Tracking transmission of details in paintings

Benoit Seguin
benoit.seguin@epfl.ch
Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Isabella di Lenardo
isabella.dilenardo@epfl.ch
Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Introduction

In previous articles (di Lenardo et al, 2016; Seguin et al, 2016), we explored how efficient visual search engines operating not on the basis of textual metadata but directly through visual queries, could fundamentally change the navigation in large databases of work of arts. In the present work, we extended our search engine in order to be able to search not only for global similarity between paintings, but also for matching details. This feature is of crucial importance for retrieving the visual genealogy of a painting, as it is often the case that one composition simply reuses a few elements of other works. For instance, some workshops of the 16th century had repertoires of specific characters (a peasant smoking a pipe, a couple of dancing, etc.) and anatomical parts (head poses, hands, etc.) ,that they reused in many compositions (van den Brink, 2001; Tagliaferro et al, 2009). In some cases it is possible to track the circulation of these visual patterns over long spatial and temporal migrations, as they are progressively copied by several generations of painters. Identifying these links permits to reconstruct the production context of a painting, and the connections between workshops and artists. In addition, it permits a fine-grained study of taste evolution in the history of collections, following specific motives successfully reused in a large number of paintings.

Tracking these graphical replicators is challenging as they can vary in texture and medium. For instance, a particular character or a head pose of a painting may have been copied from a drawing, an engraving or a tapestry. It is therefore important that the search for matching details still detects visual reuse even across such different media and styles. In the rest of the paper, we describe the matching method and discuss some results obtained using this approach.

Method

Matching patterns in a bank of images is a problem that has been extensively studied in the Computer Vision community as « Visual instance retrieval » (Sivic and Zisserman, 2003). The definition of the task is : *given a region of a query image, can we identify matching regions in other images of a large iconographic collection* ?

Historically, the most successful methods were based on feature point descriptors (like SIFT, see Lowe, 2004) used in a Bag-of-Words (Jegou et al, 2008) fashion. The global architecture can be summarized as follows:

For each image in the collection:

- Extract the feature points for the image.
- Quantize the point descriptors to a Visual-Bag-of-Words representation.

Given a query region:

- Use the Bag-of-Words signatures to rank the first N most likely candidates of the collection.
- For each candidate, re-rank them according to a spatial verification of the matching points with the query region.

In practice, such an approach works extremely well and numerous improvements (Shen et al, 2009; Crowley and Zisserman, 2014) have been brought over the years to the different steps of the procedure. Hence it is still considered state of the art for the traditional datasets which focus on building and object retrieval in photographs.

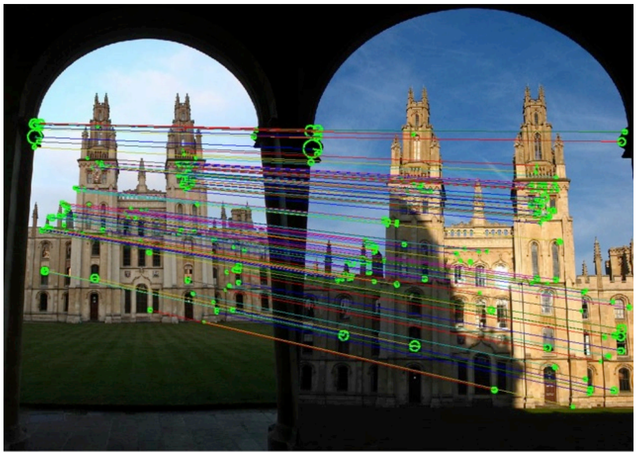

Figure 1: Working Bag-of-Words matching for buildings.

However, it was shown recently (Seguin et al, 2016; Crowley and Zisserman, 2014) that such approaches break completely when not dealing with the same physical objects and important style variations like we do in paintings. An example can be seen on Figure 2.

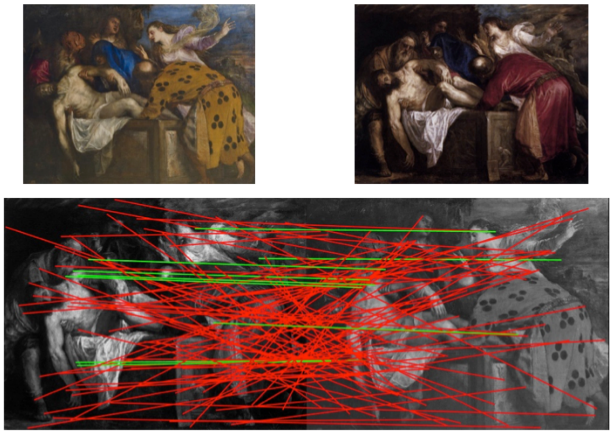

Figure 2: Feature point matching breaks when local features and style vary.

More recently, Convolutional Neural Networks (CNN) have had tremendous success in almost all areas of Computer Vision (object detection, recognition, segmentation, face identification, etc.) and CNN have established themselves over the last couple of years as an extremely powerful tool for almost any vision based problem.

A CNN is a multi-layer architecture where each layer transforms its input according to some parameters (also called weights). What makes them so powerful is that all these parameters can be learned « end-to-end » (for example in the case of object classification, just with images and their corresponding labels).

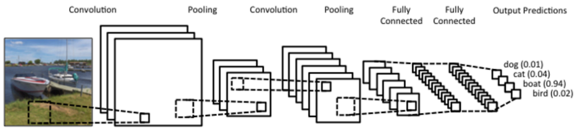

Figure 3: Simplified structure of a Convolutional Neural Network.

It has been shown in Donahue et al (2014) that CNN pretrained on very large datasets, like Deng et al (2009), for object classification tasks offer a very good abstract representation of the image information, and are thus applicable for other vision problems. They generalize much better when transferring from photographs to paintings, contrary to the traditional Bag-of-Words techniques (Seguin et al, 2016; Crowley and Zisserman, 2014). However, its application to visual instance retrieval was always hindered by the fact that they traditionally output a single global descriptor for the image, hence not directly allowing for region (sub-image) retrieval (Babenko et al, 2014).

In Razavian et al (2014), the authors proposed to just precompute the CNN descriptors for some subdivision of the image. However, such approach limits the possible granularity of the windows, and multiply the memory requirement by a huge factor.

Another more promising approach introduced in Tolias et al (2015) is to work directly on the CNN feature maps. More precisely, a common way of extracting a CNN descriptor for image retrieval is to take an image of size *(H, W, 3)* (height of *H* pixels, width of *W* pixels and 3 color channels RGB) go through all the convolutional layers of CNN, which outputs the *feature maps* : a structure of size *(H', W', F)* (*F* channels of size *H'* and *W'*). From these feature maps, taking the sum (or the max) of each channel gives a signature of size F which is (after normalization) the descriptor of the image.

Traditionally, the network used is the VGG16 (Simonyan and Zisserman, 2014) architecture which given an image of size *(H, W, 3)* creates feature maps of size *(H/32, W/32, 512).*

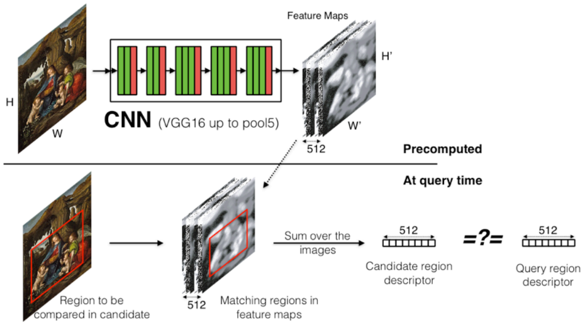

Figure 4: Region evaluation architecture.

Now, starting from the feature maps, computing the signature of a sub-part of an image is already easier, we just need to compute the sum of the corresponding region in the feature maps to obtain the descriptor of size *F*. However, evaluating many different regions would still be prohibitive from a computation point of view.

In order to alleviate the performance problem, Tolias et al (2015) proposed to use *integral images*. Given an image I, the integral image $I_\int$ is $I_\int(y,x) = \sum_{i<x,j<y} I(j,i)$. This allow for extremely quick computation of the sum of an image for a given area (y_1, y_2, x_1, x_2) (Fig.5) :

$$\sum_{x_1 \leq i < x_2, y_1 \leq j < y_2} I(j,i) = I_\int(y_2,x_2) + I_\int(y_1,x_1) - I_\int(y_1,x_2) - I_\int(y_2,x_1)$$

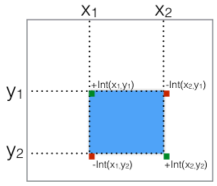

Figure 5: Integral images

This trick allows for extensive evaluation for the best matching window for the query image in the target collection. The global procedure for searching is the following, quite close to the Bag-of-Word approach:

For each image in the collection:

- Extract the feature maps of the image, and compute the corresponding integral images.
- Compute the global signature (with the whole image as window).

Given a query region:

- Use the global signatures to rank the first N most likely candidates of the collection.
- For each candidate, re-rank them according to an extensive look of the sub-windows in the images using their pre-computed integral images.

In order to greatly improve the results, we add the following improvements:

- The parameters of the network we use were fine-tuned using the Replica dataset (Seguin et al, 2016) (image retrieval in paintings). This dramatically improves the system resilience to color and style.
- We use Spatial Pooling according to Razavian et al (2014) which consists of extracting 4 blocks per evaluated region instead of 1. It makes the search roughly 4 times slower but allow for much better retrieval of complex patterns by directly encoding spatiality.

Results

The following experiment was run on the whole Web Gallery of Art collection (38'000 elements). Each image was resized so that its smaller dimension is 512 pixels, and the integral images of the feature maps computed on it. Given a query region for an image, the 300 most likely candidates are extracted from the WGA collection, and re-ranked according to the best matching window on each of them. Using 35 cores on a server machine, the complete request takes less than 4 seconds.

Examples of queries and their results are shown on Figure 6. In query 1 and 2, the starting image is Leonardo da Vinci's *Virgine delle Rocce* (Paris, Louvres), first version of this subject (1483-1486). Leornado is an interesting case study and his influence in Europe is known to be extremely important for the general compositions of paintings, character typologies and landscape patterns. In query 1, the group of Mary, the angel and the two children is selected. The first result is the version of the same subject (London, National Gallery) finished a decade after the Paris version, with the contribute of Ambrogio de Predis. The painting is different in color but has the same composition. The second and third results are other versions of the same subject by unidentified painters of the XVIth century. This constitutes typical examples of the propagation of a complex theme.

In query 2, only a detail from the landscape is chosen. Interestingly, the second result is a painting from

Bernardino de Conti, which is a variation on the same theme, reusing the landscape but without the angel and with the two children kissing.

For query 3, we use a painting by Marco d'Oggiono, a follower of Leonardo, very similar to the one by de Conti and we select only the two children kissing. Results 1 and 3 feature paintings where only the children are present, showing that this replicator has an autonomy of its own. The third result by Joos van Cleve confirms the historical migration of this subject, as autonomous, from Italy to Flanders.

These 3 simple queries illustrate how the detail matching method can easily unveil the transmission network between different series of paintings.

Perspectives

The search of matching details in large-scale databases of paintings may enable to find undocumented links and therefore new historical connections between paintings. By tracking the propagation and transformations of a replicator, it becomes possible to follow the evolution through time of repertoires of forms and view each painting as a temporary vehicle playing the role of an intermediary node in a long history of images transmissions. Although in continuation with traditional methods in Art History, such a tool opens the avenue for research at a much larger scale, searching for patterns and finding new links simultaneously in millions of digitized paintings.

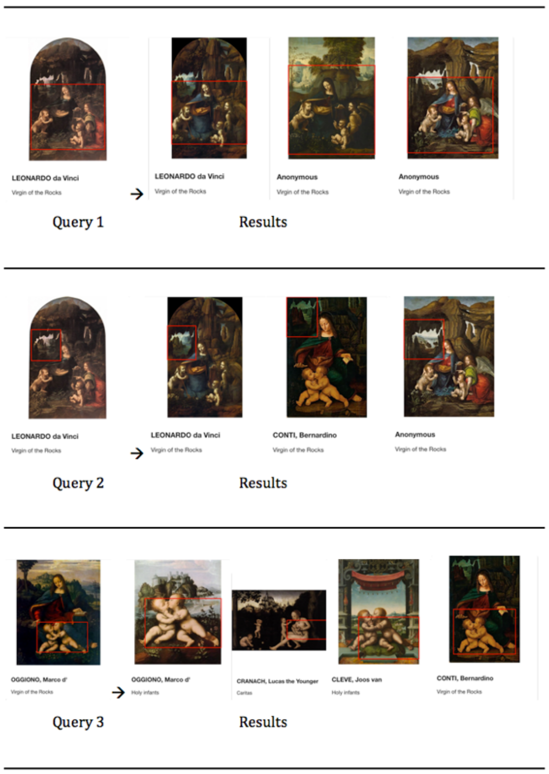

Figure 6: Examples of results of detail search.

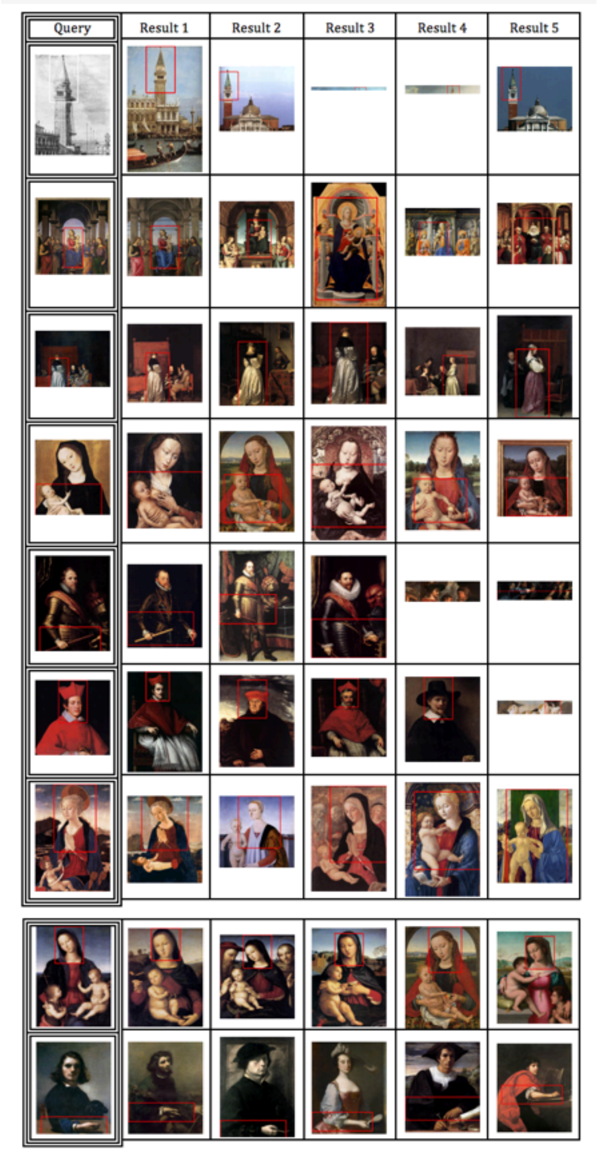

Figure 7: Additional examples of results of detail search.

Bibliography

Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. (2014) “Neural codes for image retrieval,” in *ECCV*.

Crowley, E. J., and Zisserman, A. (2014) “In search of art,” *ECCV Workshops*, 2014.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014)“DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition,” *ICML*.

Deng, J., Dong, W., Socher, R. Li, L.-J., Li, K., and Fei-Fei, L. (2009) “ImageNet: A large-scale hierarchical image database,” *CVPR*.

Jegou, H., Douze, M., and Schmid, C. (2008) “Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search ”, *ECCV*.

di Lenardo, I., Seguin, B. L. A., and Kaplan, F. (2016). Visual Patterns Discovery in Large Databases of Paintings. Digital Humanities 2016, Krakow, Polland, July 11-16, 2016.

Lowe, D. G. (2004) “Distinctive Image Features from Scale-Invariant Keypoints,” *Int. J. Comput. Vis.*, Nov. 2004.

Razavian, A. S., Sullivan, J., Maki, A., and Carlsson, S. (2014) “A Baseline for Visual Instance Retrieval with Deep Convolutional Networks,” Dec.

Seguin, B., Striolo, C., di Lenardo, I., and Kaplan, F. (2016) Visual link retrieval in a database of paintings, VISART : Where Computer Vision Meets Art, 3rd Workshop on Computer Vision for Art Analysis, October 2016, Amsterdam, The Netherlands

Shen, X., Lin, Z., Brandt, J., Avidan, S., and Wu, Y. (2012) “Object retrieval and localization with spatially-constrained similarity measure and k -NN re-ranking,” *CVPR*.

Simonyan, K., and Zisserman, A. (2014) “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv Prepr.*

Sivic, J., and Zisserman, A. (2003) “{Video Google:} A text retrieval approach to object matching in videos,” *CVPR*.

Tagliaferro, G., Aikema, B., Mancini, M., Martin, A. J.. (2009) *Le Botteghe di Tiziano Alinari*, Firenze

Tolias, G., Sicre, R., and Jégou, H. (2015) “Particular object retrieval with integral max-pooling of CNN activations,” *arXiv Prepr. arXiv1511.05879*

van den Brink, H. M., ed. (2001) L’entreprise Brueghel, , Maastricht, Bonnefantenmuseum- Bruxelles, Musée royaux des beaux-arts, Beaux-Arts Collection 2001