White Paper Report

Report ID: 109235

Application Number: HD-51735-13

Project Director: Conrad Rudolph (conrad.rudolph@ucr.edu)

Institution: University of California, Riverside

Reporting Period: 5/1/2013-5/31/2015

Report Due: 8/31/2015

Date Submitted: 5/24/2015

FACES Phase II: White paper

FACES: Faces, Art, and Computerized Evaluation Systems, Phase II

A feasibility proposal for the use of face recognition systems in the identification of unidentified works of portrait art

Conrad Rudolph, Professor, Department of the History of Art, University of California, Riverside
Amit Roy-Chowdhury, Professor, Department of Electrical Engineering, University of California, Riverside
Jeanette Kohl, Associate Professor, Department of the History of Art, University of California, Riverside
Ramya Srinivasan, Graduate Student Researcher, University of California, Riverside

Before the nineteenth century, most portraits were, almost by definition, depictions of people who were important in their own worlds. But, as a walk through almost any major art museum will show, a large number of these portraits from before the nineteenth century have lost the identities of their subjects through the fortunes of time. Traditionally, identification of many of these portraits has been limited to often quite variable personal opinion. FACES (Faces, Art, and Computerized Evaluation Systems) proposes to establish the initial parameters of the application of face recognition technology to works of portrait art--this highly subjective aspect of art history--while at the same time retaining the human eye as the final arbiter.

During this grant period, FACES began work establishing these parameters, asking such questions as: is face recognition technology, originally designed for actual (that is, photorealistic) human faces, applicable to works of portrait art, which are subject to a process of visual interpretation on the part of the artist? Which of the many different face recognition techniques should be used? Which functions of the many functions of a given technique would apply most effectively to our subjects? What culture and period would work best in this initial stage of testing? What types of portraits would best be used, sculpture (three-dimensional) or painting and drawing (two-dimensional)--or both? How will the identifying characteristics in a portrait of one sitter by an artist with a distinctive artistic style compare to a portrait of the same sitter but by a different artist who also has a distinctive style? If face recognition technology works with sculpture, will the identical process be able to be used for painting and drawing? If face recognition technology operates best with a straight-on view of the subject, how will the rate of successful tests be with three-quarter view portraits, the standard pose for portraits in early modern Western culture? For two-dimensional works, will the medium--oil painting, tempera, pencil, chalk, engraving, and so on--affect the test results differently? What about copies of portraits (for example, of famous sitters, like Isaac Newton) and copies of copies--how closely will they retain the identifying characteristics found in the original portrait? What about extreme or poor lighting in painting and drawing? What about aging as found in multiple portraits of the

same sitter made over a long period of time? By the same artist? By different artists? What about the vary artistic ability of the individual artist?

In the course of initial investigation, it gradually became clear that of all the different methods of face recognition technology, two gave the most dependable results: the computation of anthropometric distances and of local features. These two methods were part of a larger, more complex process we call the FACES algorithm (detailed below).

While the FACES algorithm was constantly developed throughout the course of this two year project, we began by testing the death mask of a known individual against an identified sculptural portrait of the same individual. That is, we tested an analogue--an unmediated image of the subject, not a work of art--against the image of a three-dimensional work of art that, in this case, physically approaches the subject in form and size but that nevertheless partakes of the subjectivity of artistic interpretation.

We then left the relative security of the analogue and work-of-art pairing, and tested paradigms of exclusively three-dimensional works of art--that is, we then tested two works both of which were now subject to the subjectivity of artistic interpretation. (We use the term paradigm here to mean a logically chosen body of related images directed toward a particular demonstrative end.) More specifically, we tested a sculptural portrait of a known individual with another sculptural portrait of the same individual, both around the same stage of the individual's life and both depicted by the same artist--in other words, we proceeded with as much control over variables as possible.

Incrementally, we broadened our tests--too involved to fully detail here--introducing a similarly controlled but wide-ranging series of systematically chosen variations extending from more controlled paradigms to less controlled (that is, more challenging) ones. These included the same stage of an individual's life but by different artists, different stages of an individual's life by the same artist, and different stages of an individual's life by different artists--all in three-dimensional imagery.

Then we tested two-dimensional imagery, first simply comparing two two-dimensional images of the same subject by the same artist, and then mixing media by testing a number of sculpture vs painting (that is, three-dimensional vs two-dimensional) paradigms, employing a systematic series of distinctions similar to those already mentioned (different ages, different artists, and so on). Finally, we tested a few identified portraits of individuals against unidentified ones.

In the first year of FACES, we established proof of concept. Practically speaking, this means that we identified the issues, established the basic methodology (even if not fully worked out yet), and applied this methodology to a particular set of paradigms.

In the second year, we developed the optimum feature set (the most effective body of identifying facial features, given the unique demands of portrait art), expanded the gallery of images with which establish non-match averages (that is, a standard with which to compare a given image under investigation), and continued to work on the problems of angle views and aging. But our work increasingly focused on the questions of the degree of influence on a portrait of the style of the individual artist. For example, a given artist generally tends to render the same detail in the same way, even in an individualized portrait. And so individual artistic style was investigated through a close and systematic study of a large number of portraits of different sitters by the same artist in order to model--that is, to teach the computer--the individual style of the artist.

We also rose to a new level of testing in the application of our newly worked out methods to a series of interesting and sometimes important "identifications." By "identification" I sometimes mean the actual identity of the subject (for example, Mary Queen of Scots) and sometimes merely the ordering of the material into "identities": group X, group Y; Lord X, Lady Y; and so on. This is not the place to go into any detail about these identification attempts, except to say that all of them were important and some of them exciting. Although all paradigms were not conclusive--sometimes for very complex reasons--some of the best known works include: what appears to be the earliest known likeness of Galileo Galilei, painted perhaps around 1590; Nicholas Hilliard's *Young Man Among Roses* (c. 1588), said to be perhaps the most famous miniature ever painted;" the only known "portrait" of Anne Boleyn, the "Moost Happi" medal; the tangled web of portraits that different proponents have claimed at one time or another portray William Shakespeare; a portrait said to represent Mary Queen of Scots (National Portrait Gallery, London, NPG 96); and a portrait that is thought by some to be of James Scott, duke of Monmouth, first duke of Buccleuch, and illegitimate son of Charles II, lying in bed with the covers pulled up to his chin, apparently in order to conceal the fact that James's head has been cut off and--at least in the painting--put back on again (NPG 1566).

The FACES algorithm

Put as succinctly as possible, the FACES algorithm works as follows.

We believe that there are not any off- the- shelf modules that can be used directly and it may be required to consider a combination of different methods. Establishing these methods necessitates a careful study of artists' renditions so as to be able to extract maximum relevant information and model their styles.

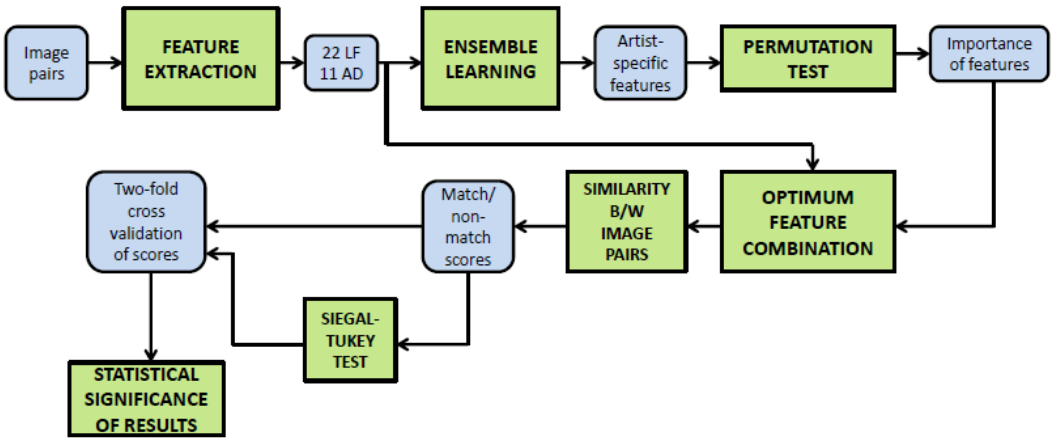

Figure 1: Overview of the Algorithm

Figure 1 provides an overview of the algorithm. Local features (LF) and anthropometric distance (AD) feature descriptors are extracted from each face. A subset of these features is identified to be characteristic of artists' styles by means of the random subspace ensemble learning method. Non-parametric statistical permutation tests give the importance of selected features. Wherever enough images of an artist are not available to learn specific style, all the features are used.

Similarity scores between image pairs are computed using the weighted features/all features to yield style-specific/general match/non-match scores as appropriate. These scores are then validated (using the robust Siegel-Tukey non-parametric statistical test for artist specific case). The learned similarity scores for the general case referred to as the Portrait Feature Space (PFS) are used for identification of unknown instances using the statistical hypothesis tests. We describe the details below.

1. Details of the Algorithm

Once we have obtained the image descriptors (local features across fiducial points and salient anthropometric distances as mentioned in Sec 1 of Report 1) that characterize a face image, we wish to learn which of these are characteristic of an artist's style. In other words, we want to learn a subset of these image descriptors characterizing an artist's renditions. Towards this we employ the random subspace ensemble learning method (This consists of several classifiers (classifier gives a label to an object and thus helps in categorization) and outputs the class based on individual classifiers). For convenience, in the rest of this report, we refer to the local feature image descriptors and the anthropometric distance image descriptors as features.

1.1 Random Subspace Ensemble Learning

The random subspace method randomly samples a subset of these features and performs training in this reduced feature space. Multiple sets (or bags) of randomly sampled features are generated, and for each bag the parameters are learned. This method is capable of handling deficiencies of learning in small sample size and has superior performance than a single classifier [1].

More specifically, we are given, say, Z training image pairs and D features. Let L be the number of individual classifiers in the ensemble. We choose $d_i<D$ (without replacement) to be the number of features to be used in i^{th} classifier. For each classifier, we determine the match and non-match scores (as appropriate) using the d_i features to obtain LF and AD similarity scores using appropriate measures as follows.

$$s(I, I') = \frac{1}{d_i} \sum_{n=1}^{d_i} S_n(I, I')$$

Where S_n (I, I') is any normalized similarity measure computed between image pairs I, I'.

In order to identify features that give the highest separation between match and non-match scores, we then compute the Fisher Linear Discriminant function for each classifier (as described in Step 3 of Sec 2 in Report 1). We choose the union of features from those classifiers that give the top k Fisher Linear Discriminant values as our style features; k chosen experimentally. It is to be noted that we select the style features separately for both local features and anthropometric distances.

1.2 Importance of the Chosen Features

Not all features identified by the above method are equally important in representing a style. In order to understand the importance of the chosen features, we consider the non-parametric permutation statistical test [2]. Permutation tests helps in assessing what features are same (in other words invariant) across all the instances belonging to a class. Thus, features which are more invariant across the instances of the class can be perceived to be more characteristic of the

class and thus be assigned greater importance. Permutation tests have been applied to determine invariant features in artworks such as in [3].

Permutation test

The null hypothesis H_0 (in statistics null hypothesis refers to a default scenario) is chosen to indicate that two image groups G1 and G2 have the same average value (μ) in a particular feature *v*; the alternate hypothesis H_1 indicating that the average value of that feature is different in the two groups. Thus,

$$H_0 : \mu_{G1} = \mu_{G2}; H_1 : \mu_{G1} \neq \mu_{G2}$$

If the null hypothesis is true, then it should not matter when this feature *v* is randomly assigned among images in the group. For instance, there is a certain way that the mouth corner looks when a person smiles. On an average, if this appearance is same across all images and across groups, then the principle behind this test is that there will not be a significant difference if the mouth tips are randomly assigned across images in the group (i.e. assigning the feature of one person to the corresponding feature of another person). Thus, if there are many images of an artist by depicting different sitters, this test essentially captures important features that are invariant across the works of the artist.

Specifically, if there are N_s images of a style class S, then we can divide these N_s images into 2 subgroups consisting of N_{s1} and N_{s2} images. Let the feature values for the first group be $[v_1, v_2, ..., v_{Ns1}]$ and in second group be $[v_{Ns1+1}, ... v_{Ns_1+s_2}]$. The two sided permutation test is done by randomly shuffling $[v_1, ..., v_{Ns}]$ and assigning the first N_{s1}values, say, $[v_{(1)}, v_{(2)}, ..., v_{(N(s_1)}]$ to the first group and the remaining N_{s2} values $[v_{N(s1+1)}, ..., v_{(Ns1+s2)}]$ to the second group.

For the original two groups we compute,

$$\delta_0 = \left| \frac{1}{N_{s_1}} \sum_{i=1}^{N_{s_1}} v_i - \frac{1}{N_{s_2}} \sum_{i=1}^{N_{s_2}} v_{N_{s_1}+i} \right|$$

δ_0 denotes the variation of the feature v as exhibited by the various image instances I_i in the 2 groups under consideration.
For any two permuted groups we compute,

$$\delta_s = \left| \frac{1}{N_{s_1}} \sum_{i=1}^{N_{s_1}} v_{(i)} - \frac{1}{N_{s_2}} \sum_{i=1}^{N_{s_2}} v_{(N_{s_1}+i)} \right|$$

δ_s denotes the variation in the feature v of style class S after assigning the feature as depicted by I_i, i=1, 2,...l to an image not necessarily of I_i

This value obtained from the permutation test referred to as the *p* value in statistics community, reflects the variation of the feature in the two groups. It is given by the number of times $\delta_s > \delta_0$. Smaller *p* denotes stronger evidence against the null hypothesis, meaning that the feature differed considerably in the two groups. If a certain feature showed no difference in the 2 groups, then it does not matter to which image this feature is associated since the average value does not change; thus it can be considered as a random assignment into any image in the pool. We compute *p* values for each feature (chosen by the random subspace method) as described above and use them as weights in computing the similarity scores between the image pairs. Thus the p-normalized similarity scores s_p (I,I') is now given by

$$s_p(I, I') = \frac{1}{M} \sum_{v=1}^{M} p_v S_v(I, I')$$

Where p_v is the p value of the feature v as determined by the permutation test and M is the number of features as chosen by the random subspace method.
Subsequently the p-normalized similarity scores from the two measures (LF/AD) are fused in an optimal manner as described in Report 1 (Sec 2).

1.3 Validation of the Style Features

Our goal here is to verify that given match/non-match scores obtained from style features of the class and given match/non-match scores obtained using all features (independent of the style), to show that there is a higher confidence associated with style-specific scores than with the latter case. In other words, we wish to show that style-specific similarity score are better representations of the style class than the similarity scores obtained using all features. Towards this, we employ a robust non-parametric statistical test called the Siegel-Tukey test that basically checks the null hypothesis that two independent score sets come from the same population (style) against the alternative hypothesis that they come from populations differing in variability or spread. If the style features are indeed good representations of the class, then there should be a higher level of confidence associated with the null hypothesis when compared with style independent features.

The principle behind this test is based on the following idea--Suppose there are two groups A and B with ***n*** observations (in our case similarity scores) for the first group and ***m*** observations for the second (so there are $N = n + m$ total observations). If all *N* observations are arranged in ascending order, it can be expected that the values of the two groups will be mixed or sorted randomly, if there are no differences between the two groups (following the null hypothesis). This would mean that among the ranks of extreme (high and low) scores, there would be similar values from Group A and Group B. If, say, Group A were more inclined to extreme values (alternate hypothesis), then there will be a higher proportion of observations from group A with low or high values, and a reduced proportion of values at the center. Thus the *p* values of this test provide a measure of the confidence of the learned style-specific similarity scores.

For artists/images where style could not be learnt, we use all the features (LF/AD) in computing the similarity scores (match/non-match). The learned similarity scores (match and non-match)

were used to construct the Portrait Feature Space (distribution of match/non-match scores). The PFS was then validated using the same procedure as mentioned in Report 1.

1.4 Identification Framework

It is to be noted that the similarity scores obtained using the style learning algorithm described above are associated with greater confidence than the ones obtained in Phase 1. Thus, identity verification is more robust. The method for identity verification is similar to that described in Report 1 and is included here for completeness.
Given the learned PFS, the question now is to verify an unknown test image against a reference image. Towards this, we employ hypothesis testing.

Hypothesis Testing

This is a method for testing a claim or hypothesis (in this case that of a match/non-match between portrait pairs) [5]. Below, we summarize it with respect to the learned PFS in arriving at the conclusion for a match.
1. Null hypothesis claims that the match distribution accounts for the test's similarity score (with reference) better than non-match distribution. The alternate hypothesis is that non-match distribution models the score better.
2. We set level of significance α (test's probability of incorrectly rejecting the null hypothesis) as 0.05, as per common practice in such problems.
3. We compute the test statistic using one independent non- directional z test [5], which determines the number of standard deviations the similarity score deviates from the mean similarity score of the learned distributions
4. We compute *p* values which are the probabilities of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If $p<\alpha$ we reject the null hypothesis.

Identity Verification

In order to examine the validity of the chosen approach, we consider similarity scores of the test image with artworks known to depict different persons other than the one depicted in reference image. We call these images as distracters. Depending on availability, we choose similar works by the same artist (artist of reference image) as distracters. If a test image indeed represents the same subject as in the reference image, not only should its score with the reference image be modeled through match distribution, but also its scores with distracter faces should be modeled by non-match distribution.

Analysis Scenarios

We computed similarity scores of test cases with corresponding reference image and with 10 distracters. Table 1 lists various hypothesis test scenarios that can arise [5] and the corresponding conclusions that one can infer. Match and non-match cases are straight forward to infer from Table 1. In cases where both match and non-match distributions are likely to account for the test data in the same way, it can be said that the learned PFS cannot accurately describe the test data (black rows in Table 1). If either match or non-match distribution is more likely to account for

both test as well as distracters (magenta rows in Table 1), it can be inferred that the chosen features do not possess sufficient discriminating power to prune outliers. Thus in these scenarios, it is not possible to reach any conclusion.

Reference		Distractors		Conclusion
Match	Non-match	Match	Non-match	
$p > \alpha$	$p \leq \alpha$	$p \leq \alpha$	$p > \alpha$	Match
$p \leq \alpha$	$p > \alpha$	$p < \alpha$	$p \geq \alpha$	Non Match
$p > \alpha$	$p > \alpha$	NA	NA	No decision
$p < \alpha$	$p < \alpha$	NA	NA	No decision
$p > \alpha$	$p \leq \alpha$	$p > \alpha$	$p \leq \alpha$	No decision
$p \leq \alpha$	$p > \alpha$	$p \leq \alpha$	$p > \alpha$	No decision

References:
[1] T. Ho, *The random subspace method for constructing decision forests*, IEEE. Trans. on Patt. Anal. and Mach. Intell., vol 20, no 8, pp 832-844, 1998.
[2] P. I. Good, *Permutation, Parametric, and Bootstrap Tests of Hypothesis*, 3rd Ed., Springer, 2009.
[3] J. Li and L. Yao and J.Wang. *Rhythmic Brushstrokes Distinguish Van Gough from his contemporaries: Findings via Automated Brushstrokes Extraction*, IEEE Trans. Patt. Anal. Mach. Intell., vol 34, no 6, pp 159-176, 2012.

Articles

We published two provisional articles on FACES during this period.

- Ramya Srinivasan, Amit Roy-Chowdhury, Conrad Rudolph, and Jeanette Kohl, "Recognizing the Royals--Leveraging Computerized Face Recognition to Identify Subjects in Ancient Artworks," *ACM International Conference on Multimedia* (2013) 581-584.
- Ramya Srinivasan, Amit Roy-Chowdhury, Conrad Rudolph, and Jeanette Kohl, "Quantitative Modeling of Artists Styles in Renaissance Face Portraiture," *Second International Workshop on Historical Document Imaging and Processing* (2013) 94-101.

Now, at the end of the second year (May 2015), two definitive articles have been written. The FACES project has been a collaboration of the humanities (art history) and the sciences (computer science). And so, we will publish one study of FACES from the point of view of the humanities (that is, how this technology generally works, what the parameters of its application to portrait art are at this time, what its advantages are, and so on), and a second study that presents the computer science basis of FACES. These two papers are meant to operate as a pair and will cross reference each other.

- Ramya Srinivasan, Conrad Rudolph, and Amit Roy-Chowdhury, "Computerized Face Recognition in Renaissance Portrait Art," *Signal Processing Magazine*, Special Issue (forthcoming July 2015).
- 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 Art" (currently under submission).

Website

We also have a website to disseminate our findings.

- http://faces.ucr.edu