIF format, we can now formalize our search system as an indexer of IIIF collections. This allows nicely separating on one side, all the digitization and processing work on the photo-collections, and on the other side, the indexation of the extracted resources.

We assume we have a set of URLs, each of them referring to the top collection of the images of an institution. Starting from each of these endpoints, one can navigate the references to the other resources in order to extract all the manifests, and the images contained in the collection. This is akin to a web-crawler which follows URLs from one page to another and extract the content it sees. In practice, we do not keep the complete structure of the IIIF Presentation format, as it is far more complex than what we need. As seen on Figure 4.3, the data model we have is only extracting the references to collections, objects (with the corresponding metadata), and the images attached to it. In addition to these immutable imported elements, the fundamental addition is that we will also have (mutable) connections between the indexed images, in order to represent the duplicate images and the morphograph (respectively physical and visual connections).

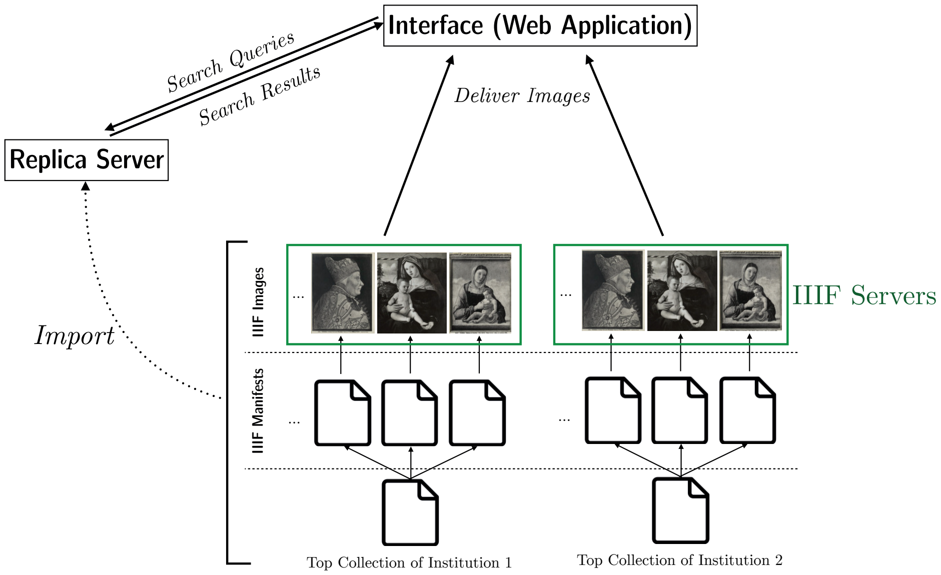


Figure 4.4: The architecture of the Replica system. IIIF collections are imported into our indexing server, which answers the search queries. The interface is based on a Web Application which displays the images of the search results directly from the infrastructure of the institutions.

In the end, for the sake of indexing the contents, we download all the metadata of the indexed objects, and a middle resolution version of the referenced images. This permits to compute the visual descriptors and index the textual information locally, such that answering the search question can be efficient. However, if an application sends a query to the search server, the answer will only contains references to the IIIF images of the institution directly (see Figure 4.4). This way, the application will display the visual contents directly from the infrastructure of the institutions, ensuring that the highest resolution document is available, and bringing proper attribution to the document.

Limitations of the current approach

This approach, despite being successful, suffers from two main drawbacks. The first one is that, because it is a static import of the collections, if a modification is done on the institution side (for instance updating the metadata, or adding new digitized document), the changes would not show on the search server. As each object and image has a unique identifier, the only solution right now is to perform a full reimport of the manifests and merge it with the already indexed data, hence updating the modified/added elements. However, that is a costly operation, especially for the institution infrastructure, as we need to perform many individual requests to gather the information of each document. Fortunately, the IIIF consortium is already working on a Discovery API [93], which would allow institutions to broadcast the changes done to their collections, effectively enabling a data aggregator to stay up-to-date.

The second pitfall of this approach for distributed aggregation is about the standardization of metadata. Indeed, we heavily rely on the “metadata” field of the IIIF manifests to acquire textual information about the images. In our case, we were mainly limiting the metadata search to a full-text search in the fields corresponding to the author attribution, and to the short description of the object. However, even this proved problematic as the corresponding fields can have a different naming depending on the institution. For instance, “Title” or “Description” for the description of the photographed artwork, or “Attribution” being used interchangeably with “Author”. In practice, when crawling each collection, we have to manually specify which one is used by each institution.

In fact, this should not come as a surprise as it is stated in the documentation of the Presentation API itself that the metadata elements are “pairs of human readable label and value to be displayed to the user” and that “there are no semantics conveyed by this information, and clients should not use it for discovery or other purposes.” Indeed, the Presentation API has no aim at bringing standardized metadata together, unlike linked data specifications (CIDOC CRM [94] for instance, or Linked Art [95]).

In the end, the distribution of visual resources through IIIF is working very well, but there are some limitations coming from the automatic discovery of new elements, and the harmonization of metadata. However, both these drawbacks can be resolved in the very near future, effectively paving the way for an efficient distributed indexation of the visual resources of these institutions.

4.3 Searching Capabilities of the System

In order to allow the maximum flexibility for the users when navigating the collection, we have implemented several search modes in the system.

4.3.1 Metadata query

The first and simplest mode is a text-only search. In that case, the results are decided from the metadata of the IIIF manifest extracted during the import and attached to the corresponding images. The indexed fields are the *author* (or attribution) and the *description* (or title) related to each image. Additionally, we also allow filtering by date. In many cases, we do not have a precise dating of the artwork, but given an attribution we can estimate a coarse time range based on the birth and death date of the corresponding artist. By combining both parameters, it allows a user to query images such as “crucifixions before 1600”, or “Titian’s Mary Magdalene”.

Something that proved quite important was to permit the text searches to match metadata in a “fuzzy” way, i.e. even if the words of the document are not exactly the words of the query. Indeed, because the Cini collection was digitized automatically with OCR, there are small errors in the extracted text and using a form a fuzzy matching permits to bypass them during the textual search. Moreover, the metadata of the documents are a mix of different languages (Cini being Italian, while the WGA is in English), it simplifies some queries (for instance, searching for “maddalenna” will retrieve “magdalene” as well).

Finally, the metadata queries are very important as they can be used as a form of pre-filtering before performing a visual search. Indeed, one might want to search similar images to a query but only among, for instance, the production of a given artist, or among the representations of madonna and child. Performing a query based on metadata before the visual search not only filter the search results as desired by the user, but it also shrinks the search space for the visual search. This is especially useful in the cases where we use a re-ranking step[2] as it is more likely that relevant images will not be missed by the first stage of the search;

4.3.2 Single and multiple images queries

In the case of a single image query, the results are computed following the procedure described at the end of Chapter 3. 1’000 candidates are retrieved based on the similarity between their visual descriptors and the descriptor of the query. Then these candidates are re-ranked according to the number of inliers between their respective feature maps. This naturally gives a matching region in the retrieved images, which we can show directly in the interface (see Figure 4.7).

However, in some situations, using a single image might not be enough to convey the intent of what we are interested in. In that case, a possible solution is to allow the user to select more than one image to be used as query. Moreover, some of these images might represent some examples of what the user does *not* want to be retrieved.

This is a standard problem in active image retrieval, where the user can iteratively add positive or negative examples to the query based on the search results, improving the quality of the search at each round. In our case, we are using a SVM classifier to rank the images. Each image is represented with its visual descriptor (fine-tuned Resnet-50, see Section 3.5), and the SVM is fitted as a standard binary classifier with the set of positive and negative image samples selected by the user as query. The obtained model is then applied to all the images and their prediction scores are used to rank them as search results. In the case where no

---

[2] This is when the visual search is performed in two steps: first a cheap search to select the most likely candidates (for instance 1000), before computing a more expensive similarity score between the query and each of the candidate.

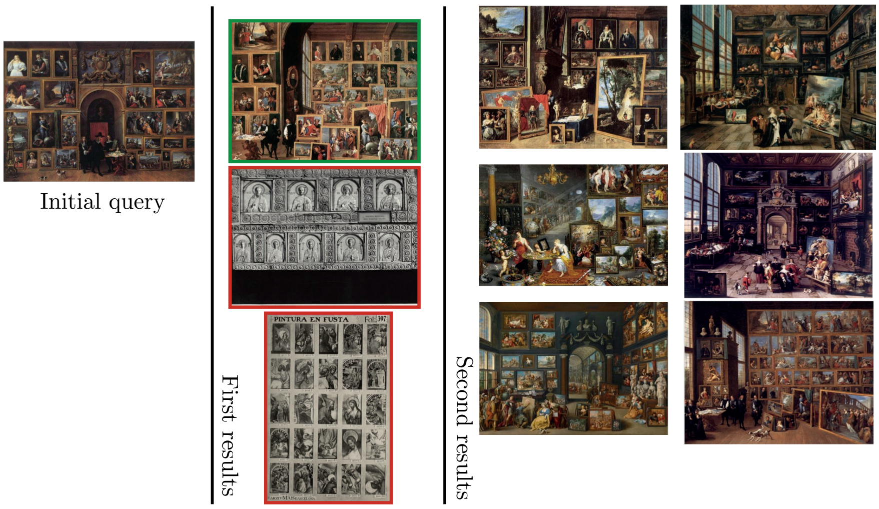


Figure 4.5: Searching visually for collectionism. With the first query, multiple wrong results presenting a regular square structure will appear (*Scenes from the life of Christ* for instance). Adding the elements as positive (green) or negative (red) allows us to search again and get better results.

negative samples were selected, we fall back to a one-class SVM model as described in [96]

In practice, we cannot use cross-validation to choose the hyper-parameters of the SVM for each query, as the number of samples given by the user is usually small. So we have to fall back to standard constant values. Because the descriptors are already normalized, we found that using a RBF kernel with $\gamma = 1$ and $C = 1$ to perform correctly.

4.3.3 Image region query

Finally, another way to make a visual search is by selecting a portion of an image as the input query. This is especially useful when the user wants to look for a specific detail, for instance singling out a character or a face from a composition.

In this scenario, we based our approach on a re-implementation of integral max-pooling[62]. This algorithm allows searching for the best matching area in a image in a efficient manner. It relies on the fact that our visual descriptors are obtained by aggregating multiple cells of the computed feature-map.

As a reminder, given a convolutional feature map $F_{j,k,l}$ (first and second dimensions are the spatial ones, and the third dimension is the feature dimension), one can compute a global descriptor of the complete image by taking the sum over the spatial dimensions of the local descriptors $\sum_{j,k} F_{j,k,:}$. However, if one is interested only in a portion of the image,

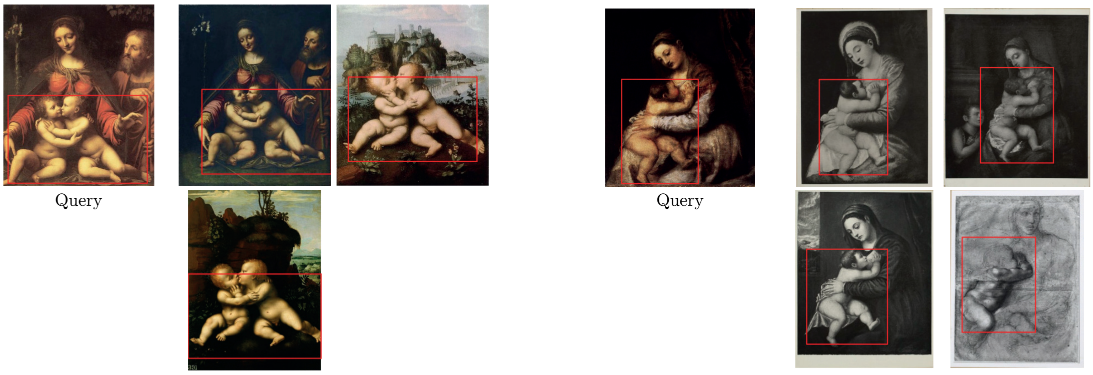


Figure 4.6: **Left**: search of Bernardino Luini, *Holy Family*. First result unattributed, the two following by Marco d'Oggiono
Right: search of Tiziano, *Madonna col Bambino* (58A_490), constrained on the text query "Madonna". The three top results are unattributed copies (58A_410, 58B_483, 58B_676), the fourth result is a sketch of Michelangelo (29A_235) estimated to have been done 40 years before the Tiziano painting.

the summation can be performed only on the relevant part of the image, which permits to compute a visual descriptor for the desired region.

In practice, given an image and a corresponding bounding box as query, we compute the region descriptor from the pre-computed feature map. This region descriptor is used as a query to retrieve 1'000 image candidates, a exhaustive search of the regions is performed for this set of candidates. Compared to the original paper, we use sum-pooling instead of max-pooling, and each local descriptor is obtained by dividing the area in a 2x2 grid. Sum-pooling is applied on each cell, giving a local descriptor of dimension $4 * d$ instead of d.

4.3.4 Other capabilities

There are also some other useful capabilities of the system that are not searching capabilities.

The first one is the ability to display, given two selected images, the blending animations presented in 3.4.1. This is a crucial tool for deciding if two images share a physical connection or not, which is necessary when trying to annotate the type of the connection.

The second one is the possibility to filter duplicate images in the results. Indeed, depending on the situation, we might want to work at the level of the photograph or at the level of the object. In the first case, we could want to look at the difference of metadata between the photographs, from which institution/photographer they are from etc. In the other case, we are more interested in the relationship between objects, and don't want the results to be cluttered with 10 images of the same popular artworks. This filtering step is relatively easy to do as we have already automatically detected the physical connections in the previous chapter. Hence, by returning at most one image per connected component of the graph defined from the physical connections, we achieve the desired output.

Finally, a last feature that is fundamental to the purpose of our system is the ability to connect images together. Indeed, one of the two main goals of the interface is to be able to annotate new connections between images in order to increase the training data for the learning algorithm presented in the last chapter.

4.4 Displaying Results

4.4.1 Interface

A capture of the interface of the web application can be seen on Figure 4.7. The main characteristic is that two selections of images (a “current” selection, and a “negative” selection) can be continuously kept and modified by the user. These selections are used as input for the several actions described previously. The view of the interface can be divided in three main parts:

- The first part at the top is where the states of the two current selections are displayed, and the actionable inputs allowing starting a new action (may it be a textual search, a visual search, an annotation action, etc.).
- The second part is where the results are displayed. It is by default a scrollable list of the retrieved images with buttons allowing adding the images to either of the two selections. Moreover, one can jump directly to the corresponding page of the institution data-silo, describing the aggregated photograph.
- Finally, on the left, a view of the current selected items stays on the side as one can scroll the result items. This proved quite important when one wants to compare the selected item with one of the search results, far in the list, as they would be located near to each other, at a relatively even size.

Map visualization

We realized that a simple list of results was not always the most practical way of displaying results, especially if there is no specific ranking. For instance, in the case of a metadata query, the results are only filtered depending on the fact that they match the textual query or not. As

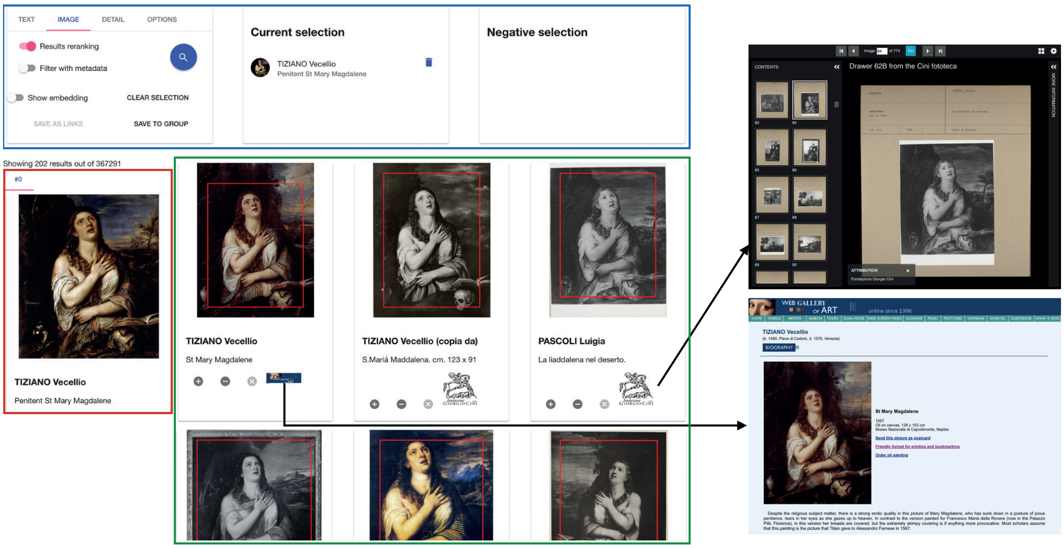


Figure 4.7: Main view of the interface system. Selections and actions at the top (blue), scrollable results list on the right (green), and current selection on the left (red). The black arrows highlights how one can always go back to the original source of the aggregated photograph, which is the raw cardboard from the photo-archive in the case of the Cini, or the corresponding webpage for the Web Gallery of Art. In both cases, this can bring additional information about the indexed document (additional notes, handwritten corrections, physical position in the archive, etc.).

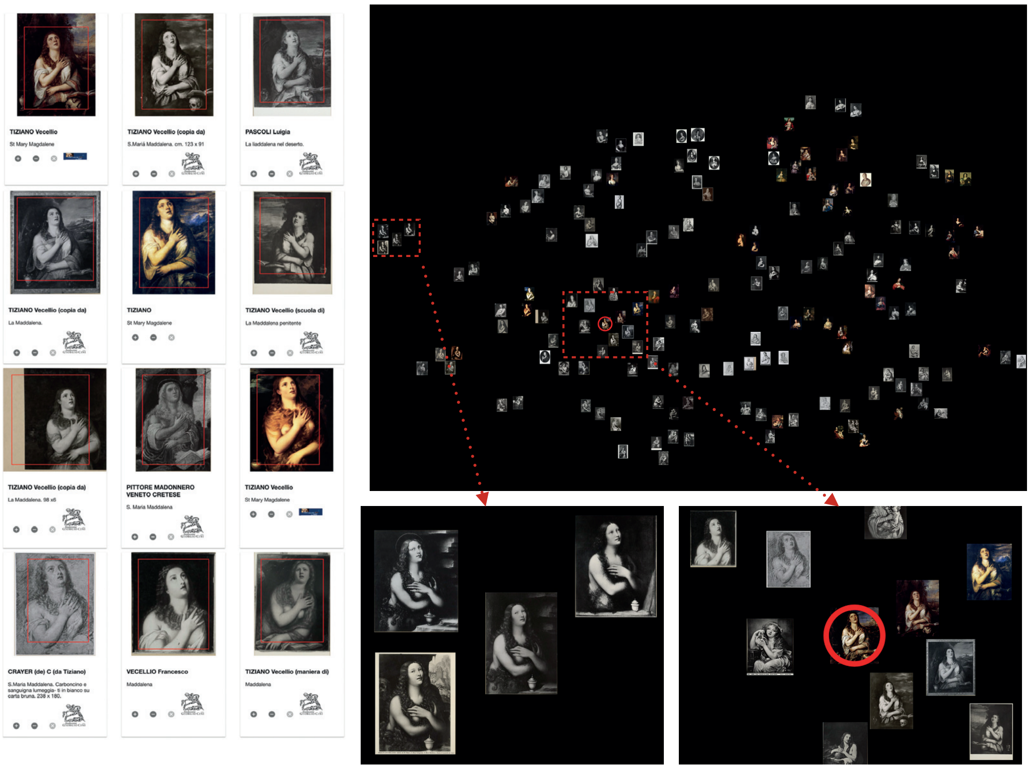

TIZIANO Vecellio
St Mary Magdalene
TIZIANO Vecellio (copia da)
S.Mariá Maddalena. cm. 123 x 91
PASCOLI Luigia
La liaddalena nel deserto.
TIZIANO Vecellio (copia da)
La Maddalena.
TIZIANO
St Mary Magdalene
TIZIANO Vecellio (scuola di)
La Maddalena penitente
TIZIANO Vecellio (copia da)
La Maddalena. 98 x6
PITTORE MADONNERO VENETO CRETESE
S. Maria Maddalena
TIZIANO Vecellio
St Mary Magdalene
CRAYER (de) C (da Tiziano)
S.Maria Maddalena. Carboncino e sanguigna lumeggia- ti in bianco su carta bruna. 238 x 180.
VECELLIO Francesco
Maddalena
TIZIANO Vecellio (maniera di)
Maddalena

Figure 4.8: The two ways of visualizing results. Left: a scrollable ordered list of results, suitable for looking at the result of a visual query. Right: the map, with the top-results being located near the initial selection (red circle), but where additional discoveries can be made. Here the highlighted left cluster is made of 4 different paintings of another version of Mary Magdalene, attributed to GIAMPIETRINO, CESARE DA CESTO, and an anonymous painter from the circle of LEONARDO DA VINCI.

such, in the case of representing an even set of images, we might want to leverage the image similarity to organize the elements in a more meaningful way.

One of the main goals of our system is to help the users find more visual connections to be added to the morphograph in order to improve the visual similarity metric. In order to help the user find pairs of visually connected elements within the subset identified through a textual or visual query, we would want two images with similar visual descriptors to be displayed near to each other, while others are displayed farther.

More precisely, given a set of images and their corresponding visual descriptors, we need to compute for each image the (x, y) coordinates where it should be displayed on the visualization. This is a standard problem of dimensionality reduction, where we need to convert our high dimensional input vectors to 2-dimension vectors, while preserving some of the distances in the original space. The most well-known algorithm for this task is probably PCA (Principal Components Analysis), but for visualization purposes t-SNE (t-distributed Stochastic Neighbor Embedding)[97], is very popular and widely regarded as a superior solution. The main strength of the t-SNE algorithm is that it is very effective at preserving the local structure from the original space to the low-dimensional space. On the other hand, if two points are “far” in the original space, then it does not really matter how “far” they will be as well in the output space.

While projecting a dataset of images to a 2-dimensional plane is already often done to visualize large corpuses of images. We think that most of the time, they fall short for two reasons. First, the two dimensional embedding is often computed only once for the whole corpus. In our case, the embedding is computed for any selection coming from the search system, which brings a much higher flexibility, and allows a more fine-grained 2-d visualization. Indeed, because search results will return only a couple hundreds images, the set is more homogeneous and its variations can be more easily represented. Second, most uses of these visualizations do not account for making the images not overlap with each other. In the worst scenario, two images have very close visual descriptors, and they will be assigned almost the same position in the visualization, making one of them totally hidden by the other. In our situation, we force all images to be displayed separately from each other.

More precisely, let us presume that we have a square viewing area of 1 by 1, and we want to draw N images, each of them with max side of S. The dimensionality reduction algorithm allows us to compute 2D positions for each image, that we can normalize so that the ranges are [0,1] in each dimension. We force image centers to be separated by at least δS (with $\delta > \sqrt{2}$ so that images are not overlapping). The size S is linked to the number of images that has to be displayed. A natural choice is $S = \sqrt{\frac{\alpha}{N}}$ so that $\alpha = N.S^2$ is a free parameter defining the ration of the total viewing space covered by images. Using the terms of [98], a small α privileges structure preservation of the image similarity while a high α places more emphasis on visibility of the images. In practice, we used a relatively low α of 0.25.

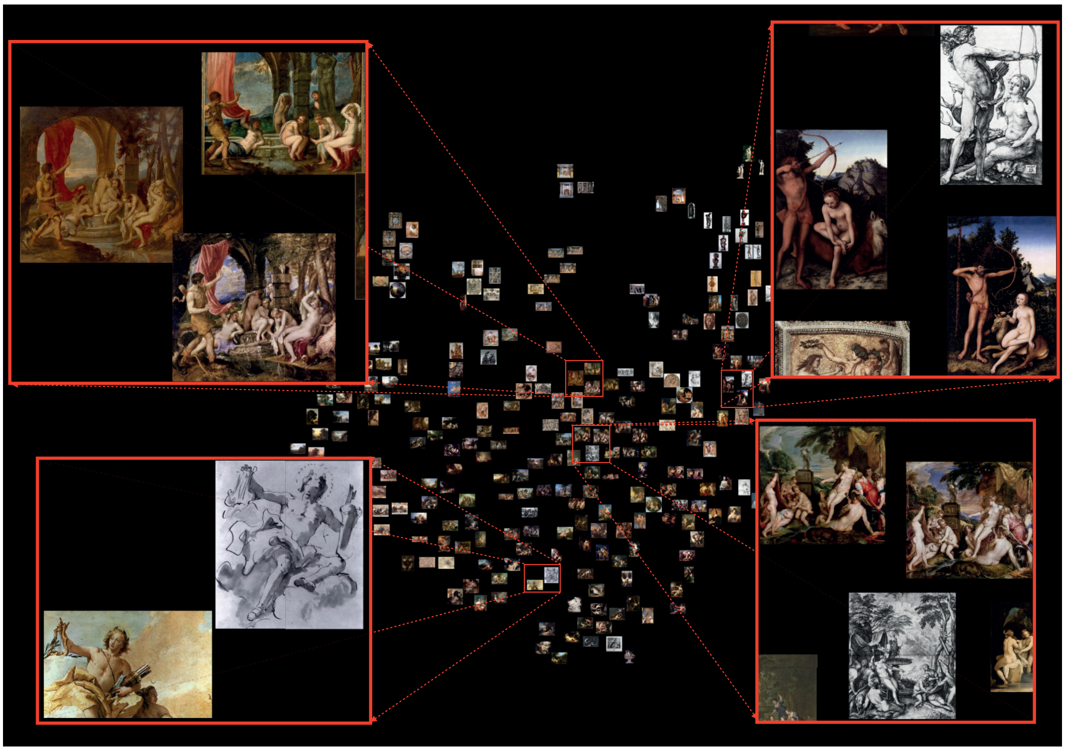


Figure 4.9: Map visualization resulting from the textual query “Diana OR Apollo” (240 images in WGA). The red highlights (added manually) show how the spatial reorganization of elements automatically create interesting connections between the images.

Users can freely drag and zoom on the map visualization interface to explore the space of visual similarity as defined by the learned visual descriptors (see Figure 4.9). They can also select elements and create connections between them directly from this view. Interestingly, this visualization does create quite different organizations, whether we filter the duplicate images (images sharing a physical connection) or not. Indeed, duplicate images will appear as a form of visual clusters, which is helpful when one wants to work at the level of same/different object in a photograph, but hinder the visualization at the object level (see Figure 4.10). This highlights again the difference between working at the photograph level, or at the object level.

4.4.2 Visual link discovery experiment

One of the main goals of the system is to be able to discover new visual connections to be added in the morphograph. Which in turn, allows learning better visual descriptors, hopefully easing the task of finding even more connections. Thus, a positive feedback loop is created.

However, finding such connections is a difficult task, as it is by itself a quadratic problem.

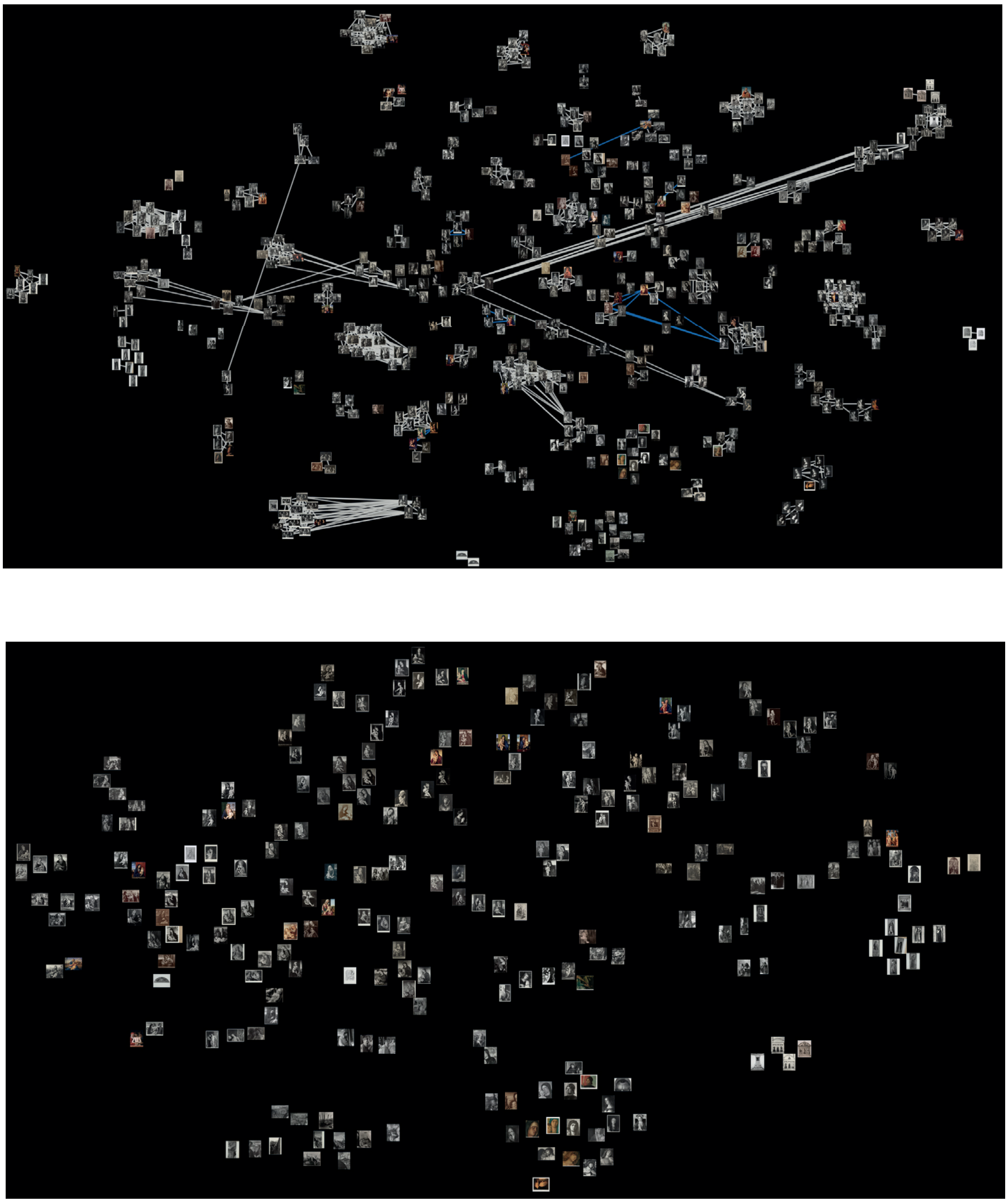


Figure 4.10: Visualization when the number of duplicate images is large and not filtered, here in the case of the 552 images coming from the textual query “Giovanni Bellini madonna”. The physical connections are shown in grey, and we can see that the space get separated by these physical clusters, which makes us not able to visualize the more abstract visual similarity between objects. When the images get filtered from the physical connections, only 237 elements are shown (one per object), and a much more continuous visual space is displayed.

Indeed, the most naive version would be to look at every pair of images, an incredibly inefficient strategy however. Of course, with the multiple searching capabilities of the system, the user can make educated guesses in order to search for a specific image, and/or filter from the metadata of the images. But at the end of these searches, there is often a set of resulting photographs to go through in a naive way, in order to find relationships between them.

However, as we have seen, if images are spatially positioned, we can consider that a user only have to “mentally compare” each image with its immediate neighbors, in order to find strong visual correlation. Hence, the order of magnitude of pairs of images examined grows only linearly with the number of images displayed, as the spatial positioning (which is a surrogate for the visual descriptor distance) provides an efficient heuristic to pre-select the pairs that should be examined.

Based on these insights, we devised a small experiment to validate two hypothesis: (1) a system that is based on a metric trained by annotations of visual connections (henceforth a “refined” system) allows users to find more connections in a new corpus of images than a standard system which was not trained on annotations. (2) spatial organization of the images in the visualization is instrumental in the search for visual connections (i.e. if the visualization does not already place images close to each other, the users rarely find and annotate these visual connections).

Protocol

Task definition: The task consisted of searching, finding, and annotating visual links in a corpus of paintings.

The control group (the standard system) was a system relying on a visual metric based on a pre-trained Resnet-50 architecture, but not trained using manually annotated visual connections. The experimental group (the refined system) was a system using the same, pre-trained Resnet-50 architecture which was then further trained on 3’381 visual connections, using the approach described in the previous chapter. The sets of images used in our experiments to test these systems were naturally not part of the training data.

Nine participants, of which five were women and four men, (Age: M=25.1, SD=4.4), none of whom had art historical training, tested the system. The experiment was conducted simultaneously with all participants. Users used equivalent personal laptops that contained similar mouse track pads.

Pre-experiment: Participants all received the same introductions demonstrating the features and functions of the search engine and describing the given task. The demonstration included an explanation of the visualization and the definition of the concept of visual links. Users were then shown a demonstration of the system and were allowed to familiarize themselves with the interface and ask questions.

Experiment: The experiment consisted of two tasks. In the first task (A) participants were presented with a map interface upon which was plotted the corpus of paintings[3] and drawings produced by the artist GIOVANNI ANTONIO CANAL (known as CANALETTO). For the second task (B) participants were presented with a map interface upon which was plotted the corpus of paintings and drawings created by the artist TIZIANO VECELLIO (known as TITIAN). These two examples were chosen because these particular painters are known to have reused similar cartoons or compositions across multiple paintings, easing the search for visual connections between paintings for a non-specialist audience.

Each participant performed one of the two tasks using a map visualization generated by the standard system algorithm, while the other task was done using a map generated by the refined system. Participants were randomly assigned to the standard or refined system, so that in the course of the two trials they performed one task using the standard system and one on the refined. Participants were not aware of which system they were using.

To start the search, participants were given personal links for the predefined query and charged with annotating visual connections. To do so, two or more paintings had to be selected, followed by the option to 'save as links'. This would result in lines being drawn between the two elements indicating their connection. The goal was to find as many visual connections between artworks as possible. Participants were allowed to keep exploring the visualization until they felt that they had discovered all connections (which usually occurred after around 8-10 minutes for each task).

Evaluation: The performance of each participant across each query was recorded, including the connections created. At the end of the experiment, participants were asked to complete a short evaluation of the system's usability. This was modeled on the SUS usability scale [99], which has been proven to be effective and remains an industry favourite [100]. Three additional questions, modeled upon the ResQue survey [101], were also added to find out specifically about the novelty of the given results, the suitability of the interface layout, and ease in getting acquainted with the system. Questions are listed in Table 4.1.

Results

The experiment showed that even users who were not art historians quickly learned how to use the interface and identified a large number of visual connections. This is further supported by the answers of the usability survey. The questions that showed the highest positive score on

---

[3] Only the images from the Web Gallery of Art were used in this experiment.

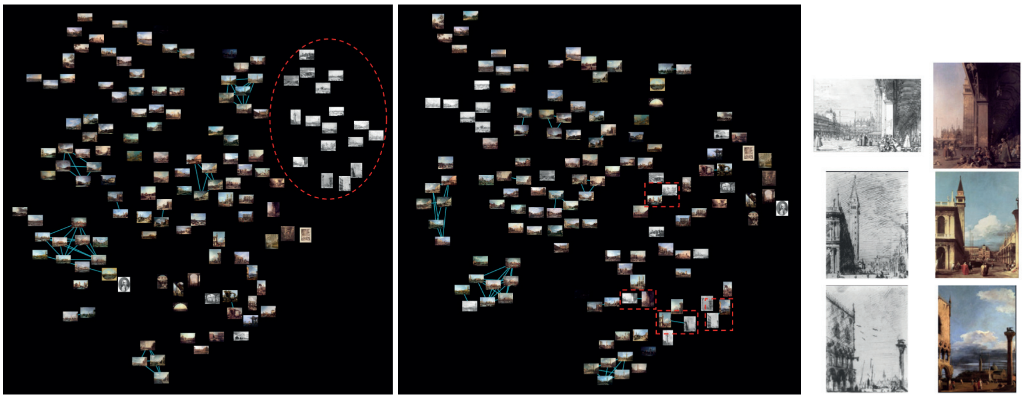


Figure 4.11: Final annotated connections' map of two participants at the end of Task A. Left visualization was created with a standard system and center with a refined one. Notice that, in both cases, no connections between distant elements were annotated. The difference between the two systems is apparent in this case, illustrating that the similarity function learned to be invariant between drawings and paintings: all the drawings are grouped with the standard metric (red circle, left), but the refined metric allowed the user to connect multiple pairs (red squares, center), which are preparatory sketches (right). It is interesting that the metric managed to group these elements, especially in the Canaletto corpus, which consisted of paintings/drawings of similar views of Venice.

the Likert scale (1.33, and 1.56 respectively) were in answer to the ease of using the system (see question 7 and 13 in Table 4.1). Overall the user survey also attested to the system's usability (with a score of 70/100 on the SUS scale) [100].

Table 4.2 shows the number of images correctly linked for each task and for each system. As we hypothesized, the number of properly connected images is higher for both tasks when users were presented with a refined system, answers which show a strong statistical significance. This attests to the generalization power of the learning algorithm and its usefulness in aiding in the navigation of the data. As importantly, it indicates that a refined system, learns from previously inputted annotations, and assists users in finding more visual connections.

A second body of results is shown in Figure 4.12, which highlights the number of connections created by users with respect to the placement of images, or distance between images, in the visualization. This shows that users tended to be efficient at connecting images that were in the vicinity of each other. But on the other hand, unless a pair of images was 'pre-selected' by the visualization algorithm to appear close together, it was extremely unlikely that users would find and annotate the connection, as we see a clear cut-off as the distance between images increases. This highlights the fact that we do not remember well images seen before during our exploration, and strengthens the importance of helping the user in finding these

	Question	Average Score
1	I think that I would like to use this system frequently.	0.56
2	I found the system unnecessarily complex.	-0.89 (r)
3	I thought the system was easy to use.	0.89
4	I think that I would need the support of a technical person to be able to use this system.	-0.89 (r)
5	I found the various functions in this system were well integrated.	0.78
6	I thought there was too much inconsistency in this system.	-0.78 (r)
7	I would imagine that most people would learn to use this system very quickly.	1.33
8	I found the system very cumbersome to use.	-0.78 (r)
9	I felt very confident using the system.	0.56
10	I needed to learn a lot of things before I could get going with this system.	-0.56 (r)
11	The system showed me new and interesting images from which I was able to learn new and relevant information.	0.89
12	The layout of the system interface is attractive and adequate.	0.33
13	I became familiar with the system very quickly.	1.56

Table 4.1: Survey questions and the average score on a 5-point Likert scale. Questions for which answer score is negated are tagged with 'r'.

		Images linked		
Task	# Images	Standard	Refined	p-value
A	138	34.3 ± 3.3	42.0 ± 3.6	0.0211
B	250	42.0 ± 5.4	55.3 ± 2.4	0.0050

Table 4.2: Number of correctly annotated images for each task. The last column corresponds to the p-value of the corresponding ANOVA test between the two systems.

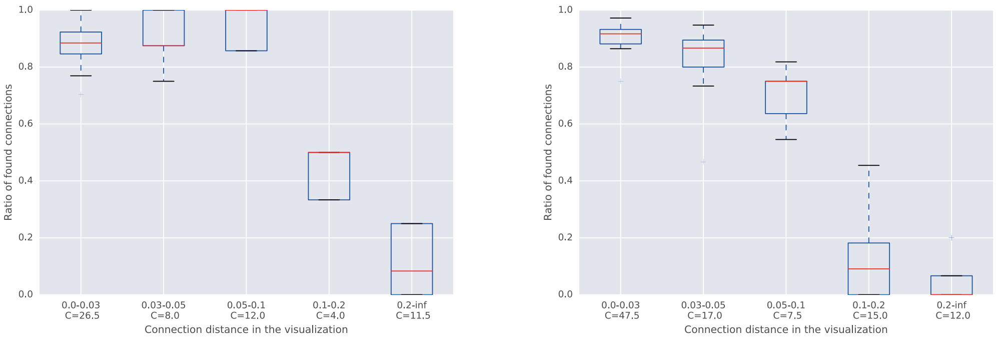


Figure 4.12: Proportions of found connections relative to their distances in the visualization. Task A is on the top, and task B at the bottom. For each task, results are aggregated between the standard and refined system. The value C corresponds to the average number of ground-truth connections falling into each distance bin. It is clear on both tasks that connections placed further apart in the visualization are more rarely found.

connections.

A related area of inquiry pertains to how well the refined system would continue to learn after its initial training. Indeed, it would seem logical that the more refined the system, the harder it would be for it to keep on improving. A difficulty would arise from the fact that easy visual connections, such as copies of two artworks, already have a very small distance between their visual descriptors, and are then placed in the visualisation in the vicinity of one another. Therefore, annotating such connections might not bring any new information to the system in terms of training and further refining the metric. On the other hand, *informative connections* may be difficult for the user to find even with the help of the interface, as they would not be recognized by the metric and would be placed further apart in the visualization.

Shortly put, an *informative connection* is a connection which will effectively modify the parameters of the function computing the visual descriptors. Given our learning framework presented in the previous chapter, an *informative connection* is defined as a pair of elements such that adding them as a connection in the database allows for the creation of training triplets with a non-zero loss. For instance, annotating a group of elements which are already nearest neighbors to each other does not foster the discovery of hard negatives useful for training, and are thus not deemed informative.

Table 4.3 represents the proportion of *informative connections* found by users in each of the two tasks, across the standard and refined systems. One can first note that the total number of informative connections to be found diminishes as the system is trained. This is what one would expect, as relatively easy connections are not training the system any more as the metric

Task	Standard	Refined
A	50.8% ± 3.3 *(out of 33)*	56.1% ± 6.6 *(out of 31)*
B	42.5% ± 8.7 *(out of 72)*	56.5% ± 3.8 *(out of 50)*

Table 4.3: Aggregated percentage of success of participants in finding informative connections for each task. The total number of possible informative connections to be found for each situation is stated in *italic*. Note that the number of informative connections available and found depends on the quality of the system.

improves. Hence, because a part is already mastered, one could assume that the number of informative connections would decrease for a refined system. On the contrary, on these two case studies, the experimental samples represented in Table 4.3, reveal that the proportion of found *informative connections* actually increases with a refined system. This is an encouraging result that seems to hint that the learning is not plateauing, and that the positive feedback loop we were aiming for is indeed working.

Conclusion

In this chapter, we have built a system that would allow us satisfying the requirements of a searching system and an annotation system.

By leveraging the latest standards for online publishing (IIIF), we show that it is now possible for a system to act as a global indexer of all art related image collections published online by institutions, with the latter keeping complete control on the data they put online. Additionally, we identified the remaining hurdles in making this process easier and sustainable.

We also described and implemented the different features such a search system should have. As a research interface, we indeed believe the users should be granted a set of tools to explore such a large collections of images. Different modes of searching, both textual and visual, with multiple ways of looking at the results, have emerged as a natural strategy.

Finally, we evaluated some of the effectiveness of the implemented system in being able to discover interesting connections between images. It allowed us to validate the fact that as the system learns; the easier it is for users to unravel new discoveries in the indexed collections.

Conclusion

Summary of the Thesis

In this work, we presented the complex layered process that went into making a large photo-archive searchable.

In the first chapter, we showed how we could efficiently digitize a photo-collection archive, using the case study of the Cini Foundation. We presented the digitization effort, which allowed for the transformation to digital format of 330'000 photographs at an unprecedented speed. Then, to attain the genericity necessary to process other photo-archives, we introduced a deep-learning framework for segmenting visual elements in historical documents, showing its effectiveness on our Venetian case, as well as on a variety of other tasks. Additionally, we demonstrated a flexible approach to align artist names with an historical knowledge database.

In the second chapter, we started by proposing a formal organization of the space of images for the purpose of "form" similarity. Based on a graph of connections between the photographs, this complex formalization distinguishes the relation of sameness (i.e. two images representing the same physical object, what we referred to as *physical connections*), compared to connections encoding relative visual similarity (*visual connections* and *morphograph*). We then presented computational methods allowing the generalization of this information. First, we used "traditional" geometric computer vision to detect duplicate images in our collection, unraveling more than 1'200 artworks with multiple attributions in the Cini photo-archive alone. Second, by combining deep-metric-learning and our graph formalization, we showed how we could learn a similarity function generalizing the visual closeness encoded by the connections of the morphograph.

In the third chapter, we presented how the latest developments in web technologies (IIIF) can be effectively leveraged in order to have independent collections of images indexed together in a single central system. Then, we detailed our approach at building a complex interface system allowing to search/navigate the large collection of photographs in various ways, as well as modifying the graph of connections between the images. This system has been instrumental in finding and annotating more than 6’200 visual connections, between 3’000+ artworks, creating without a doubt the largest dataset of its kind.

In the introduction, we stated that we were trying to answer two fundamental questions: *how to make these photo-collections accessible?* and *how do we help Art Historians navigate such large iconographic collections?*. How what we produced in this work contributes in tackling these issues?

When it comes to the first question, the answer we provided is three-fold. First, the use of a specific scanner allowed the digitization of the raw documents extremely efficiently. This permitted us to tackle the large scale of data while keeping the manual operations to the minimum (Section 1.1). Second, being able to process the scanned documents (semi-)automatically to convert them to structured documents. As photo-archives are collections of diverse semi-structured documents, being able to separate the visual elements (Section 2.2, 2.4.1) and identify the textual parts (Section 2.3) in a flexible way is crucial if we want to make the data searchable. Finally, even in the case of having clean processed data, making it available to the general public is key to its accessibility. In Section 4.2, we highlighted how recent web standards are an effective way of ensuring these collections are used to their full capacities by the greater community. The combination of these three components is what allowed us to convert the large physical (and mostly unused) stack of papers to a structured and accessible digital end-point.

About the issue of being able to navigate such large collections, the basic building block is the specific visual similarity function we learn (Section 3.5), which powers the visual search (Section 4.3) and the map visualization (Section 4.4). However, we realized how important it is to acknowledge the dual nature of the photograph (as a representation of an artwork, and as an object by itself), as different research questions will prefer one view or the other (for instance, looking at artworks with respect to each other, as opposed to comparing the multiple descriptions/authorship photographers attributed to the same object). As such, this distinction is at the basis of the formalization (Section 3.2), and appears as well in the interface as a way to filter search results (Section 4.3). Finally, while the original goal is more to use computer vision as a searching mechanism, we found using the textual information (extracted from the OCR of Section 2.2) to be quite beneficial as an additional searching cue, especially to contribute to the morphograph.

Apart from these two original questions we started with, a transverse observation throughout the project is that we rely a lot on continuous user feedback, as it clearly appears on Figure 4.13. The most important aspect is to be able to refine a system from the added inputs of the users. It might be most obvious for the similarity learning (Section 3.5) which relies on the edition of the morphograph done through the interface system (Chapter 4), but the name alignment process (Section 2.3) is very similar, as we simply iteratively improve the matching performance by

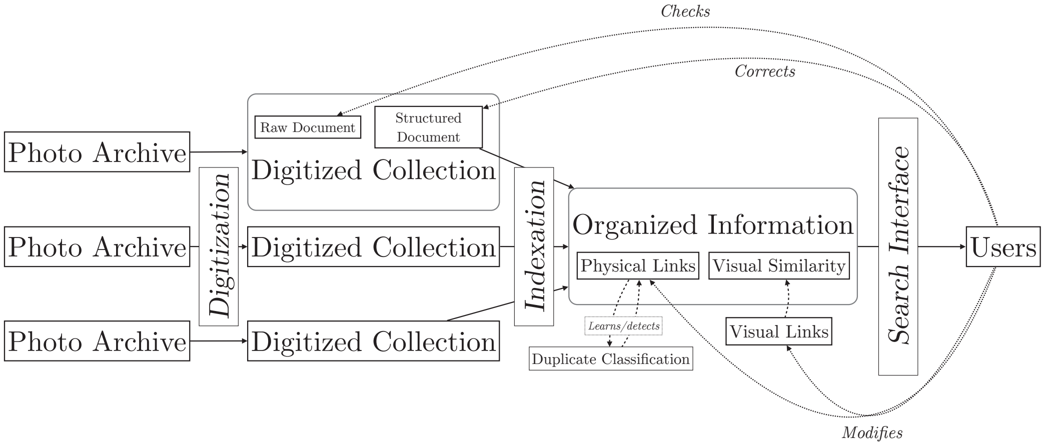


Figure 4.13: Pipeline of the project, which is much more detailed version of the simpler linear version of the introduction.

manually adding additional information (disambiguation cases, additional names, etc.). This "human-in-the-loop" approach also appears as we always give the ability for the user to go back in the pipeline, and check that the automated extraction of the document was done correctly (through IIIF, Section 4.2). We feel that this is a common characteristic of Digital Humanities projects, where there is a constant requirement for the scholar to always be able to view the primary sources, especially when an imperfect automated process (like OCR) is used.

Limitations and Future Work

This work is also opening new avenues which are waiting to be explored further. The most obvious one being that we hopefully paved the way for other institutions to undertake as well the task of digitizing their collections. By using the detection of physical connections, it is possible in the near future to automatically align the images coming from all the world's photo-collections, even without waiting for the standardization of the metadata between the institution' systems. Also as one tries to find connections between objects, the more we have (i.e. the bigger the coverage) the more likely we are to find interesting connections.

For instance, the Cini collection contains little of the works produced by the Dutch painters, and connecting the Venetian archive with some institutions in the Netherlands would allow interesting research in studying the transmission of motifs between the two regions.

Even for Italian Art, the data we have acquired is still very far from being an exhaustive coverage of the Art production that reached us. For instance, just on Wikipedia, one can find an example about multiple variations based on a design of LEONARDO DA VINCI (Figure 4.14), but in our set of 330’000 we only find two of the six examples shown on the collaborative encyclopedia[4]. This example highlights how necessary it is to keep the digitization effort going, for instance by tackling the rest of the photo-collection at the Cini (the 700’000 remaining photographs).

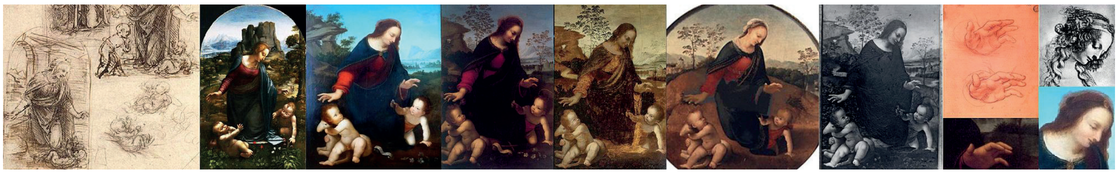


Figure 4.14: Nativity by various followers of LEONARDO DA VINCI - SALAI, CESARE DA SESTO, FERNANDO YANEZ DE LA ALMEDINA and Anonymous. (retrieved from the Wikipedia “Leonardeschi” page)

Another avenue for improvement is a more flexible visual similarity system. In this work, we focused on a metric tailored for a specific research problem (propagation of motifs, hence a more “shape-based” visual similarity), but scholars might want to be able to perform a more flexible search, for instance based on colors, focusing on the style, or a tunable combination of all of the above. For instance, as a partly texture-invariant metric, our system is not very suited for navigating a modern art collection, where the textural component might be the most important visual characteristic.

A natural extension of this project would also be the automatic discovery of these reoccurrences of pattern. Even if we defined visual connections as a relative strength in term of visual similarity, these seem in many situations to hold a certain stability as the manifestation of a strong visual correlation between two images. As such, it would be interesting to see how the machine, simply given a large collection of images, could be able to detect these correlations. Preliminary tests, only possible thanks to the acquired training data, showed some feasibility. For instance, we were partly successful in automatically finding engraving/painting pairs, but many of these connections were uninteresting relations between similar canonical representations, for instance grouping together all the female face portraits, or the crucifixes. For such an approach to work, a global view of the collection is necessary in order to assess the saliency of a connection with respect to the visual space of the artworks produced throughout history.

---

[4] The position in the Cini for them are: 48C_333 for one attributed to CESARE DA CESTO, and the second one has two photographs 136A_298 (attributed to BERNARDINO LUINI) and 136B_626 (attributed to CESARE DA CESTO).

To conclude, we are aware that aligning all the Art History photographic collections of the world is much more than a technical problem. To successfully reach this goal, many open challenges regarding copyrights, economic interests, private-public partnerships, or legal standardization, need to find a sustainable solution. Nevertheless, we hope that this thesis constitutes a significant step demonstrating the feasibility for a global search engine for Art History collections.

A “Discoveries” in the collection

In this appendix, we present a couple of interesting “discoveries” that were done navigating the photo collection of the Cini (+WGA) with our system. To clarify, we do not pretend these relations to be new to the Art History community, and by “discovery”, we only mean cases where we naturally unraveled these relations through our search engine. Also, we are just showing them as interesting strong visual correlation, and do not consider any form of influence between them.

As such, these examples should just be considered as examples of the possibilities such a search system can provide for the study of pattern propagation in the artistic production.

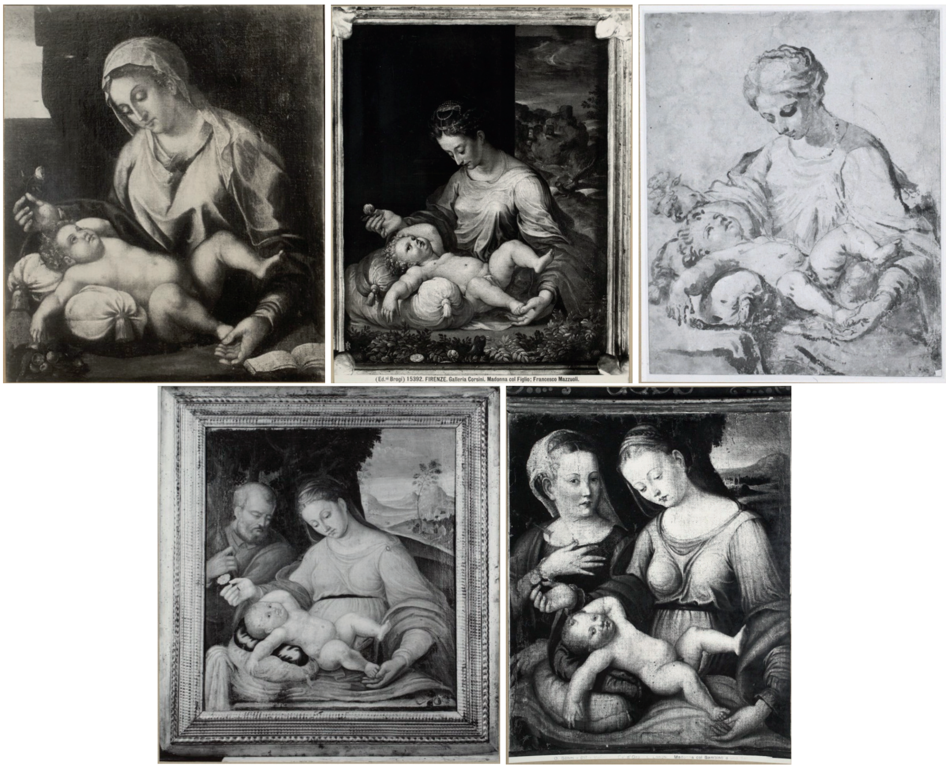

(Ed.ni Brogi) 15392. FIRENZE. Galleria Corsini. Madonna col Figlio; Francesco Mazzuoli.

Figure A.1: *Madonna col bambino*.
ANTONIO BADILE (ATTR), (46C_62), PARMIGIANINO, (139C_239), GIROLAMO BEDOLI, (20C_192) BARBARA LONGHI, (62A_489), LUCA LONGHI, (73B_405).

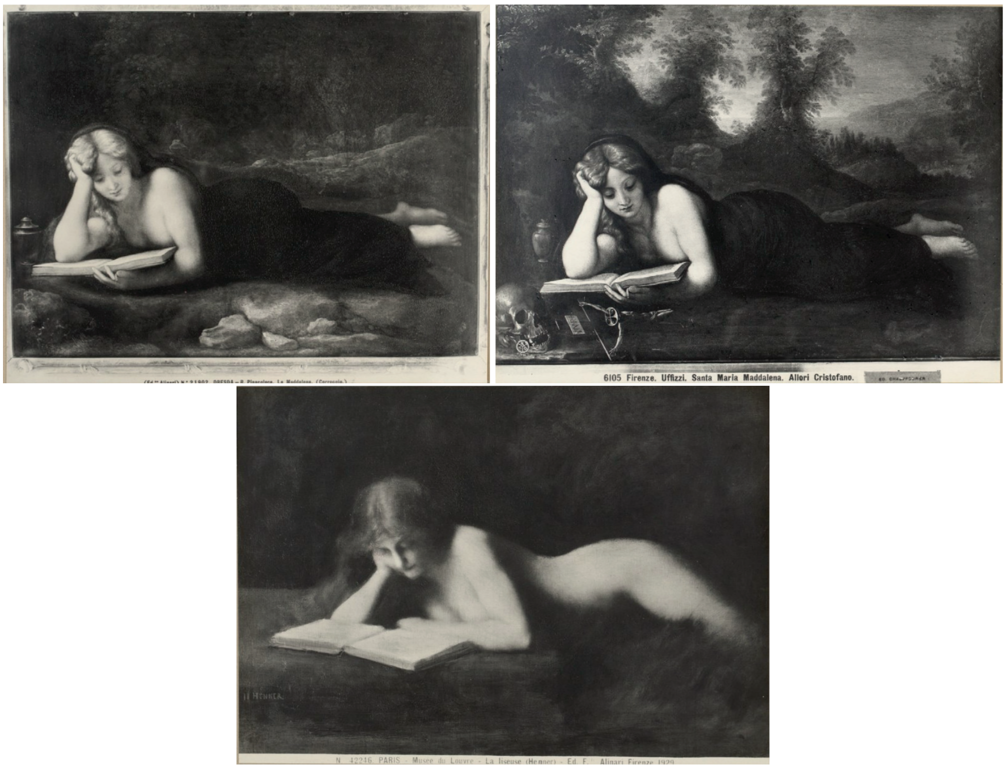

6105 Firenze. Uffizi. Santa Maria Maddalena. Allori Cristofano.
JJ Henner
N. 42246. PARIS - Musee du Louvre - La liseuse (Henner) - Ed. F.lli Alinari Firenze 1929

Figure A.2: *Maddalena.* il Correggio, (158A_560), Cristofano Allori, (134A_218) *Woman reading*, Jean-Jacques Henner, (156C_37).

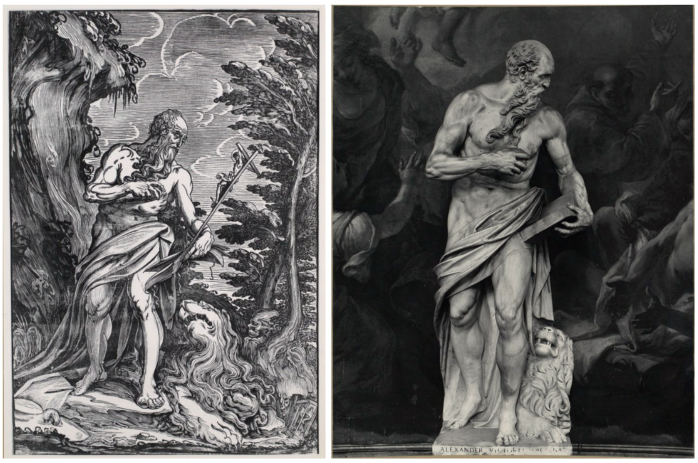


Figure A.3: *San Girolamo.*
Left: Giuseppe Scolari, (44B_979).
Right: Alessandro Vittoria, (85C_410).

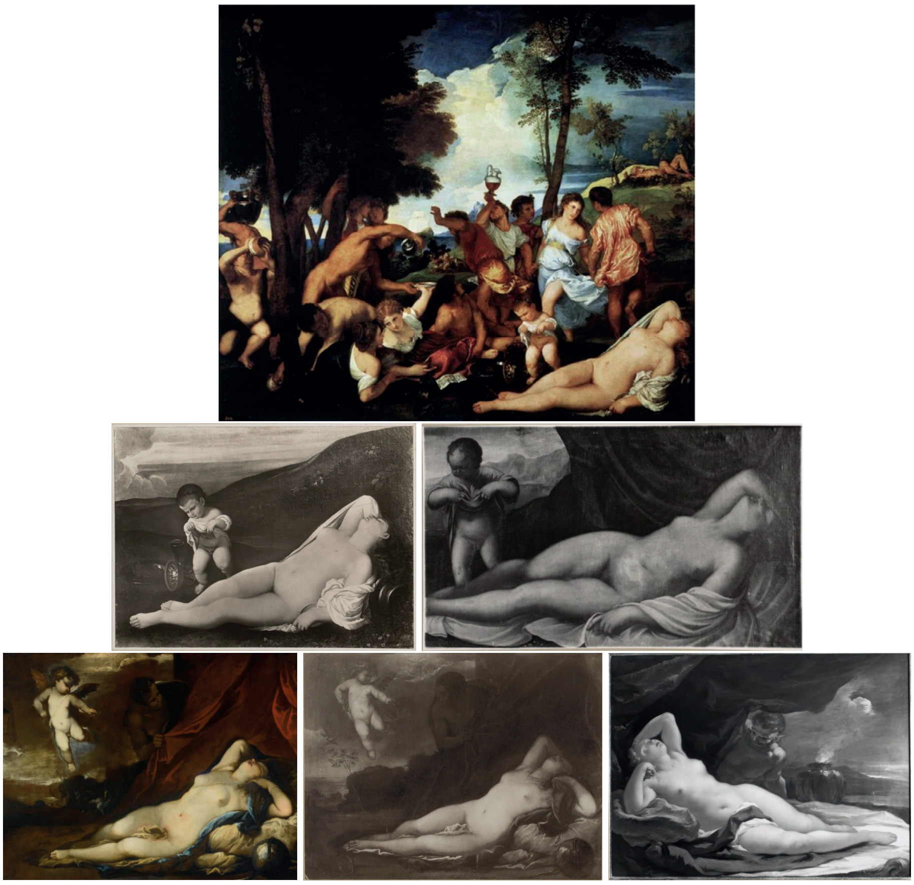


Figure A.4: *Bacchanal of the Andrians.* TIZIANO (two other known copies, 158C_245 and 58B_674).
Venere dormiente, (58B_672), *Baccante e putto* (60A_156), both wihtout attribution.
Jupiter and Antiope, ALESSANDRO GHERARDINI. *Ninfa e Satiro,* LUCA GIORDANO (SCUOLA DI) (51A_329). *Venere e Satiro,* GIOVANNI PELLEGRINI (161B_648).

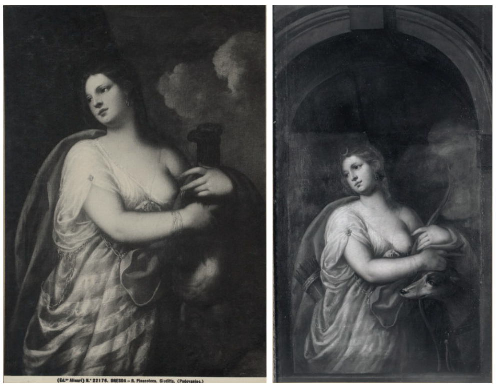

(Ed.ne Alinari) N.° 22176. DRESDA – R. Pinacoteca. Giuditta. (Padovanino.)

Figure A.5: *Giuditta,* PADOVANINO (158A_732). *Diana,* IMITATORE DI PADOVANINO (158C_466).

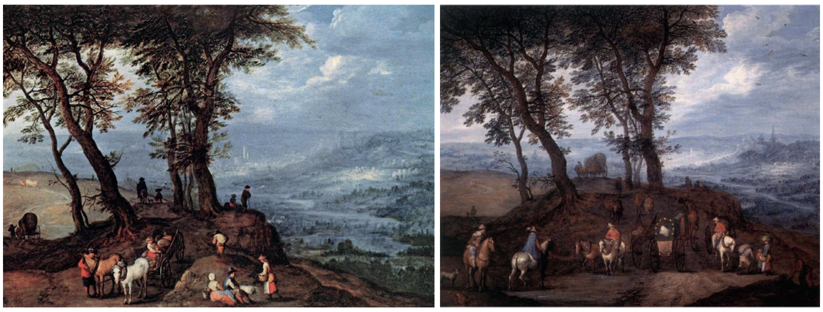


Figure A.6: *Going to the Market* and *Travellers on the Way*, JAN BRUEGHEL, THE ELDER. Example of transmission of a background with the pattern of the trees.

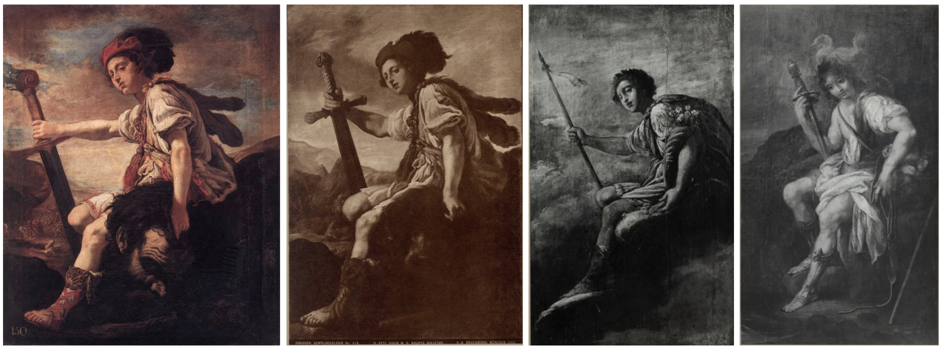


Figure A.7: *Davide*, except the third image which is an *Allegory*, DOMENICO FETTI, except the last image by ANDREA CELESTI. (First image from WGA, then 50A_333, 50A_385 and 48C_168)

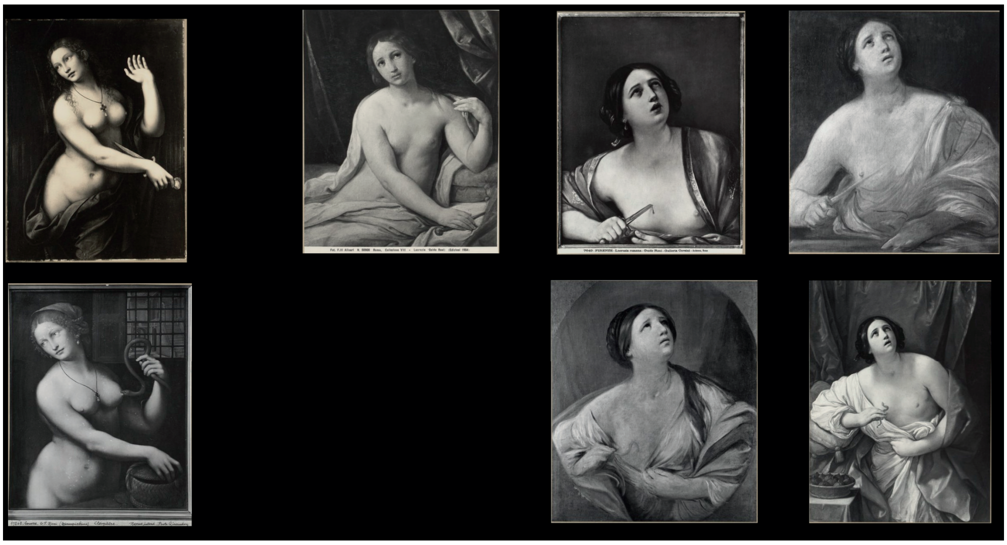


Figure A.8: Transition between the figure of two suicides by female characters: *Lucretia* (top row) and *Cleopatra* (bottom row). The two left images are by GIAMPIETRINO, all the others by GUIDO RENI.

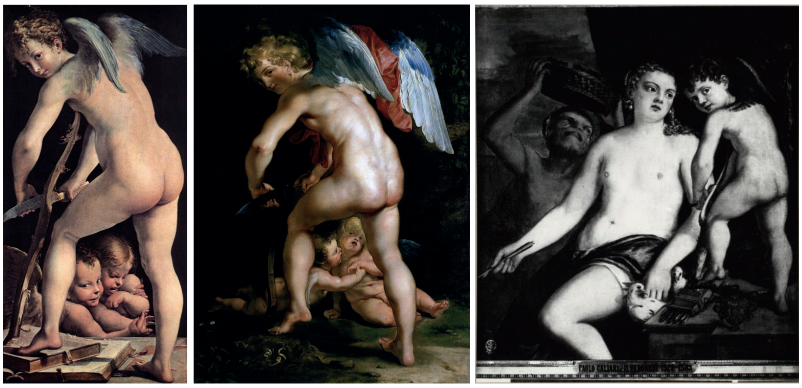


Figure A.9: *Cupido.* Parmignianino, Rubens. *Venere e satiro*, Paolo Veronese (scuola di), (111C_240). (at least 5 other anonymous copies of the Parmignianino version are known).

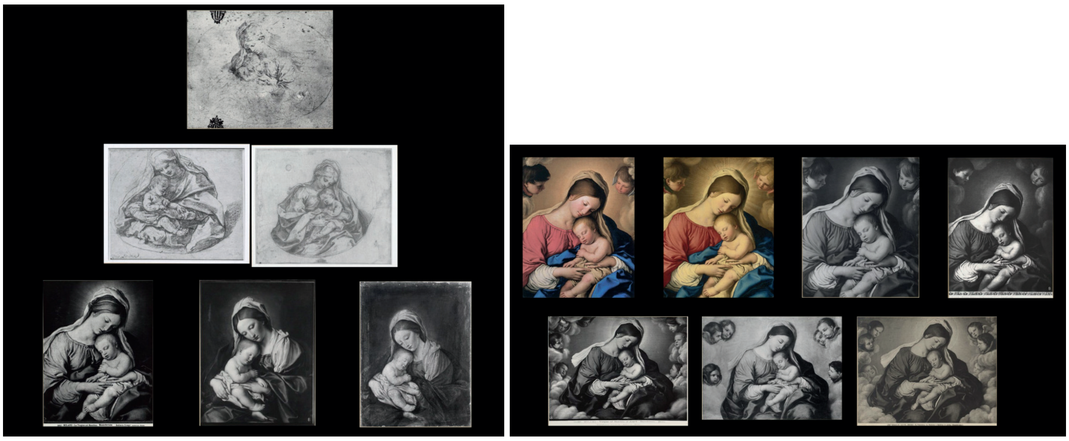


Figure A.10: *Madonna col bambino.* On the left, top drawings attributed to Guido Reni, bottom row by Sassoferrato and Anonymous Flemish. On the right, 7 unique paintings, all attributed to Sassoferrato.

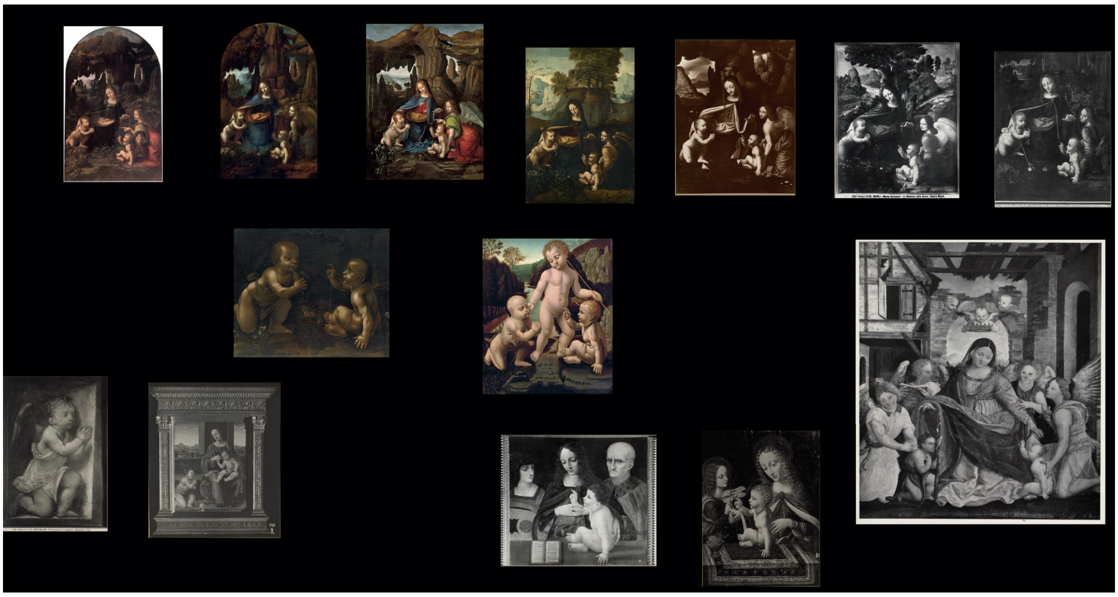


Figure A.11: Variations on the *Virgin and the rocks* of LEONARDO DA VINCI. The color images were known in advance, the black-and-white were found in the Cini. Notice how the original representations of St-John and Jesus can be found in other artworks. Bottom line: *Angelo in adorazione*, by BERNARDINO LUINI (124C_744) ; *Madonna in trono e S. Giovannino*, by LORENZO DI CREDI (BOTTEGA DI) (52C_144) ; *Sacra famiglia e Angelo musicante*, by GIAN ANTONIO BOLTRAFFIO (76B_365) ; *Madonna con hambino e Angeli musicante*, byBERNARDINO DEI CONTI (ATTR) (47B_497). On the bottom right, the image displayed is *L'incoronazione di Maria Lempera* by FRANCESCO DA MILANO (67C_207), one can notice that the global gesture of the Virgin Mary, as well as the same configuration for Jesus and St-John are reappearing, despite treating a different subject.

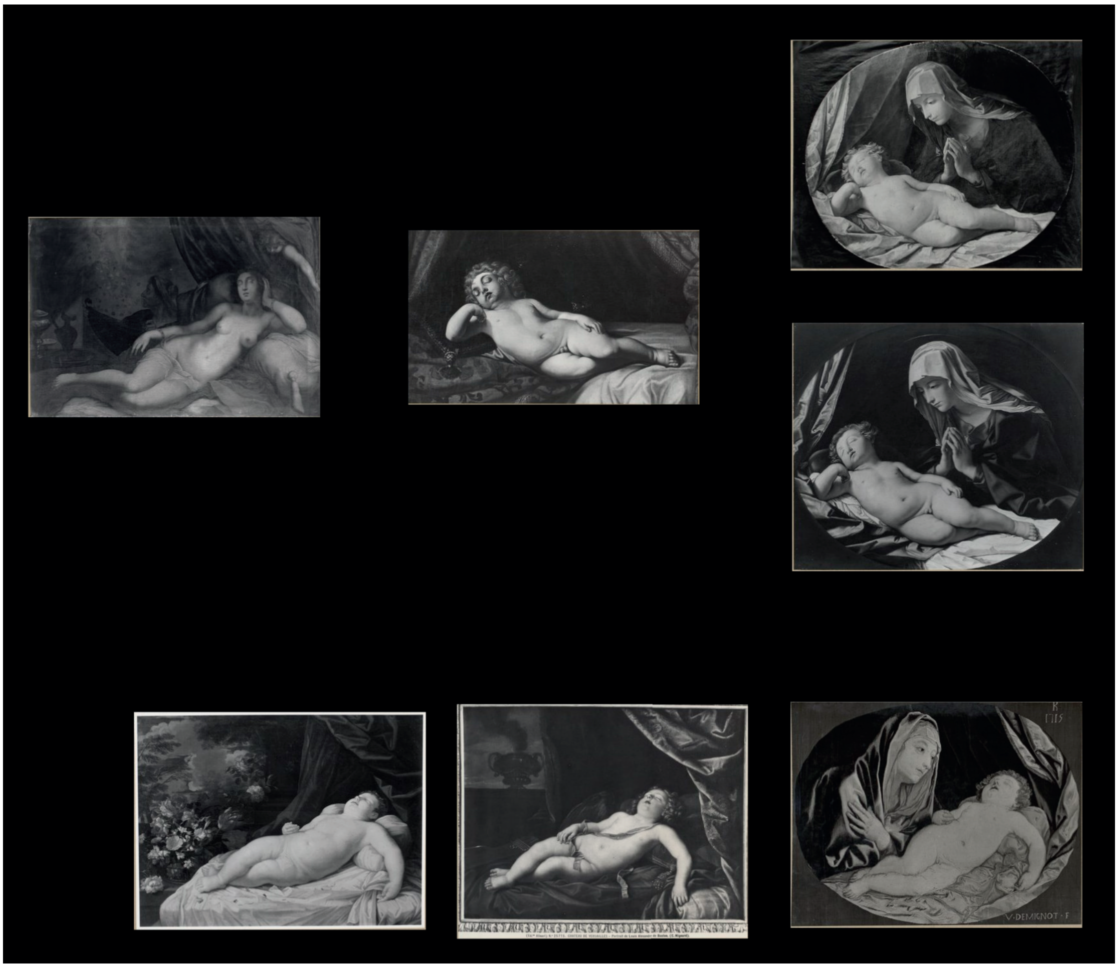


Figure A.12: The same pattern of a sleeping baby is reused as a stand-alone element for *Amore* or *Cupido*, (middle column both by PIERRE MIGNARD), and as baby Jesus for the *Madonna col bambino* on the right. Top right are actually two artworks, one attributed to SASSOFERRATO and the other to GUIDO RENI. Another remark is how the figure of the sleeping baby often show the same variations as the sleeping venus or the sleeping danae, as is highlighted by the painting by PALMA IL GIOVANE on the top left.

Bibliography

[1] B. Berenson, *Three Essays in Method.* 1930.

[2] C. Caraffa, *Photo Archives and the Photographic Memory of Art History.* 2009.

[3] “Iconclass.” www.iconclass.nl.

[4] “PHAROS: The International Consortium of Photos Archives.” http://pharosartresearch.org/.

[5] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[6] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in *CVPR*, vol. 1, pp. 886–893, IEEE, 2005.

[7] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, “Object Detection with Discriminatively Trained Part Based Models,” *PAMI*, pp. 1–20, 2009.

[8] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, nov 2004.

[9] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[10] R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object retrieval,” in *CVPR*, 2012.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” in *ECCV*, 2006.

[12] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” *Proc. CVPR*, 2003.

[13] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation Applied to Handwritten Zip Code Recognition,” *Neural Comput.*, 1989.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, 1998.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *NIPS*, pp. 1097–1105, 2012.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *CVPR*, pp. 2–9, 2009.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” apr 2017.

[18] V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” *ICML*, 2010.

[19] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” *AISTATS ’11 Proc. 14th Int. Conf. Artif. Intell. Stat.*, vol. 15, pp. 315–323, 2011.

[20] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” 2012.

[21] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv Prepr.*, pp. 1–10, 2014.

[22] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” feb 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” dec 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Identity Mappings in Deep Residual Networks,” mar 2016.

[25] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015.

[26] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” may 2015.

[27] Y. Xu, W. He, F. Yin, and C.-L. Liu, “Page segmentation for historical handwritten documents using fully convolutional networks,” in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1, pp. 541–546, IEEE, 2017.

[28] C. Tensmeyer, B. Davis, C. Wigington, I. Lee, and B. Barrett, "Pagenet: Page boundary extraction in historical handwritten documents," in *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing*, pp. 59–64, ACM, 2017.

[29] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, "Convolutional neural networks for page segmentation of historical document images," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1, pp. 965–970, IEEE, 2017.

[30] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1717–1724, 2014.

[31] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," *Int. Conf. Mach. Learn.*, pp. 647–655, 2014.

[32] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of Transferability for a Generic ConvNet Representation," jun 2014.

[33] C. Richard Johnson and Ella Hendriks, "Image processing for artist identification," *IEEE Signal Process. Mag.*, vol. 25, no. 4, pp. 37 – 48, 2008.

[34] J. M. Hughes, D. J. Graham, and D. N. Rockmore, "Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 4, pp. 1279–1283, 2010.

[35] "WikiArt." www.wikiart.org.

[36] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing Image Style," *Eccv*, pp. 1–20, 2014.

[37] Y. Bar, N. Levy, and L. Wolf, "Classification of Artistic Styles using Binarized Features Derived from a Deep Neural Network," 2014.

[38] B. Saleh and A. Elgammal, "Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature," p. 21, may 2015.

[39] A. Elgammal and B. Saleh, "Quantifying Creativity in Art Networks," jun 2015.

[40] D. Picard, P.-H. Gosselin, and M.-C. Gaspard, "Challenges in Content-Based Image Indexing of Cultural Heritage Collections," *IEEE Signal Process. Mag.*, no. july 2015, pp. 95–102, 2015.

[41] T. Mensink and J. V. Gemert, "The Rijksmuseum Challenge : Museum-Centered Visual Recognition," pp. 2–5, 2014.

[42] N. Van Noord, E. Hendriks, and E. Postma, “Toward Discovery of the Artist’s Style: Learning to recognize artists by their artworks,” *IEEE Signal Process. Mag.*, vol. 32, no. 4, pp. 46–54, 2015.

[43] S. Ginosar, D. Haas, T. Brown, and J. Malik, “Detecting people in cubist art,” in *ECCV Work.*, 2014.

[44] “ArtUK.” www.artuk.org.

[45] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, 2010.

[46] “VGG Paintings Dataset.” http://www.robots.ox.ac.uk/~vgg/data/paintings/.

[47] E. J. Crowley and A. Zisserman, “The State of the Art : Object Retrieval in Paintings using Discriminative Regions,” 2014.

[48] E. J. Crowley and A. Zisserman, “In search of art,” *ECCV Work.*, p. 16, 2014.

[49] E. J. Crowley and A. Zisserman, “The art of detection,” in *ECCV Work.*, vol. 9913 LNCS, pp. 721–737, 2016.

[50] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *CVPR*, 2016.

[51] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, “Data-driven visual similarity for cross-domain image matching,” *ACM Trans. Graph.*, vol. 30, p. 1, dec 2011.

[52] T. Malisiewicz, A. Gupta, and A. a. Efros, “Ensemble of exemplar-SVMs for object detection and beyond,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 89–96, 2011.

[53] M. Aubry, B. C. Russell, and J. Sivic, “Painting-to-3D model alignment via discriminative visual elements,” *ACM Trans. Graph.*, vol. 33, no. 2, pp. 1–14, 2014.

[54] H. Jegou, M. Douze, and C. Schmid, “Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search,” *Eur. Conf. Comput. Vis.*, 2008.

[55] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” *CVPR*, 2007.

[56] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, “Total recall: Automatic query expansion with a generative feature model for object retrieval,” *Proc. IEEE Int. Conf. Comput. Vis.*, 2007.

[57] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Lost in quantization: Improving particular object retrieval in large scale image databases,” *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008.

[58] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," in *2006 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Vol. 2*, vol. 2, pp. 2161–2168, IEEE.

[59] S. Wang and S. Jiang, "INSTRE : a New Benchmark for Instance-Level Object Retrieval and Recognition," vol. 9, no. 4, 2010.

[60] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features off-the-shelf : an Astounding Baseline for Recognition," *CVPR*, pp. 512–519, 2014.

[61] A. Babenko and V. Lempitsky, "Aggregating Local Deep Features for Image Retrieval," 2015.

[62] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," *arXiv Prepr. arXiv1511.05879*, pp. 1–11, 2015.

[63] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *ECCV*, vol. 8689 LNCS, pp. 584–599, Springer Verlag, 2014.

[64] A. Bellet, A. Habrard, and M. Sebban, "A Survey on Metric Learning for Feature Vectors and Structured Data," *arXiv Prepr. arXiv1306.6709*, 2013.

[65] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, vol. 2, pp. 1735–1742, 2006.

[66] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *CVPR*, vol. 07-12-June, pp. 815–823, 2015.

[67] F. Radenović, G. Tolias, and O. Chum, "CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples," in *ECCV*, 2016.

[68] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end Learning of Deep Visual Representations for Image Retrieval," oct 2016.

[69] "Factum Arte." http://www.factum-arte.com/.

[70] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," dec 2014.

[71] S. Ioffe, "Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models," feb 2017.

[72] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, 1996.

[73] J. A. Hanley and A. Lippman-Hand, "If nothing goes wrong, is everything all right? Interpreting zero numerators.," *JAMA*, vol. 249, pp. 1743–5, apr 1983.

[74] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold, “Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts,” in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pp. 471–476, IEEE, 2016.

[75] F. Simistira, M. Bouillon, M. Seuret, M. Würsch, M. Alberti, R. Ingold, and M. Liwicki, “Icdar2017 competition on layout analysis for challenging medieval manuscripts,” in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1, pp. 1361–1370, IEEE, 2017.

[76] “DIVA Layout Analysis Evaluator.” https://github.com/DIVA-DIA/DIVA_Layout_Analysis_Evaluator.

[77] F. Junker, “Extraction of ornaments from a large collection of books,” Master’s thesis, EPFL, 2017.

[78] T. Grüning, R. Labahn, M. Diem, F. Kleber, and S. Fiel, “Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents,” *arXiv preprint arXiv:1705.03311*, 2017.

[79] M. Diem, F. Kleber, S. Fiel, T. Gruning, and B. Gatos, “cbad: Icdar2017 competition on baseline detection,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, pp. 1355–1360, Nov. 2017.

[80] S. Ares Oliveira and F. Kaplan, “Comparing Human And Machine Performances In Transcribing 18th Century Handwritten Venetian Script,” in *Digit. Humanit. Conf.*, 2018.

[81] M. Weidemann, J. Michael, T. Grüning, and R. Labahn, “READ Report D7.8: HTR Engine Based on NNs P2,” tech. rep., 2017.

[82] “Web Gallery of Art.” www.wga.hu.

[83] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision.* Cambridge University Press, 2003.

[84] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, pp. 381–395, jun 1981.

[85] J. S. Chung, R. Arandjelovi, G. Bergel, A. Franklin, and A. Zisserman, “Re-presentations of Art Collections,” in *ECCV Work.*, pp. 1–16, 2014.

[86] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, pp. 273–297, sep 1995.

[87] T. Sapatinas, “The Elements of Statistical Learning,” *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 2004.

[88] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” feb 2017.

[89] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, “From Generic to Specific Deep Representations for Visual Recognition,”

[90] Y. Collet, “Zstandard - Fast real-time compression algorithm,” 2015.

[91] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, “Object retrieval and localization with spatially-constrained similarity measure and k -NN re-ranking,” *CVPR*, pp. 1–8, 2012.

[92] “International Image Interoperability Framework.” https://iiif.io.

[93] “IIIF Discovery API.” https://iiif.io/api/discovery.

[94] “CIDOC Conceptual Reference Model.” http://www.cidoc-crm.org/.

[95] “Linked Art.” https://linked.art/.

[96] Y. Chen, X. Zhou, and T. S. Huang, “One-Class SVM for Learning in Image Retrieval,” in *ICIP*, pp. 1–4, 2001.

[97] L. Van Der Maaten and G. Hinton, “Visualizing high-dimensional data using t-sne,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[98] G. P. Nguyen and M. Worring, “Interactive access to large image collections using similarity-based visualization,” *J. Vis. Lang. Comput.*, vol. 19, no. 2, pp. 203–224, 2006.

[99] J. Brooke, “SUS - A quick and dirty usability scale Usability and context,” in *Usability Eval. Ind.*, pp. 189–194, London: Taylor and Francis, 1996.

[100] A. Bangor, P. T. Kortum, and J. T. Miller, “An empirical evaluation of the system usability scale,” *Int. J. Hum. Comput. Interact.*, pp. 575–594, 2008.

[101] P. Pu, L. Chen, and R. Hu, “A User-Centric Evaluation Framework for Recommender Systems,” *Proc. fifth ACM Conf. Recomm. Syst.*, pp. 157–164, 2011.

Benoit Seguin

Personal Data

Address: Place du Tunnel 9, 1005 Lausanne, Switzerland
Phone: +41 78 910 98 15
email: seg.benoit@gmail.com
Citizenship: French
Webpage: seguinbe.github.io

Professional Experience

Current
Sept 2014

PhD Student, DHLAB, EPFL
Making large-scale art history collections searchable: A deep learning approach, with Prof. Kaplan

Use of modern computer vision and image analysis techniques in order to allow art historians and archivists to digitize and navigate large iconographic collections.

Aug 2014
Sept 2013

Scientific assistant CVLAB EPFL
FastScan Project, with Prof. Fua

Implemented a fast multi-threaded prediction algorithm for mitochondria segmentation in SEM images. A prototype of integration directly with the software of a Microscope showed promising result in accelerating the scanning of biological tissues.

Feb-Aug 2013

Master Thesis at IBM Research, Zurich
Estimating VLSI pattern sensitivity with respect to variability in optical lithography printing, with Dr. Gabrani

Developed an automatic analysis tool for the success and the variability of the lithography printing process for a specific pattern (based on image analysis of SEM images and error evaluation). Showed how VLSI patterns react differently according to variations in the printing conditions.

Apr-Sept 2011

Internship at Carnegie Mellon University, Pittsburgh
Unsupervised object detection with an eye-tracking system, with Prof. Hebert

Skills

Areas: Machine Learning, Computer Vision, Image Processing.
Programming: Python, C++, Tensorflow, UNIX systems, basic web-programming.

Education

2011-2013 Master of Science in Computer Science, **EPFL**, Lausanne
Very High Honours, Gpa: 5.53/6.0

2008-2013 Diplôme d'Ingénieur, **École Polytechnique ParisTech**, Palaiseau
Gpa: 3.5/4.0

2006-2008 Preparatory Classes, **Lycée du Parc**, Lyon
Gpa: 3.92/4

2006 Scientific Baccalaureate, **Lycée Charles Nodier**, Dole
Very High Honours

LANGUAGES

French: Mothertongue
English: Fluent, TOEFL IBT 106/120, prior to a 5 months stay in the USA.
Japanese: Basic Knowledge, JLPT N4 (equivalent of CEFR A2). Two months stay in 2010.

EXTRA CURRICULAR ACTIVITIES

Piano: *Certificat de fin d'étude*, awarded with very high honors in 2005.
Choir: Has been part of multiple choruses, in Paris and Lausanne. Member of the organizing team of the Lausanne's University Choir from 2013 to 2017. Main organizer of a classical concert attended by 2'000+ persons in 2017.
Robotics: In 2009, as the vice-chairman of the robotics association of the École Polytechnique, led a team of 12 persons to the French Robotics Cup for a top-15% finish.

AWARDS

- Qualified for the final round of Google Hashcode 2016 (top-50 out of 1000+ teams)
- Best Demonstration Award at the Research Days of the CS Faculty of EPFL in 2017.

PUBLICATIONS

M. Gabrani, B. Seguin, H. Saab Estimating pattern sensitivity to the printing process for varying dose/focus conditions for RET development in the sub-22nm era, in *Metrology, Inspection, and Process Control for Microlithography XXVIII*, 2014

I. diLenardo, B. Seguin, F. Kaplan Visual Patterns Discovery in Large Databases of Paintings, in *Digital Humanities Conference* 2016, Krakow

B. Seguin, C. Striolo, I. diLenardo, F. Kaplan Visual Link Retrieval in a Database of Paintings, in *VISART Workshop at European Conference of Computer Vision* 2016, Amsterdam.

B. Seguin, I. diLenardo, F. Kaplan Tracking Transmission of Details in Paintings, in *Digital Humanities Conference* 2017, Montréal.

W. Haaswijk*, E. Collins*, B. Seguin*, M. Soeken, S. Süsstrunk, F. Kaplan, S. De Micheli Deep Learning for Logic Optimization, in *International Workshop on Logic & Synthesis* 2017.

B. Seguin The Replica Project: Building a visual search engine for art historians, in *ACM XROADS Magazine* Spring 2018.

B. Seguin, L. Costiner, I. diLenardo, F. Kaplan New Techniques for the Digitization of Art Historical Photographic Archives—the Case of the Cini Foundation in Venice, in *Archiving* 2018, Washington DC.

B. Seguin, L. Costiner, I. diLenardo, F. Kaplan Extracting and Aligning Artist Names in Digitized Art Historical Archives, in *Digital Humanities Conference* 2018, Mexico.

W. Haaswijk*, E. Collins*, B. Seguin*, M. Soeken, S. Süsstrunk, F. Kaplan, S. De Micheli Deep Learning for Logic Optimization Algorithms, in *International Symposium on Circuits and Systems* 2018.

S. Ares Oliveira*, B. Seguin*, F. Kaplan dhSegment: A generic deep-learning approach for document segmentation, in *International Conference on Frontiers in Handwriting Recognition* 2018, Niagara Falls.