International Journal of Geo-Information

MDPI

Article

POSE-ID-on—A Novel Framework for Artwork Pose Clustering

Valerio Marsocci [1,*,†] and Lorenzo Lastilla [1,2,†]

[1] Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG), Sapienza University of Rome, 00185 Rome, Italy; lorenzo.lastilla@uniroma1.it
[2] Sapienza School for Advanced Studies, 00161 Rome, Italy
* Correspondence: valerio.marsocci@uniroma1.it
† These authors contributed equally to this work.

Abstract: In this work, we focus our attention on the similarity among works of art based on human poses and the actions they represent, moving from the concept of *Pathosformel* in Aby Warburg. This form of similarity is investigated by performing a pose clustering of the human poses, which are modeled as 2D skeletons and are defined as sets of 14 points connected by limbs. To build a dataset of properly annotated artwork images (that is, including the 2D skeletons of the human figures represented), we relied on one of the most popular, recent, and accurate deep learning frameworks for pose tracking of human figures, namely OpenPose. To measure the similarity between human poses, two alternative distance functions are proposed. Moreover, we developed a modified version of the K-Medians algorithm to cluster similar poses and to find a limited number of poses that are representative of the whole dataset. The proposed approach was also compared to two popular clustering strategies, that is, K-Means and the Nearest Point Algorithm, showing higher robustness to outliers. Finally, we assessed the validity of the proposed framework, which we named POSE-ID-on, in both a qualitative and in a quantitative way by simulating a supervised setting, since we lacked a proper reference for comparison.

Keywords: pose clustering; pose tracking; digital humanities; image retrieval; Warburg

Citation: Marsocci, V.; Lastilla, L. POSE-ID-on—A Novel Framework for Artwork Pose Clustering. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 257. https://doi.org/10.3390/ijgi10040257

Academic Editors: Susana Del Pozo and Wolfgang Kainz

Received: 3 February 2021
Accepted: 5 April 2021
Published: 11 April 2021

1. Introduction

Between the end of the 19th and the beginning of the 20th century, the German art historian Aby Warburg developed the concept of *Pathosformel* to explain the survival of figurative expressions of profound experiences that re-emerge in works of art, even very distantly in time. In other terms, the *Pathosformel* describes the portrayal or communication of emotion, movement, and passion through a repeatable visual paradigm or formula [1]. This concept culminated in the *Bilderatlas Mnemosyne* [2], which consisted of a work-in-progress series of wooden panels covered with black Hessian, on which he pinned clusters of images (photographic reproductions, photos, diagrams and sketches, postcards, and various kinds of printed materials, including advertisements and newspaper clippings) and which he developed in several stages [3]. The relationships between the images on Warburg's *Mnemosyne Atlas* are not chronological; in fact, images belonging to the same cluster deliver similar emotions and make visible how antique imagery disappears and re-emerges at later times in other contexts [4]. This figurative *Atlas* combined images for similarities and relationships, setting up a method that is increasingly recognized as a profitable instrument of analysis, even outside the history of art [5,6], as proven by the diffusion of tools such as *imgs.ai* [7], a dataset-agnostic deep visual search engine for digital art history based on neural network embeddings and approximate k-Nearest Neighbors (k-NN) algorithms to deliver fast search results, even for very large datasets in low-resource environments.

In this work, we move from the concept of *Pathosformel* in Aby Warburg, and, in particular, we focus our attention on the similarity among works of art based on human poses [8]

and the actions they represent, which are strongly related to the emotions that the same artworks can deliver [1,9] (see Figures 1 and 2). This form of similarity is investigated through an innovative methodology, which is called POSE-ID-on (standing for POSE IDentification and recalling the name of the ancient Greek god of the sea), which performs a pose clustering of the human poses—modeled as 2D skeletons, that are defined as sets of 14 points (or joints) connected by limbs.

Figure 1. *Venere Capitolina* and *La nascita di Venere* (detail) by Sandro Botticelli.

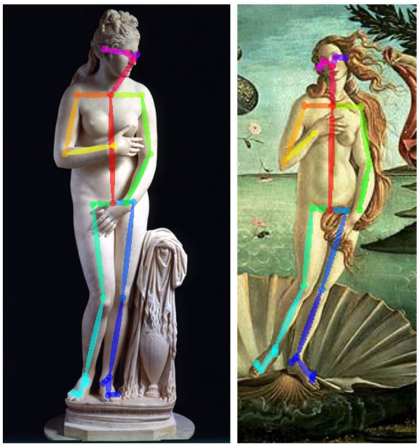

Figure 2. Human poses detected by OpenPose [10] for the two artworks.

Since no annotated artwork datasets are currently available, even if similar approaches have been proposed in the literature [11–13] (they will be discussed more in depth in Section 2), we tested one of the most popular, recent, and accurate deep learning frameworks for pose tracking of human figures, namely OpenPose [10,14], on a dataset exclusively consisting of photographs of artworks collected from the Museum of Classical Art of Sapienza University of Rome (which contains several plaster casts of Classical sculptures) and from the Scala [15] and ArtResource [16] archives. OpenPose was tested without previous fine tuning. As a result, we were both able to assess the efficiency of this pose tracking framework—which was trained on real human pose datasets—with respect to an artwork dataset and to produce a dataset with the required properties for future fine tuning. The 2D poses obtained in this way underwent a preliminary selection process (whose criteria are discussed in Section 4).

Two metrics for pose comparison (described in Section 3) were defined and tested. The first metric is based on a limb-per-limb comparison of the two poses involved in the distance computation. The second one, instead, based on a completely different methodology, relies on the previous computation of a feature vector for each pose; hence,

we measure the pose similarity in the feature space. Both the proposed metrics are invariant to changes in scale, rotation, translation, and horizontal flip.

Having defined a quantitative way to measure the similarity or distance between different poses, we set up a modified version of the K-Medians algorithm to cluster the selected poses and to compute the centroids for each cluster (code available at the following link: https://github.com/L9L4/POSE-ID-on, accessed on 1 February 2021), thus preserving the invariance to shift, rotation, and scale. However, the clustering approach establishes a powerful and innovative solution to the pose retrieval task (having been suggested only by [1,17] in the available literature), both in the formulation and the possibility to generate a limited number of centroids, which would significantly speed up a pose retrieval system.

The lack of external knowledge or a reference to compare to the results of our clustering methodology was faced, essentially, in two different ways:

1. First of all, we carried out a performance evaluation of the clustering approach based on simulated data according to suitable metrics, such as the Adjusted Rand Index (ARI), the Normalized Mutual Information (NMI), and the Adjusted Mutual Information (AMI);
2. Secondly, the obtained clusters and the respective centroids were evaluated in a qualitative way to assess the coherence among the human poses with respect to the Warburgian concept of *Pathosformel*.

Furthermore, we compared K-Medians with two popular clustering algorithms, namely K-Means and a Hierarchical Clustering Algorithm [18] (more precisely, the Nearest Point Algorithm), which are both based on one of the two previously defined metrics. This comparison proved the higher robustness to outliers of the K-Medians approach with respect to the other ones, which is a desirable property when dealing with data characterized by high variability, such as those analyzed in this paper.

The remainder of this paper is divided into the following sections: Section 2 is dedicated to a review of the most relevant works in the literature; in Section 3, we briefly introduce the methodology (which is discussed in more detail in Appendix A) and the different components of the pose clustering pipeline; in Section 4, the datasets for pose tracking and pose clustering are described, the tests performed are briefly outlined, and the main results and the accuracy assessment are presented (with further analyses in Appendix B); in Section 5, the main conclusions and considerations are drawn.

2. Related Works

2.1. Theoretical Background

Any attempts to quantify the relationships between human figures represented in works of art must answer to some fundamental questions: what does it mean to understand a work of art from a quantitative point of view? What are the mechanisms underlying this relationship? Is it possible to rely on the human pose as a quantitative element of comparison?

According to [19], the artistic image understanding process consists of receiving an artistic image as input and producing a set of global, local, and pose annotations; moreover, the same process involves the retrieval of images given a query containing an artistic keyword. Based on this definition, the authors provided a quantitative assessment of several annotation and retrieval algorithms on a dataset of monochromatic artistic images (including the task of pose annotation).

As for the key elements underlying the relationships between works of art, the authors of [20,21] pointed out the mechanism of intericonicity, which is based on the phenomena and processes of imitation, repetition, and similarity, in a dialogic and dynamic path between images and words, between gestures and communication, and between emotion and feeling. The authors of [11], instead, differentiated three types of similarity in visual art (paintings, drawings, prints, and frescoes): the physical link, replication, and composition transfer.

Coming to the third question, the author of [17] made important theoretical and contextual considerations. The first was that the relative positions of the human limbs are a quantitative parameter, and are therefore useful in the digital humanities, even though analyses based on this criterion are rare in the exclusive historical art domain. Therefore, the digitization of big archives and the increasing performances in automatic pose annotation are accelerating this trend. Taking up [22], the author introduced the concept of operationalization, which is "the process of taking a concept from history, literary, theory, etc. and turning it into a sequence of quantitative operations". In [22], this concept was applied precisely on Warburg's *Atlas*. Moreover, the importance of defining a coherent similarity criterion was investigated. To do so, the authors defined the differences between form and formula, arguing that the latter is a form that "has learned to replicate itself", and therefore assumed that the poses represent the latter, like in the Warburg conception. In particular, they asserted that if *Pathos* is expressed by "outward movements", then the body angles could quantify it. Concerning this choice about angles as a similarity criterion, it was asked if they should be measured with respect to the horizontal axis or relative to each other. Looking for this statistical correspondence between *Pathos* and *Formel*, they applied both a K-Means and an Agglomerative Hierarchical Clustering with a non-Euclidean distance on more than 1600 annotated images. The same author of [17], taking up [23], investigated the temporality in gospel stories, focusing on the Madonna and Gabriel poses in the biblical episode of the Annunciation. In this work, the poses moved not through art-historical time, but across narrative time. To capture this different kind of movement, the authors, after a clustering phase, built a temporal network of gestures from the identified poses by using the distances among them.

2.2. Similar Approaches

Some relevant assumptions and intuitions were discussed in [1], which provided one of the first attempts to find meaningful clusters of related poses that concentrate only on the relative angles of limbs based on a dataset of manually annotated 2D human poses (even though human pose search and retrieval is a long-standing issue [24,25]).

A similar approach was also proposed by [8], where a framework for interactive feature-based retrieval and visualization of human statues (modeled as 3D skeletons) was presented. Our first assumption—modeling human poses through 2D skeletons—has the main advantage, with respect to its 3D version, of allowing for the comparison and retrieval of poses from both statues and paintings. The choice of relying on a two-dimensional representation also allows the exploitation of the capabilities of pose tracking deep learning frameworks based on Convolutional Neural Networks (CNNs), such as OpenPose, which have been widely used to obtain reliable local observations of body parts and have significantly boosted the accuracy of body pose estimation [10,26–28].

The choice of OpenPose as a pose detector for artworks, however, is not new. The authors of [13], indeed, relied on this framework to detect the pose keypoints of the protagonists of an image with the aim of performing a pose-based segmentation of the foreground and background of that image and favoring the generation of the global action lines and action regions.

The approach proposed in [11], instead, addressed the discovery of composition transfer in artworks based on their visual content, and specifically based on the similarity of human figures depicted in images (pose retrieval). The first stage of the pipeline consists of the detection of all the human figures represented in an image through OpenPose. Secondly, the poses are matched through a specific mirror-invariant distance function based on the computation of the cosine similarity between poses. Finally, a geometric transformation between the images is performed to simultaneously align all of the figures represented.

Even if the methodology described in [11] shares the initial stage of pose detection through OpenPose with our pipeline, there are several differences that are worth mentioning: first of all, while this approach explicitly aims at addressing the problem of similarity based on composition transfer (thus considering a dataset of paintings with repeating

motifs and/or associated drawings and engravings, which are characterized by strong similarities among each other, and focusing on a limited search space for rotated poses), we instead suggest two alternative distance metrics that are completely independent from rotation, which makes them more flexible with respect to datasets characterized by higher variability. The same distance functions also consider the search for turned poses. Furthermore, and most importantly, our methodology includes the possibility of defining a feature vector for each pose, clustering a set of poses based on this new representation, and, finally, estimating the centroid (archetypal pose) of each cluster, which potentially allows for speeding up the pose retrieval task, especially for big datasets.

Finally, the workflow presented in [12] paved the way toward further progress in the field of human pose estimation in artwork collections by improving the performances of the existing pose tracking methods in generalizing across different domains and styles. As for the pose retrieval stage of their pipeline, they relied on the Object Keypoint Similarity (OKS) evaluation metric.

3. Methodology

3.1. OpenPose

OpenPose [10] is a real-time multi-person system for jointly detecting human body, hand, facial, and foot keypoints (in total, 135 keypoints). It was proposed by the Perceptual Computing Lab of Carnegie Mellon University, and it was released in the form of Python code, a C++ implementation, and a Unity Plugin. For the purposes of our study, it was applied on artwork images for single-person pose extraction.

OpenPose is a deep learning algorithm composed of several stages. In the first step, the image is passed through a CNN. To extract the feature maps, the first ten layers of the VGG-19 architecture are chosen. The feature maps are then processed in a multi-stage CNN pipeline to generate Confidence Maps (CMs) for Part Detection and Part Affinity Fields (PAFs) for Part Association. Part Affinity Field $\mathbf{L_c}$ is a set of 2D vector fields that encodes the location and orientation of the joints of different people in the image. It encodes the data in the form of pairwise connections between body parts. A Confidence Map $\mathbf{S_j}$ is a 2D representation of the belief that a particular body part can be located in any given pixel. Figure 3 shows the multi-stage CNN steps.

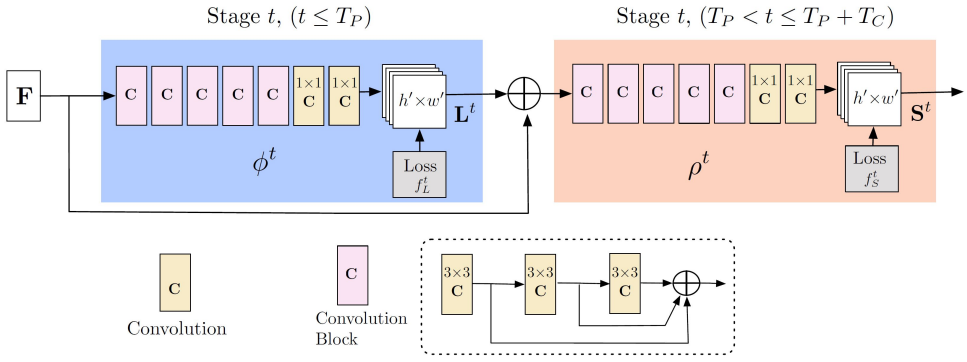

Figure 3. Multi-stage Convolutional Neural Network (CNN) [10].

The first set of stages predicts the PAFs. The second set uses the PAFs given as output from the previous layers to refine the Confidence Map detection. Finally, the generated CMs and PAFs are processed by a greedy bipartite matching algorithm to obtain the poses for each person in the image. Specifically, the loss function of the PAF branch at stage t_i and the loss function of the CM branch at stage t_k are, respectively:

$$f_L^{t_i} = \sum_{c=1}^{C} \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot \left\| \mathbf{L}_c^{t_i}(\mathbf{p}) - \mathbf{L}_c^*(\mathbf{p}) \right\|_2^2 \quad (1)$$

$$f_S^{t_k} = \sum_{j=1}^{J} \sum_{p} \mathbf{W}(\mathbf{p}) \cdot \left\| \mathbf{S}_j^{t_k}(\mathbf{p}) - \mathbf{S}_j^*(\mathbf{p}) \right\|_2^2. \quad (2)$$

The overall loss is:

$$f = \sum_{t=1}^{T_P} f_{\mathbf{L}}^t + \sum_{t=T_P+1}^{T_P+T_C} f_S^t, \quad (3)$$

where $\mathbf{L}_\mathbf{c}^*$ is the ground-truth PAF, $\mathbf{S}_\mathbf{j}^*$ is the ground-truth CM, and $\mathbf{W}$ is a binary mask with $\mathbf{W}(\mathbf{p}) = 0$ when the annotation is missing at the pixel $\mathbf{p}$.

3.2. Pose Comparison

By means of the pose tracking approach explained previously, the human figures represented in artworks can be easily modeled as 2D skeletons. However, how do we measure the similarity between two poses? Taking up the difference between the typology of measurable angles raised in [17], two strategies to fulfill this task are presented. First, a key advantage of our proposed methods is that no pre-processing is needed for the poses to be compared; they can have different scales and orientations. The latter property is an important novelty with respect to other approaches proposed in the literature [11].

Before the comparison stage, just for convenience purposes, according to the conventions subsequently adopted, the 2D skeleton coordinate system is changed: the joint coordinates, which are defined in a left-handed coordinate system (usually adopted for images), are converted into a right-handed coordinate system. This first transformation is followed by a reference system translation to deal with positive coordinates only.

When coming to pose comparison, given a query pose, we measure its similarity with respect to another one by means of a loss function. This loss function quantifies the distance between two poses; the lower the loss, the closer the poses. We defined two different losses—one per method. Based on these loss functions, it is possible to rank a set of poses according to their similarity to a given one.

Moreover, both the approaches include the possibility of computing the distance between the query pose and the mirrored and turned versions of the compared pose, thus expanding the intuition of [11], which was limited to mirroring (Figures 4 and 5 show, respectively, the aforementioned projections, while Figure 6 shows an example of the mirrored and turned versions of the *David* of Michelangelo). The mirroring and turning transformations are described in more detail in Appendix A.3.

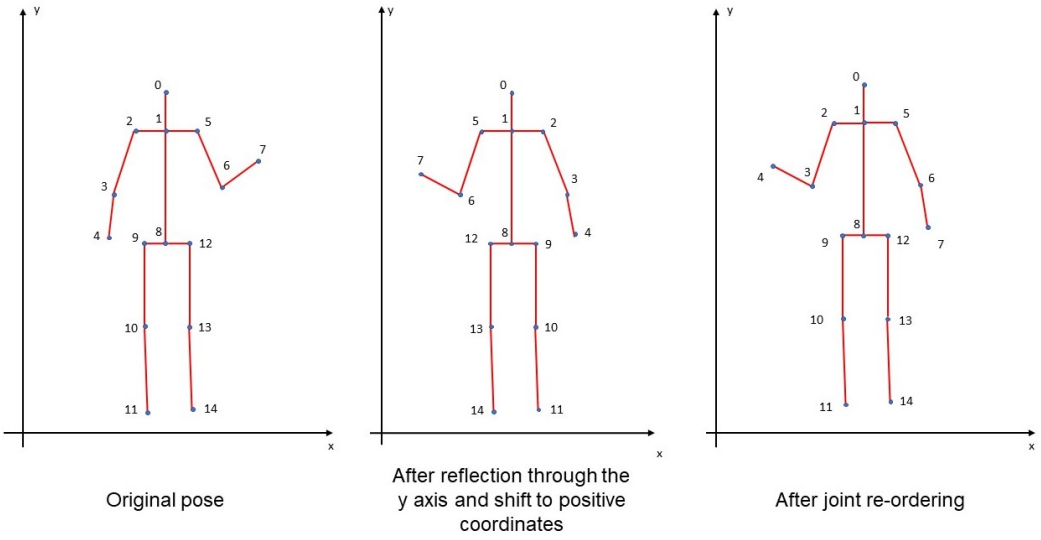

Figure 4. Mirroring of the pose.

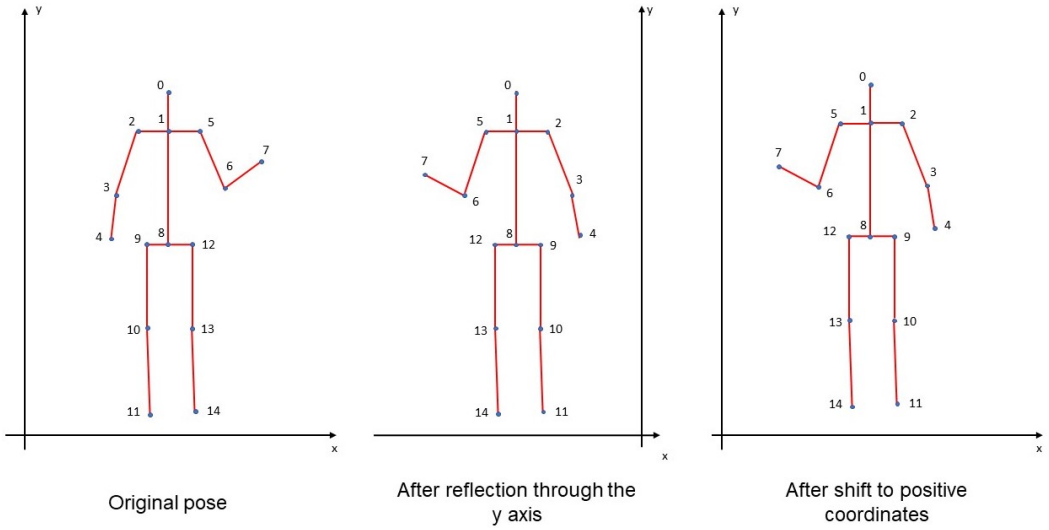

Figure 5. Turning of the pose.

Figure 6. Mirrored image of the *David* of Michelangelo (**left**), front view (**center**), and turned version or back view (**right**).

Every pose is composed of 15 pairs of coordinates, $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_{15}) \in Mat(15, 2, \mathbb{R})$, with $\mathbf{p}_j = (p_j^x, p_j^y)$ representing the j-th keypoint or joint.

By connecting two consecutive joints, we obtain a link. Hence, the same pose $\mathbf{P}$ can be represented by means of a set $\mathbf{\Lambda}$ of links or limbs.

$$\mathbf{\Lambda} = (\mathbf{l}_1, ..., \mathbf{l}_{14}) \in Mat(14, 2, \mathbb{R}) \quad with \quad \mathbf{l}_i = (l_i^x, l_i^y) \tag{4}$$

We can also define a matrix $\mathbf{B} \in Mat(14, 15, \mathbb{Z})$ that encodes the relationship between $\mathbf{\Lambda}$ and $\mathbf{P}$.

$$\mathbf{\Lambda} = \mathbf{B} \cdot \mathbf{P} \tag{5}$$

$$\mathbf{B} = \{b_{ij} \in \{-1, 0, 1\} \quad | \quad i = 1, \ldots, 14 \quad \wedge \quad j = 1, \ldots, 15\} \tag{6}$$

Now that the conventions have been defined, we present the proposed methodology.

3.2.1. First Method

The first method (shown in Figure 7) is based on a simple assumption: once two poses have been oriented in an optimal relative configuration through a relative rotation ω, their distance can be quantified by observing how much the directions of the corresponding limbs differ. The relative orientation of the poses allows one to find the most authentic similarity between them. Furthermore, in this way, we do not need to make hypotheses on

the standard orientation of the poses (by orienting the torso or the neck along a principal axis, for example [1,11]).

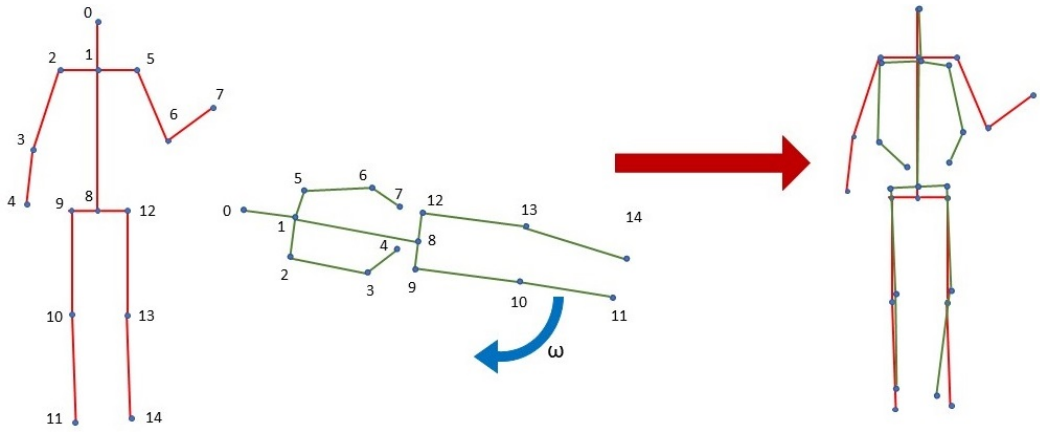

Figure 7. First method for pose comparison based on a preliminary rotation ω to find the optimal relative configuration of the poses.

Given two poses, A and B, we compute the angle α_i^{AB} between the i-th limbs $l_{A,\,i}$ and $l_{B,\,i}$ with $i = 1, ..., 14$. Each angle depends on $\omega \in [0, 2\pi]$.

Finally, we iteratively optimize the loss function (7), which is the sum of the 14 angles, with respect to ω.

$$L^{AB} = \min_{\omega} \sum_{i=1}^{14} \alpha_i^{AB}(\omega) \tag{7}$$

The same operation can be repeated after respectively turning and mirroring the pose B to be able to grasp all the similarities that were not captured by statically comparing the starting poses: in the end, the best matching version of pose B is selected.

In Appendix A.1, we present the methodology for defining Equation (7).

3.2.2. Second Method

As for the second method (shown in Figure 8), we follow a different approach. In fact, we consider all the angles that each limb of a given pose forms with all the other limbs of the same pose—not only the consecutive ones—ending up with a 91-feature vector.

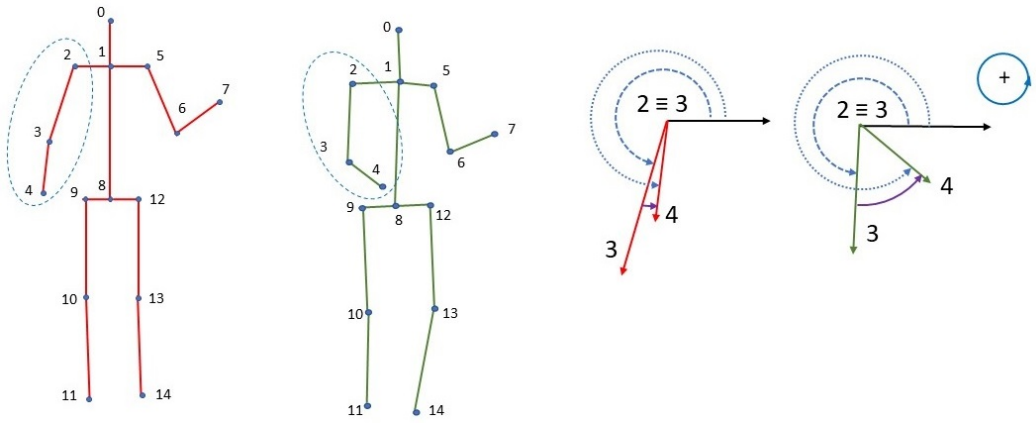

Figure 8. Second method for pose comparison based on the computation of all the angles that each limb of a given pose forms with all the other ones of the same pose.

Based on the obtained feature vectors, we can compare two poses through a specific loss function.

To calculate the angle between two links, we take two steps: first, we compute, for each link l_i, its angle θ_i with respect to the x-axis.

Then, the difference δ_{ik} between the angles θ_i and θ_k (corresponding to links i and k) is computed, with $i = 1, ..., 14 \wedge k = 1, ..., 14 \wedge i \neq k$.

If we compute this difference for each combination, we obtain the feature vector $\mathbf{\Delta} = (\delta_1, \ldots, \delta_{91})$.

Having obtained $\mathbf{\Delta}$, based on this representation, we can compare different poses (A and B) by computing the loss function (8).

$$L^{AB} = \frac{1}{91} \sum_{n=1}^{91} \left(1 - \cos\left(\delta_n^A - \delta_n^B\right)\right) \tag{8}$$

Even in this case, the loss quantifies the distance between two poses. So, the smaller it is, the more similar the poses are.

In Appendix A.2, we show the methodological framework proposed for obtaining Equation (8).

It is worth noticing that this loss is a non-Euclidean distance, since we are dealing with circular quantities, following the line already adopted in [17].

By including all the angles formed by a link with the other ones, a higher relevance is attributed to the position of the lower limbs, thus highlighting the poses with peculiar orientations of such parts of the body (indeed, the upper limbs are usually characterized by higher variability than the lower ones [1,17], as also shown by Figure A1 in Appendix B.1).

Furthermore, the method is completely invariant to scale transformations, rotations, and translations, and, just like the first one, we can choose to compare pose A with the mirrored or turned version of pose B, thus also addressing the horizontal flip invariance.

In addition to this, this approach can be considered as a baseline, since it includes all the possible angles within a human pose, and no assumptions are made on the most relevant features. Because of this, it can be adapted according to the purpose that we want to achieve—for example, by selecting and decreasing the number of features. As for the time efficiency, it is faster than the first approach (for more details, see Section 4.2.1).

3.3. Pose Clustering

The second method, which was described in the previous section, is based on the definition of a mapping function $f : Mat(15, 2, \mathbb{R}) \rightarrow \mathbb{R}^{91}$, which maps the generic pose $\mathbf{P}$ into the feature space as a feature vector $\mathbf{\Delta}$.

Thus, we rely on this mapping function to cluster the poses of our dataset in the feature space. The clustering algorithm overcomes the limits of the one-to-one comparison, and allows the subdivision of the artwork dataset into macro-groups and the identification of the respective archetypal figures (that is, the group centroids). To solve this clustering task, we make use of a modified version of the K-Medians algorithm, which is more robust to the outliers compared to the K-Means one. The K-Medians algorithm is an iterative procedure that involves the following stages:

1. Selection of the number of clusters m, which is a hyperparameter of the problem, with $1 < m < N$, where N is the number of poses in the dataset;
2. Initialization of the algorithm: this is a crucial aspect of our pipeline, since the K-Medians algorithm suffers from initial starting condition effects [29]; we opted for a Forgy approach [30], which chooses m instances of the dataset (seeds), $\mathbf{z_j}, j = 1, ..., m$, at random as starting centroids;
3. Assignment of the instances to the closest centroid: the same Forgy approach assigns the rest of the instances to the cluster represented by the nearest seed according to the distance presented before in Equation (8);
4. Update of the positions of the centroids: once all the instances have been assigned to a cluster, the new positions of the centroids are updated. The new centroid $\overline{\mathbf{\Delta}}_\mathbf{j}$ is the median of the instances assigned to the cluster that the centroid represents;

5. Iteration of the process: the steps 3 and 4 are repeated until convergence—that is, until the distance between the j-th centroid at step t and the corresponding one at step $t-1$ is lower than a given threshold (fixed to 0.0001 in our tests) for all the centroids.

Because we are dealing with circular quantities, each instance is represented as a set of 91 points on the unit circle, instead of a set of angles—that is, $\mathbf{\Delta} = \{(\cos\delta_i, \sin\delta_i), \quad i = 1, ..., 91\}$. Thereafter, each coordinate of the centroid $\overline{\boldsymbol{\delta}}_{\boldsymbol{ij}}$ is computed as follows:

$$\overline{\boldsymbol{\delta}}_{\boldsymbol{ij}} = \text{atan2}(\text{median}(\mathbf{C}_\mathbf{i}^\mathbf{j}), \text{median}(\mathbf{S}_\mathbf{i}^\mathbf{j})) \tag{9}$$

where $\mathbf{C}_\mathbf{i}^\mathbf{j}$ is the set of cosines of the i-th feature of the instances assigned to the j-th cluster, and $\mathbf{S}_\mathbf{i}^\mathbf{j}$ is the set of sines of the i-th feature of the instances assigned to the j-th cluster.

It is worth noticing that, based on this process, the obtained centroids do not properly represent a human pose: indeed, any relationships between the angles formed by limb pairs are lost. Nevertheless, it was still possible to retrieve, for each cluster, a reconstructed pose based on the centroids features, since the reconstruction error was negligible (as proven in Section 4.3).

4. Experiments

4.1. Dataset

The proposed methodology was applied on a corpus that we built ad hoc. The corpus was made up of more than 1400 images collected from three repositories: Scala [15], ArtResource [16], and the dataset of the Museum of Classical Art. The prerequisite of each image was that it had to represent a pose in its entirety. In fact, our algorithm, even if still functioning, is not meant to work with poses that lack one or more joints, at least for the moment. However, the choice of including poses with missing keypoints presents some risk because the range of possible matching poses for a given query becomes wider, and the most similar matches are not always truly relevant.

4.2. Tests Performed and Results

Although most of the images represented sculptures, our dataset had great variability. In fact, it included works from different eras (e.g., from Ancient Greece to Modern Art), different styles (e.g., from Classicism to Baroque), different techniques (e.g., bas-reliefs or statue), different materials (e.g., bronze, wood, or marble), and with different backgrounds (e.g., uniform red background or composed of other statues). Moreover, the quality of the images also varied from low-resolution black and white (B/W) to high-resolution images. Each of these factors had some influence on the automatic annotation phase. We applied OpenPose as the pose tracking algorithm on these images.

In Figure 9, we can observe some examples of failure. We were able to correctly annotate 618 images out of the whole starting dataset through OpenPose. If we recall the main reasons for failure highlighted in [10], we can easily understand the 44% success rate. Indeed, OpenPose tends to fail for non-typical poses and upside-down examples; moreover, body occlusion can also lead to false negatives and high localization error, as for keypoint estimation (this problem is inherited from the dataset annotations, in which occluded keypoints are not included). As a consequence, pose tracking was much more difficult for the artworks with uncommon poses, for images with bad quality, and, most of all, for figures covered, for example, with veils or shields. Finally, it is worth mentioning another reason for failure, which was not considered by the authors of OpenPose, and which is clearly evident in Figure 9: the algorithm tends to fail for human figures with missing limbs (probably because they were not included in the training dataset). Despite all of the aforementioned limitations, however, we believe that our test proved a good efficiency of the algorithm, which was tested with no previous fine tuning, making it an interesting tool for speeding up the automatic labeling of artworks, waiting for proper datasets to be released, and the diffusion and improvement of approaches like the one described in [12],

which relies on style-transfer learning to enhance the human pose recognition framework capability of generalizing across domains.

Figure 9. Some examples of failure.

4.2.1. Pose Comparison

We applied the techniques described in Section 3 on a dataset made of the joints of the correctly annotated images. As far as the comparison stage is concerned, for each pose, which we can call the query, we selected and present here five relevant elements: the four most similar poses (which are the ones with the smallest losses with respect to the query) and the most dissimilar one (that is, the one with the highest loss with respect to the query). We applied the first method without mirroring and turning (Figure 10), the first method with mirroring and turning (Figure 11), and the second method (Figure 12). In particular, we present the results for the subset of images from the Museum of Classical Art because it is sufficiently homogeneous in style and refers to a precise epoch, thus avoiding forced interpretations. It is surprising to note how, even for a subset of only about 125 elements, we can notice interesting connections among the artworks.

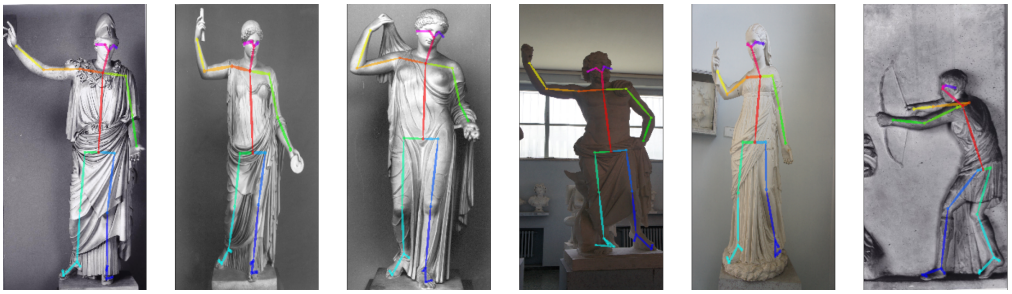

Figure 10. Comparison—method 1 (query—four best matching poses—worst matching pose).

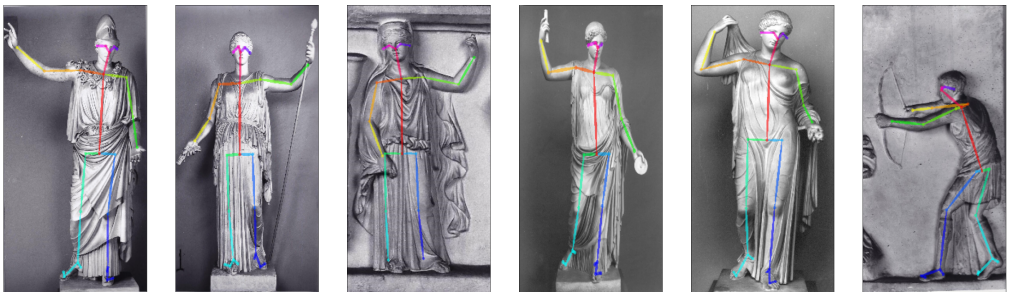

Figure 11. Comparison—method 1 with turning and mirroring options enabled (query—four best matching poses—worst matching pose).

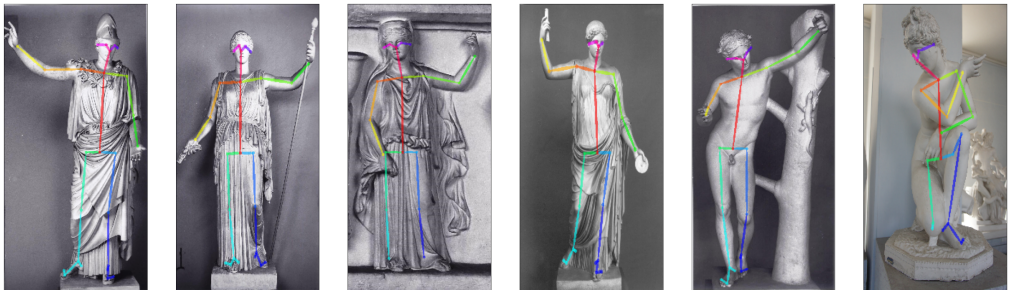

Figure 12. Comparison—method 2 with turning and mirroring options enabled (query—four best matching poses—worst matching pose).

Although the statues in Figures 10–12 represent different subjects in different activities, the pose similarity is evident. In particular, we can identify a limited number of common features across the artworks. In general, the female characters and the *Statua dell'Apollo Borghese* share a graceful and balanced posture. The same pose is associated with control, power, and authority, as we can see in the query image (*Atena di Velletri*), and in the images of the goddess Demeter (holding a spear) and of Poseidon. On the contrary, the farthest artworks (the archer—part of a scene from the myth of the Niobids—and the statue of *Afrodite accovacciata*) are crouching figures; therefore, they were ranked as the most distant from the query. Despite the slight differences between the first and second methods, the results were quite consistent, and the introduction of mirroring and turning helped to enrich the similarity survey.

To assess the consistency of our methods, we checked that the pose most similar to the query was always the query itself, with a loss value equal to 0. To do this, we just re-inserted the query itself among the poses to compare with the query.

For our experiments, we used a machine with 64 GB of RAM, Intel(R) Core(TM) i7-9700 CPU @ 3.00 GHz. As to the time efficiency of our methodologies, the computational time is hereinafter reported (by iteration, we mean the step by which a query pose is compared to all the other poses of the dataset):

- 3.2 s per iteration for the first method without mirroring and turning options enabled;
- 10.0 s per iteration for the first method with mirroring and turning options enabled;
- 2.2 s per iteration for the second method, which we only ran with mirroring and turning options enabled.

It is immediately evident how the second method outperforms the first one in terms of time efficiency and computational cost. Indeed, if the mirroring and turning options are enabled for the first method, three optimizations have to be performed for each pose-to-pose comparison; this is the reason for why the processing time triples when both the options are enabled.

4.2.2. Pose Clustering

As far as the clustering is concerned, we applied our clustering algorithm on the features associated with the poses according to the representation described in Section 3.2.2.

We did several tests with a variable number of clusters, from 2 to 50. From a qualitative point of view, the optimal number of clusters fell in the range of 10–15 (a quantitative accuracy assessment of the clustering approach is presented in Section 4.3). This is related, of course, to the features of our dataset—in particular, the number and the types of poses.

In the following discussion, we will show two cases of clustering with 5 and 10 clusters because these values include the elbow of the curve visible in Figure 13. In Figures 14 and 15, we can see a cluster of poses deriving from a clustering with, respectively, 10 and 5 clusters (each pose is accompanied by the respective distance from the centroid).

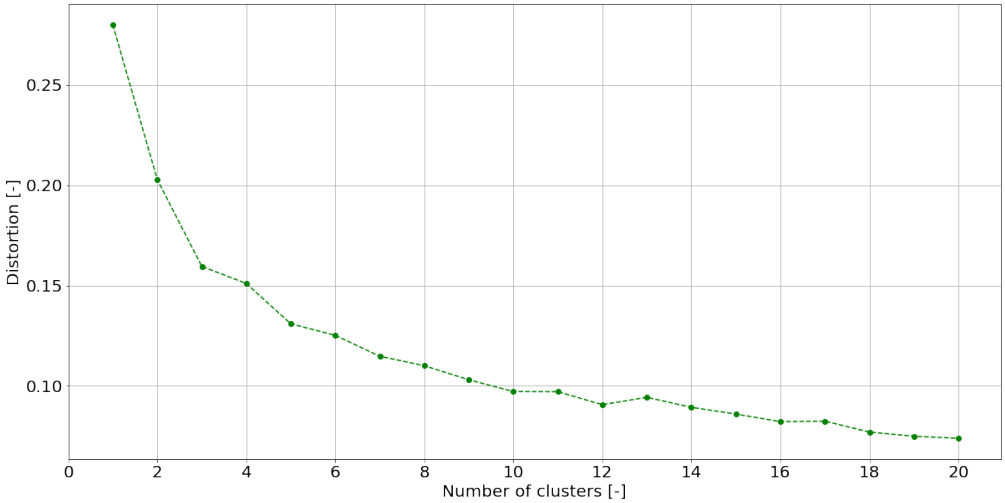

Figure 13. Elbow method for picking the optimal number of clusters.

We can discuss some things based on these images. First of all, with fewer clusters, there is a higher variability in the same cluster, as expected (see, for example, how the right arm direction varies across the artworks in Figure 15). Second, we can observe how, in both cases, the legs are less relevant than the arms in the definition of a specific cluster. Finally, Figure 15 is particularly useful for appreciating how the distance increases as the upper limbs progressively deviate from the median pose. The higher impact of the variation of the upper limb position on the distance is confirmed by the ablation study shown in Appendix B.1.

In our experiments, we initialized the clustering algorithm with casual centroids, leading to a satisfactory result. However, the same methodology would be useful for searching for specific poses within a given dataset, too; for instance, if we had some *a priori* information, such as a dataset of images annotated with the emotions that they can convey to the observer, we could search for all the possible emotions that a given pose can convey.

As previously anticipated, after the clustering stage, we were able to reconstruct the poses associated with each centroid to make it clear the median pose around which the elements of the cluster aggregated. Two examples of reconstructed poses are visible in the previously mentioned Figures 14 and 15. To compute the joints of the reconstructed poses, we started by choosing among the 91 features of the centroids only those that represented angles between consecutive limbs (e.g., the angle between the arm and forearm). Then, we calculated the average length of each limb on the basis of the poses of our dataset, which were previously scaled through the respective radius of inertia (10).

$$\bar{l}^j = \frac{1}{N} \sum_{i=1}^{N} \frac{l_i^j}{\rho_i}, \tag{10}$$

where j refers to the j-th limb, N to the number of the poses in the dataset, ρ_i to the radius of inertia of the i-th pose, and l_i^j to the length of the j-th limb of the i-th pose.

By fixing the coordinates of two joints of the centroid pose in the desired coordinate system, and through the average lengths previously calculated, we can progressively compute the position for all the joints. The initial choices ensure the representation of the reconstructed pose with positive coordinates only.

The choice to exclude the majority of the features from the reconstruction of the pose leads to a small error, which is evaluated in Section 4.3. Future updates may provide for a more precise reconstruction approach that takes into account all of the features for the calculation of each joint, and thus aims at minimizing the reconstruction error.

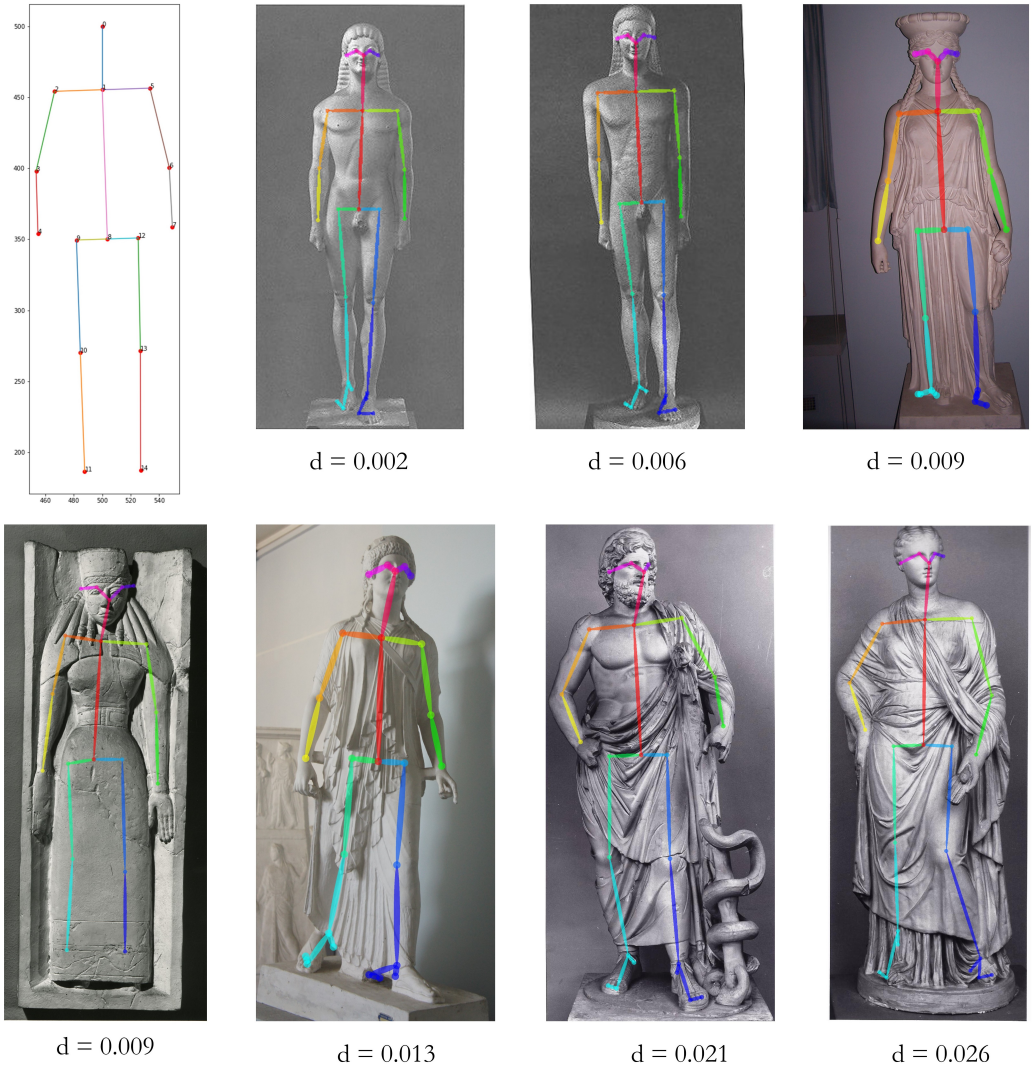

Figure 14. Example of a cluster of poses (together with the median pose and the respective distance from the centroid) for clustering with n = 10.

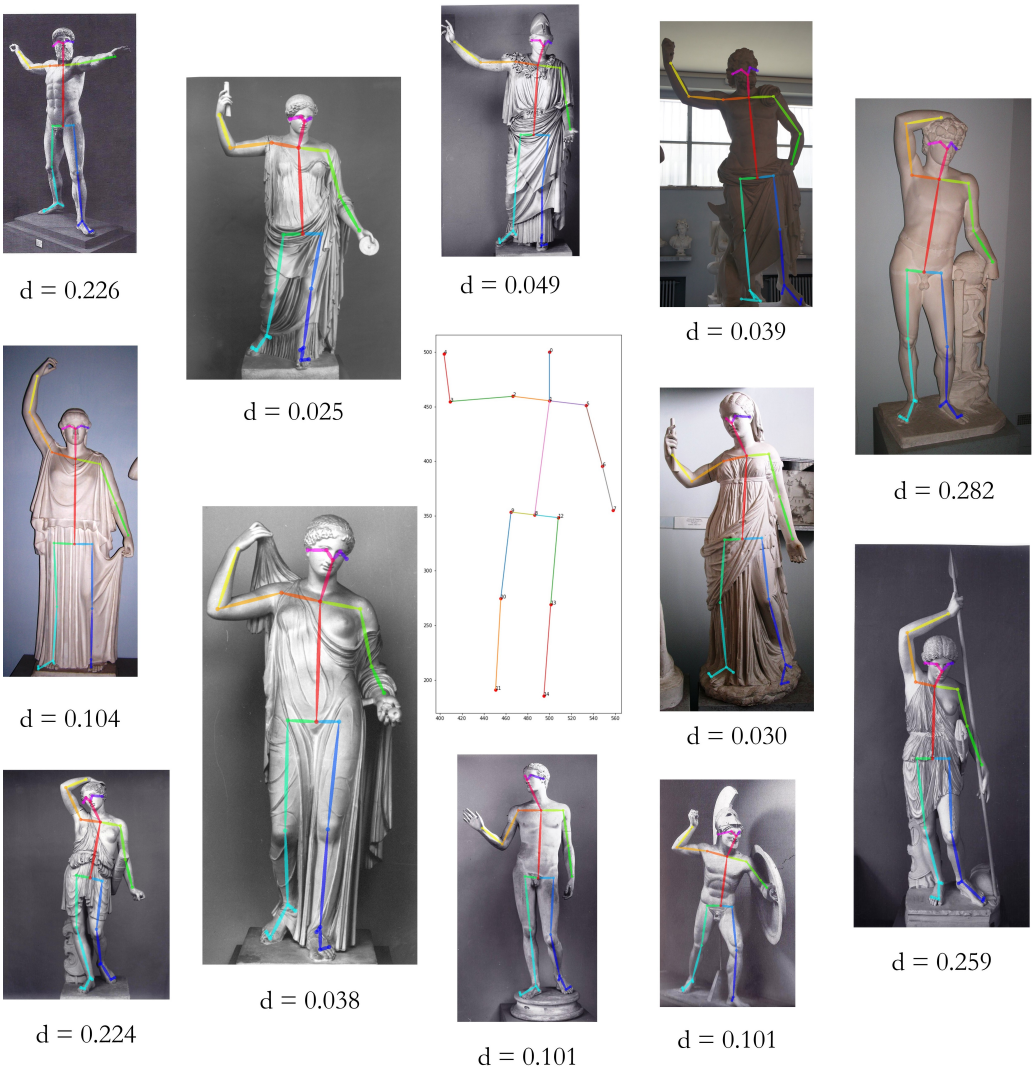

Figure 15. Example of a cluster of poses (together with the median pose and the respective distance from the centroid) for clustering with n = 5.

4.3. Validation

The quantitative assessment of the proposed methodology was made difficult by several factors; indeed, we lacked a proper reference for comparison and/or a benchmark dataset (for example, a dataset of labeled poses according to specific actions) because the similar approaches in the literature faced slightly different tasks on private collections of images. Nevertheless, we carried out some experiments to highlight the validity of our methodology.

First of all, as previously anticipated in Section 4.2.2, the reconstruction of the pose from the 91 centroid features inherited some errors from the clustering stage; indeed, since each feature was computed as the median of a circular quantity (according to Equation (9)), the geometric constraints among the features were lost. Because of this, the poses associated with the cluster centroids could only be computed in an approximate way.

To quantify the error deriving from this approximation, we defined the reconstruction error e as:

$$e = \frac{1}{91}\sum_{i=1}^{91}\min(|\delta_i^{(r)} - \delta_i^{(c)}|, |\delta_i^{(r)} - \delta_i^{(c)} - 2\pi|, |\delta_i^{(r)} - \delta_i^{(c)} + 2\pi|), \tag{11}$$

where $\delta_i^{(r)}$ and $\delta_i^{(c)}$, for $i = 1, ..., 91$, are the features of the reconstructed pose $\mathbf{\Delta_r}$ and of the centroid $\mathbf{\Delta_c}$, respectively. For a given clustering with m clusters to learn, the mean reconstruction error $\bar{e}$ is the average of the reconstruction errors for the m centroids obtained $\bar{e} = \frac{1}{m}\sum_{j=1}^{m} e_j$.

We analyzed the relationship between the reconstruction error and the number of clusters n for $n \in [2, 50]$. For each n, we repeated the test 10 times to make the experiment more consistent from a statistical point of view. In Figure 16, the results obtained are shown: the average error fell in the range of 2.5°–4.5°, which means that, on average, the difference between each feature of a centroid $\delta_i^{(c)}$ and the respective feature of the reconstructed pose $\delta_i^{(r)}$ falls in that range, which represents a good approximation for the pose encoded by that centroid. Moreover, we can notice that the mean reconstruction error did not seem to be dependent on the number of clusters n, with the exception of $n = 2$, but its variance was progressively reduced as n increased. This can be explained with the high variability of the clusters when n was low, which highly affects the median value and, thus, the discrepancy between $\mathbf{\Delta_c}$ and $\mathbf{\Delta_r}$.

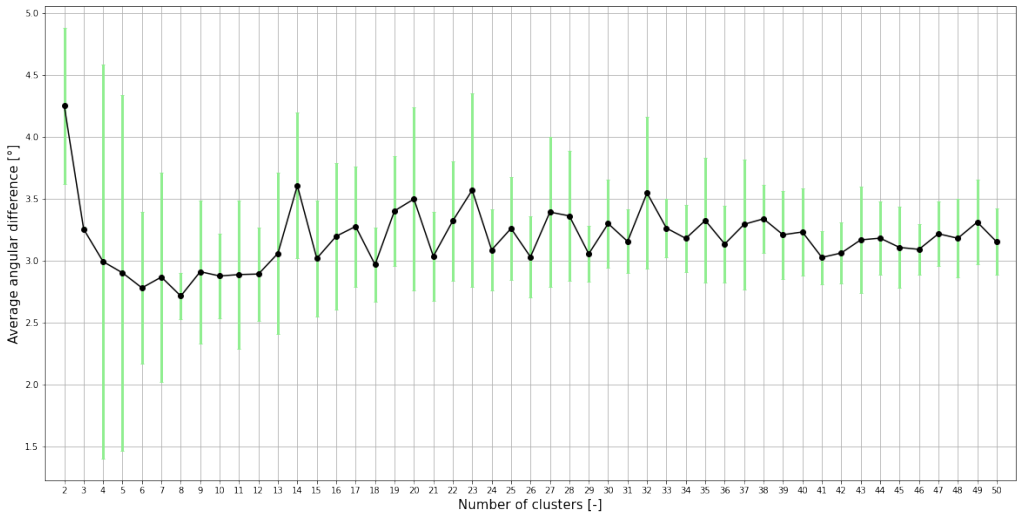

Figure 16. Error for pose reconstruction from centroids for a number of clusters $n \in [2, 50]$ (one standard deviation of uncertainty is represented).

Coming to the second relevant factor, which hinders the possibility of a performance evaluation of our pose clustering strategy, that is, the absence of a labeled dataset, we carried out the following simulation:

- We chose m representative or archetypal poses from our dataset;
- For each pose, we generated 100 samples by adding random noise $\sim \rho \cdot N(0, \sigma^2)$ to the archetypal pose keypoints, where ρ is the radius of inertia of the archetypal pose and σ is the standard deviation of the Gaussian distribution;
- We obtained, as a result, a labeled dataset of $m \cdot 100$ samples;
- Based on this synthetic dataset, we were able to evaluate the clustering performance by means of proper functions and indicators.

We carried out the simulation for $m \in \{5, 10\}$ and for 10 values of σ that were evenly spaced on a base-10 scale in the range $[0.01, 1]$. In Figure 17, the impact of the noise insertion on the keypoints is shown for different values of σ.

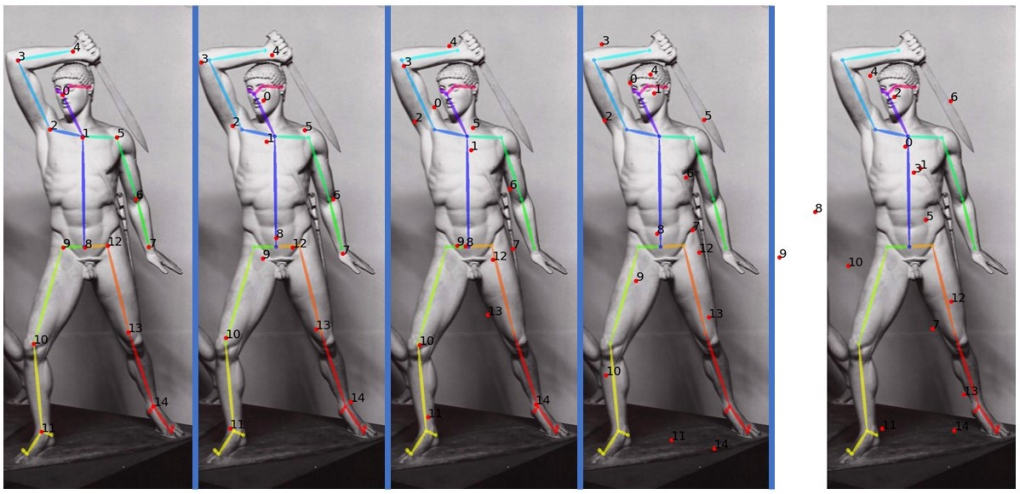

Figure 17. Example of perturbed poses with an increasing σ.

For each value of σ, we repeated the test 20 times to add statistical consistency to our experiment (the results of these tests seem, however, quite stable with respect to the number of iterations, as shown in Appendix B.2). As far as the clustering performance evaluation is concerned, we must recall that the functions and indicators involved in this task had the main aim of assessing if the clustering defined data separations similarly to some ground-truth data or if it satisfied the assumption that items of the same cluster are more similar than items of different clusters. Therefore, we made use of the following ones:

1. Adjusted Rand Index (ARI), which measures the similarity between the ground-truth assignment and the clustering, ignoring permutations and with chance normalization [31];
2. Two different normalized versions of the Mutual Information function (which measures the agreement of the two assignments, ignoring permutations), namely the Normalized Mutual Information (NMI) and the Adjusted Mutual Information (AMI), where the latter is normalized against chance [32].

The same tests were carried out with both K-Medians and two alternative approaches, namely K-Means and the Hierarchical Clustering known as the Nearest Point Algorithm. All of the algorithms made use of the distance function defined in Equation (8).

In Figures 18–23, the results obtained from this simulation are shown. In particular, the mean values of the evaluation functions and the error bars (based on 20 repetitions and representing one standard deviation of uncertainty) are shown.

As for K-Medians, the pose clustering approach showed good and stable results based on the three functions ($\sim$0.9 for AMI, $\sim$0.9 for NMI, and $\sim$0.7 for ARI), with a negligible effect of the number of clusters. The performance began decreasing, instead, when the standard deviation of the Gaussian noise exceeded 0.25, which was equivalent to 25% of the pose radius of inertia, proving the stability of the proposed approach with respect to considerable noise insertions (see Appendix B.1 for more details).

Compared to the other two approaches, it might seem that K-Means (for σ in the range $[0.01, 0.2]$) and the Hierarchical Clustering (for σ in the range $[0.01, 0.1]$) outperform K-Medians. Nonetheless, if we consider the most relevant range for a clustering algorithm that aims to gather human poses on the basis of mutual similarity, thus dealing with heterogeneous and variable datasets, i.e., the trait $[0.1, 0.5]$—which can be deduced by looking at Figure 17—the conclusion is completely overturned. The Hierarchical Clustering method, indeed, is highly inefficient, since its performance drops to 0 even before the aforementioned range; as for K-Means, although it is higher than K-medians in the initial part of the range of interest, it is progressively overcome in terms of performance, as is clearly visible in Figures A2–A4 and in Appendix B.1, which focus on the trait $[0.1, 0.5]$.

This comparison proves, in conclusion, the higher robustness to outliers of the K-Medians approach with respect to the other ones.

Finally, it is worth noticing how the variability of the performance (as measured by the three indicators) is inversely proportional to the standard deviation of the Gaussian noise and to the number of clusters, as expected.

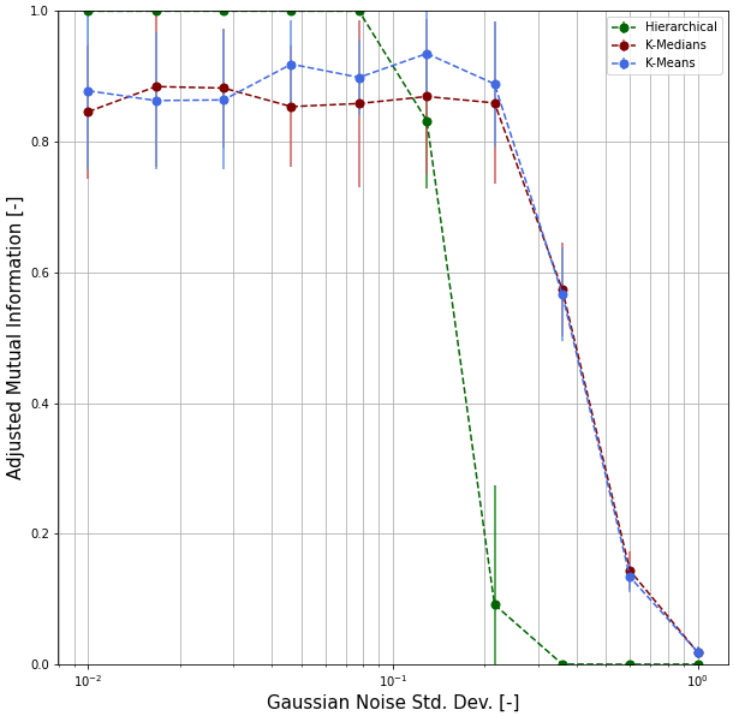

Figure 18. Adjusted Mutual Information (AMI) for the simulation with 5 clusters.

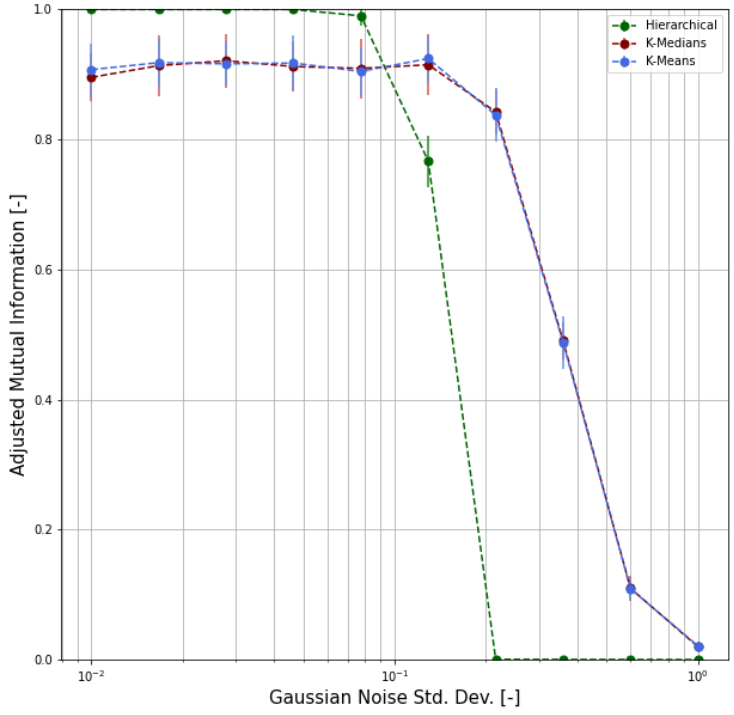

Figure 19. AMI for the simulation with 10 clusters.

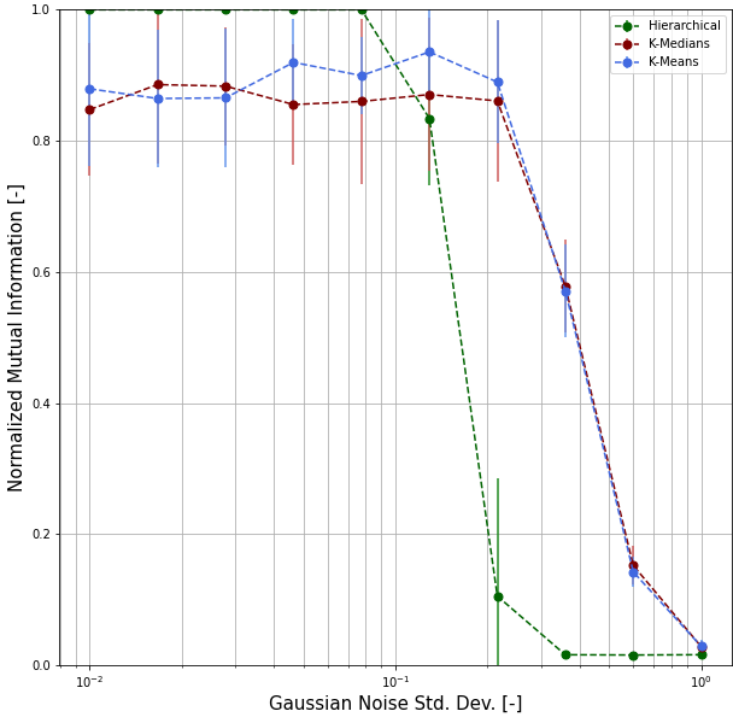

Figure 20. Normalized Mutual Information (NMI) for the simulation with 5 clusters.

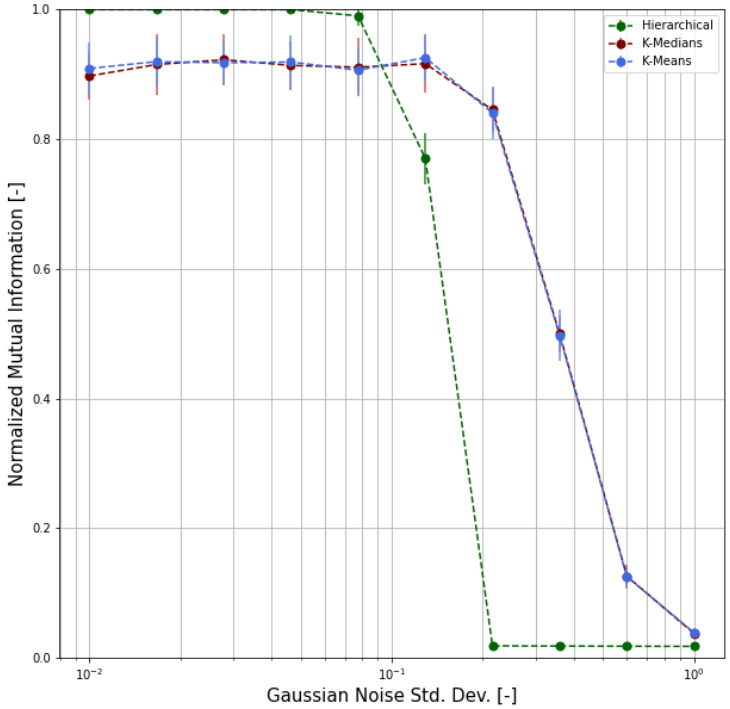

Figure 21. NMI for the simulation with 10 clusters.

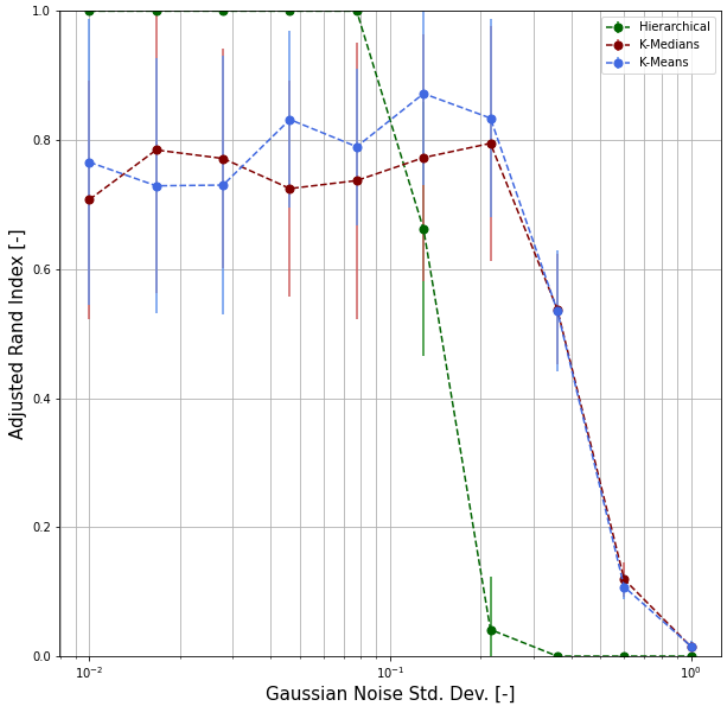

Figure 22. Adjusted Rand Index (ARI) for the simulation with 5 clusters.

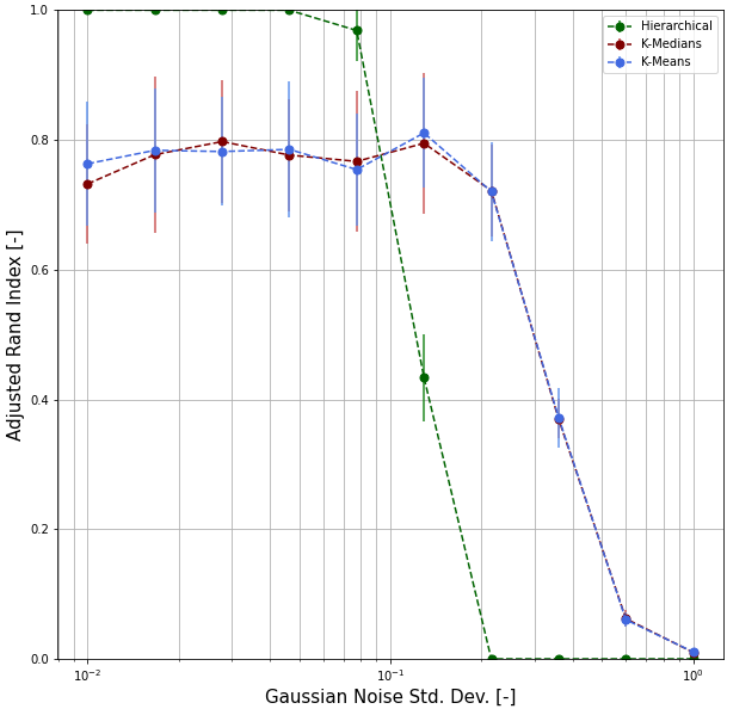

Figure 23. ARI for the simulation with 10 clusters.

5. Conclusions and Outlook

In this paper, we proposed a novel method with the aim of clustering the human poses represented in artworks, starting from Aby Warburg's concept of the *Pathosformel*. The German art historian developed this concept to explain the survival of figurative expressions that are found in artworks, even those that are very distant in time. Therefore, we focused our attention on the similarity among works of art based on human poses, the actions that they represent, and, in some sense, to the emotions delivered by the artworks.

So, our aim was to offer a fruitful tool for further research in different domains, such as art history and archaeology. This form of similarity was investigated by modeling human poses as 2D skeletons, which were defined as sets of 14 points (or joints) connected by limbs. For the dataset, we automatically annotated the images with OpenPose, testing its efficiency with respect to a heterogeneous artwork dataset.

We proposed a robust and consistent system for pose comparison, which consisted of two methods of pairwise comparison, each based on a specific distance function, and a method of clustering. The methodology was also validated on several clustering algorithms in a quantitative way by simulating a supervised setting, which proved the ability of the algorithm to find relevant clusters, even for noisy datasets.

As far as the outlooks are concerned, further analyses can be carried out to test new experiments and investigations on specific clusters and specific poses.

It is important to stress how we could also consider gestures of the hands as specific categories of poses themselves, which is an increasingly promising line of research.

The outcomes can be fruitful: this could be a way to explore the repositories of body poses and gestures in art by applying new filters that can confirm or overturn previous knowledge about periods, techniques, iconographies, and intericonicity.

From an algorithmic perspective, there are two research lines that could hopefully be deepened more: focusing on the different importances of the features and investigating other different clustering strategies, such as the rotational ones.

In addition to this, the analysis of multiple viewpoints of the same statue could add meaningful insights. The study of ancient sculptures that are missing limbs and their reconstruction through POSE-ID-on could produce useful hypotheses of virtual restoration without neglecting a dialogue about the contemporary production of bodies, which shows sensitivity to disabilities.

As far as museum applications are concerned, POSE-ID-on can be an interesting starting point for developing itineraries through collections based on suggested recognition of specific poses and gestures.

Moreover, another perspective is the cooperative creation of a new dataset based on high-quality annotated images that are enhanced with tags and codified annotations regarding the actions and emotions represented by the artworks.

Our intent, however, is not only limited to artworks; indeed, we believe that the potentialities of POSE-ID-on could be also extended to any kinds of tasks that require the clustering of human poses.

Author Contributions: Conceptualization, Valerio Marsocci and Lorenzo Lastilla; methodology, Valerio Marsocci and Lorenzo Lastilla; software, Valerio Marsocci and Lorenzo Lastilla; validation, Valerio Marsocci and Lorenzo Lastilla; formal analysis, Valerio Marsocci and Lorenzo Lastilla; investigation, Valerio Marsocci and Lorenzo Lastilla; resources, Valerio Marsocci and Lorenzo Lastilla; data curation, Valerio Marsocci and Lorenzo Lastilla; writing—original draft preparation, Valerio Marsocci and Lorenzo Lastilla; writing—review and editing, Valerio Marsocci and Lorenzo Lastilla; visualization, Valerio Marsocci and Lorenzo Lastilla All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by two Ph.D. fellowships granted to Valerio Marsocci and Lorenzo Lastilla by Sapienza University of Rome, Italy—Ph.D. course in Data Science.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Acknowledgments: The present paper stems from a series of meetings under the title "Recognizing emotions: a dynamic path between works of art and memes", which were aimed at the investigation and analysis of fruitful exchange and cooperation between the Data Sciences and Humanities, more specifically in the field of visual culture, and organized by the Ph.D. program in Data Science at Sapienza University of Rome by Stefano Leonardi. The authors wish to thank Antonella Sbrilli and Davide Nadali for their fruitful lectures and for their constant supervision of this work, the Museum of Classical Art of Sapienza University of Rome for the possibility of processing their images (par-

ticularly Marcello Barbanera, Mariateresa Curcio, and Raffaella Bucolo), and Mattia Crespi for the supervision of the technical aspects of this paper. Finally, we wish to thank the reviewers for their valuable insights.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Methodology

In this section, from a mathematical point of view, we illustrate the path that led us to the definition of the losses for each method, as previously described. We will follow the same conventions adopted in Section 3.2.

Appendix A.1. First Method

For the first method, the path to defining the distance function is the following.

Given two poses, A and B, we compute the angle α_i^{AB} between the i-th limbs $l_{A,\,i}$ and $l_{B,\,i}$ with $i = 1, ..., 14$. Each angle depends on $\omega \in [0, 2\pi]$ according to (A1).

$$\alpha_i^{AB}(\omega) = \arccos\left(\frac{d_i}{f_i}\cos\omega + \frac{e_i}{f_i}\sin\omega\right) \tag{A1}$$

with

$$\begin{aligned} d_i &= l_{i,A}^x \cdot l_{i,B}^x + l_{i,A}^y \cdot l_{i,B}^y \\ e_i &= -l_{i,A}^x \cdot l_{i,B}^y + l_{i,B}^x \cdot l_{i,A}^y. \\ f_i &= \|l_{i,A}\| \cdot \|l_{i,B}\| \end{aligned} \tag{A2}$$

Finally, we iteratively optimize the loss function (Equation (7)), which is the sum of the 14 angles α_i^{AB}, with respect to ω.

Appendix A.2. Second Method

The mathematical framework of the second method can be summarized as presented below. To calculate the angle between two links, we take two steps: first, we compute, for each link l_i, its angle θ_i (A3) with respect to the x-axis.

$$\theta_i = \begin{cases} \text{atan2}\left(\frac{l_i^y}{l_i^x}\right), & \text{if atan2}\left(\frac{l_i^y}{l_i^x}\right) > 0 \\ \text{atan2}\left(\frac{l_i^y}{l_i^x}\right) + 2\pi, & \text{otherwise} \end{cases} \tag{A3}$$

Then, the difference δ_{ik} between the angles θ_i and θ_k (corresponding to links i and k) is equal to (A4), with $i = 1, ..., 14 \wedge k = 1, ..., 14 \wedge i \neq k$.

$$\delta_{ik} = \begin{cases} \theta_i - \theta_k, & \text{if } \theta_i > \theta_k \\ \theta_i - \theta_k + 2\pi, & \text{otherwise} \end{cases} \tag{A4}$$

If we compute this difference for each possible combination, we obtain the feature vector $\mathbf{\Delta} = (\delta_1, \ldots, \delta_{91})$.

Having obtained $\mathbf{\Delta}$, based on this representation, we can compare different poses (A and B) by computing the loss function (Equation (8)).

Even in this case, the loss quantifies the distance between two poses. So, the smaller it is, the more similar the poses are.

Appendix A.3. Mirroring and Turning

To grasp all of the similarities that are not captured by statically comparing the query pose A with a given pose B, the distance computation, according to the metrics presented in Sections 3.2.1 and 3.2.2, is repeated after respectively turning and mirroring pose B.

To do so, first, we define **T** (A5), which corresponds to a reflection of the original pose with respect to the y-axis.

$$\mathbf{T} = \left\{ \mathbf{t_i} \in \mathbb{R}^2 \mid t_i^x = -p_i^x + \max p^x \quad \wedge \quad t_i^y = p_i^y \quad \wedge \quad i = 1, \ldots, 14 \right\} \tag{A5}$$

Then, in the case of the turned pose, we get the turned links $\mathbf{T_\Lambda}$.

$$\mathbf{T_\Lambda} = \mathbf{B} \cdot \mathbf{T} \tag{A6}$$

For the mirrored pose, we have to change the order of the joints belonging to the arms and the legs. We define a new matrix $\mathbf{B_M}$, which represents the new order of the joints.

$$\mathbf{M_\Lambda} = \mathbf{B_M} \cdot \mathbf{T} \tag{A7}$$

Once these new poses have been generated, the same comparison based on either the first or the second metric can be repeated.

Appendix B. Ablation Studies

In this section, we report some ablation studies to validate the choices that led to the results presented in the main text.

Appendix B.1. Focus on the Standard Deviation Range

In addition to what is presented in Section 4, we wanted to deepen the behavior of the clustering algorithms with respect to some datasets with particular variability. To identify the most interesting σ values, we used both quantitative and qualitative criteria. As for the latter, we refer to Figure 17, from which it can be seen that for σ values greater than 0.5, the pose is not at all attributable to a human figure, while for values that are too small, there is no perturbation so as to create significant variability. From a quantitative point of view, we studied the circular variance of our dataset and reported the results in Figure A1. The circular variance, as shown in [33], is calculated as follows:

$$Var(\delta_j) = 1 - R_j/n, \tag{A8}$$

where n = number of poses in our dataset, $j = 1, ..., 91$, and:

$$R_j^2 = \left(\sum_{i=1}^{n} \cos \delta_{i,j} \right)^2 + \left(\sum_{i=1}^{n} \sin \delta_{i,j} \right)^2 \tag{A9}$$

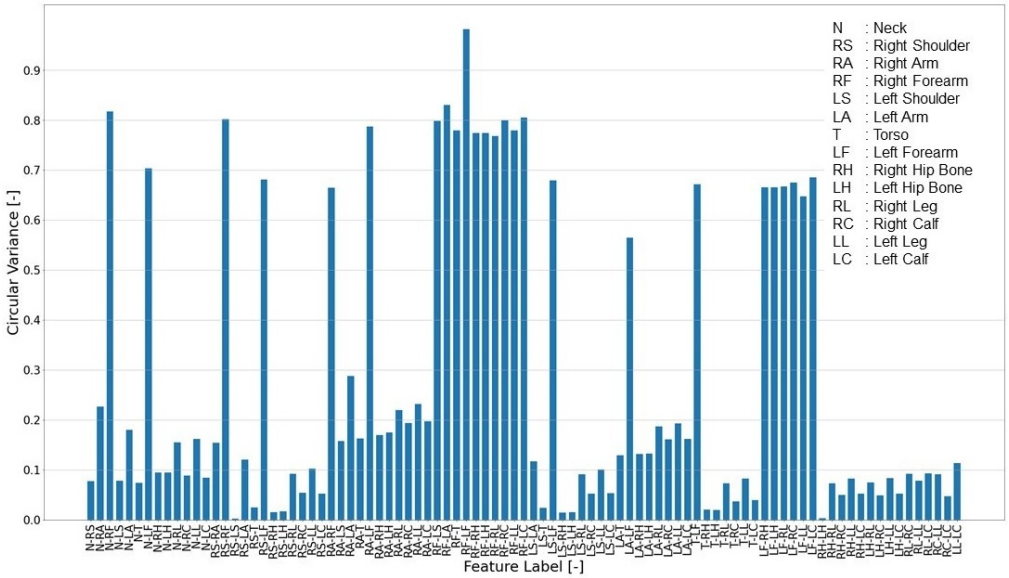

Figure A1. Variance of the 91 features in our dataset.

From Figure A1, we can argue how a dataset like ours is not properly described by the range of variability represented by $\sigma < 0.1$. In light of this, we can also observe that, as already done by [17], there is a greater variance in the angles that are formed by at least one upper limb than in those in which the upper limbs are absent. However, this property of our dataset does not undermine the choice of adding a uniform noise to the archetypal poses for the simulation because the noise effect is more distorting for the lower limbs, which are characterized by a lower variance.

In addition, for these reasons, in this section, we show the results for K-Means and K-Medians only, which are the best-performing clustering algorithms in a σ range that is more representative of a real scenario. In particular, we chose 20 σ values in the range [0.1, 0.5]. Thus, with each σ, we carried out 20 experiments, always running a 10-cluster clustering on the dataset, which was obtained by perturbing the starting poses with a noise equal to the chosen σ. Below, Figures A2–A4 demonstrate how, for each metric, the K-Medians performs better, as it is more stable against the presence of outliers. In particular, the mean values of the evaluation functions and the error bars are shown.

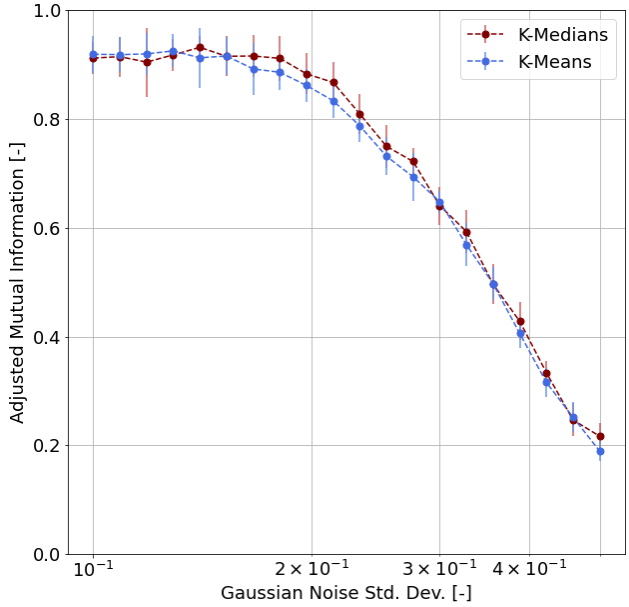

Figure A2. AMI for the simulation with 20 sigmas.

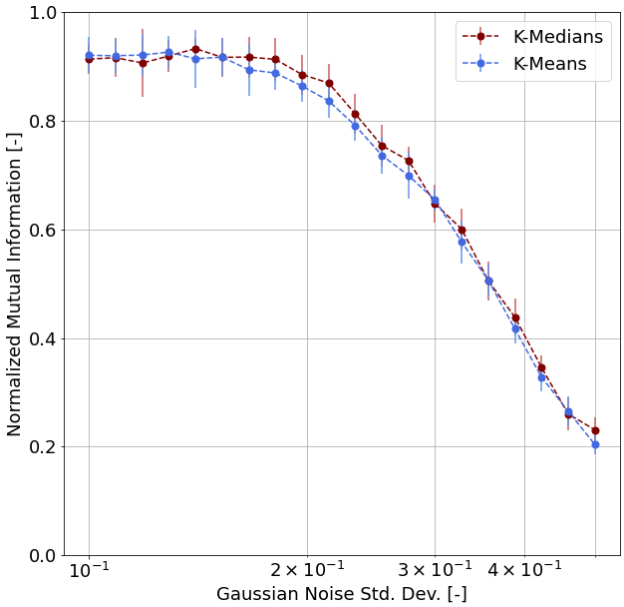

Figure A3. NMI for the simulation with 20 sigmas.

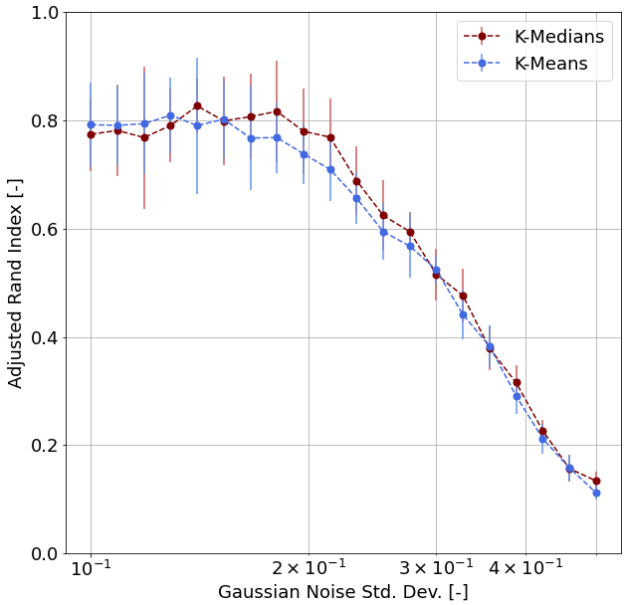

Figure A4. ARI for the simulation with 20 sigmas.

Appendix B.2. Number of Iterations

To justify the number of simulations conducted for each experiment, we ran trials with lower (10) and higher (30) repetitions. We conducted these experiments with a 10-cluster clustering. The results are shown in Figures A5–A7 and in Figures A8–A10, respectively.

In particular, the mean values of the evaluation functions and the error bars (based on 10 and 30 repetitions and representing one standard deviation of uncertainty) are shown.

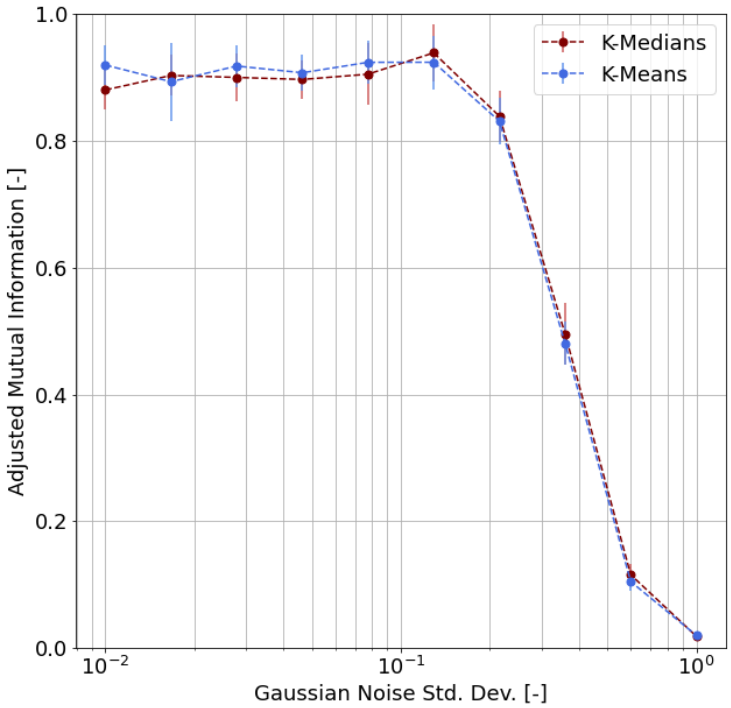

Figure A5. AMI for the simulation with 10 iterations.

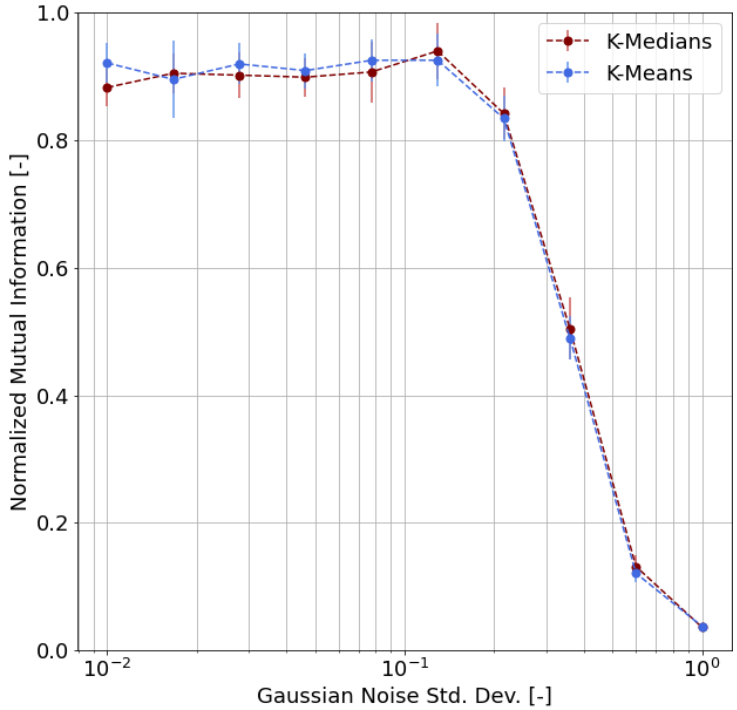

Figure A6. NMI for the simulation with 10 iterations.

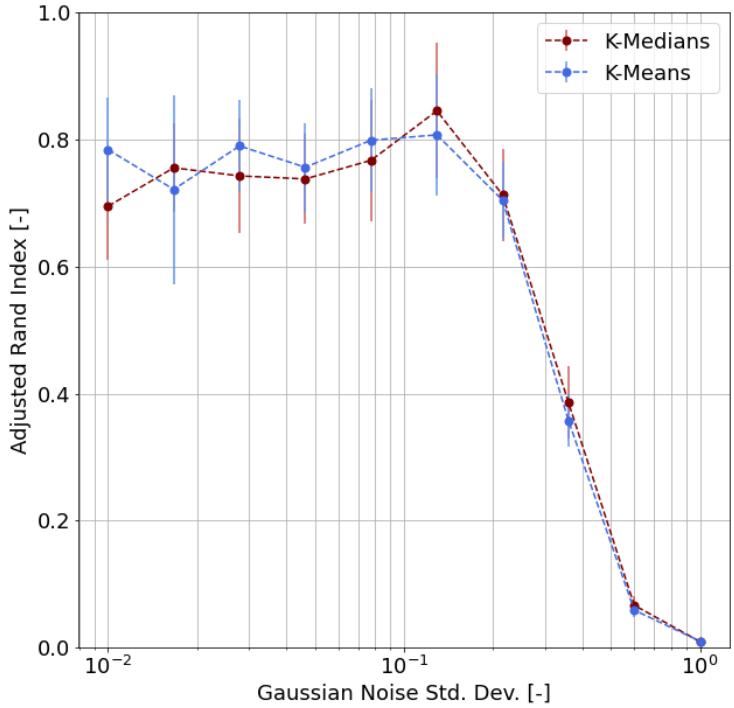

Figure A7. ARI for the simulation with 10 iterations.

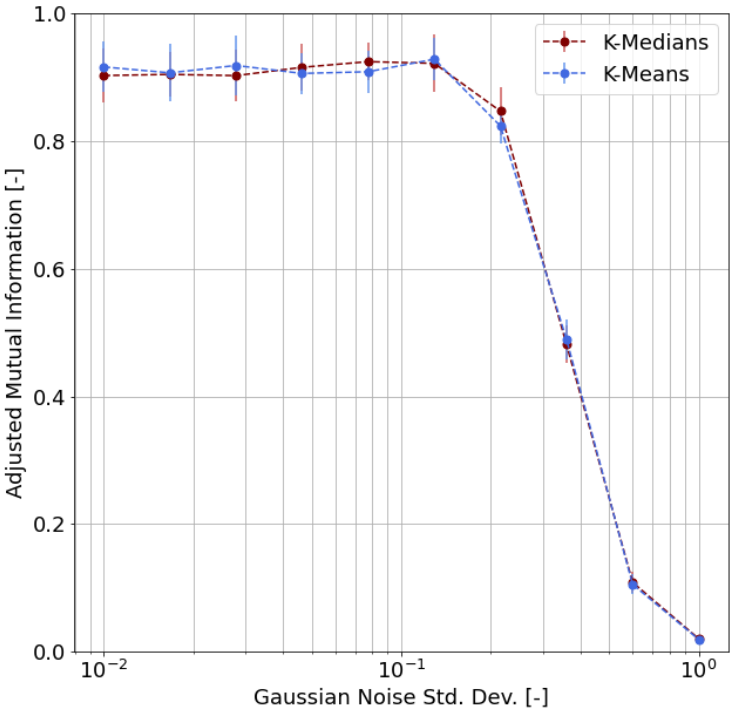

Figure A8. AMI for the simulation with 30 iterations.

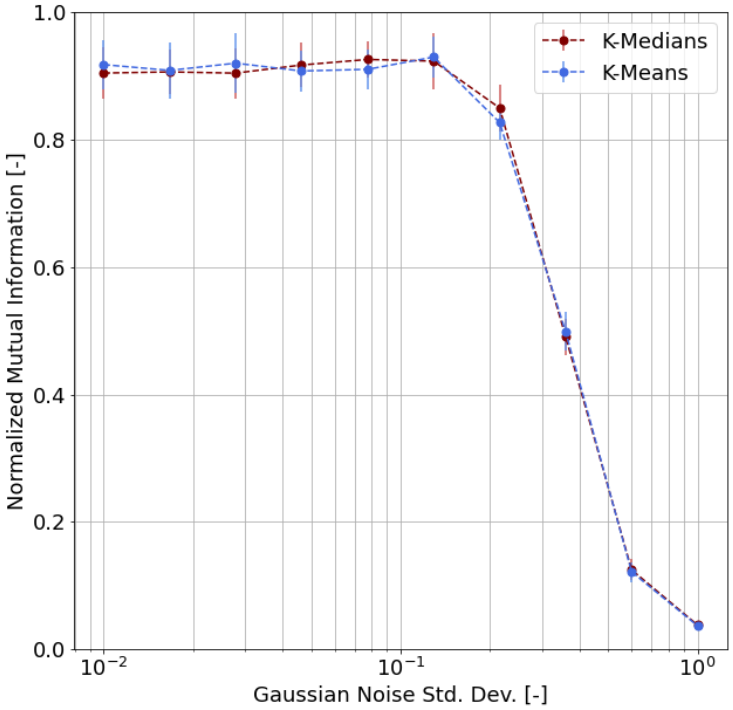

Figure A9. NMI for the simulation with 30 iterations.

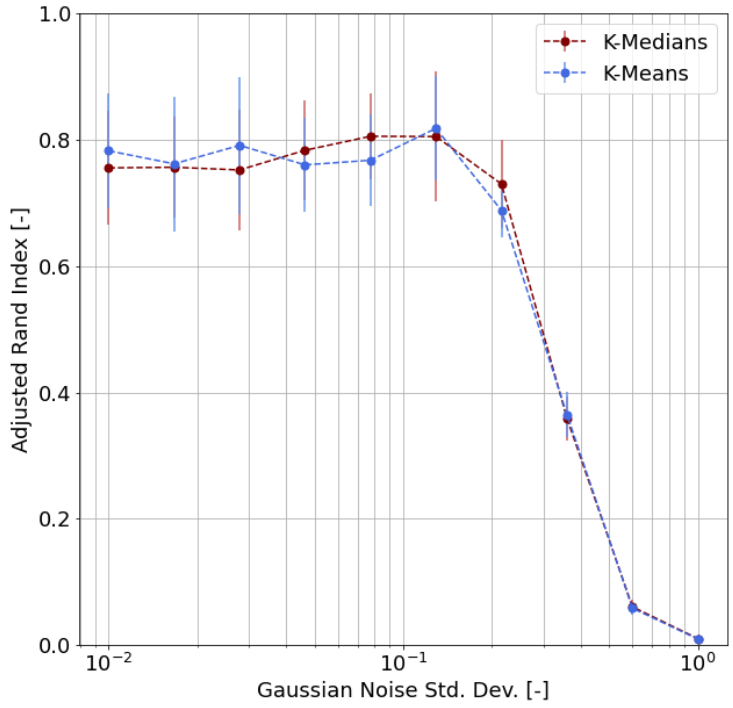

Figure A10. ARI for the simulation with 30 iterations.

As we can see, in both cases, the reference values do not differ considerably from those shown in Figures 19, 21 and 23. In particular, Figures A8–A10 have a lower error bar, but the metric values are perfectly comparable to those reported in the main text, thus amply justifying the choice of 20 as the number of simulations for each experiment.

References

1. Impett, L.; Süsstrunk, S. Pose and Pathosformel in Aby Warburg's Bilderatlas. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 888–902.
2. Aby Warburg Mnemosyne Atlas. Available online: http://www.engramma.it/eOS/core/frontend/eos_atlas_index.php (accessed on 3 February 2021).
3. The Warburg Institute. The Warburg Institute Archive. 2018. Available online: https://warburg.sas.ac.uk/library-collections/warburg-institute-archive (accessed on 26 August 2020).
4. le Fevre Grundtmann, N. Digitising Aby Warburg's *Mnemosyne Atlas*. *Theory Cult. Soc.* **2020**, *37*, 3–26. [CrossRef]
5. Didi-Huberman, G. *L'image Survivante Histoire de l'Art et Temps des Fantômes Selon aby Warburg*; Les Éditions de Minuit: Paris, France, 2002.
6. Becker, C. Aby Warburg's Pathosformel as methodological paradigm. *J. Art Historiogr.* **2013**, *9*, CB1.
7. imgs.ai. Available online: http://imgs.ai/ (accessed on 20 March 2021).
8. Barmpoutis, A.; Bozia, E.; Fortuna, D. Interactive 3D Digitization, Retrieval, and Analysis of Ancient Sculptures, Using Infrared Depth Sensors for Mobile Devices. In *International Conference on Universal Access in Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 3–11.
9. Freedberg, D.; Gallese, V. Motion, emotion and empathy in esthetic experience. *Trends Cogn. Sci.* **2007**, *11*, 197–203. [CrossRef] [PubMed]
10. Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; Sheikh, Y.A. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef] [PubMed]
11. Jenicek, T.; Chum, O. Linking Art through Human Poses. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 1338–1345.
12. Madhu, P.; Villar-Corrales, A.; Kosti, R.; Bendschus, T.; Reinhardt, C.; Bell, P.; Maier, A.; Christlein, V. Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-grounded Style Transfer Learning. *arXiv* **2020**, arXiv:2012.05616.
13. 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 *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 109–125.
14. Hidalgo, G. Openpose. Available online: https://github.com/CMU-Perceptual-Computing-Lab/openpose/ (accessed on 1 May 2020).
15. Scala Archives. Available online: http://www.scalarchives.com/ (accessed on 1 September 2020).
16. Art Resource. Available online: https://www.artres.com/ (accessed on 1 September 2020).
17. Impett, L. Analyzing Gesture in Digital Art History. In *The Routledge Companion to Digital Humanities and Art History*; Brown, K., Ed.; Routledge: Oxfordshire, UK, 2020.
18. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. *arXiv* **2011**, arXiv:1109.2378.
19. Carneiro, G.; Da Silva, N.P.; Del Bue, A.; Costeira, J.P. Artistic Image Classification: An Analysis on the PRINTART Database. *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 143–157.
20. Isekenmeier, G. *Interpiktorialität: Theorie und Geschichte der Bild-Bild-Bezüge*; Transcript Verlag: Bielefeld, Germany, 2014; Volume 42.
21. Heydemann, N.; Dhabi, A. The Art of Quotation: Forms and Themes of the Art Quote, 1990–2010—An Essay. *Vis. Past* **2015**, *2*, 11–64.
22. Impett, L.; Moretti, F. Totentanz. *Operationalizing Aby Warburg's Pathosformeln*; Technical Report; Stanford Literary Lab: Stanford, CA, USA, 2017.
23. Bell, P.; Impett, L. Ikonographie und Interaktion. Computergestützte Analyse von Posen in Bildern der Beilsgeschichte. *Das Mittelalter* **2019**, *24*, 31–53. [CrossRef]
24. Ferrari, V.; Marin-Jimenez, M.; Zisserman, A. Pose search: Retrieving people using their pose. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1–8.
25. Eichner, M.; Marin-Jimenez, M.; Zisserman, A.; Ferrari, V. 2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images. *Int. J. Comput. Vis.* **2012**, *99*, 190–214. [CrossRef]
26. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
27. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
28. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 483–499.
29. Pena, J.M.; Lozano, J.A.; Larranaga, P. An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recogn. Lett.* **1999**, *20*, 1027–1040. [CrossRef]
30. Forgey, E. Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications. *Biometrics* **1965**, *21*, 768–769.
31. Hubert, L.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
32. Vinh, N.X.; Epps, J.; Bailey, J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.* **2010**, *11*, 2837–2854.
33. Fisher, N.I. *Statistical Analysis of Circular Data*; Cambridge University Press: Cambridge, UK, 1995.