DEEP LEARNING BASED IDENTITY VERIFICATION IN RENAISSANCE PORTRAITS

Akash Gupta, Niluthpol C. Mithun, Conrad Rudolph, Amit K. Roy-Chowdhury

University of California, Riverside, CA-92521, USA
{agupta@ece, nmithun@ece., conrad.rudolph@, amitrc@ece.}ucr.edu

ABSTRACT

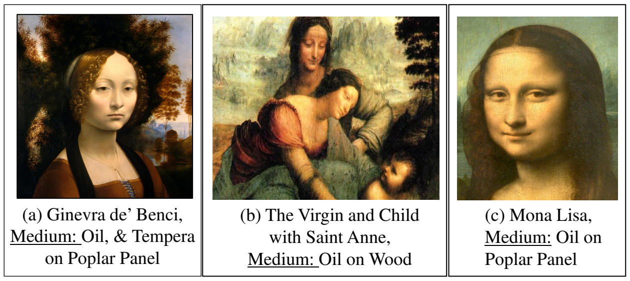
The identity of subjects in many portraits has been a matter of debate for art historians that relied upon subjective analysis of facial features to resolve ambiguity in sitter identity. Developing automated face verification technique has thus garnered interest to provide a quantitative way to reinforce the decision arrived at by the art historians. However, most existing works often fail to resolve ambiguities concerning the identity of the subjects due to significant variation in artistic styles and the limited availability and authenticity of art images. To these ends, we explore the use of deep Siamese Convolutional Neural Networks (CNN) to provide a measure of similarity between a pair of portraits. To mitigate limited training data issue, we employ CNN based style-transfer technique that creates several new images by recasting an art style to other images, keeping original image content unchanged. The resulting system thereby learns features which are discriminative and invariant to changes in artistic styles. Our approach shows significant improvement over baselines and state-of-the-art methods on several examples which are identified by art historians as being very challenging and controversial.

Index Terms— Face Recognition, Art Images, Style Transfer, Siamese Network, CNN, Hypothesis Testing

1. INTRODUCTION

The history of portraiture can be dated since the time people have known art. At least for the first few thousand years, most of the portraits, whether those portraits were drawn, painted, sculpted or cast into death masks, were a depiction of the important people of their time. Apart from being used for a variety of dynastic and commemorative purposes, they were used to depict individuals often to convey an aura of power, beauty or other abstract qualities [1]. The most common subjects for these artworks were the wealthy — mostly royals and nobels— religious and historical figures [2]. However, due to fortunes of time, many portraits tend to lose the identities of their subjects.

From the perspective of the art historian, it is of vital importance to identify the subjects in portraits, as analyzing these portraits can offer significant insight into the personal, social and political aspect of the subject and their period. However, identifying a subject in art portraits is a very complex task since such portraits are usually subject to social and artistic conventions that construct the sitter as a type of their time [3], and results in high ambiguity of the subject identity in many of these portraits. Traditionally, the identification of the subjects in these portraits are limited to the opinion of experts, which is quite often contradictory and it is impossible to resolve disagreements in many cases.



(a) Ginevra de' Benci, Medium: Oil, & Tempera on Poplar Panel

(b) The Virgin and Child with Saint Anne, Medium: Oil on Wood

(c) Mona Lisa, Medium: Oil on Poplar Panel

Fig. 1: Variation in Artistic Style of Leonardo da Vinci. Early 20^{th}-century scholars were vociferous in their disagreement about (a) and (b). (c) is universally accepted as authentic

In this regard, developing computerized face verification technique for art portraits has garnered interest to provide a quantitative measure of similarity evidence to aid the art historians in answering questions regarding subject identity. Although some success has been achieved by these techniques, most of these methods fail to generalize well across portraits in resolving ambiguities due to significant variations in artistic style and the limited availability of authentic images. Recently, deep CNN based approaches have shown remarkable performance in face recognition problems. However, apart from the typical challenges associated with face recognition systems such as variations in pose, expression, illumination, etc., face recognition in portraits comes with additional challenges like large variations in artistic style and degradation of image quality over years [4]. For example, many art portraits might not have visually distinctive features and the visual features of images of same person may be significantly different between image of different style (e.g., a oil on wood portrait compared to a death mask portrait). A few variations in artistic style of Leonardo da Vinci is shown in Fig. 1. Hence, we

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

observe that applying such a CNN trained on natural images for face verification in art images shows poor performance. Moreover, we do not have the luxury of a large database of authentic images to train a CNN directly from scratch. It is extremely challenging task to gather authentic images with the certainty of the subject identity of early modern period as most of the artworks have lost their identity. Thus, let alone training, even fine-tuning a pre-trained model is an uphill battle. As an example, we were able to gather about 400 images of unquestioned authenticity, which comprise an the average of 3 images per subject.

The above challenges prevent traditional CNN-based face recognition systems to achieve state-of-the-art accuracy in art images. In this regard, we train deep Siamese network to learn features which are discriminative and invariant to changes in artistic styles. To mitigate limited training data issue, we employ CNN-based style-transfer technique for data augmentation that creates several new images by recasting a style to other artworks, keeping original image content unchanged. Based on the similarity scores, we perform hypothesis testing for statistical validation. Our approach shows significant improvement over baselines and state-of-the-art methods on several examples which are identified by art historians as being very challenging and controversial.

2. RELATED WORK

Deep convolutional network embedding for face representation is considered the state-of-the-art method for face verification, face clustering, and recognition [5, 6, 7]. The deep convolutional network maps the face image, typically after a pose normalization step, into an embedding feature vector such that features of the same person have a small distance while features of different individuals have a higher distance. Various face recognition techniques have been employed in surveillance and entertainment applications.

Analysis of paintings using sophisticated computer vision tools has gained popularity in recent years [8]. A recent work has explored the application of CNN-based facial image analysis to find a close match of a celebrity image from a database of portrait images [9]. In this work, authors encode both celebrity natural images and art portraits using CNN encoder. Using this encoding, a CNN classifier learns the embedding between the features and returns the top matching results from the retrieval portrait database.

There has also been some work using hand-crafted features for face recognition in art images. It is evident from [10] that while drawing a human body, a lot of emphasis was laid upon maintaining the proportions of various parts. The importance of anthropometric ratios/distances was preserved even during the Renaissance era. According to Da Vinci, in a well-proportioned face, the size of the mouth equals the distance between the parting of the lips and the edge of the chin, whereas the distance from chin to nostrils, from nostrils to eyebrows, and from eyebrows to hairline are all equal, and the height of the ear equals the length of the nose [11].

Authors in [4, 12], exploit this knowledge by using the local features (LF) and anthropometric distance (AD) to learn a feature space, which they term Portrait Feature Space (PFS). This feature space is optimized and subjected to hypothesis testing. However, hand-crafted features have not been able to achieve performance similar to the state-of-the-art CNN methods in other applications, and it is natural to explore their applicability in the domain of art image.

In [13], authors have used cross-spectral hallucination to match NIR (near infrared) to VIS (visible light) face images. This problem is challenging due to the difference in the light spectrum in which the images are taken. The implementation in this work is to learn the mapping of NIR images to VIS images and train a network to generate VIS equivalent image of a NIR image. Using this method, data augmentation is performed to train the classifier. Some researchers have also used cross spectral face recognition to compare images taken in heterogeneous environments [14].

These methods are not applicable to our study since the images in the present scenario are obtained from museums across the world, and we have no control on the kind of sensors used to capture them. Also, we do not have privilege of thousands of authentic art images from early modern period.

3. METHODOLOGY

The aim of this work is to aid art historian to solve lingering ambiguities in work of art by providing a probabilistic measure of similarity by means of state-of-the-art methods. With this work, we want to demonstrate the efficiency of our fine-tuned model, VGG-Art, and compare it with the VGG-16 base model.

3.1. Overview of the Approach

We provide a probabilistic measure of similarity given an image pair, one test image, and another reference, to identify the subject in question. To this extent, we leverage upon Siamese network architecture based on pre-trained model to generate feature vectors for each of the images. The overview of our methodology is depicted in Fig. 2 and Fig. 4. The image pairs are represented as $\{I, I'\}$ pairs which consist of original and style transfer images. The portraits for which there is ambiguity in the subject identity are, henceforth, referred to as the "test images". The artworks for which the subject identity is known are referred to as reference images. Note that the images are considered reference images only if there is absolute confidence in subject identity. To ensure that images are authentic, deliberate efforts have been made while procuring the portraits to train the network.

We learn the similarity metric by fine-tuning the Siamese network over our image pairs. Similarity scores using these

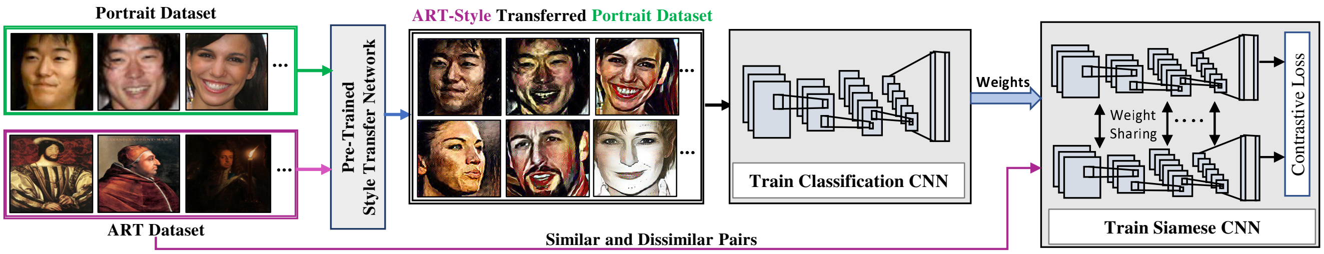


Fig. 2: Brief illustration of our proposed training framework using Siamese Network with Style transfer module.

features are computed for similar and dissimilar pairs. These scores are used to generate Gaussian Distributions for similarity and dissimilarity. We call these distributions as Portrait Feature Space (PFS). The similarity score between the test and the reference image, as indicated in Fig. 4, is analyzed with respect to the learned feature space to derive conclusions of similar or dissimilar image pairs. If both similar and dissimilar score happens to be likely, no decision can be drawn.

3.2. Data Collection and Data Augmentation

For the purpose of creating ART dataset for training, we collected only those art images for which there was absolute confidence of the sitter identity. Authenticity of the images is critical for this application as art historians can rely on our similarity score to solve long-standing ambiguity about the identity of the sitter in many portraits. In such cases, noisy labels while training the network may degrade the performance and may result in high error rate. With deliberate efforts we were able to collect about authentic 400 images from various sources like museums and art historians.

As discussed in section 1.2, variation in artistic style for one sitter by various artists and a limited number of authentic images conflicts with the basic requirement of a large dataset to train CNN. To overcome this hindrance, we employ CNN based style-transfer technique, as discussed in [15], to recast the style of authentic images on an image dataset. We generate about 20k style-transferred images taking 20k face images from VGG dataset and applying styles of our dataset consisting of 400 images. Precautions have been taken to cast the style of all the portraits in our training dataset. Examples of application of style transfer algorithm for a portrait image is presented in Fig. 3.

CNN-based style transfer works by learning the Gram matrix of the style image and content image and minimizing the content loss and style loss by back-propagating the total loss. The tensor which we back-propagate into is the stylized image we wish to achieve, which is called pastiche from here on out. Content loss contains information on how close the pastiche is in content to the content image, and the style loss contains information on how close the pastiche is in style to the style image. The content loss and style loss are added and the total loss is back-propagated through the network to reduce this loss by getting a gradient on the pastiche image. It iteratively changes the content image to look more and more like a stylized content image[15].

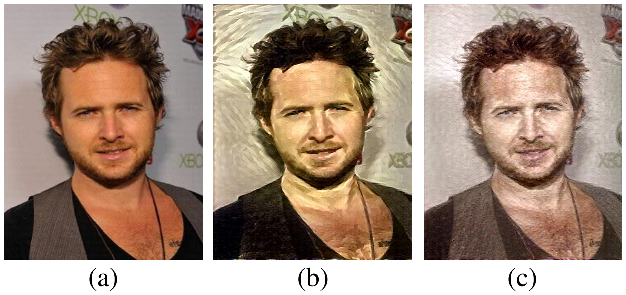

(a) (b) (c)

Fig. 3: Examples of Style Transfer for Data Augmentation. (a) shows an image from the portrait dataset. (b) and (c) shows two augmented version of image in (a), styled by two different art images created by John Singleton Copley.

3.3. Training the Art Image Verification Network

We follow a two-step learning approach for training our network. First, we fine-tune last two layer in VGG-16 Face CNN classifier on 20k styled images of VGG dataset [5]. This is done so that network now learns about different artistic styles. Optimization is done using Stochastic Gradient Decent (SGD) using mini-batches of 64 image pairs and momentum coefficient of 0.9. This model is regularized using dropout ration of 0.5 and weight decay set to 5×10^{-4}. The learning rate was initially set to 10^{-3} and then decreased by factor of 10 when the validation set accuracy stopped increasing. Final model at 45000 iterations is used as base model for the Siamese network.

We ensure that network learns style specific features by using style transfer images on VGG dataset. Since, the pre-trained network base model has knowledge about the original dataset, fine-tuning it with style transferred images learns style specific details related to these images. Second, a Siamese network is trained using contrastive loss in Eq. (1) and margin (α) of 1 is used to learn the similarity metric between the pair of portrait images we gathered.

$$E = \frac{1}{N}\sum_{n=1}^{N}(y)d^2 + (1-y)max(\alpha - d, 0)^2 \qquad (1)$$

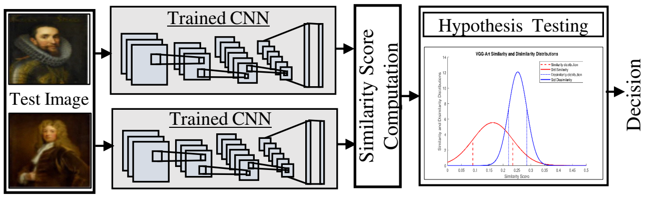


Fig. 4: Our Identification framework for Test Images. Similarity score calculation based on trained CNN, followed by hypothesis testing for final similarity measure. Distributions are drawn with portrait dataset by computing similarity using features from trained model.

where, f_1 and f_2 are feature vectors of the Siamese image pairs, $d = \|f_1 - f_2\|^2$, provides Euclidean distance between the feature vectors and α is the margin value for the contrastive loss. N is the number of image pairs.

For an image pair, we get two feature vectors, one for each image. These feature vectors are then subjected to hypothesis testing. The similarity and dissimilarity distribution obtained over training samples are described in Section 4.

3.4. Similarity and Dissimilarity Score Computation

We make a reasonable assumption that each element in the difference of feature vector is Gaussian. We call the difference of the two feature vectors as Portrait Feature Space projections. By definition, the sum of the square of each element is a Chi-square random variable. Thus we compute Chi-distance as a measure of similarity between two images. Chi-square distance between the images is calculated using the Eq. (2).

$$\chi^2(f_1, f_2) = \sum_{i=1}^{M} \frac{(f_1[i] - f_2[i])^2}{f_1[i] + f_2[i]} \tag{2}$$

where, f_1, f_2 are feature vectors of the Siamese image pair and M is the length of the feature vector.

Intuitively, an image pair whose feature values are very close for many different dimensions are more likely to be the same person. A low value of Chi-distance between images pairs corresponds to high likelihood of image pairs belonging to the same person and a high value implies low likelihood of the image pair to be of same person.

Using the procedure described above, we compute similarity scores between portrait pairs that are known to depict same sitters and different sitters to get similar and non-similar scores respectively. The resulting set of similarity and dissimilarity scores, computed across various artists and sitters, are modeled as two Gaussian distributions (one for similar scores and another for dissimilar scores). The mean and standard deviations of these distributions are estimated from training data. We refer to these similarity and dissimilarity distributions as the "Portrait Feature Space" (PFS).

3.5. Hypothesis Testing

Hyothesis testing is a method for verifying a claim or hypothesis about a parameter in a population [16, 4]. The need of hypothesis testing arises as we need to define how close the similar images and how far the dissimilar pairs are in terms of the feature distance. We cannot guarantee that similar images will always yield zero distance and dissimilar images will be furthest apart. Thus, we resort to Neyman-Pearson hypothesis testing to provide a probabilistic measure of the specific claim — the image pairs are similar or not. Below, we describe it with respect to the learned PFS. Table 1 summarizes Neyman-Pearson hypothesis testing for our model.

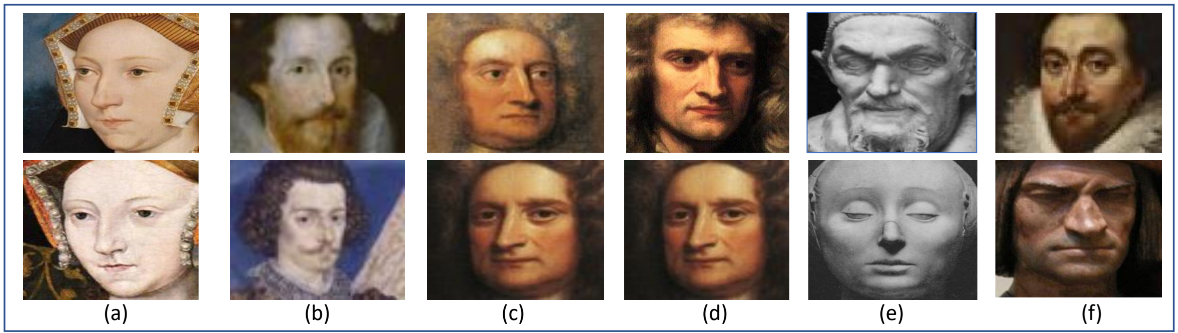
1. Null hypothesis ($\mathcal{H}_0$) claims that the similarity distribution accounts for the image pair test score should be better than dissimilarity distribution.
$$\mathcal{H}_0 : \mu = \mu_S \text{ and } \sigma = \sigma_S$$
2. The alternate hypothesis ($\mathcal{H}_1$) is that dissimilarity distribution models the score better.
$$\mathcal{H}_1 : \mu = \mu_{\bar{S}} \text{ and } \sigma = \sigma_{\bar{S}}$$
3. We calculate the mean and the variance of the training set for similar and non-similar image pair from their distributions.
4. Assuming the distribution to be Gaussian, we compute the likelihood of the Chi-square distance of the image pair for each distribution.
5. We reject the null hypothesis if the likelihood of similarity to dissimilarity is less than 1+ δ, where δ is margin of uncertainty.
$$\rho = \frac{g(x, \mu_S, \sigma_S)}{g(x, \mu_{\bar{S}}, \sigma_{\bar{S}})} < 1 - \frac{\delta}{2}$$
where, $g(x, \mu, \sigma)$ is Normal Distribution with mean = μ, standard deviation = σ and x is Chi-distance.

Table 1: Hypothesis Testing on Portrait Feature Space

Hypothesis Test	Decision
$\rho > 1 + \delta/2$	Similar Pair
$\rho < 1 - \delta/2$	Non-Similar Pair
$1 - \delta/2 < \rho < 1 + \delta/2$	No Decision

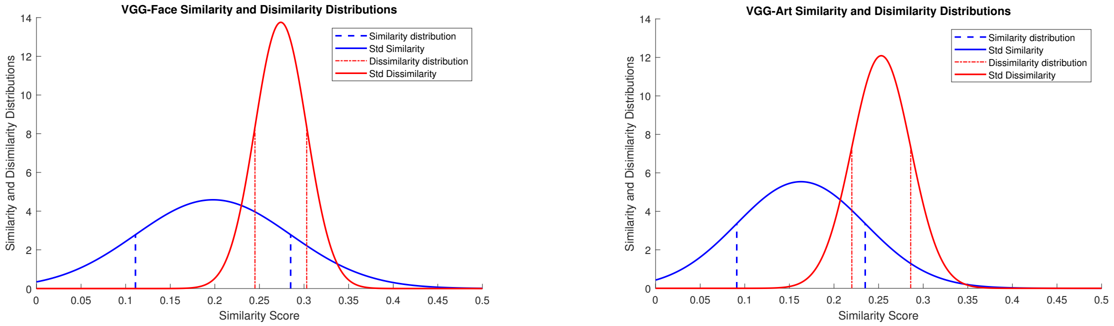
4. RESULTS

We train our model combining collected actual and style transferred image pairs. We randomly chose 80% pairs for training and 20% pair for testing. Using the training data, we get the the similarity and dissimilarity distribution using VGG-Art and VGG-Face as shown in Fig. 6. The mean and standard deviation for both the models is listed in the Table 2



GT	Match	Match	Match	Match	Non-Match	Non-Match
Decision	Match	Match	Match	Non-Match*	No Decision	Non-Match
Score	0.866	0.825	0.729	0.26	0.52	0.228

Fig. 5: The figure shows 6 pairs from the test set of ART dataset and the decision obtained by our VGG-Art Model. GT indicates the ground truth decision. We also report similarity scores and decisions obtained by our method. It can be seen from the figure that our system was successful in arriving at correct decision most of the cases. *(d) shows one of our failure cases - it is a pair of Newton's portraits with about 10 years of age difference.



(a) VGG-Face: Similarity and Dissimilarity Distributions

(b) VGG-Art: Similarity and Dissimilarity Distributions

Fig. 6: VGG-Face and VGG-Art: Similarity and dis-similarity distributions obtained from our ART dataset. It is evident from the figures that ours VGG-Art model has better separation between the distributions than state of the art VGG-Face. Hence, our network is more reliable in verifying faces in art images.

Fig. 5 shows qualitative results for some similar and dissimilar pairs from ART dataset based on VGG-Art model. As shown in Fig. 5, we are able to predict the similar images well. In case (c) and (d) of the Fig. 5, sitter in both the portrait pairs is Sir Issac Newton, however, our method classify case (c) correctly but case (d) incorrectly. It may be due to the fact the portraits have a age difference of about 10 years.

Table 2: Mean and Standard Deviation for VGG-Face and VGG-Art Distributions

	VGG-Face		VGG-Art	
Distributions	Mean	Std	Mean	Std
Similarity	0.198	0.087	0.163	0.072
Dissimilarity	0.274	0.029	0.253	0.033

In both the models for VGG-Face and VGG-Art, we can see that there is some overlap between the similarity and dissimilarity distributions. However, the overlap in VGG-Art distributions is significantly less as compared to VGG-Face. Hypothesis testing on portrait validation dataset with VGG-Art model has shown accuracy of 91.253 % whereas the accuracy of VGG-Face was found to be 87.29%. This is significant improvement over the portrait image dataset and can help art historians to solve many long-standing ambiguity of sitter identity in some of the portraits. The comparison of Similarity Score for similar and dissimilar for VGG-Art and VGG-Face is given in the Fig. 5. A complete list of similarity scores for image pairs in training, validation and test datasets will be posted on the our project site.

5. CONCLUSIONS

We present a work that focuses on developing method for face verification in art images. Due to limited availability of authentic art images, it is not possible to directly train deep CNNs using these images. In this regard, we employ a style transfer network which generates a large pool of pseudo art images from portraits by transferring style of art images. Utilizing these style transferred images, we start training our art face verification network. Finally, our network is fine-tuned using the art images. Subsequently, the similarity metric for similar and dissimilar pairs is learned using Siamese network and Chi-distance is computed for similarity score of the image pair. Similarity and dissimilarity distributions are derived from the training set of 400 portrait images and hypothesis testing is done on validation and test image set. Our fine-tuned network out-performs the state-of-the-art VGG-Face model on the art image data set by a significant margin. We believe that our approach can be used to provide a source of complementary evidence to the art historians in addressing questions such as verifying the identity of uncertain subjects in art images.

6. ACKNOWLEDGMENT

This work was supported by a grant from the Samuel H. Kress Foundation.

7. REFERENCES

[1] Shearer West, *Portraiture*, Oxford University Press, 2004.

[2] Paul Heaston, “The history of portraiture,” 2013.

[3] Jia Li, Lei Yao, Ella Hendriks, and James Z Wang, “Rhythmic brushstrokes distinguish van gogh from his contemporaries: findings via automated brushstroke extraction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1159–1176, 2012.

[4] Ramya Srinivasan, Conrad Rudolph, and Amit K Roy-Chowdhury, “Computerized face recognition in renaissance portrait art: A quantitative measure for identifying uncertain subjects in ancient portraits,” *IEEE Signal Processing Magazine*, vol. 32, no. 4, pp. 85–94, 2015.

[5] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al., “Deep face recognition.,” in *Proceedings of the British Machine Vision Conference*, 2015, vol. 1, p. 6.

[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[7] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[8] Christopher W Tyler, William AP Smith, and David G Stork, “In search of leonardo: computer-based facial image analysis of renaissance artworks for identifying leonardo as subject,” in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 82911D–82911D.

[9] Elliot J Crowley and Andrew Zisserman, “In search of art,” in *Workshop at the European Conference on Computer Vision*. Springer, 2014, pp. 54–70.

[10] Alexander Perrig, “Drawing and the artists basic training from the 13th to the 16th century,” *The Art of the Italian Renaissance. Architecture, Sculpture, Painting, Drawing, Cologne/Germany*, pp. 416–441, 2007.

[11] Florine Vegter and J Joris Hage, “Clinical anthropometry and canons of the face in historical perspective,” *Plastic and reconstructive surgery*, vol. 106, no. 5, pp. 1090–1096, 2000.

[12] Conrad Rudolph, Amit Roy Chowdhury, Ramya Srinivasan, and Jeanette Kohl, “Faces: Faces, art, and computerized evaluation systems–a feasibility study of the application of face recognition technology to works of portrait,” *Artibus et historiae: an art anthology*, , no. 75, pp. 265–291, 2017.

[13] José Lezama, Qiang Qiu, and Guillermo Sapiro, “Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 6807–6816.

[14] Nathan D Kalka, Thirimachos Bourlai, Bojan Cukic, and Lawrence Hornak, “Cross-spectral face recognition in heterogeneous environments: A case study on matching visible to short-wave infrared imagery,” in *Proceedings of the International Joint Conference on Biometrics*. IEEE, 2011, pp. 1–8.

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, “A neural algorithm of artistic style,” *arXiv preprint arXiv:1508.06576*, 2015.

[16] Malcolm O Asadoorian and Demetrius Kantarelis, *Essentials of inferential statistics*, University Press of America, 2005.