DEEP LEARNING BASED IDENTITY VERIFICATION IN RENAISSANCE PORTRAITS

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

University of California, Riverside

ABSTRACT

The identity of subjects in many portraits has been a matter of debate by art historians that relied on subjective analysis of facial features. Developing automated face verification technique has thus garnered interest to provide a quantitative way to reinforce the decision arrived 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 a style to other artworks, 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 in Art Images, Style Transfer, Siamese Network, 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 royalty and nobility— religious and historical figures [2]. However, due to fortunes of the time, many portraits tend to lose the identities of their subjects.

From an art historians perspective, it is of vital importance to identify the subject 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.

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 variation in artistic style and the limited availability of authentic images. Recently, deep Convolutional Neural Network (CNN) based approaches have shown remarkable performance in face recognition problem, which require millions of training images to train CNNs. However, apart from the typical challenges associated with face recognition systems such as variations in pose, expression, illumination, etc., face recognition in portraits come with additional challenges [4]. Hence, we observe that applying such a CNN trained on natural images for face verification in art images shows poor performance due to large variation in artistic style in art images and degradation of image quality over years etc. 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 variation in artistic style of Leonardo da Vinci is shown in Fig. 1. On the other hand, we do not have the luxury of 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 Renaissance period as most of the artworks have lost their identity. Thus, let alone training but even fine-tuning a pre-trained model is an uphill battle. With our efforts, we were able to gather about 400 images, which comprise of the average of 3 images per subject.

The above challenges prevent traditional CNN face recognition system 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 considerable distance. Various face recognition techniques have been employed in surveillance and entertainment applications. However, these methods can surpass human performance only for the images under constrained environment.

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 find a close match of a celebrity image from a dataset of portrait images [9]. In this work, authors encode both celebrity natural images and paintings 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 dataset. The task of image retrieval from a known set of portrait images is not challenging as compared to face verification problem in portrait images, most of which may not be part of our training set.

There has also been some work using the handcrafted feature for face recognition in art images. It is evident from [10] that while drawing a human body, a lot of emphases 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], exploits this knowledge by using the local features (LF) and anthropometric distance (AD) to learn feature space, as they call it Portrait Feature Space (PFS). This feature space is optimized and subjected to hypothesis testing. However, with the state-of-the-art CNN methods, the use of handcrafted feature becomes obsolete due to full automation and high reliability of CNN.

Some researchers have used cross spectral face recognition to compare images taken in heterogeneous environments [13]. These methods are not applicable to our study since the images in the present scenario are obtained from museums across the world, we have no control on the kind of sensors used to capture them.

In [14], 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 work in this paper follows the same principle to learn the style of one portrait and recast it on another.

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

The aim of this work is to provide a probabilistic measure of similarity between 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 VGG-Face descriptor model to generate feature vectors for each of the images. The overview of our methodology is depicted in Fig. 2. 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 portrait to training the network.
We learn the similarity metric by fine-tuning the Siamese network over our image pairs. Similarity scores using these 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 by the green dotted line in testing part of Fig. 2, 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, then no decision can be drawn.

### 3.2. Data Collection and Data Augmentation

Authors have been cautious while collecting art images. This is very important for our application as it is critical for art historians who can rely on our similarity score to solve long-standing ambiguity about the identity of the sitter in many portraits. We were able to collect about 400 images from various sources, which were authenticated by an art historian on our team. 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 a face 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 onto face dataset. Application of style transfer algorithm for a 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 when the total loss is back-propagate through the network to reduce this loss by getting a gradient on the pastiche image, it iteratively change the content image to look more and more similar.
like a stylized content image[15].

3.3. Fine-tuning the VGG-16 and Siamese network

We follow a two-step learning approach. First, we fine-tune last four two layer in VGG-16 classifier on 20k styled images of VGG dataset. 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 103 and then decreased by factor of 10 when the validation set accuracy stopped increasing. Final model at 45000 iteration is used as base model for Siamese. We ensure that network learns style specific features by using style transfer images on VGG dataset. Since, the VGG network base model has knowledge about the original VGG dataset, fine-tuning it with style transferred images learns style specific details related to these images. Then a Siamese network is trained using contrastive loss Eq. (1) and margin of 1 is used to learn the similarity metric between the pair of portrait images we gathered. For an image pair, we get two feature vectors one for each image. The size of each feature vector is 4096. These feature vectors are then subjected to hypothesis testing. The similarity and dissimilarity distribution obtained over training samples are described in Section 4.

\[ E = \frac{1}{N} \sum_{n=1}^{N} (y)d^2 + (1 - y)\max(margin - d, 0)^2 \]  

where, \( d = \|a - b\|^2 \).

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. Also, we computed the rank-1 count similarity as described in [16]. Rank-1 count (\( R \)) is the number of feature dimensions in which the two images are closer in value than the first image is to any of a set of reference images. In our case, we do not have a set of reference images. Since we are dealing with face verification problem, we computed mean of the PFS projections for similar pairs and dissimilar pair. The two mean PFS projections are considered references for computing Rank-1 count similarity as described by Eq. (2). Intuitively, an image pair whose feature values are very close for many different dimensions are more likely to be the same person.

\[ R = \sum_{i=1}^{4096} I[S_i + \sigma S_i < |A_i - B_i| < S_i + \sigma S_i - \sigma S_i] \]

where, \( I[\cdot] \) is indicator function, \( A, B \) are image pairs, \( S \) and \( S \) are mean of similarity and dissimilarity of PFS projection, respectively and \( \sigma \) is standard deviation.

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

This is a method for testing a claim or hypothesis about a parameter in a population [17, 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. Below, we summarize it with respect to the learned PFS. Tabular summary of the same is described in Table 1.

<table>
<thead>
<tr>
<th>Image Pair</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity Score</td>
<td>Dissimilarity Score</td>
</tr>
<tr>
<td>( \rho &gt; \alpha )</td>
<td>( \rho &lt; \alpha )</td>
</tr>
<tr>
<td>( \rho &lt; \alpha )</td>
<td>( \rho &gt; \alpha )</td>
</tr>
<tr>
<td>( \rho &gt; \alpha )</td>
<td>( \rho &gt; \alpha )</td>
</tr>
<tr>
<td>( \rho &lt; \alpha )</td>
<td>( \rho &lt; \alpha )</td>
</tr>
</tbody>
</table>

1. Null hypothesis claims that the match distribution accounts for the tests similarity score with reference better than non-match distribution. The alternate hypothesis is that non-match distribution model the score better.

2. We set the level of significance \( \alpha \), i.e., the tests probability of incorrectly rejecting the null hypothesis, as 0.05, as per behavioral research standard.

3. We compute the test statistic using one independent non-directional \( z \) test [15], which determines the number of standard deviations the similarity score deviates from the mean similarity score of the learned match/non-match distributions.
Fig. 4. VGG-Face and VGG-Art : Similarity Score for Image pairs in Test Set. (a),(b) - Score for Similar and Dissimilar Image pair for VGG-Face and (c),(d) - Score for Similar and Dissimilar Image pair for VGG-Arts. We can compare the number of incorrect classifications (red in case of Similar pairs and blues in case of Dissimilar Pairs) in (a)-(d) and see that VGG-Arts performs better than VGG-Face. Accuracy of VGG-Art is 92% as compare to 87% accuracy in VGG-Face.

4. We compute p values which are the probabilities of obtaining the test statistic that was observed, assuming that the null hypothesis is true. If $\rho < \alpha$, we reject the null hypothesis.

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. Fig. 4 describes the similarity score for similar and dissimilar pairs and indicates how state-of-the-art VGG-Face and VGG-Art performs. All red points are incorrectly classified, blue are correctly classified and no decision can be made for green pairs. List of images and label attached in supplementary materials.

Using the training data, we get the similarity and dissimilarity distribution using VGG-Art and VGG-Face model as shown in Fig. 5 and Fig. 6. The mean and standard deviation for both the models is listed in the Table 2.

| 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 distribution, we can see that there is some overlap between the similarity and dissimilarity distributions. However, the overlap in VGG-Art distributions is smaller as compare to VGG-Faces. Hypothesis testing on portrait validation dataset with VGG-Art model have 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 historian 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. 4.

5. CONCLUSIONS

We presented a work that builds upon state-of-the-art face recognition system for face verification in art images. After
fine-tuning the VGG-16 network using images generated by style transfer the network learns features specific to artistic style. Subsequently, the similarity metric for similar and dissimilar pairs is learned using Siamese network algorithm and chi-distance is computed for similarity score of the image pair. Similarity and Dissimilarity distribution 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-16 model on the portrait data set. We believe that these results can serve as a source of complementary evidence to the art historians in addressing questions such as verifying authenticity, recognition of uncertain subjects in art images.

6. REFERENCES


[9] Elliot J Crowley and Andrew Zisserman, “In search of art.,” in *ECCV Workshops (1)*, 2014, pp. 54–70.


