Similar paintings retrieval from individual and multiple poses

Adrien Deliege[1] and Maria Giulia Dondero[1,2]

[1] University of Liège, Belgium
[2] F.R.S.-FNRS, Belgium

Abstract. This paper introduces an approach for retrieving similar paintings in terms of poses of the characters depicted within them. The key contributions are a method to extract and normalize individual character poses, the development of several criteria to compare groups of poses across paintings, and the integration of these criteria into a unified ranking system to identify the most similar artworks for a given query. The proposed techniques are demonstrated on a corpus of religious paintings, showing their effectiveness in retrieving visually and semantically analogous artworks. The findings suggest that this methodology could prove helpful for extensive analyses of large image datasets.

Keywords: digital humanities · pose estimation · computer vision

1 Introduction

Advancements in the last decade in computer vision have opened up new avenues for the analysis and exploration of artworks [3, 16, 18, 19]. Deep learning-based techniques have enabled the automated extraction of rich visual features from paintings, sculptures, and other forms of figurative art. This has unlocked opportunities for large-scale, data-driven studies that were previously infeasible through manual inspection alone, such as classification [8, 15, 23], retrieval [7, 12, 20, 25, 26] or even visual question answering [1, 9].

In the specific domain of artwork retrieval [7, 12, 20, 25, 26], the ability to quantify and compare the poses of characters depicted in paintings holds significant promise [2,10,12,14,15,17,27]. Identifying visually and semantically similar artworks based on their figurative elements can aid art historians, curators, and enthusiasts in tasks like collection management, provenance research, and visual analysis. In particular, Impett and Moretti's formalization of gestures [10, 11] in Warburg's Atlas Mnemosyne panels, has attempted to measure and compare the movements of bodies represented in various images, by reducing the body to a set of angles formed by the joints of the character skeletons. The attempt is remarkable because it does so in order to arrive at the formal description of Pathosformel, the forms of pathos in gestures. Nevertheless, only individual skeletons are considered, not groups, and they are extracted from a relatively small corpus. Jenicek and Chum [12] consider a larger database and all the characters of the paintings, displaying exciting results, although no flexibility seems to be given to the user on the characters to focus on, and no global clustering visualization is provided. These works used the pose estimator OpenPose [4].

The present paper goes in that direction as well. It also joins the current efforts of visual semiotics in studying temporality and rhythm in still images [6]. According to Warburg, the fabrication of an image may be considered as a potential energy storage that may be activated and discharged by other images in succeeding periods, resulting in the revival of motifs or patterns [6]. In a similar vein, this paper aims at tracing poses and providing specialists with qualitative and quantitative tools to analyze gestures, and other kinds of movement and dynamics of forces in still images such as paintings and photographs.

Let us note that, instead of separating the method from the experimental results, we provide and discuss results over the course of the explanation of our method, as we believe that this benefits the reader, given the rather long story that we aim to tell. We start by detailing the pose estimation algorithm used in this paper, then we focus on a qualitative similarity retrieval based on the PixPlot software, before moving on to a more quantitative approach, first for individual poses, then for groups of poses. This latter part presents many challenges that we explain, but lies at the heart of the novelties of this paper.

Contributions. All in all, the main contributions of this paper are the following: 1) A technique to extract and normalize individual character poses from paintings, allowing for scale and position-(in)dependent comparisons, 2) the identification of 7 configurable criteria that enable unique ways of comparing groups of poses across paintings, including analyzing the configuration, shape, scale, and relative positions of the poses, 3) the development of a ranking system that aggregates these 7 criteria into a single score, enabling the retrieval of the most similar paintings for a given query image and set of poses of interest.

2 Pose estimation

Model used. In this work, we use the MMPose library [22] with the RTMPose model [13] to estimate the pose of characters in paintings, which we found sufficiently accurate and fast in our preliminary experiments.

Skeletons and keypoints. As any pose estimation model, RTMPose considers that a human skeleton is composed of a collection of keypoints, localized at the joints between limbs or at salient important body parts. RTMPose produces skeletons articulated around 17 keypoints, which it aims to detect: one keypoint for the nose, then two (left and right) keypoints for the eyes, ears, shoulders, elbows, wrists, hips, knees, ankles. Only for visual representations, these keypoints are linked together as appropriate to produce a proper skeleton-looking figure: ankles with knees, knees with hips, hips together, hips with shoulders, shoulders together, shoulders with elbows, elbows with wrists, shoulders with ears, ears with eyes, eyes together, and eyes with nose.

A two-stage operation: human detection then keypoint detection. RTMPose is composed of two modules. Given an image to analyze, the first module aims at detecting the characters depicted on the image, by providing a tight bounding box around each of them separately. Then, the image is cropped along each bounding box, and each crop is passed to the second module. The second module aims at detecting the coordinates of each of the 17 keypoints of the single character represented in the crop. These coordinates are then transformed back to coordinates relative to the original (not the cropped) image.

Confidence scores for keypoints. The second module always outputs estimated coordinates for the 17 keypoints that compose a human skeleton, even if some of them are not directly visible on the image (occlusion, close-up portrait, *etc.*), in which case the model provides its best approximation while trying to respect human body proportions. To indicate the confidence that the model has in its predictions, it also outputs a confidence score for each keypoint. This way, the model can indicate when uncertainty arises by providing low confidence scores to keypoints that are presumably not accurately estimated. For the model used in this work, we use the recommended default value stating that keypoints with a confidence score above 30% are sufficiently reliably detected.

What if no human is present? In the case of an image where no character is represented, the first module will provide a single bounding box that corresponds to the whole image itself. The second module then still outputs a single set of 17 keypoints, but all of them very likely have a low confidence score. If there are characters on the image but the first module misses them all, then the whole image is once again provided to the second module, which may or may not output keypoints actually belonging to one or several of the previously missed characters, with potentially varying degrees of confidence.

Images analyzed. For this study, in a first phase, we consider images from "religious paintings" of WikiArt[3]. We downloaded the 11,980 images of this category available when we started this work. We ran MMPose on each of them to extract human poses. We found that 5,269 images contain at least one pose whose 17 keypoints have a sufficiently high confidence score, and there are a total of 8,599 such individual poses present on these images. These images and these individual poses now constitute our corpus of interest.

3 Qualitative similarity retrieval with PixPlot

PixPlot. The first way to perform the similarity retrieval in paintings is to do it visually, from an educated representation of the dataset at hand, where images deemed similar according to the criteria of interest are clustered together, far away from images deemed dissimilar regarding said criteria. A suitable tool for this task is PixPlot[4]. Given a collection of images, PixPlot computes an embedding per image (that is, a numerical representation of the image as a list of *e.g.* 2,048 values) with the neural network InceptionV3 [24] trained by deep learning on the generic dataset ImageNet [5]. Then, PixPlot uses the UMAP algorithm [21] to project all these embeddings in a 2-dimensional plane, by trying to maintain as much as possible the distances between the embeddings, that is,

[3] `https://www.wikiart.org/`

[4] `https://dhlab.yale.edu/projects/pixplot/`

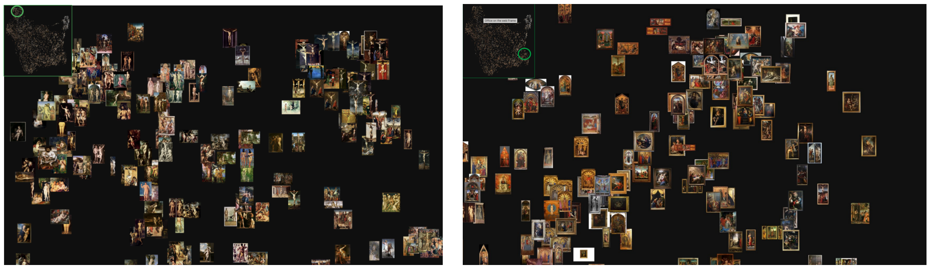

Fig. 1: Examples of clusters on the default PixPlot (top left), with apparently naked bodies and Jesus on Cross (left), and frames around the paintings (right).

embeddings that are far away (respectively close) initially should remain distant (respectively close) from each other in the plane. Finally, in a web browser, PixPlot places each initial image at its corresponding position in the plane, and the browser's functionalities allow navigating through this large meta-image.

Lack of interpretability. Applying the default PixPlot to our corpus of images is interesting but yields hardly interpretable results (see that PixPlot here: `https://bit.ly/3xyiJN3` and a static image in supplementary material). Indeed, it is not always possible to "guess" why some images are located close to each other. It might be because of some elements of the content of the images (bodies, long clothes,*etc.*), the color palette of the images, the presence of particular shapes, or some combinations of multiple factors that are not easy to explain. For example, in a corpus of religious paintings, it might be observed that images of Jesus on his cross are scattered across the meta-image (some of them are clustered though), depending on the other elements on the images, as shown in Fig. 1. Given the limitations of this approach, we would like to have the possibility to focus on one modality only; in our case, the poses of the characters.

Translation to skeleton images. The natural step to focus on poses only is to transform our images into skeleton images, *i.e.* blacked out images where characters are replaced by their skeletons. However, applying the default PixPlot on a corpus of such skeleton images is moderately interesting, because, unsurprisingly, PixPlot can only differentiate images by roughly their number of colored pixels. Therefore, images are clustered more or less according to their number of characters (see this Pixplot `https://bit.ly/3RO13DR` -increase image sizes in the settings- and a static image in supplementary material). The exact poses, understood as articulations of the skeletons, barely play a role in this representation. Also, the result depends on the choice made for the color representation of the skeletons and the thickness of the lines, which is a useless dependency that does not correspond to any bodily resemblance.

Separation into individual skeleton images. As a consequence of the previous observation, we decided to analyze poses individually, *i.e.* each pose of the corpus is represented in its own image. Nevertheless, to make poses comparable to each other, we need to get rid of the position of the character in the image, and of its size. Therefore, each skeleton is drawn in a square image of 512×512

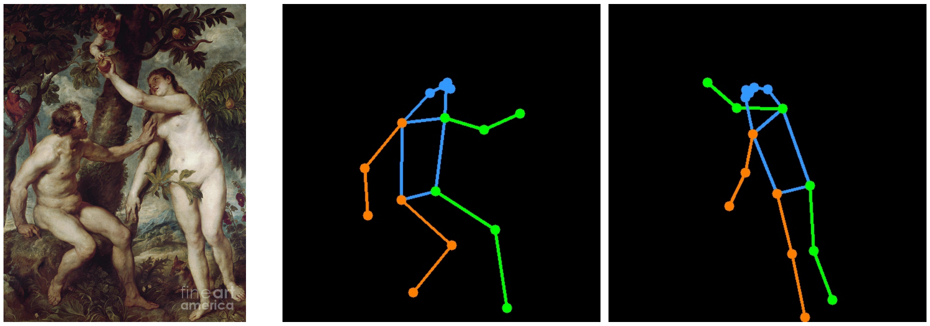

Fig. 2: Individual poses extracted, normalized and plotted in their own image.

pixels, such that the center of gravity computed across the keypoints is located at the center of the image. To achieve scale-independence, the skeleton is enlarged or compressed such that the largest distance between the center of the image and the keypoints equals 256 pixels. This forces to inscribe the pose within a circle centered at the center of the image, and of radius of 256 pixels. This is examplified in Fig. 2. This gives yet another corpus of images that can be passed to PixPlot, which yields interesting results that will now tend to cluster similar poses together (see this PixPlot `https://bit.ly/4cly42r` and a static image in supplementary material). However, we still noticed a lot of variability that can be observed within poses placed close to each other. This might be due to some artefacts present in the embeddings, over which we have so far no control. Indeed, these embeddings are still computed with a pretrained model applied to our pose images. This means that the keypoints coordinates that we have at our disposal, which constitute a rich and accurate source of information, are not used explicitly in the current process. This is the change that we will make next.

Customizing the pose distance for individual skeletons. The individual pose images are just a visual representation of the keypoints extracted by the model. These images use specific colors and trait thickness that influence how the neural network sees the image and generates the embeddings. So, instead, let us consider the list of normalized keypoints as the embeddings themselves, so that we remove the non-interpretable neural network from the equation. Our embedding dimension is thus $2 \times 17 = 34$ instead of $2,048$. Besides, we need to circumvent the default distance metric used by PixPlot to cluster similar images together. The default metric, namely the cosine similarity, is well-suited for embeddings that originate from neural networks. In our case, we decided to use a metric directly related to our keypoints. Given two sets of 17 keypoints, belonging to two characters, we compute their distance as the sum of the euclidean distances between the pairs of corresponding keypoints (distance between noses + distance between left eyes + *etc.*). Re-wiring PixPlot (the UMAP part actually) with these two modifications (keypoints as embeddings and pose distance), makes it produce a new meta-image where the focus is completely on the pose itself, as shown in this PixPlot `https://bit.ly/3VLLm1m` and in Fig. 3.

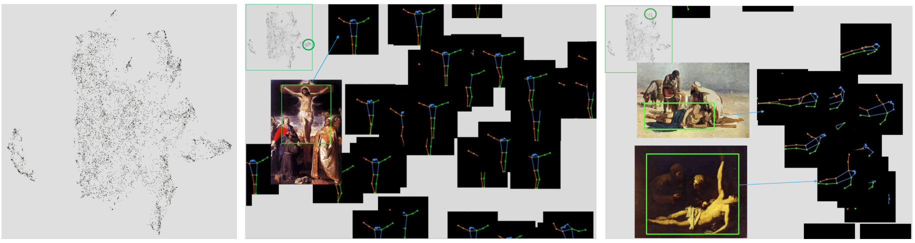

Fig. 3: Pixplot of the individual poses and examples of clusters retrieved visually.

Analysis of the pose clusters. An in-depth analysis of this meta image reveals interesting clusters of similar poses, as desired. The center of the PixPlot tends to regroup poses that are relatively neutral, depicting a person standing, facing the viewer. As we move away from the center, the poses continuously vary, reaching completely different poses in the corners of the meta-image, such as characters lying down, sitting, falling, *etc.* A cluster of Jesus on his cross is also visible, which is a common pose among religious paintings. We can also spot a cluster where the character is seen from the back, which is completely, and rightfully, dissociated from the rest of images. Let us also note that a body lying down with the head on the left is a completely different pose (according to the metric used) than if the head is on the right. Such opposite poses are also placed in opposite parts of the meta-image, namely top and bottom in this case. Finally, as for every large-scale automated analysis, some unfiltered errors sneaked through this visualization; in this case as a small cluster of poses with legs cut at the knees, corresponding to characters that are not completely shown on the images.

Application: Cross-domain comparison and retrieval. An interesting side application that can be derived from our method is the retrieval of similar poses in very different domains. For example, in a first attempt to compare modern fashion photography and religious paintings in terms of poses, we downloaded Artsy's fashion photography catalog[5] (12000 images, 5000 complete individual poses) and produced a PixPlot of its poses, which can be visualized here `https://bit.ly/3RNZaXM`. For an easier comparison, we produced a PixPlot combining Artsy's poses and WikiArt's religious paintings poses here `https://bit.ly/4eE9Im9`. The user can toggle between both corpuses with the pre-established clusters on the left. Overall, it seems that Artsy's poses are more present at the edge of the PixPlot, with more acrobatic/artistic/extreme poses than regular standing characters, more characteristic of religious paintings and more present in the center of the PixPlot. Another preliminary interesting result is the presence of a couple of Artsy images in the cluster of Jesus on his Cross produced by religious paintings, at the very top of the representation, as shown in Fig. 4. Searching for a specific pose of a corpus in another corpus would be difficult without our method, given the large number of images to browse.

[5] `https://www.artsy.net/gene/fashion-photography`

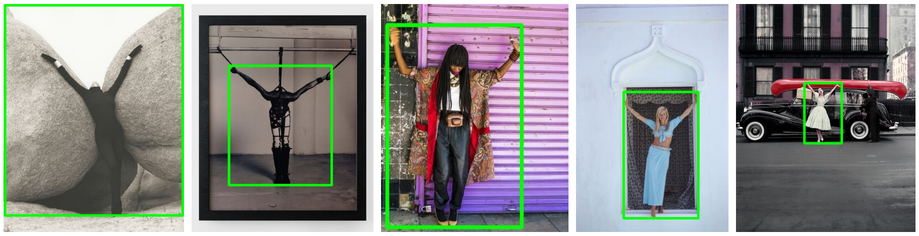

Fig. 4: Artsy images displaying characters with “Jesus on his cross”-like poses.

4 Quantitative similarity retrieval: individual poses

Similar single pose retrieval with respect to a query image. PixPlot is a nice visualization tool, useful for exploring a dataset and figuring out its underlying structure, with respect to either generic or specific features. However, in its current state, it does not allow one to submit a query image, which might or might not belong to the corpus, and retrieve the images of the corpus that are the closest to it, again according to generic or specified features. Besides, even for a query image from the corpus, its neighbors in the PixPlot visualization might not be exactly the ones that are the closest in term of computed distance, and their ranking is uncertain, which is not convenient as it might be hard to grasp a sense of which images are the closest in densely populated areas of the meta-image. There is indeed a loss of information when projecting high-dimensional embeddings into a 2-dimensional plane, which results in possibly flawed interpretations of what a close neighbor is when looking at the PixPlot image only. As a consequence of these observations, we also developed a protocol to define a query image and a pose of interest of a character in this query image, which outputs the list of closest poses among the corpus. This is our second way of performing the similarity retrieval in paintings, in a more quantitative way.

Filtering images to rank. The most direct way of retrieving similar poses to a query pose is to compute the distances between the poses of the corpus and said query pose, and to output the poses sorted by increasing distance, as shown in the top row if Fig. 5 for the query image *Bergström over Paris* of Helmut Newton. Nevertheless, some retrieved images are not satisfactory. For instance, the first one, which is estimated as the closest religious image to Newton’s in term of pose of the main woman, displays a very small character. The next two images barely represent anything and accidentally end up being highly ranked. Only a couple of images are relevant: those where a human body is lying down to the right (which is often the Christ). To remove such unwanted results, we set up 4 filtering parameters on the box and keypoints features:

1. Valid box threshold: threshold above which the bounding box confidence score of the first module of RTMPose should be to accept the box as a valid candidate. We simply select it as 0, such that images are discarded from the search only when no human character is detected on the images (this eliminates *e.g.* the second-ranked image in the Figure).

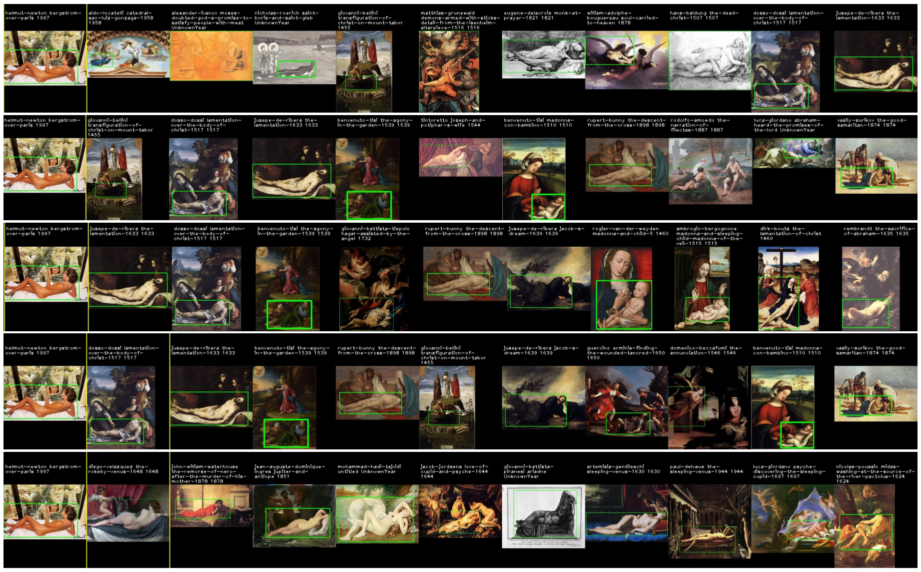
helmut-newton bergstrom-over-paris 1997
aldo-locatelli catedral-sao-luiz-gonzaga-1958 1958
alexander-ivanov moses-doubted-god-s-promise-to-satisfy-people-with-meat UnknownYear
nicholas-roerich saint-boris-and-saint-gleb UnknownYear
giovanni-bellini transfiguration-of-christ-on-mount-tabor 1455
matthias-grunewald demons-armed-with-sticks-detail-from-the-isenheim-altarpiece-1516 1516
eugene-delacroix monk-at-prayer-1821 1821
william-adolphe-bouguereau soul-carried-to-heaven 1878
hans-baldung the-dead-christ-1507 1507
dosso-dossi lamentation-over-the-body-of-christ-1517 1517
jusepe-de-ribera the-lamentation-1633 1633
helmut-newton bergstrom-over-paris 1997
giovanni-bellini transfiguration-of-christ-on-mount-tabor 1455
dosso-dossi lamentation-over-the-body-of-christ-1517 1517
jusepe-de-ribera the-lamentation-1633 1633
benvenuto-tisi the-agony-in-the-garden-1539 1539
tintoretto joseph-and-potiphar-s-wife 1544
benvenuto-tisi madonna-con-bambino-1510 1510
rupert-bunny the-descent-from-the-cross-1898 1898
rodolfo-amoedo the-narration-of-filectas-1887 1887
luca-giordano abraham-heard-the-promises-of-the-lord UnknownYear
vasily-surikov the-good-samaritan-1874 1874
helmut-newton bergstrom-over-paris 1997
jusepe-de-ribera the-lamentation-1633 1633
dosso-dossi lamentation-over-the-body-of-christ-1517 1517
benvenuto-tisi the-agony-in-the-garden-1539 1539
giovanni-battista-tiepolo hagar-assisted-by-the-angel 1732
rupert-bunny the-descent-from-the-cross-1898 1898
jusepe-de-ribera jacob-s-dream-1639 1639
rogier-van-der-weyden madonna-and-child-5 1460
ambrogio-bergognone madonna-and-sleeping-child-madonna-of-the-veil-1515 1515
dirk-bouts the-lamentation-of-christ 1460
rembrandt the-sacrifice-of-abraham-1635 1635
helmut-newton bergstrom-over-paris 1997
dosso-dossi lamentation-over-the-body-of-christ-1517 1517
jusepe-de-ribera the-lamentation-1633 1633
benvenuto-tisi the-agony-in-the-garden-1539 1539
rupert-bunny the-descent-from-the-cross-1898 1898
giovanni-bellini transfiguration-of-christ-on-mount-tabor 1455
jusepe-de-ribera jacob-s-dream-1639 1639
guercino erminia-finding-the-wounded-tancred-1650 1650
domenico-beccafumi the-annunciation-1546 1546
benvenuto-tisi madonna-con-bambino-1510 1510
vasily-surikov the-good-samaritan-1874 1874
helmut-newton bergstrom-over-paris 1997
diego-velazquez the-rokeby-venus-1648 1648
john-william-waterhouse the-remorse-of-nero-after-the-murder-of-his-mother-1878 1878
jean-auguste-dominique-ingres jupiter-and-antiope 1851
mohammad-hadi-tajvidi untitled UnknownYear
jacob-jordaens love-of-cupid-and-psyche-1644 1644
giovanni-battista-piranesi ariadne UnknownYear
artemisia-gentileschi sleeping-venus-1630 1630
paul-delvaux the-sleeping-venus-1944 1944
luca-giordano psyche-discovering-the-sleeping-cupid-1697 1697
nicolas-poussin midas-washing-at-the-source-of-the-river-pactolus-1624 1624

Fig. 5: Example of retrieval results among the corpus of religious paintings for a query image (left). The poses compared are indicated by green boxes. Top row: unfiltered nearest religious paintings in term of estimated pose. Second row: filtered nearest paintings in term of scale-independent estimated pose. Third row: filtered nearest paintings in term of scale-dependent estimated pose. Fourth row: filtered nearest paintings in term of average combination of scale-dependent and scale-independent rankings. Fifth row: same as fourth, but retrieval performed among mythological paintings.

2. Area threshold: threshold above which the area of the bounding box relatively to the image size must be to accept the box as a valid candidate. We set it to 0.05, which means that characters occupying less than 5% of the image are considered too small to be interesting enough in the search for similar poses (this eliminates *e.g.* the top ranked image in the Figure).
3. Keypoint confidence threshold: threshold above which the confidence score of the keypoints produced by the second module of RTMPose are considered valid. As done for the PixPlot visualizations, we set this threshold to 0.3.
4. Number of valid keypoints: a pose is valid when it has at least a certain number of valid keypoints. For instance, to allow one uncertain keypoint, we set this value to 16 (this removes *e.g.* the third-ranked image in the Figure).

As a result, the updated list of retrieved images is represented in the second row of Fig. 5, which is more satisfying than the previous one. We still compare normalized scale-independent poses, which means that only the pose itself matters, regardless of its size (provided that is is large enough to fulfill the second filtering). If we do not normalize the poses and want to keep its initial relative size with respect to the image into account, we obtain the ranking in the third row of Fig. 5. The top result of the second row is not displayed anymore, but

new images that might not be relevant are now well-ranked, such as Van der Meyden's *Madonna and Child*, because the Child's occupancy of the space is very similar to Bergstrom's, and the poses are somewhat "similar".
Combining multiple rankings. As each image of the retrieval corpus has two rankings (where none of them is right or wrong, it all depends on the researcher's criteria in the search for similarities), one for the scale-independent retrieval, the other for the scale-dependent one, we can also average those rankings to get the best of both worlds and to obtain a balanced ranking, that takes the scale of the pose into account but that still leaves room for poses of very different sizes if they are similar enough to the query pose. This is represented in the fourth row of Fig. 5 and might well be the most relevant ranking of all, in this case. Let us note that, in none of these considerations, the position of the pose within the image was taken into account. In other words, the poses are compared as if they were all centered, even if they are actually located in different parts of their respective images. We noticed that taking the localization into account may worsen the rankings by favoring too much poses that are located almost at the same spot as the query pose while completely disregarding the pose itself, which could be a person standing, sitting, praying, *etc*. Hence, our rankings can be qualified as translation-invariant. Finally, it is well-known that *Bergström over Paris* is inspired by Velasquez's *The Rokeby Venus*, which is a mythological painting. Therefore, looking for similar images in religious painting might not be an appropriate choice. When looking after WikiArt's mythological paintings (whose PixPlot solo skeletonized visualization can be found here `https://bit.ly/3L1qPAX`), we have the result displayed in the fifth row of Fig. 5 (after filtering and combination of scale-dependent and independent rankings). We can see that Velasquez's painting appears first in the ranking, and that the retrieved images better align with the query image in term of pose.

5 Quantitative similarity retrieval: multiple poses

From individual poses to groups of poses. When comparing images depicting multiple characters, it might be interesting to take into account all the poses together, rather than individual poses, such that images are considered close to each other when human bodies form the same shapes globally on the image. In the same spirit, in some cases, some characters are just part of the crowd and do not play any specific role, thus might not need to be considered in this group-level analysis, and we might want to focus only on a sub-group of characters. Therefore, we also worked on the comparison of groups of poses, which presents some extra difficulties, as explained hereafter.
Combinatorial explosion. The first obstacle, not the least, is the combinatorial explosion that arises when many characters are represented on a painting. Indeed, if N characters (at least two) are detected, then the total number S of different subgroups of at least two characters that can be formed from these N detections is $S = 2^N - N - 1$. While this remains manageable for small values of N, *e.g.* for $N = 2$ then $S = 1$, for $N = 3$ then $S = 4$, for $N = 4$ then

$S = 11$, this number grows exponentially with N (it roughly doubles for each extra character detected). Thus, with values of N that are not so large and not so rare in paintings, for example $N = 10$ characters, we can form $S = 1,013$ distinct subgroups of at least two characters, most of them being presumably not so much interesting to examine. For that purpose, in the following, when an image has more than 3 characters, we only keep the 3 most confidently detected ones, determined via their bounding box confidence score.

Various ways to compare two groups of poses. Another difficulty that arises when comparing multiple poses is answering the question: what do we compare? First, we need to compare groups (or subgroups, but we will write groups in the following) with the same number of characters. So, each image of interest yields groups of poses, and we can analyze altogether all the groups of two characters, three, four, etc., but we cannot compare a group of two characters with a group of three. So, let us assume that we have two groups of characters from different paintings to compare. How do we compare them? We identified 7 criteria that allow each a unique way of comparing the poses. In detail, 3 of them focus on the configuration of the group of poses: the poses are reduced to their average point (*i.e.* center of gravity), these average points altogether give the configuration of the group of poses. We analyze these configurations by looking at (1) its average relative position in the image, (2) its shape (*e.g.* triangle pointing upwards) independently of its scale, that is, for example, an upwards equilateral triangle of side 2 and another one of side 5 are both upwards equilateral triangles and are thus equivalent for the analysis, (3) its shape dependently of its scale, such that the two upwards equilateral triangles are not considered equivalent anymore. The next 4 criteria focus on the poses themselves and not on the configuration of the whole group of poses. Again, we can either keep (4)(5) or remove (6)(7) the dependency to the scale of the poses, by letting them as is or normalizing them, just as discussed for the configuration of the group of poses. Besides, when comparing two groups of equally-numbered poses, we must decide which pose of the first group is compared with which pose of the second group. The matching between the poses of the two groups can be done either by matching the poses by their appearance (which pose of the second group is the most similar, as in the individual poses analysis, to which pose in the first group?), or by their localization (which one of the second group is the closest to which one of the first group in term of relative position in the image?). We name the first matching "best pose matching" (4)(6) and the second one "location-based matching" (5)(7). Hence, our 4 pose-based criteria for comparing groups of poses are defined by the 2×2 combinations of scale dependence/independence $\times$ best pose matching/location-based matching. In summary, the 7 criteria available are:

1. Configuration comparison: localization of the group of poses in the image. The distance between groups is the distance between their localization point.
2. Configuration comparison: shape of the group of poses, scale-dependent. The distance between the groups is the distance between their shapes, computed as the smallest distance obtained by summing the distances between pairs of vertices among all permutations that match a vertex of the first shape with

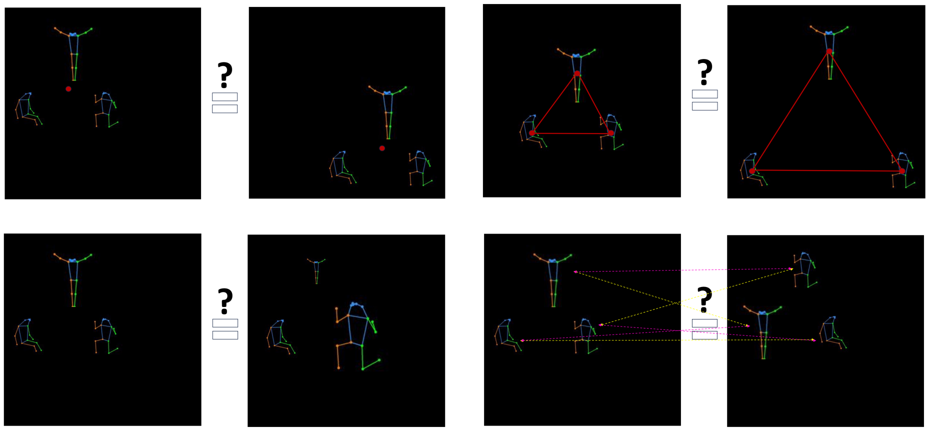

Fig. 6: Criteria for comparing groups of poses. Top left: (1) Configuration comparison: should we take into account the localization (red dot) of the group of poses within the image? Top right: (2) and (3) Configuration comparison: should we take into account (2) or not (3) the scale of the shape formed by the group of poses (in this case the red triangle)? Bottom left: (4), (5), (6), (7) Poses comparison: should we take into account (4, 5) or not (6 ,7) the scale of the individual poses? Bottom right: (4), (5), (6), (7) Poses comparison: should we compare poses that are the most similar (best pose matching, yellow arrows, 4, 6) or that are located at the most similar place in the group configuration (location-based matching, pink arrows, 5, 7)?

a vertex of the other (*i.e.* the sum of the distances of the matched vertices after a location-based matching).

3. Configuration comparison: shape of the group of poses, scale-independent. The distance between the groups is the same as for item 2.
4. Poses comparison: scale-dependent, best pose matching. The distance between groups is the sum of the distances between matched poses, each computed as in the individual case (sum of distances of corresponding keypoints).
5. Poses comparison: scale-dependent, location-based matching. The distance between the groups is the same as for item 4.
6. Poses comparison: scale-independent, best pose matching. The distance between the groups is the same as for item 4.
7. Poses comparison: scale-independent, location-based matching. The distance between the groups is the same as for item 4.

These choices are embodied in Fig. 6. Let us note that this naturally extends to the case of individual poses comparisons, but criteria 2 and 3 then become irrelevant since there is no such thing as “shape of the group of poses”, and the matching does not matter anymore since there is only one pose to compare with another one, thus criteria 4 and 5 are equivalent, as well as 6 and 7. Thus, one needs to compute only criteria 1, 4, 6 in the case of individual poses, as already discussed and implemented implicitly previously.

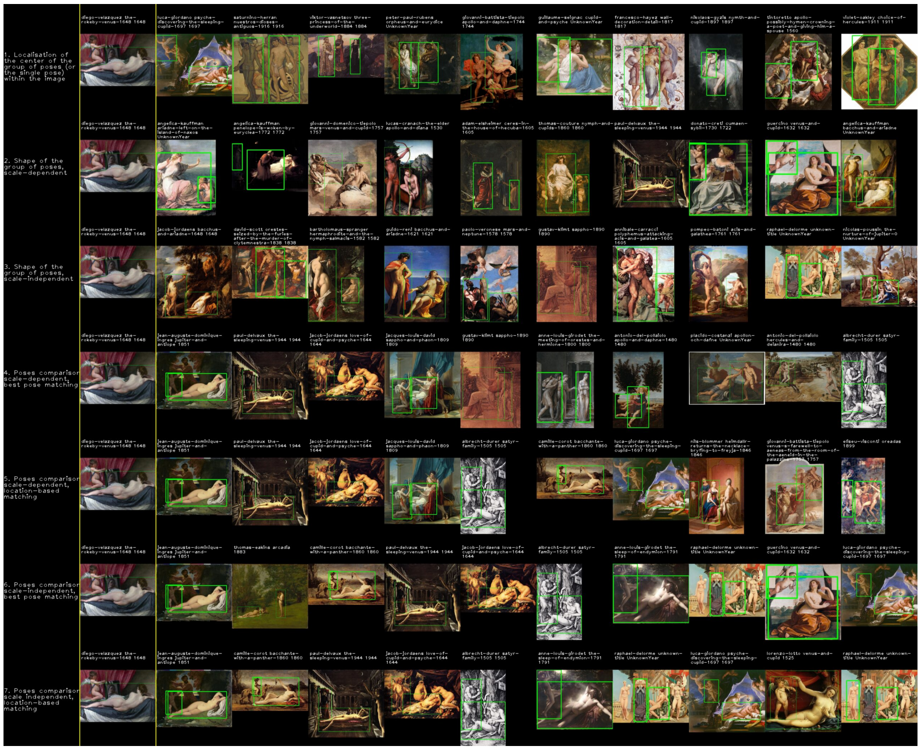
1. Localisation of the center of the group of poses (or the single pose) within the image
2. Shape of the group of poses, scale-dependent
3. Shape of the group of poses, scale-independent
4. Poses comparison scale-dependent, best pose matching
5. Poses comparison scale-dependent, location-based matching
6. Poses comparison scale-independent, best pose matching
7. Poses comparison scale independent, location-based matching

Fig. 7: Retrieval result of Diego Velasquez's *The Rokeby Venus* according to the 7 rankings for the pair composed of Venus and the Angel.

Too many rankings? As soon as at least two poses are considered, the 7 types of criteria can be used to rank the retrieval corpus with respect to the query image, which will amount to a lot of rankings. For example, let us consider a query image with only 3 characters and let us list the rankings that can be computed. We can consider individual query poses, thus 3 of them, and for each, we can compute a ranking for criteria 1, 4, 6, which makes already 9 rankings. Then, we can consider the 3 pairs of poses. For each of them, we can compute the 7 rankings, which makes 21 additional rankings. Finally, we can consider the whole group of the 3 poses, for which we can compute 7 extra rankings. In total, we can thus compute 9+21+7 = 37 rankings for this image only (which contains only 3 detected characters). This already gives a hint on the difficulty to retrieve very similar images when many characters are present: the combinatorial explosion and the choice of what to rank give plethora of results, that need to be further refined by combining these rankings to compress all that information in a unified view. We show an example of the 7 rankings for the pair Venus-Angel for Velasquez's *The Rokeby Venus* as query image in Fig. 7.

Side note. Technically speaking, we can compute a PixPlot visualization for each set of images containing a subgroup of any given number of poses, for each of the 7 criteria. That is, there can be 7 PixPlots for the set of images focusing on subgroups of 2 characters, then 7 again for the set of images focusing on subgroups of 3 characters, 7 again for 4 characters, *etc*. This becomes quickly hard to track and analyze, albeit being technically feasible, if the number of images remains limited (as explained before, a combinatorial explosion makes a complete analysis challenging). In our case, we did not compute all those PixPlots. Moreover, it goes without saying that, generally speaking, the bigger the group of characters considered, the larger the variability in the representations, which means that finding very similar images becomes increasingly difficult, if not irrelevant. On the PixPlot visualizations, this materializes by scattered meta-images, where it becomes tricky to determine clusters (if there are any clusters at all).

Combining criteria with image rankings for query/retrieval, not with distances. For a given number of characters of interest in the subgroups of poses, it might be worth combining the 7 criteria into one single summarizing criterion that measures some kind of "global distance" between groups. This combination can hardly be done at the distances-level, as we might risk comparing apples and oranges: the distance between configurations of poses is not at the same scale as the distance between poses themselves. To mitigate this issue, let us again consider the task of having a query image, with some number of characters detected, and retrieving its most similar-looking images, in terms of groups of poses. Each image in the retrieval dataset can be compared with the query image according to each of the 7 criteria, and thus has a ranking among the images of the dataset with respect to these criteria. Contrary to distances, which were assimilated to apples and oranges, rankings can be combined, as they are all "rankings" in their respective categories, thus they have the same range and scale. Consequently, we can aggregate the rankings of each image in the dataset to obtain, for each such image, a unique "score", representing some kind of average ranking, with respect to the query image. Finally, we can output the images that are the closest to the query image based on this aggregated metric.

How to aggregate rankings? Each criterion can be assigned a weight, and the aggregation of the 7 rankings of an image is simply a weighted sum of its rankings. While many choices can be made, it seemed to us that, in the case of individual poses, a combination of equal weights of criteria 4 and 6 (pose comparisons, with scale-dependence and independence, equivalent to 5 and 7) is appropriate, as already mentioned previously. This prevents putting too much emphasis on poses that look very different but have the same scale as the query image (which acts thus in its favor in the ranking while being not relevant) as well as putting too much emphasis on poses that look very similar to the query pose but are way too different in term of scale/prominence in the image (which may thus convey a different interpretative meaning). Besides, criterion 1 (the position of the pose in the image) does not seem that useful to us, as we believe desirable to consider as similar poses that are the same, with or without the same scale, but potentially located at different positions in the image. In the case of multiple poses, the same motivations regarding the scale of the poses lead us to consider criteria 5 and 7, with a location-based matching preferred over a pose-

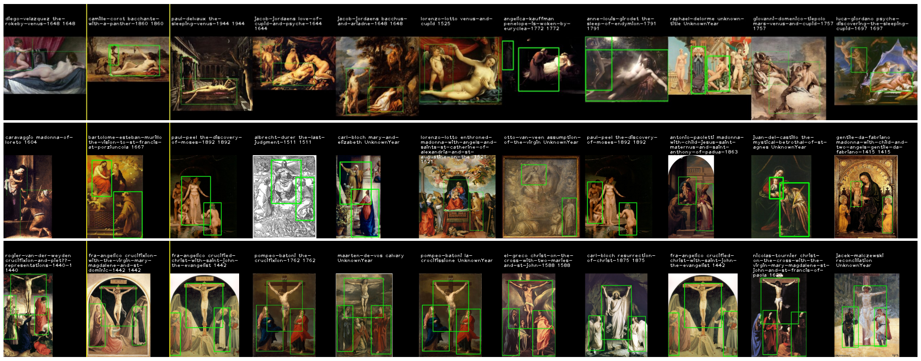
diego-velazquez the-rokeby-venus-1648 1648
camille-corot bacchante-with-a-panther-1860 1860
paul-delvaux the-sleeping-venus-1944 1944
jacob-jordaens love-of-cupid-and-psyche-1644 1644
jacob-jordaens bacchus-and-ariadne-1648 1648
lorenzo-lotto venus-and-cupid 1525
angelica-kauffman penelope-is-woken-by-euryclea-1772 1772
anne-louis-girodet the-sleep-of-endymion-1791 1791
raphael-delorme unknown-title UnknownYear
giovanni-domenico-tiepolo mars-venus-and-cupid-1757 1757
luca-giordano psyche-discovering-the-sleeping-cupid-1697 1697
caravaggio madonna-of-loreto 1604
bartolome-esteban-murillo the-vision-to-st-francis-at-porziuncola 1667
paul-peel the-discovery-of-moses-1892 1892
albrecht-durer the-last-judgment-1511 1511
carl-bloch mary-and-elizabeth UnknownYear
lorenzo-lotto enthroned-madonna-with-angels-and-saints-st-catherine-of-alexandria-and-st-augustine-on-the-1521 1521
otto-van-veen assumption-of-the-virgin UnknownYear
paul-peel the-discovery-of-moses-1892 1892
antonio-paoletti madonna-with-child-jesus-saint-maternus-and-saint-anthony-of-padua-1863
juan-del-castillo the-mystical-betrothal-of-st-agnes UnknownYear
gentile-da-fabriano madonna-with-child-and-two-angels-gentile-da-fabriano-1415 1415
rogier-van-der-weyden crucifixion-and-piet??-representations-1440-1 1440
fra-angelico crucifixion-with-the-virgin-mary-magdalene-and-st-dominic-1442 1442
fra-angelico crucified-christ-with-saint-john-the-evangelist 1442
pompeo-batoni the-crucifixion-1762 1762
maarten-de-vos calvary UnknownYear
pompeo-batoni la-crocifissione UnknownYear
el-greco christ-on-the-cross-with-two-maries-and-st-john-1588 1588
carl-bloch resurrection-of-christ-1875 1875
fra-angelico crucified-christ-with-saint-john-the-evangelist 1442
nicolas-tournier christ-on-the-cross-with-the-virgin-mary-magdalene-st-john-and-st-francis-of-paola 1628
jacek-malczewski reconciliation UnknownYear

Fig. 8: Retrieval results of several query images according to our combination of rankings 2,3,5,7. Groups of 2 (rows 1 and 2) and 3 (row 3) people are considered.

based matching. Indeed, we believe that, if the same poses are permuted within the pose configuration, this yields a different image, which is not captured by the pose-based matching. We also consider criteria 2 and 3 in the mix, to incorporate the shape aspect of the configuration in the final comparison, with both of them equally weighted for similar reasons as those motivating the choice of 5 and 7. We neglect criterion 1 again for the same reason as previously. For *The Rokeby Venus*, this aggregation for the pair Venus-Angel gives the result displayed at the first row in Fig. 8. The combined ranking for the group of 3 poses (Venus, Angel, and Venus' reflection in the mirror) as well as for Venus alone and Angel alone are given in supplementary material. It looks that most of these rankings make sense to some extent, and that the proposed methodology could prove helpful for extensive analyses of large image corpuses. This is also supported by the additional results provided the second and third rows in Fig. 8.

6 Conclusion

We proposed a methodology for retrieving similar paintings based on character poses which shows promising results on a corpus of religious and mythological artworks. By extracting and normalizing individual poses, defining multiple comparison criteria, and integrating them into a unified ranking system, the approach allows for an effective identification of qualitatively and quantitatively similar paintings. While the combinatorial explosion of possible pose configurations presents challenges for large-scale analysis, the techniques developed in this work demonstrate the potential for data-driven and computer vision-based exploration and comparison of art collections. Further research is needed to extend the applicability of these methods to broader artistic domains and investigate their utility for digital humanities and art history.

Acknowledgments

A. Deliege is funded by the F.R.S.-FNRS research project T.0065.22 - R.FNRS.5453 entitled “Pour une généalogie des formes visuelles. Approches sémiotique et computationnelle de larges collections d’images — Towards a Genealogy of Visual Forms: Semiotic and Computer-Assisted Approaches to Large Image Collections” of M.G. Dondero, who is funded permanently by the F.R.S.-FNRS. Besides, both authors of the paper are affiliated with the University of Liège, Belgium.

References

1. Bai, Z., Nakashima, Y., Garcia, N.: Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation. In: ICCV (2021)
2. Bernasconi, V., Cetinic, E., Impett, L.: A Computational Approach to Hand Pose Recognition in Early Modern Paintings. Journal of Imaging **9**(6) (2023)
3. Brey, A.: Digital art history in 2021. History Compass **19**(8) (2021)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In: CVPR (2017)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
6. Dondero, M.G.: The Language of Images. Springer (2020)
7. Garcia, N., Renoust, B., Nakashima, Y.: Context-Aware Embeddings for Automatic Art Analysis. In: International Conference on Multimedia Retrieval (2019)
8. Garcia, N., Vogiatzis, G.: How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval. In: ECCVW (2018)
9. Garcia, N., Ye, C., Liu, Z., Hu, Q., Otani, M., Chu, C., Nakashima, Y., Mitamura, T.: A Dataset and Baselines for Visual Question Answering on Art. In: ECCVW (2020)
10. Impett, L.: The Routledge Companion to Digital Humanities and Art History, chap. Analyzing gesture in digital art history. Routledge (2020)
11. Impett, L., Moretti, F.: Totentanz : Operationalizing Aby Warburg’s Pathosformeln. New Left Review **107**, 68–97 (2017)
12. Jenicek, T., Chum, O.: Linking Art through Human Poses. In: International Conference on Document Analysis and Recognition (ICDAR) (2019)
13. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. arXiv (2023)
14. Ju, X., Zeng, A., Wang, J., Xu, Q., Zhang, L.: Human-Art: A Versatile Human-Centric Dataset Bridging Natural and Artificial Scenes. In: CVPR (2023)
15. Kutrzyński, M., Król, D.: Deep learning-based human pose estimation towards artworks classification. Journal of Information and Telecommunication (2024)
16. Lang, S., Ommer, B.: Transforming Information Into Knowledge: How Computational Methods Reshape Art History. Digital Humanities Quarterly **15** (2021)
17. Madhu, P., Marquart, T., Kosti, R., Bell, P., Maier, A., Christlein, V.: Understanding Compositional Structures in Art Historical Images Using Pose and Gaze Priors. In: ECCVW (2020)
18. Manovich, L.: Data Science and Digital Art History. International Journal for Digital Art History **1** (2015)
19. Manovich, L. (ed.): Cultural Analytics. MIT Press (2020)

20. Masclef, T., Scuturici, M., Bertin, B., Barrellon, V., Scuturici, V.M., Miguet, S.: A Deep Learning Approach for Painting Retrieval Based on Genre Similarity. In: Image Analysis and Processing - ICIAP 2023 Workshops (2023)
21. McInnes, L., Healy, J., Saul, N., Großberger, L.: UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software **3**(29) (2018)
22. MMPose Contributors: OpenMMLab Pose Estimation Toolbox and Benchmark. `https://github.com/open-mmlab/mmpose` (2020)
23. Sabatelli, M., Kestemont, M., Daelemans, W., Geurts, P.: Deep Transfer Learning for Art Classification Problems. In: ECCVW (2018)
24. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. CVPR (2016)
25. Ufer, N., Lang, S., Ommer, B.: Object Retrieval and Localization in Large Art Collections Using Deep Multi-style Feature Fusion and Iterative Voting. In: ECCVW (2020)
26. Yemelianenko, T., Tkachenko, I., Masclef, T., Scuturici, M., Miguet, S.: Learning to Rank Approach for Refining Image Retrieval in Visual Arts. In: ICCVW (2023)
27. Zhao, S., Salah, A.A.A., Salah, A.A.: Automatic Analysis of Human Body Representations in Western Art. In: ECCVW (2022)