

Visual Link Retrieval in a Database of Paintings

Benoit Seguin^(✉), Carlotta Striolo, Isabella diLenardo, and Frederic Kaplan

DHLAB, EPFL, Lausanne, Switzerland
{benoit.seguin, carlotta.striolo, isabella.dilenardo,
frederic.kaplan}@epfl.ch

Abstract. This paper examines how far state-of-the-art machine vision algorithms can be used to retrieve common visual patterns shared by series of paintings. The research of such visual patterns, central to Art History Research, is challenging because of the diversity of similarity criteria that could relevantly demonstrate genealogical links. We design a methodology and a tool to annotate efficiently clusters of similar paintings and test various algorithms in a retrieval task. We show that pre-trained convolutional neural network can perform better for this task than other machine vision methods aimed at photograph analysis. We also show that retrieval performance can be significantly improved by fine-tuning a network specifically for this task.

Keywords: Paintings · Visual search · Visual similarity

1 Introduction

In Art History, comparing paintings and finding relations between them is the basic block of many (if not most) analysis. The example of the painting of the *Virgin of the rocks*, by Leonardo da Vinci (Fig. 1), exemplifies how some painters were exposed in one way or another to the work of other’s, and how the masterpiece represents the final culmination of several visual references and the starting point for other interpretations of a specific theme or *formula* that we can summarize with the name “pattern”. These *visual links* are essential for studying the propagation of patterns and understanding the genesis of a single work of art, its reception and the history of a school of painting, and its influences, through centuries in Art History.

In order to study these visual links, art historians are often required to spend a lot of time in the few libraries which have acquired, across the years, the necessary amount of collections of photos to perform these analysis. Collecting and analyzing images is the starting point of the method for Art History. Starting by examining images of masterpieces was the approach that has characterized the largest schools of art criticism. It is clear that in order to define a set of homogeneous works attributed to a single author, or to the same painting school, historians made use of large photos datasets which helped them in cataloging and creating *corpora* [10]. In practice, however, scholars are still required to go

manually over thousands of physical photos with limited metadata to navigate through them.

With the increasing efforts of digitization of artworks in various institutions, we have an unprecedented access to large iconographic databases of the past, with hundreds of thousands of images. However, art historians are in need of tools to navigate through such large collections of images other than just using text-queries.

In this work, we acquired a dataset encoding pairs of images which are considered as *visually linked* by art historians. We investigated the challenges of making a visual retrieval system, which from one painting could retrieve elements which share a visual link with the query. For this purpose, we compare various visual encoding methods. Finally, we propose a way to improve the retrieval accuracy by specializing our method to the task at hand.

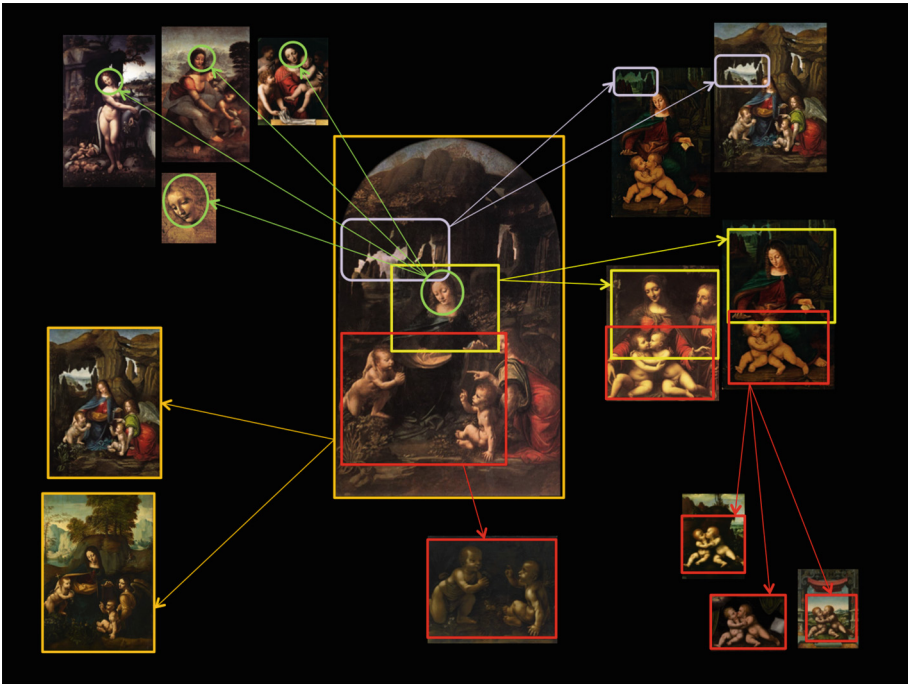


Fig. 1. Examples of visual links between artworks. The center image is the *Virgin of the rocks* by Leonardo da Vinci. It is easy to see how the global composition was reused by other painters (followers of Leonardo) on the bottom left. On the top left, other compositions by da Vinci himself reusing the same face. On the right, various sub-elements reused in other paintings. (*Best viewed in color*) (Color figure online)

2 Related Work

As far as analysis of paintings is concerned, most of the previous work actually comes from the Image Processing world with analysis such as brush-stroke extractions and image statistics to perform authorship ([20] for instance). But the goals and methods are not related to our project.

With the emergence of some online paintings datasets, some experiments trying to have automatic classification of style and/or artists have been done. Using CNN features [22] or combinations of them [9], the authors built classifiers to predict the painting style, genre or artist. In [31], they went slightly further by learning a metric to represent these classifications and used the learned metric to evaluate the “influence” of paintings [18].

In [12], the authors show that modern object classification frameworks based on convolutional neural network perform relatively well on paintings data. That way, they can have the user search for an object category in large collection of paintings from a simple text query.

Image retrieval is of course a well established field, with very powerful traditional methods based on local descriptors [21, 26, 32], and more recent methods using pre-trained CNN as global image descriptors with good performances [7, 8, 30]. However, the main benchmarks for image retrieval are *always* photographs, either of the same place (*Oxford5k*, *Paris6k*, *Holidays*) or of the same object (*UKB*). The closest dataset for our problem is probably the PRINTART database [11] but they only consider labels of scenes and not a fine grained visual similarity.

Since the signal of a painting image is different than a photograph. Applying methods that perform well on traditional datasets is not always straightforward. To our knowledge, there are only limited experiments for visual searches in paintings. Because of the extreme variety in style, working with them leads to tackling the issue of cross-domain matching. Previous work was mainly based on HoG [15] features used in a computationally expensive fashion to link paintings/sketches with photographs of the same scene [33] or with the 3D-model of the area [5]. The use of discriminant regions was also evaluated in [13].

3 Dataset Creation

Our first contribution is the creation of a dataset tackling the problem of visual links retrieval in paintings. Given a set of images of works of art P we consider two paintings $x, y \in P$ to be linked if an expert consider them to have a visual relation with each other. Each one of these links can actually be considered as an edge, building a graph linking elements of the dataset with each other.

Annotating such information is difficult in practice because it is a N-to-N problem. Unlike tasks like classification or prediction, an expert can not look at one image and give the complete ground-truth. In order to get the complete ground-truth, one would need to look at all pair-wise relationships ($O(N^2)$) which is impossible for a large N .

Hence, building the whole graph is intractable in practice, but our goal was to build a subset of it for evaluation purposes. Some of these visual links are actually known by the art history community, but often scattered in multiple books and separate analysis.

In fact, a fairly common approach is monographic. It is related to analyze a painter in particular, his artistic career, trying to track down all of his works, and that of his workshop. Rarely this analysis leaves the geographical boundaries and a specific time of diffusion [36, 37].

Another main approach, however, seeks to analyze the dimensions and diffusion of the transmission of visual knowledge through several criteria. The images are used to understand the cultural contexts in which some elements, some patterns, have been taken, reformulated, and have been successful. This way takes into account different implications such as the “geography of art”: the propagation of relationships through countries and cultures [14]. The spread of a particular pattern in an author and his commercial success are related to the history of collecting and to the history of the taste, both aspects being relevant in order to explain the propagation [28, 39].

In all these approaches finding the links between the images has a key role. For this our task was then to transfer this knowledge to a digitized format.

3.1 Choice of the Base Corpus

In order for experts to draw links between elements, we needed a base corpus of images. The fact that the migration of patterns in paintings is mainly important in the Modern Period (1400–1800) is an important factor in choosing our base *corpus*. As far as online catalogs of paintings are concerned, a few candidates are possible:

- **Google Art Project** [2]: large collection extracted unfortunately mainly from American museums, with poor coverage of the Renaissance.
- **BBC YourPaintings** [1]: British effort of categorization and labeling of the British museums collections. Mainly focused on British oil paintings of the 19th century. Used in [12, 13] for object classification.
- **RKD Challenge** [25]: coming from the Rijksmuseum, this benchmark was created for scientists to test their algorithms on artists identification, labelling of materials and estimating the creation year. Boasting 112k elements, only 3’600 are actual paintings.
- **BnF Benchmark** [27]: created for the work in [27]. This benchmark coming from the Bibliotheque Nationale de France is made of 4’000 images with the goal of label propagation. Additionally, the diversity of mediums is high (paintings, drawings, illuminations, maps etc.).
- **Wikiart** [4]: large collection of images (126k) of paintings. Because it associates each painting with a style and a genre, it is the basis of various algorithms trying to predict these characteristics [9, 18, 22, 31]. It was one our two main candidates.

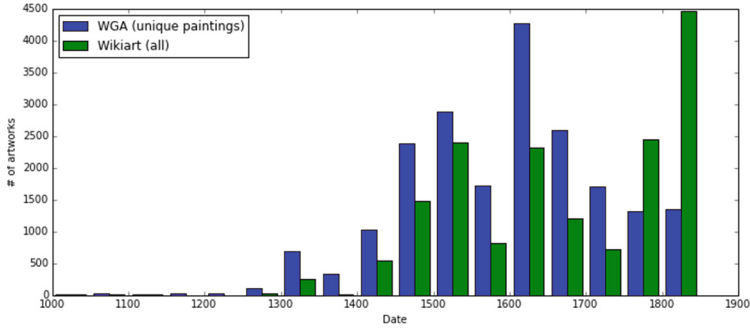


Fig. 2. Distribution of the artworks over time (until 1850) for two different datasets.

- **Web Gallery of Art [3]:** the Web Gallery of Art (WGA) is a smaller collection of almost 40k images. After taking out images which are not related to our analysis (sculpture, architecture...) and removing the images which are details of others, we get around 28k elements.

For the Wikiart and WGA datasets, we plotted the distribution of artworks over time on Fig. 2. It is obvious here that despite having less elements overall the WGA is a better choice for our analysis on the 1400–1800 period, making it our base corpus later in this work.

3.2 Gathering Method

We designed a web-based annotation tool that had three characteristics: the user can easily navigate through the database and compare images, the user can upload new images to the database and the user can make connections between entries of the database.

With this tool, an expert could find visual links by navigating the data through educated guesses and create a connection. Or if he knows about specific links (through the art-history literature and/or experience), he could transcribe the information to the system, either by finding the elements back in the database, or uploading the missing ones.

In practice, we realized it was impractical for the experts to annotate the links one by one. More precisely, in the examples we find, it is more common to find some “cluster-like” or group structure, where all the elements are linked with each other. Examples of such groups can be seen on Fig. 4. Most of these groups consist of a set of paintings (mostly between 2 and 7 elements) sharing a common pattern. In the end, we had users annotate these groups directly that we later translate to fully-connected clusters in the graph.

3.3 Data Gathered

Over the course of a month, an art historian was able to annotate 217 different groups of images. The numbers of images per group is variable and the distrib-

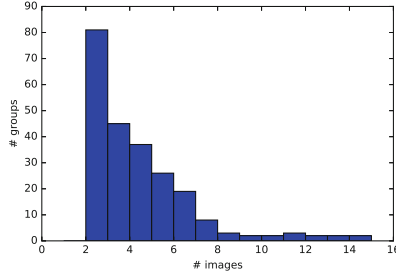


Fig. 3. Distribution of the number of images per annotated cluster.

ution can be seen on Fig. 3. This translates to 1’280 edges in the graph of visual links between 845 different images. 461 images were extracted from other sources and manually added when they were not found in the base corpus.

The extracted data provides us with a challenging benchmark as seen in Fig. 4. Variability in medium, style, and reuse of details is unique, and gives us a unique case of cross-domain visual matching.

4 Algorithms Evaluated

4.1 Bag-of-Words Methods

The main class of algorithms used very successfully in the problem of visual instance retrieval are based on local visual descriptors (mainly SIFT [24]). From the first Bag-of-Words representation for image retrieval [35], various improvements were proposed ranging from better clustering [21], spatial verification [21, 26] or query expansion [32].

However, previous works on cross-domain matching [5, 12, 33] have shown that while these methods perform well on photographs, the performance of SIFT across domains drops drastically. Still, to support our claim, we implemented a version of the algorithm described in [26].

We computed the SIFT descriptors for every image of the dataset. We used 10M descriptors extracted from 5’000 randomly chosen images as our training data for our dictionary. Using K-means we clustered it in 100k visual words. Re-ranking is done by evaluating a simple scale + rotation transformation.

4.2 CNN Methods

In the recent years, deep Convolutional Neural Networks (CNN) [23] trained on very large corpus [16] have been shown to perform very well in almost every area of computer vision. For instance, reusing the first layers of a network have been shown to be an extremely good base representation of the visual information [17, 29]. More specifically, applications of pre-trained CNN to the problem of visual instance retrieval have been studied in [6, 8, 30] on the classic *Oxford5k*, *Paris5k* and *Holidays* benchmarks.



Fig. 4. Examples of portions of annotated groups. **First row:** *Leda and the swan* different mediums (RUBENS, Peter Paul: painting; CORT, Cornelis: engraving; MICHELANGELO Buonarroti: drawing) **Second row:** similar composition (MASSYS, Quentin *The Moneylender and his Wife*; REYMERSWAELE, Marinus van *The Banker and His Wife*) **Third 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)

Building on these analysis, we use the VGG16 CNN architecture [34] as our base network (see Fig. 5). We extracted the activation of the *fc6* and *fc7* layers, almost mimicking [8], and the last convolutional layer activations *pool5*, inspired by [30].

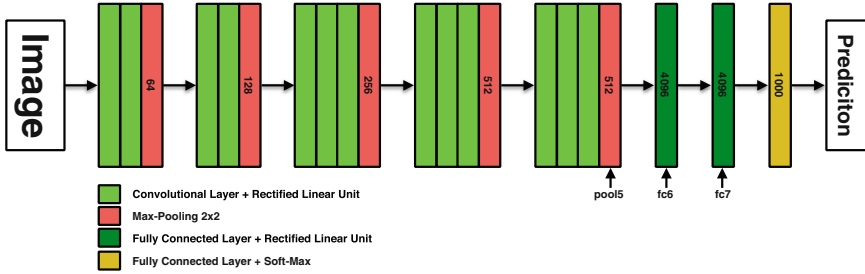


Fig. 5. The VGG16 architecture trained on the ImageNet competition. It is made by successively stacking two or three 3×3 convolutional layers, then using a max-pool layer to downsize the spatial resolution. Three fully connected layers are finally used, giving the class prediction scores. The number of feature maps at some layers is displayed. In order to use the fully-connected layers though, the result of pool5 is supposed to be 7×7 spatially, which forces the input image to be a 224×224 square.

In order to extract the fully-connected features (*fc6* and *fc7*), we need to give a square input of size 224×224 to the network. Because of the variable image ratio, we tried either extracting the center of the image or warping the image to a square. The feature vectors are then l_2 -normalized.

For the convolutional features (*pool5*), the image was isotropically resized for its smaller dimension to be equal to 256. Then we take a global sum-pool or max-pool operation (following [7] or [30] respectively) on the obtained feature-maps. We also experimented with *spatial-pooling* (SP) [30], which consists of performing the pooling operation separately on the four quadrants of the feature maps, hence multiplying the dimension of the feature vector by four. Finally l_2 -normalization is also applied. A schematic of this pipeline can be seen on Fig. 6.

Searches are then performed by using the l_2 distance between the image descriptors in a nearest neighbour fashion.

4.3 Fine-Tuning the Network

On the one hand, the visual variations across elements are high: the image can be grayscale or a sketch, the colors might be completely different, etc. On the other hand, the visual features we used were pre-trained on ImageNet which is only a collection of photographs of objects with their labels. It then makes sense to hope for improvements in the retrieval performance by fine-tuning the network.

A related approach was taken in [8] where they train a classification CNN on locations in cities, and then use the learned filters trained on this dataset instead of ImageNet, showing an improvement. Here, we want to learn the visual representation directly.

Our visual search is performed by doing nearest neighbour in our feature space from a query. To that regard, our feature extraction pipeline can be seen as a function embedding an image to a point in the feature space. Our goal

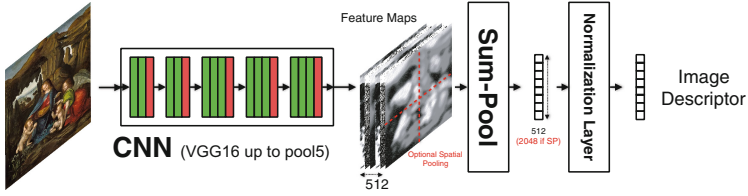


Fig. 6. Feature extraction pipeline.

would be to improve this embedding such that for two images to be close in this embedding would mean a high probability to share a visual connection.

In order to learn an embedding with a neural network, two approaches are possible. The first approach consists of submitting pairs of training images (X, Y) to the network, telling it if they are similar or not [19]. The second approach is to use triplets of images (A, B, C) telling the network that $d(f(A), f(B))$ should be smaller than $d(f(A), f(C))$ (where d is a distance function and f the embedding function) [38]. Since we are interested in making a ranking system, the order of proximity is what is important to us and the second approach then better suited.

In practice, we start with the feature extraction pipeline described above and represented on Fig. 6. Using some part of our dataset, we generate training queries. Each query $(Q_i, \{T_{i,j}\})$ consists of an image Q_i , and a set of images $\{T_{i,j}\}$ which all have a visual link with Q . Then we perform some hard-negative mining: we first run the query Q_i using the feature representations computed with our initial model, then we can easily generate interesting learning triplets by outputting $(Q_i, T_{i,j}, N_{i,j,k})$ where $N_{i,j,k}$ is an image not sharing a visual link with Q_i but is highly ranked if we search from Q_i in the original feature space.

From these triplets, we use a similar learning approach as [38]. If we consider the output of our network to be the function $f(\cdot)$ then the loss we try to minimize is the Hinge loss:

$$\max(0, d(Q_i, N_{i,j,k}) - d(Q_i, T_{i,j}) - \delta)$$

In our case d is the l_2 distance. Also, unlike [38] we did not use a regularization term, the l_2 norm of the parameters was actually almost not varying during training.

Training was done with Stochastic Gradient Descent with momentum (learning rate: 10^{-5} , momentum term: 0.9) and took around 50 epochs to converge. Batches are slightly tricky to make as we need each part of the triplet to have similar sized images (i.e. all the Q_i of the batch to have size s_1 , all the $T_{i,j}$ to have size s_2 etc.). Because of this, we had to discard a small portion of the data to make batches with a minimal size of 5 (and forced the maximum size to be 10). In the end, we used around 25k triplets for training and 5k for validation purposes.

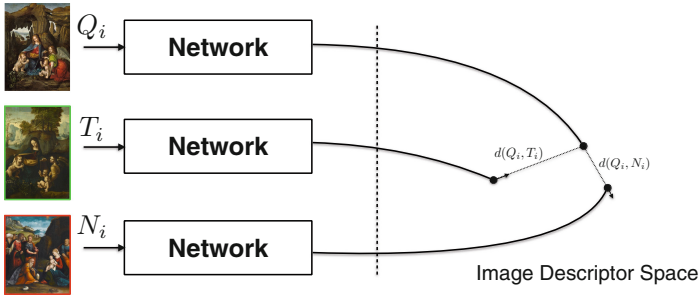


Fig. 7. Triplet learning framework.

5 Evaluation

Our goal is to make a search system to help art historians navigating through large collection of images. Hence, the main scenario is the user submitting an image as query, and we want to evaluate how well the system can give back the elements linked to it in our visual links graph. The metric we used is then the recall at certain ranks in the search results.

We divided our dataset into separate sub-graphs. 50% of the data was kept for training, 25% for validation purposes and 25% for actual testing. The testing set was made of 199 images.

Given a ranking algorithm F that given an image input Q outputs an ordered list of images O_i , we want to evaluate its performance. Every image I of the testing set defines a query $(I, \{T_i^I\})$ where $\{T_i^I\}$ is the set of images sharing a visual connection with I . The recall at rank n for a single query is:

$$R^I[n] = \frac{|\{T_i^I\} \cap \{O_i\}_{i \leq n}|}{|\{T_i^I\}|}, \text{ where } \{O_i\} = F(I) \text{ and } |\cdot| \text{ is the cardinal of a set.}$$

Computing the recall for the whole testing set is then just an aggregation of the recall for single queries:

$$R[n] = \sum_I w(I) \cdot R^I[n]$$

However, choosing the weights $w(\cdot)$ to balance the influence of each query in the final result is a bit arbitrary. If we choose $w(I) = 1$, then all the queries would be considered equivalent, even if some have a higher number of visual connections than other. If we choose $w(I) = |\{T_i^I\}|$, then every visual connection is considered equally influent, which is not desirable either. Indeed, if we have a group of N elements which are close variations of each other, we have $\frac{N(N+1)}{2}$ separate links but which mainly encode the same visual relation. Taking this case as a basis, we want the weight of a fully connected group to be proportional to the square root of the number of visual links it represents. This gives us

the weight function: $w(I) = \sqrt{\frac{|\{T_i^I\}|}{|\{T_i^I\}|+1}}$. In practice, the choice of $w(\cdot)$ is not so important as it seems to have little impact on the ranking of the different methods.

6 Results

We evaluated the algorithms on the 199 queries of the testing set, using the whole WGA (38’500 images) as our search space. In Table 1, we are displaying various values of the *Recall* metric described in the previous section. We did not include results concerning the *fc7* layer because they perform poorly compared to layer *fc6* (this is in accordance with previous research of CNN features transferring for image retrieval).

The first observation from the results is the confirmation of our intuition that the Bag-of-Words method is not performing very well, even with a geometrical re-ranking step. The extreme variability in patterns, style and colors seems to be too strong for a dictionary of SIFT descriptor to handle.





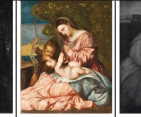

As far as the output of the first fully-connected layer is concerned (*fc6*), it seems that extracting the squared-center of the image performs better than warping the image to a square. This seems to imply it is better to use only a sub-part of the image unmodified rather than using all of it, even if distorted.








Table 1. Recall metrics for the evaluated methods. D specifies the dimension of each representation.

Method	D	R[20]	R[50]	R[100]	R[200]
BoW	-	7.8	11.6	13.9	15.8
BoW + Geometrical Reranking	-	11.3	13.0	14.3	15.2
fc6 layer + Warp Extraction	4096	33.4	42.0	46.6	53.8
fc6 layer + Center Extraction	4096	37.2	43.1	50.1	57.7
fc6 layer + Center Extraction + PCA	2048	40.2	48.8	54.9	61.6
pool5 layer + max-pool	512	33.5	41.1	46.1	53.5
pool5 layer + sum-pool	512	36.4	43.0	51.7	58.1
pool5 layer + 2 × 2-sum-pool	2048	46.1	49.9	54.6	59.8
pool5 layer + 2 × 2-sum-pool + PCA	1024	46.5	51.4	56.4	62.5
pool5 layer + sum-pool + fine-tuning	512	45.3	53.4	60.3	68.3
pool5 layer + 2 × 2-sum-pool + fine-tuning	2048	47.5	55.5	60.8	68.3
pool5 layer + 2 × 2-sum-pool + fine-tuning + PCA	1024	48.2	57.5	63.6	70.8







When we use the output of the last convolutional layer (*pool5*), we do not need to crop or warp the image but we need to aggregate the activations of this layer. As already hinted by [7], using the *sum* operation instead of *max* during




Table 2. Example of queries of the testing set, and the retrieval rank of their respective linked images. Here *fc6*, *pool5* and *fine-tuned* represent respectively *fc6 layer + Center Extraction + PCA*, *pool5 layer + 2 × 2-sum-pool + PCA* and *pool5 layer + 2 × 2-sum-pool + fine-tuning + PCA* in the result table. For each table, the first image is the query and the others are the targets of the query.

					
fc6	>1000	>1000	1	>1000	3
pool5	504	716	1	764	3
fine-tuned	32	52	1	74	4

						
fc6	1	186	3	>1000	>1000	>1000
pool5	1	17	13	951	>1000	>1000
fine-tuned	2	3	1	91	813	968

				
fc6	>1000	238	53	>1000
pool5	>1000	4	76	92
fine-tuned	>1000	1	>1000	35

					
fc6	317	536	330	52	487
pool5	>1000	964	598	14	11
fine-tuned	>1000	>1000	126	1	27

		
fc6	449	>1000
pool5	633	>1000
fine-tuned	365	652

the pooling phase improves the results. Also, the spatial-pooling proposed by [30] (referred as $2x2$ -*-pool in the table) allows a very efficient way to incorporate some structure in the image descriptor, improving the $R[20]$ score by 10%. Although, it is probable this step greatly helps for similar global composition link (i.e. easy cases), but might hurt for links only defined by a detail.

After fine-tuning our convolutional filters through our triplet-learning procedure, we can observe a dramatic improvement in performance. The *pool5 + sum-pool* method improves by 8.9% and 10.2% for the $R[20]$ and $R[200]$ scores respectively. Comparatively speaking, the improvement in the case of spatial-pooling is smaller, especially for the first elements of the ranking (Table 2).

From a qualitative point of view, some examples of queries are displayed in Fig. 2. The first two queries are typical cases where fine-tuning the convolutional filters allow the retrieval system to better handle variations (color \leftrightarrow grayscale, style,...). In the third row, we can see the improvement in rankings for the second and fourth target, but the actual loss of precision because of a mirroring composition for the third element. Finally, the last rows describes very difficult cases, either because the similarity is almost more semantic than local (fourth row), or because the medium is very different (fifth row).

7 Conclusion

In this paper, we interested ourselves in the retrieval of visual links in databases of paintings. Using a specific dataset created for this purpose, we showed that traditionally efficient methods based on Bags-of-Words fall short on this specific problem. However, recent methods based on pre-trained CNN perform favorably. Finally, we demonstrated how using some initial knowledge as training can dramatically improve the performance of the CNN descriptors at little cost.

References

1. BBC your paintings. www.artuk.org/discover/artworks
2. Google art project. www.google.com/culturalinstitute
3. Web gallery of art. www.wga.hu
4. WikiArt. www.wikiart.org
5. Aubry, M., Russell, B.C., Sivic, J.: Painting-to-3D model alignment via discriminative visual elements. *ACM Trans. Graph.* **33**(2), 1–14 (2014)
6. Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: From generic to specific deep representations for visual recognition
7. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval (2015)
8. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014*. LNCS, vol. 8689, pp. 584–599. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-10590-1_38](https://doi.org/10.1007/978-3-319-10590-1_38)

9. Bar, Y., Levy, N., Wolf, L.: Classification of artistic styles using binarized features derived from a deep neural network. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8925, pp. 71–84. Springer, Heidelberg (2015). doi:[10.1007/978-3-319-16178-5_5](https://doi.org/10.1007/978-3-319-16178-5_5)
10. Berenson, B.: Venetian Painters of the Renaissance (1894)
11. Carneiro, G., Silva, N.P., Bue, A., Costeira, J.P.: Artistic image classification: an analysis on the PRINTART database. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 143–157. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-33765-9_11](https://doi.org/10.1007/978-3-642-33765-9_11)
12. Crowley, E.J., Zisserman, A.: In search of art. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8925, pp. 54–70. Springer, Heidelberg (2015). doi:[10.1007/978-3-319-16178-5_4](https://doi.org/10.1007/978-3-319-16178-5_4)
13. Crowley, E.J., Zisserman, A.: The State of the Art: Object Retrieval in Paintings using Discriminative Regions (2014)
14. Da Costa Kaufmann, T.: Toward a Geography of Art. The University of Chicago Press Books, Chicago (2004)
15. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE (2005). <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1467360>
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2–9 (2009)
17. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: a deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning, pp. 647–655 (2014). <http://arxiv.org/abs/1310.1531>
18. Elgammal, A., Saleh, B.: Quantifying creativity in art networks, June 2015. <http://arxiv.org/abs/1506.00711>
19. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1735–1742 (2006)
20. Hughes, J.M., Graham, D.J., Rockmore, D.N.: Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. Proc. Natl. Acad. Sci. U.S.A. **107**(4), 1279–1283 (2010)
21. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008). doi:[10.1007/978-3-540-88682-2_24](https://doi.org/10.1007/978-3-540-88682-2_24)
22. Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H.: Recognizing image style. In: ECCV, pp. 1–20 (2014). <http://arxiv.org/abs/1311.3715>
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004). <http://link.springer.com/10.1023/B:VISI.0000029664.99615.94>
25. Mensink, T., Gemert, J.V.: The Rijksmuseum Challenge: Museum-Centered Visual Recognition, pp. 2–5 (2014)

26. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with largevocabularies and fast spatial matching. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2007)
27. Picard, D., Gosselin, P.H., Gaspard, M.C.: Challenges in content-based imageindexing of cultural heritage collections. *IEEE Signal Process. Mag.* 95–102 (2015). <https://hal.archives-ouvertes.fr/hal-01164409>
28. Pomian, K.: *Collectionneurs, amateurs, et curieux. XVIe - XVIIIe siècle.* Paris, Venise (1987)
29. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S., Sharif, A., Hossein, R., Josephine, A., Stefan, S., Royal, K.T.H.: CNN features of-the-shelf: an astounding baseline for recognition. In: CVPR, pp. 512–519 (2014)
30. Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: A baseline for visual instance retrieval with deep convolutional networks, December 2014. <http://arxiv.org/abs/1412.6574>
31. Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: learning the right metric on the right feature, p. 21, May 2015. <http://arxiv.org/abs/1505.00855>
32. Shen, X., Lin, Z., Brandt, J., Avidan, S., Wu, Y.: Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2012)
33. Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A.: Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph.* **30**(6), 1 (2011). <http://dl.acm.org/citation.cfm?id=2070781.2024188>
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scaleimage recognition. arXiv Preprint, pp. 1–10 (2014). <http://arxiv.org/abs/1409.1556>
35. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of CVPR (ICCV), pp. 2–9 (2003)
36. Giorgio, T., Bernard, A., Mancini Matteo, M.A.J.: *Le botteghe di Tiziano.* Alinari, Florence (2009)
37. van Hout, N., Merlu du Bourg, A., Gruber, G., Galansino, A., Howarth, D.: *Rubens and His Legacy.* Royal Academy of Arts, London (2014)
38. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: CVPR, pp. 1386–1393 (2014)
39. Warnke, M.: *Bilderatlas Mnemosyne.* Akademie, Berlin (2000)