Image Collation: Matching illustrations in manuscripts

Ryad Kaoua[1], Xi Shen[1], Alexandra Durr[2], Stavros Lazaris[3], David Picard[1], and Mathieu Aubry[1] (✉)

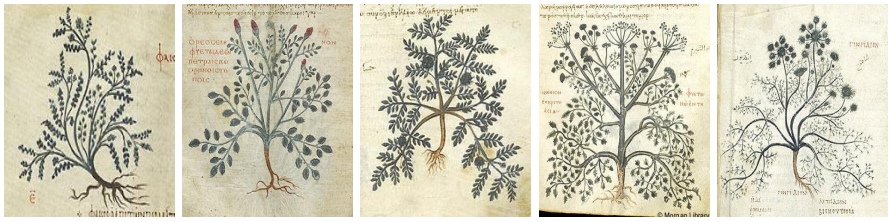
[1] LIGM, Ecole des Ponts, Univ. Gustave Eiffel, CNRS, Marne-la-Vallée, France
[2] Université de Versailles-Saint-Quentin-en-Yvelines, France
[3] CNRS (UMR 8167), France

Abstract. Illustrations are an essential transmission instrument. For an historian, the first step in studying their evolution in a corpus of similar manuscripts is to identify which ones correspond to each other. This image collation task is daunting for manuscripts separated by many lost copies, spreading over centuries, which might have been completely re-organized and greatly modified to adapt to novel knowledge or belief and include hundreds of illustrations. Our contributions in this paper are threefold. First, we introduce the task of illustration collation and a large annotated public dataset to evaluate solutions, including 6 manuscripts of 2 different texts with more than 2 000 illustrations and 1 200 annotated correspondences. Second, we analyze state of the art similarity measures for this task and show that they succeed in simple cases but struggle for large manuscripts when the illustrations have undergone very significant changes and are discriminated only by fine details. Finally, we show clear evidence that significant performance boosts can be expected by exploiting cycle-consistent correspondences. Our code and data are available on http://imagine.enpc.fr/~shenx/ImageCollation.

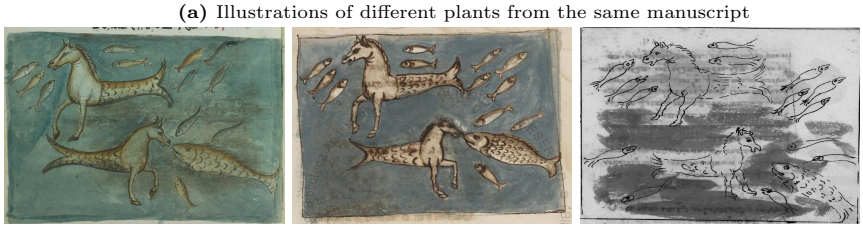
1 Introduction

Most research on the automatic analysis of manuscripts and particularly their alignment, also known as collation, has focused on text. However, illustrations are a crucial part of some documents, hinting the copyist values, knowledge and beliefs and are thus of major interest to historians. One might naively think that these illustrations are much easier to align than text and that a specialist can identify them in a matter of seconds. This is only true in the simplest of cases, where the order of the illustrations is preserved and their content relatively similar. In harder cases however, the task becomes daunting and is one of the important limiting factor for a large scale analysis.

As an example, the "De materia Medica" of Dioscorides, a Greek pharmacologist from the first century, has been widely distributed and copied between the 6th and the 16th century. Depending on the versions, it includes up to 800 depictions of natural substances. In particular the manuscripts we study in this paper contain around 400 illustrations of plants. They have been re-organized



(a) Illustrations of different plants from the same manuscript



(b) Same illustration in three manuscripts with different styles

Fig. 1: The illustration alignment task we tackle is challenging for several reasons. It requires fine-grained separation between images with similar content (a), while being invariant to strong appearance changes related to style and content modifications (b). The examples presented in this figure are extracted from the two groups of manuscript we include in our dataset: (a) the *De Materia Medica* of Dioscoride; (b) the *Physiologus*.

in different orders in the 17 different known illustrated versions of the text, for example alphabetically or depending on their therapeutic properties. The changes in the illustrations and their organizations hints both at the tradition from which each manuscript originates and at the evolution of scientific knowledge. However, the important shifts both in the illustrations appearance and in the order in which they appear makes identifying them extremely time consuming. While the text could help, it is not always readable and it is sometime not next to the illustrations.

From a Computer Vision perspective, the task of retrieving corresponding illustrations in different versions of the manuscripts present several interesting challenges, illustrated in Figure 1. First, we are faced with a fine-grained problem, since many illustrations correspond to similar content, such as different plants (Figure 1a). Second, the style, content and level of details vary greatly between different versions of the same text (Figure 1b). Third, we cannot expect relevant supervision but can leverage many constraints. On the one hand the annotation cost is prohibitive and the style and content of the illustrations vary greatly depending on the manuscripts and topics. On the other hand the structure of the correspondences graph is not random and could be exploited by a learning or optimization algorithm. For example, correspondences should mainly be one on one, local order is often preserved, and if three or more versions of the same text are available correspondences between the different versions should be cycle consistent.

Table 1: The manuscripts in our dataset come from two different texts, have diverse number of illustrations and come from diverse digitisations. In total, it includes more than 2000 illustrations and 1 200 annotated correspondences.

name	code	number of folios	folios' resolution	number of illustrations	annotated correspondences
Physiologus	P1	109	1515x2045	51	P2: 50 - P3: 50
"	P2	176	755x1068	51	P1: 50 - P3: 51
"	P3	188	792x976	52	P1: 50 - P2: 51
De Materia Medica	D1	557	392x555	816	D2: 295 - D3: 524
"	D2	351	1024x1150	405	D1: 295 - D3: 353
"	D3	511	763x1023	839	D1: 524 - D2: 353

In this paper, we first introduce a dataset for the task of identifying correspondences between illustrations of collections of manuscripts with more than 2 000 extracted illustrations and 1 200 annotated correspondences.

Second, we propose approaches to extract such correspondences automatically, outlining the crucial importance both of the image similarity and its non-trivial use to obtain correspondences. Third, we present and analyze results on our dataset, validating the benefits of exploiting the problem specificity and outlining limitations and promising directions for future works.

2 Related work

Text collation. The use of mechanical tools to compare different versions of a text can be dated back to Hinman's collator, an opto-mechanical device which Hinman designed at the end of the 1940s to visually compare early impressions of Shaekspeare's works [30]. More recently, computer tools such as CollateX [13] have been developed to automatically compare digitised versions of a text. The core idea is to explain variants using the minimum number of edits, or block move [5], to produce a variant graph [26]. Most text alignement methods rely on a transcription and tokenization step, which is not adapted to align images. Methods which locally align texts and their transcriptions e.g., [16,17,14,25,10], are also related to our task. Similar to text specific approaches, we will show that leveraging local consistency in the alignments has the potential to improve results for our image collation task.

Image retrieval in historical documents. Given a query image, image retrieval aims at finding images with similar content in a database. Classic approaches such as Video-Google [29] first look for images with similar SIFT [20] features then filter out those which features cannot be aligned by a simple spatial transformation. Similar bag of words approaches have been tested for pattern spotting in manuscripts [9]. However, handcrafted features such as SIFTs fail in the case

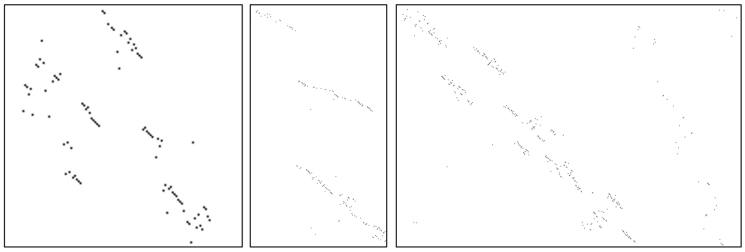


Fig. 2: Structure of the correspondences for the "De Materia Medica". From left to right we show crops of the full correspondence matrices for D1-D2, D1-D3 and D2-D3. The black dots are the ground truth annotations. While the order is not completely random, the illustrations have been significantly re-ordered. Best viewed in electronic version.

of strong style changes [27] which are characteristic of our problem.
Recent studies [23,12] suggest that directly employing global image features obtained with a network trained on a large dataset such as ImageNet [7] is a strong baseline for image retrieval. Similarly, [31] leverages features from a RetinaNet [18] network trained on MS-Coco [19] for pattern spotting in historical manuscripts and shows they improve over local features. If annotations are available, the representation can also to be learned specifically for the retrieval task using a metric learning approach, i.e., by learning to map similar samples close to each other and dissimilar ones far apart [12,22,24]. Annotations are however rare in the case for historical data.
Recently, two papers have revisited the Video-Google approach for artistic and historical data using pre-trained deep local features densely matched in images to define an image similarity: [27] attempts to discover repeated details in artworks collections and [28] performs fine-grained historical watermark recognition. Both papers propose approaches to fine-tune the features in a self-supervised or weakly supervised fashion, but report good results with out-of-the box features.

3 Dataset and task

We designed a dataset to evaluate image collation, i.e., the recovery of corresponding images in sets of manuscripts. Our final goal is to provide a tool that helps historians analyze sets of documents by automatically extracting candidate correspondences. We considered two examples of such sets originating from different libraries with online text access to digitized manuscripts [1,2,3,4], which characteristics are summarized in Table 1 and which are visualized in Figure 1 and 3:

– The "Physiologus" is a christian didactic zoological text compiled in Greek during the 2nd century AD. The three manuscripts we have selected contain illuminations depicting real and fantastic animals and contain respectively 51, 51 and 52 illustrations which could almost all be matched between versions.

(a) "Physiologus" **(b)** "De Materia Medica"

Fig. 3: Examples of annotated image triplets in our two sets of manuscripts. Note how the depiction can vary significantly both in style and content.

In the three versions we used the order of the illustrations is preserved. The content of the illuminations is similar enough that the correspondences could be easily identified by a human annotator. The style of the depictions however varied a lot as can be seen Figures 1b and 3a.

– "De Materia Medica" was originally written in greek by Pedanius Dioscorides. The illustrations in the manuscript are mainly plants drawings. We consider three versions of the text with 816, 405 and 839 illustrations and annotated 295, 353 and 524 correspondences in the three associated pairs. Finding corresponding images in this set is extremely challenging due to three difficulties: many plants are visually similar to each other (Figure 1b), the appearance can vary greatly in a matching image pair (Figure 3b) and the illustrations are ordered differently in different manuscripts (Figure 2).

Note that we included three manuscripts in both sets so that algorithms and annotations could leverage cycle-consistency.

Illustrations annotations. We ran an automatic illustration extraction algorithm [21] and found it obtained good results, but that some bounding boxes were inaccurate and that some different but overlapping illustrations were merged. To focus on the difficulty of finding correspondences rather than extracting the illustrations, we manually annotated the bounding boxes of the illustrations in each manuscript using the VGG Image Annotator [8]. The study of joint detection and correspondence estimation is left for future work.

Correspondences annotations. For the manuscripts of the Physiologus, annotating the corresponding illustration was time-consuming but did not present any significant difficulties. For the De Materia Medica however, the annotation presented significant challenge. Indeed, as explained above and illustrated in

Figure 2, the illustrations have been significantly re-ordered, modified, and are often visually ambiguous. Since the manuscripts contain hundreds of illustrations, manually finding correspondences one by one was simply not feasible. We thus followed a three step procedure. First, for each illustration we used the image similarity described in Section 4.1 to obtain its 5 nearest neighbors in each other manuscript. Second, we provided these neighbors and their context to a specialist who selected valid correspondences and searched neighboring illustrations and text to identify other nearby correspondences. Third, we used cycle consistency between the three manuscripts to validate the consistency of the correspondences identified by the specialist and propose new correspondences. Interestingly, during the last step we noticed 51 cases where the captions and the depictions were not consistent. While worth studying from an historical perspective, these cases are ambiguous from a Computer Vision point of view, and we removed all correspondences leading to such inconsistencies from our annotations.

Evaluation metric. We believe our annotations to be relatively exhaustive, however the difficulty of the annotation task made this hard to ensure. We thus focused our evaluation metric on precision rather than recall. More precisely, we expect algorithms to return a correspondence in each manuscript for each reference image and we compute the average accuracy on annotated correspondences only. In our tables, we report performances on pairs of manuscripts $M_1 - M_2$ by finding correspondences in both directions (finding a correspondence in M_2 for each image of M_1, then a correspondence in M_1 for each image in M_2) and averaging performances. Note that there is a bias in our annotations in the De Materia Medica since we initially provided the annotator with the top correspondences using our similarity. However, while it may slightly over-estimate the performance of our algorithm, qualitative analysis of the benefits brought by our additional processing remains valid.

4 Approach

In this section, we present the key elements of our image collation pipeline, visualized in Figure 4. Except when explicitly mentioned otherwise, we focus on studying correspondences in a pair of manuscripts. First, we discuss image similarities adapted to the task. Second, we introduce different normalizations of the similarity matrix associated to a pair of manuscripts. Third, we present a method to propagate information from confident correspondences to improve results. Finally, we give some implementation details.

4.1 Image similarities

We focus on similarities based on deep features, following consistent observations in recent works that they improve over their classical counterparts for historical image recognition [27,31]. Since we want our approach to be directly applicable to new sets of images, with potentially very different characteristics, we use

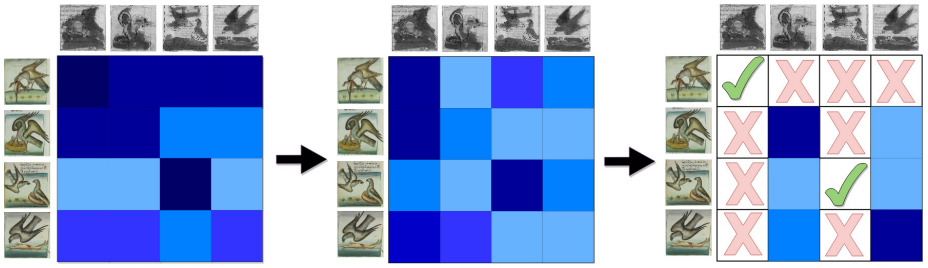


Fig. 4: Overview of our approach. We first compute a similarity score between each pair of image, which we visualize using darker colors for higher similarity. We then normalize the similarity matrix to account for images that are similar to many other, such as the first line of our example. Finally, we propagate signal from confident correspondences which are maxima in both directions (green marks) to the rest of the matrix.

off-the-shelf features, without any specific fine-tuning. More precisely we used ResNet-50 [15] features trained for image classification on ImageNet [7], which we found to lead to better performances (see Section 5).

Raw features. Directly using raw features to compute image similarity is a strong baseline. Similar to other works [27,27], we found that using *conv4* features and averaging cosine similarity of these features at the same location consistently performed best. More formally, given two images I_1 and I_2 we consider their *conv4* features $f_k = (f_k^i)_{i \in \{1, \ldots, N\}}$, where $k = 1$ or 2 is the image ID, i is the index of the spatial location in feature map and N is the size of the feature map. We define the feature image similarity as:

$$S_{features}(I_1, I_2) = \frac{1}{N} \sum_{i=1}^{N} \frac{f_1^i}{\|f_1^i\|} \cdot \frac{f_2^i}{\|f_2^i\|} \tag{1}$$

where $\cdot$ is the scalar product. Note that the normalization is performed for each local feature independently and this similarity can only be defined if the two images are resized at a constant size. We used 256×256 in our implementation.

Matching-based similarity. The feature similarity introduced in the previous paragraph only considers the similarity of local features at the same spatial location and scale, and not their similarity with other features at other locations in the image. To leverage this information, [28] proposed to use a local matching score. Each feature f_k^i of a source image I_k is matched with the features extracted at several scales in a target images I_l. Then, each of the features of the target image is matched back in the source image and kept only if it matches back to the original feature, i.e. if it is a cycle consistent match. Finally, the best cycle consistent match among all scales of the target image $m_{k,l}(f_k^i)$ is identified. Writing $x_k^i \in \mathbb{R}^2$ the position of the feature f_k^i in the feature map and $x_{k,l}(f_k^i) \in \mathbb{R}^2$

the position of its best match $m_{k,l}(f_k^i)$ (which might be at a different scale), we define the similarity between I_1 and I_2 as:

$$S_{matching}(I_1, I_2) = \frac{1}{2N}\sum_{i=1}^{N} e^{-\frac{\|x_1^i - x_{1,2}(f_1^i)\|^2}{2\sigma^2}} \frac{f_1^i}{\|f_1^i\|} \cdot \frac{m_{1,2}(f_1^i)}{\|m_{1,2}(f_1^i)\|} + \frac{1}{2N}\sum_{i=1}^{N} e^{-\frac{\|x_2^i - x_{2,1}(f_2^i)\|^2}{2\sigma^2}} \frac{f_2^i}{\|f_2^i\|} \cdot \frac{m_{2,1}(f_2^i)}{\|m_{2,1}(f_2^i)\|}$$

where $\cdot$ is the scalar product and σ is a real hyperparameter. This score implicitly removes any contribution for non-discriminative regions and for details that are only visible in one of the depictions, since they will likely match to a different spatial location and thus have a very small contribution to the score. It will also be insensitive to local scale changes. Note that [28] considered only the first term of the sum, resulting in a non-symmetric score. On the contrary, our problem is completely symmetric and we thus symmetrized the score.

Transformation dependent similarity. While the score above has some robustness to local scale changes, it assumes the images are coarsely aligned. To increase robustness to alignment errors, we follow [27] and use RANSAC [11] to estimate a 2D affine transformation between the two images. More precisely, keeping the notations from the previous paragraph, we use RANSAC to find an optimal affine transformation $\mathcal{T}_{k,l}$ between image I_k and I_l:

$$\mathcal{T}_{k,l} = \arg\max \sum_{i=1}^{N} e^{-\frac{\|\mathcal{T}_{k,l}x_k^i - x_{k,l}(f_k^i)\|^2}{2\sigma^2}} \frac{f_k^i}{\|f_k^i\|} \cdot \frac{m_{k,l}(f_k^i)}{\|m_{k,l}(f_k^i)\|} \tag{2}$$

Note this is slightly different from [27] which only uses the RANSAC to minimize the residual error in the matches to optimize the transformation. We found that maximizing the score instead of the number of inliers significantly improved the performances. Considering again the symmetry of the problem, this leads to the following score:

$$S_{trans}(I_1, I_2) = \frac{1}{2N}\sum_{i=1}^{N} e^{-\frac{\|\mathcal{T}_{1,2}x_1^i - x_{1,2}(f_1^i)\|^2}{2\sigma^2}} \frac{f_1^i}{\|f_1^i\|} \cdot \frac{m_{1,2}(f_1^i)}{\|m_{1,2}(f_1^i)\|} + \frac{1}{2N}\sum_{i=1}^{N} e^{-\frac{\|\mathcal{T}_{2,1}x_2^i - x_{2,1}(f_2^i)\|^2}{2\sigma^2}} \frac{f_2^i}{\|f_2^i\|} \cdot \frac{m_{2,1}(f_2^i)}{\|m_{2,1}(f_2^i)\|}$$

This score focuses on discriminative regions, is robust to local scale changes and affine transformations. We found it consistently performed best in our experiments, outperforming the direct use of deep features by a large margin.

4.2 Normalization

Let us call S the similarity matrix between all pairs of images in the two manuscripts, $S(i, j)$ being a similarity such as the ones defined in the previous section between the ith image of the first manuscript and the jth image of the second manuscript. For each image in the first manuscript, one can simply predict

Table 2: Row-wise and Column-wise normalizations.

Normalization	$R(i,j)$	$C(i,j)$
$sm(\lambda S)$	$\exp(\lambda S(i,j))/\sum_k \exp(\lambda S(i,k))$	$\exp(\lambda S(i,j))/\sum_k \exp(\lambda S(k,j))$
$S/\operatorname{avg}(S)$	$R_{\text{avg}} = S(i,j)/\operatorname{avg}_k S(i,k)$	$C_{\text{avg}} = S(i,j)/\operatorname{avg}_k S(k,i)$
$S/\max(S)$	$R_{\max} = S(i,j)/\max_k S(i,k)$	$C_{\max} = S(i,j)/\max_k S(k,i)$
$sm(\lambda S/\operatorname{avg}(S))$	$\exp(\lambda R_{\text{avg}}(i,j))/\sum_k \exp(\lambda R_{\text{avg}}(i,k))$	$\exp(\lambda C_{\text{avg}}(i,j))/\sum_k \exp(\lambda C_{\text{avg}}(k,j))$
$sm(\lambda S/\max(s))$	$\exp(\lambda R_{\max}(i,j))/\sum_k \exp(\lambda R_{\max}(i,k))$	$\exp(\lambda C_{\max}(i,j))/\sum_k \exp(\lambda C_{\max}(k,j))$

the most similar image in the second one as a correspondence, i.e. take the maximum over each row of the similarity matrix. This approach has however two strong limitations. First, it does not take into account that some images tend to have higher similarity scores than other, resulting in rows or columns with higher values in the similarity matrix. Second, it does not consider the symmetry of the problem, i.e., that one could also match images in the second manuscript to images in the first one.

To account for these two effects, we propose to normalize the similarity matrix S along each row and each column resulting in two matrices R and C.

We experimented with five different normalization operations using softmax (sm), maximum (max) and average (avg) operations either along the rows (leading to R) or the columns (leading to C), as shown in Table 2.

We then combine the two matrices R and C into a final score: we experimented with summing them or using element-wise (Hadamard) multiplication. Both performed similarly, with a small advantage from the sum, we thus only report those results. We found in our experiments that the max normalization performed best, without requiring an hyper-parameter. As such, our final normalized similarity matrix N_S is defined as:

$$N_S(i,j) = \frac{S(i,j)}{\max_k S(i,k)} + \frac{S(i,j)}{\max_k S(k,j)} \tag{3}$$

4.3 Information propagation

While the normalized score N_S obtained in the previous section includes information about both directions of matching in a pair of manuscripts, it does not ensure that correspondences are 2-cycle consistent, i.e. that the maxima in the rows of N_S correspond to maxima in the columns. If one has access to more than 2 manuscripts, one can also check consistency between triplets of manuscripts and identify correspondences that are 2 and 3-cycle consistent. Correspondences that verify such cycle-consistency are intuitively more reliable, as we validated in our experiments, and thus can be used as anchors to look for other correspondences in nearby images. Indeed, while the order of the images is not strictly preserved in the different versions, there is still a clear locally consistent structure as can be seen in the ground truth correspondence matrices visualized in Figure 2.

Many approaches could be considered to propagate information from confident correspondences and an exhaustive study is beyond the scope of our work. We

(a) Query and 5 nearest neighbors according to our similarity score (Section 4.1)

(b) Query and 5 nearest neighbors according to our normalized score (Section 4.2)

(c) Query and 5 nearest neighbors after information propagation (Section 4.3)

Fig. 5: Query (in blue) and 5 nearest neighbors (ground truth in green) after the different steps of our method. Despite not being in the top-5 using the similarity, the correct correspondence is finally identified after the information propagation step

considered a simple baseline as a proof of concept. Starting from an initial score N_S and a set of confident correspondences $\mathcal{C}^*$ as seeds (e.g., correspondences that verify 2 or 3 cycle consistency constraints), we define a new score after information propagation P_S as:

$$P_S(i,j) = N_S(i,j) \prod_{(k,l)\in\mathcal{C}^*} \left(1 + \alpha \exp\left(\frac{-||(i,j)-(k,l)||^2}{2\sigma_p^2}\right)\right) \tag{4}$$

where σ_p and α are hyperparameters. Note that this formula can be applied with any definition of $\mathcal{C}^*$, and thus could leverage sparse correspondence annotations.

Implementation details In all the experiments, we extract *conv4* features of a ResNet50 architecture [15] pre-trained on ImageNet [7]. To match illustrations between different scales, we keep the original aspect ratios and resize the source image to have 20 features in the largest dimension and the target image to five scales such that the numbers of features of the largest dimension are 18, 19, 20, 21, 22. We set σ in Equation 2 and 3 to $\frac{1}{\sqrt{50}}$ times the size of the image and the number of iterations in the RANSAC to 100. For the information propagation, we find that $\sigma_p = 5$ and $\alpha = 0.25$ performs best. With our naive Pytorch implementation, computing correspondences between D1 (816 illustrations) and

Table 3: Percentage of accuracy of the correspondences obtained using $S_{features}$ with different conv4 features for all manuscripts pairs.

Pairs	ResNet18	MoCo-v2	ArtMiner	ResNet50
P1-P2	78.0	75.0	**92.0**	84.0
P1-P3	75.0	62.0	**78.0**	73.0
P2-P3	99.0	99.0	98.0	**100.0**

Pairs	ResNet18	MoCo-v2	ArtMiner	ResNet50
D1-D2	31.9	31.2	35.3	**35.4**
D1-D3	42.0	35.6	43.7	**46.1**
D2-D3	27.6	26.6	31.7	**34.1**

Table 4: Accuracy of the correspondences obtained using the different similarities explained in Section 4.1, as well as the similarity used in [27], which is similar to S_{trans} but uses the number of inliers instead of our score to select the best transformation.

Pairs	$S_{features}$	$S_{matching}$	[27]	S_{trans}
P1-P2	84.0	98.0	99.0	**100.0**
P1-P3	73.0	94.0	**98.0**	**98.0**
P2-P3	**100.0**	98.0	**100.0**	**100.0**

Pairs	$S_{features}$	$S_{matching}$	[27]	S_{trans}
D1-D2	35.4	54.6	56.3	**61.7**
D1-D3	46.1	69.8	71.3	**77.7**
D2-D3	34.1	51.8	51.7	**60.1**

D2 (405 illustrations) takes approximately 80 minutes, 98% being spent to compute similarities between all the 330,480 pairs of images.

5 Results

In this section, we present our results. In 5.1 we compare different features similarities. In 5.2 we show the performance boost by the different normalizations. In 5.3 we demonstrate that the results can be improved by leveraging the structure of the correspondences. Finally, in 5.4 we discuss the failure cases and limitations.

To measure the performance, as described in Section 3, we compute both the accuracy a_1 obtained by associating to each illustration of the first manuscript the illustration of the second manuscript which maximizes the score and the accuracy a_2 by associating illustrations of the second manuscript to illustrations of the first manuscript. We then report the average of these two accuracies $\frac{a_1+a_2}{2}$. Because using a good image similarity already led to almost perfect results on the Physiologus, we focus our analysis on the more challenging case of the De Materia Medica. The benefits of the three steps of our approach are illustrated in Figure 5, where one can also assess the difficulty of the task.

5.1 Feature similarity

We first compare in table 3 the accuracy we obtained using the baseline score $S_{features}$ with different conv4 features: ResNet18 and ResNet50 trained on ImageNet, MoCo v2[6], and the ResNet18 features fine-tunned by [27]. The feature from [27] achieve the best results on Physiologus manuscripts. However, on the challenging De Materia Medica manuscripts, the ResNet50 features perform best, and we thus use them in the rest of the paper.

In table 4, we compare the accuracy using the different similarities explained in Section 4.1. The results obtained using S_{trans} leads to the best performances on all pairs, and we consider only this score in the rest of the paper. Note in particular that optimizing the function of Equation (2) leads to clearly better

Table 5: Accuracy of the correspondences obtained using the different normalizations explained in Section 4.2 and in Table 2.

Pairs	S	$sm(\lambda S/\text{ avg}(S))$	$sm(\lambda S/\text{ max}(s))$	$sm(\lambda S)$	$S/\text{ avg}(S)$	$S/\text{ max}(S)$
D1-D2	61.7	68.3	**70.5**	67.8	67.1	**70.5**
D1-D3	77.7	83.5	**85.3**	83.1	82.5	**85.3**
D2-D3	60.1	66.3	66.7	66.1	65.3	**69.0**

Fig. 6: Examples of correspondences recovered only after the information propagation. These examples are with many local appearance changes.

result than using the number of inliers as in [27]. Since results on the Physiologus, where illustrations are fewer and more clearly different, are almost perfect, we only report the quantitative evaluation on the more challenging De Materia Merdica in the following sections.

5.2 Normalization

In table 5, we compare the accuracy we obtained using the different normalizations presented in Section 4.2 and in Table 2. For the softmax-based normalizations that include an hyperparamter, we optimized it directly on the test data, so the associated performance should be interpreted as an upper bound. Interestingly, a simple normalization by the maximum value $S/max(S)$ outperforms these more complexe normalization without requiring any hyper-parameter tuning. It is also interesting that all the normalization schemes we tested provide a clear boost over the raw similarity score, outlining the importance of considering the symmetry of the correspondence problem.

5.3 Information propagation

We analyze the potential of information propagation in Table 6. Using the normalized similarity score, we can compute correspondences (column 'all'). Some of these correspondences will be 2-cycle or 3-cycle consistent. These correspondences will be more reliable, as can be seen in the 'only 2-cycle' and 'only 3-cycle'

Table 6: The left part of the table details the accuracy of the correspondences with the normalized score N_S on different subsets of the annotated correspondences: all (the measure used in the rest of the paper), the correspondences obtained with N_S are 2-cycle consistent and and the correspondences obtained with N_S are 2 and 3-cycle consistent. The number in parenthesis is the number of correspondences. The right part of the table present the average accuracy obtained when performing information propagation from either the 2-cycle or the 3-cycle consistent correspondences.

Pairs	N_S			P_S - $\mathcal{C}^*$:2-cycles	P_S - $\mathcal{C}^*$:3-cycles
	all	only 2-cycle	only 3-cycle	all	all
D1-D2	70.5 *(295)*	83.5 *(224)*	99.2 *(118)*	**82.5**	82.0
D1-D3	85.3 *(524)*	90.2 *(457)*	98.3 *(118)*	88.5	**88.6**
D2-D3	69.0 *(353)*	78.5 *(279)*	96.6 *(118)*	**81.7**	79.3

columns, but there will be fewer (the number of images among the annotated ones for which such a cycle consistent correspondence is found given in parenthesis). In particular, the accuracy restricted to the 3-cycle consistent correspondences is close to 100%. Because the accuracy of these correspondences is higher, one can use them as a set of confident correspondences $\mathcal{C}*$ to compute a new score P_S as explained in Section 4.3. The results, on all annotations, can be seen in the last two column. The results are similar when using either 2 or 3-cycle consistent correspondences for $\mathcal{C}^*$ and the improvement over the normalized scores is significant.

This result is a strong evidence that important performance boost can be obtained by leveraging consistency. Qualitatively, the correspondences that are recovered are difficult cases, where the depictions have undergone significant changes, as shown in Figure 6.

5.4 Failure cases, limitations and perspectives

Figure 7 shows some typical examples of our failure cases. As expected they correspond to cases where the content of the image has been significantly altered and where very similar images are present. For such cases, it is necessary to leverage the text to actually be able to discriminate between the images. Extracting the text and performing HTR in historical manuscripts such as ours is extremely challenging, and the text also differs considerably between the different versions. However, a joint approach considering both the text and the images could be considered and our dataset could be used for such a purpose since the full folios are available.

Conclusion

We have introduced the new task of image collation and an associated dataset. This task is challenging and would enable to study at scale the evolution of illustrations in illuminated manuscripts. We studied how different image similarity measures perform, demonstrating that direct deep feature similarity is

Fig. 7: Examples of failure cases. We show the queries, predicted matches and ground truth correspondences in the first, second and third line respectively.

outperformed by a large margin by leveraging matches between local features and modeling image transformations. We also demonstrated the strong benefits of adapting the scores to the specificity of the problem and propagating information between correspondences. While our results are not perfect, they could still speed-up considerably the manual collation work, and are of practical interest.

Acknowledgements: This work was supported by ANR project EnHerit ANR-17-CE23-0008, project Rapid Tabasco, and gifts from Adobe. We thank Alexandre Guilbaud for fruitful discussions.

References

1. https://www.wdl.org. 4
2. https://www.themorgan.org. 4
3. https://digi.vatlib.it. 4
4. http://www.internetculturale.it. 4
5. Julien Bourdaillet and Jean-Gabriel Ganascia. Practical block sequence alignment with moves. In LATA, 2007. 3
6. Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv, 2020. 11
7. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009. 4, 7, 10
8. Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In ACM Multimedia, 2019. 5
9. Sovann En, Caroline Petitjean, Stephane Nicolas, and Laurent Heutte. A scalable pattern spotting system for historical documents. Pattern Recognition, 2016. 3

10. Daniel Stökl Ben Ezra, Bronson Brown-DeVost, Nachum Dershowitz, Alexey Pechorin, and Benjamin Kiessling. Transcription alignment for highly fragmentary historical manuscripts: The dead sea scrolls. In ICFHR, 2020. 3
11. Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 1981. 8
12. Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. IJCV, 2017. 4
13. Ronald Haentjens Dekker, Dirk Van Hulle, Gregor Middell, Vincent Neyt, and Joris Van Zundert. Computer-supported collation of modern manuscripts: Collatex and the beckett digital manuscript project. DSH, 2015. 3
14. Tal Hassner, Lior Wolf, and Nachum Dershowitz. Ocr-free transcript alignment. In ICDAR, 2013. 3
15. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 7, 10
16. John D Hobby. Matching document images with ground truth. IJDAR, 1998. 3
17. E Micah Kornfield, R Manmatha, and James Allan. Text alignment with handwritten documents. In DIAL, 2004. 3
18. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In ICCV, 2017. 4
19. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014. 4
20. David G Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004. 3
21. Tom Monnier and Mathieu Aubry. docExtractor: An off-the-shelf historical document element extraction. In ICFHR, 2020. 5
22. Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. TPAMI, 2018. 4
23. Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. MTA, 2016. 4
24. Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In ICCV, 2019. 4
25. Gil Sadeh, Lior Wolf, Tal Hassner, Nachum Dershowitz, and Daniel Stökl Ben-Ezra. Viral transcript alignment. In ICDAR, 2015. 3
26. Desmond Schmidt and Robert Colomb. A data structure for representing multi-version texts online. International Journal of Human-Computer Studies, 2009. 3
27. Xi Shen, Alexei A Efros, and Mathieu Aubry. Discovering visual patterns in art collections with spatially-consistent feature learning. In CVPR, 2019. 4, 6, 7, 8, 11, 12
28. Xi Shen, Ilaria Pastrolin, Oumayma Bounou, Spyros Gidaris, Marc Smith, Olivier Poncet, and Mathieu Aubry. Large-scale historical watermark recognition: dataset and a new consistency-based approach. In ICPR, 2020. 4, 7, 8
29. Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In ICCV, 2003. 3
30. Steven Escar Smith. " the eternal verities verified": Charlton hinman and the roots of mechanical collation. Studies in Bibliography, 2000. 3
31. Ignacio Úbeda, Jose M Saavedra, Stéphane Nicolas, Caroline Petitjean, and Laurent Heutte. Pattern spotting in historical documents using convolutional models. In HIP, 2019. 4, 6