Identification of Raphael’ s paintings from the forgeries

Liang Zhicong

School of Mathematics, Sun Yet-Sen University
liangzhc3@mail2.sysu.edu.cn

Abstract

Given 28 paintings, 11 of which are known as Raphael’ s and 10 are known as non-Raphael’ s, our task is to predict the authorship of another 7 disputed paintings. This project can be divided into two parts: (1) feature extraction and (2) classification.

For the first part, I applied the geographic tight frame method [1]. For the second part, I try different methods :(1) Distance Discriminant Analysis [1] and (2) Neural Network and (3) my Combined Method.

After looking into the misclassified paintings of distance discriminant analysis, I surprisingly found that all the misclassified paintings belong to Raphael, which means that we got a pretty large False Negative Rate (FNR). According to this fact, after applying Distance Discriminant Analysis, I applied Neural Network to re-predict the “outliers”. If a painting is predicted as outlier but has a probability large than 70% to be Raphael in Neural Network, we still classified it as Raphael’ s. We named this method as Combined Model.

In leave-one-out cross validation test, Distance Discriminant Analysis reaches an accuracy of 80%, Neural Network reaches about 92.5% while Combined Model reaches about 96.5%.

And we are confident that the 7/10/25/26th are Raphael’ s , 20/23rd are non-Raphael’ s while the 1st is still under uncertainty.

All the codes of this project could be found on my Github repository: Identification-of-Raphael’ s-Paintings [2].

Processing Data

Firstly, I transform all the 27 RGB paintings into gray-scale images by $pixel = (R * 299 + G * 587 + B * 114)\ //1000$ and then cut out the edges of them. For some reasons, the 28th painting could not be read in, so it is abandoned.

Feature Extraction

I applied the 18 geometric tight frame filters [1], using a neighbor size of 5*5:

$$\tau_0 = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \quad \tau_1 = \frac{1}{16}\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \quad \tau_2 = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix},$$

$$\tau_3 = \frac{\sqrt{2}}{16}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -1 \end{bmatrix}, \quad \tau_4 = \frac{\sqrt{2}}{16}\begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix}, \quad \tau_5 = \frac{\sqrt{7}}{24}\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix},$$

$$\tau_6 = \frac{1}{48}\begin{bmatrix} -1 & 2 & -1 \\ -2 & 4 & -2 \\ -1 & 2 & -1 \end{bmatrix}, \quad \tau_7 = \frac{1}{48}\begin{bmatrix} -1 & -2 & -1 \\ 2 & 4 & 2 \\ -1 & -2 & -1 \end{bmatrix}, \quad \tau_8 = \frac{1}{12}\begin{bmatrix} 0 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 0 \end{bmatrix},$$

$$\tau_9 = \frac{1}{12}\begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad \tau_{10} = \frac{\sqrt{2}}{12}\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \tau_{11} = \frac{\sqrt{2}}{16}\begin{bmatrix} -1 & 0 & 1 \\ 2 & 0 & -2 \\ -1 & 0 & 1 \end{bmatrix},$$

$$\tau_{12} = \frac{\sqrt{2}}{16}\begin{bmatrix} -1 & 2 & -1 \\ 0 & 0 & 0 \\ 1 & -2 & 1 \end{bmatrix}, \quad \tau_{13} = \frac{1}{48}\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix}, \quad \tau_{14} = \frac{\sqrt{2}}{12}\begin{bmatrix} 0 & 0 & 0 \\ -1 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\tau_{15} = \frac{\sqrt{2}}{24}\begin{bmatrix} -1 & 2 & -1 \\ 0 & 0 & 0 \\ -1 & 2 & -1 \end{bmatrix}, \quad \tau_{16} = \frac{\sqrt{2}}{12}\begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \quad \tau_{17} = \frac{\sqrt{2}}{24}\begin{bmatrix} -1 & 0 & -1 \\ 2 & 0 & 2 \\ -1 & 0 & -1 \end{bmatrix},$$

Then we got 18 feature-images of each painting. We calculate 3 statistics of each feature-image:(1) the mean (2) the standard deviation (std) (3) and the proportion of pixels that are more than one standard deviation away from the mean (tail). So, we have 18*3=54 features for every painting.

Distance Discriminant Analysis

Here, we assume that Raphael had **a certain consistence in his painting style,** that is, his paintings will concentrate around a center in a certain high-dimensional Euclidean space.

Before distance discriminant analysis, we will first select some most discriminating features. We hope that under these features, Raphael' s paintings concentrate in a circle around their mean while the non-Raphael' s lay as outliers. Mathematically, **we define the best feature subset as the one with largest AUC.**

What' s more, since it is unpractical to enumerate all the subsets of 54 features, we use forward step-wise approach to maximize AUC.

In experience, I found that normalizing the features or not will give different results. If we normalize the features first, then the best choice would be [6,33,2], which represent the mean of τ_2, τ_{12} and the tail of τ_0, respectively. If not normalize, then the best choice would be [18,24], the mean of τ_6 and τ_9.

Under the same hypothesis, we can use **standard deviation rather than AUC** to select features, since we are going to distinguish the two sets based on their consistence on some features.

In this case, we do not perform normalization $(x - mean(x))/std(x)$ but scale $x_{ij}/{}^{max}_{i}|x_{ij}|$, on each column of the features, trying not to lose the information of deviation. Then we calculate the standard deviation of each features of the authentic set and forgery set, denoted by std_T and std_N, respectively. Then we calculate the ratio between these two sets: $std_{ratio} = std_T/std_N$, $std_{ratio} \in R^{54}$. Finally, we choose a certain number of features with the smallest std_{ratio}. And the features we select in this model is [33,45], that is, the mean of τ_{12}, and τ_{16}. Finally, we choose to use ellipse as our boundary, by multiplying a weight to every component of the distance:

$$distance = \sum exp(-i) * (x_i - center_i)^2$$

The idea of this selection standard is that we hope the features we choose could be those have a relatively small standard deviation in authentic set while a large one in forgery set, which means they are consistent in authentic samples while diffuse in forgery samples. And the smaller deviation a feature has, the larger weight it contributes to the distance.

Main result of Distance Discriminant Analysis

	Features	Accuracy	1st	7th	10th	20th	23th	25th	26th
Normalized	[6,33,2]	80%	×	×	✓	×	×	✓	✓
Non-Normalized	[18,24]	80%	✓	✓	✓	✓	×	✓	✓
Scale	[33,45]	80%	✓	×	✓	✓	×	✓	✓

Note: 1st, 7th, ……26th represent the 7 disputed paintings, × means it is forgery while ✓ means it is authentic. "Accuracy" is the accuracy of the leave-one-out cross validation test.

Misclassified Paintings in Cross Validation Test

	Features	Misclassification
Normalized	[6,33,2]	2(R), 5(R), 22(R), 27(R)
Non- Normalized	[18,24]	4(R), 11(N), 13(N), 19(N)
Scale	[33,45]	3(R), 4(R), 9(R), 27(R), 19(N)

"R" means this painting is Raphael's while "N" means non-Raphael's

According to the two tables above, a very important point should be proposed is that all the misclassified paintings in normalized model and 4 out of 5 in scale model belong to Raphael, which means these two models both have a large False Negative Rate (FNR) while a perfect True Positive Rate (TPR). We can improve them by re-predict those paintings that are classified as outliers and try to lower the FNR, which will be discussed in section "Combined Model".

Four Paintings Misclassified as Forgeries in normalized model

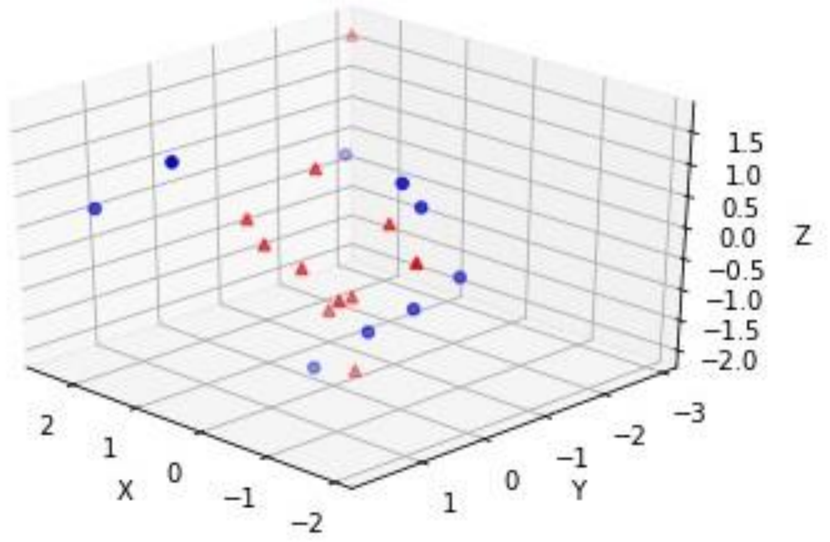
2nd (Raphael)

5th (Raphael)

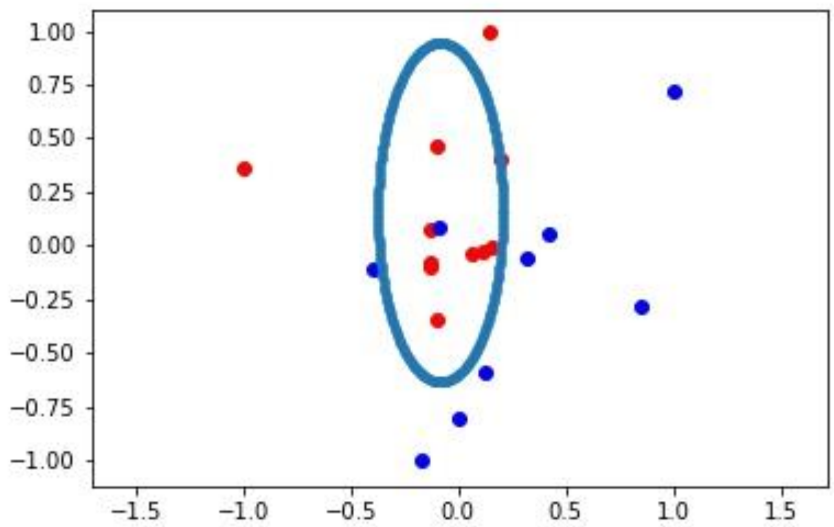
22nd (Raphael)

27th (Raphael)

Here, we will visualize the normalized and scale models. The boundary of the second figure is the best threshold of classification on all the data (instead of leave-one-out cross validation):



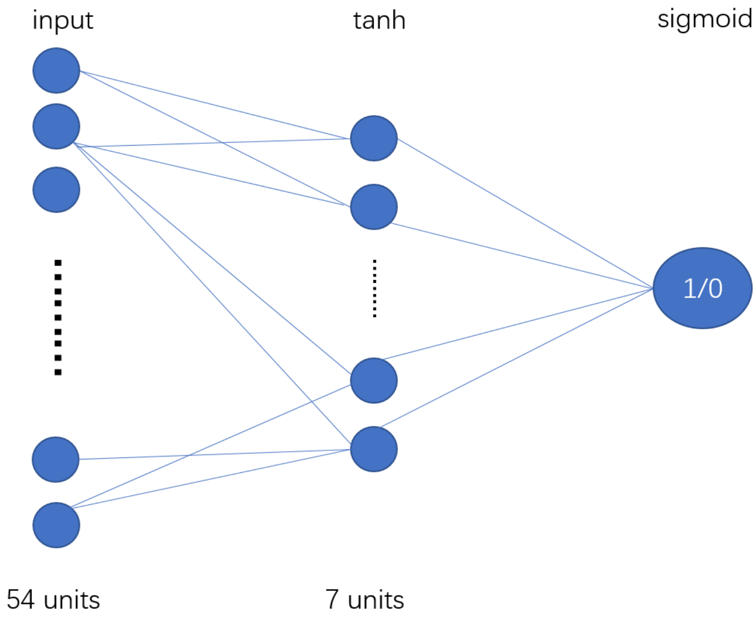
Normalized Model: Red points are Raphael’s while blue ones are Non-Raphael’s.
As we can see, a few Raphael’s paintings spread as outliers, too.



Scale Model: Red points are Raphael’s while blue ones are Non-Raphael’s
The ellipse is the classification boundary.

Neural Network

I use a neural network, with 54 input units (all features used) and one hidden layer of 7 units, as my second classification method and reach an accuracy of about 92.5%. Model summary is as follow:



Activation_1 is tanh and Activation_2 is sigmoid.

There are 393 parameters totally.

Main result of Neural Network

Features	**Average Accuracy**	**1st**	**7th**	**10th**	**20th**	**23th**	**25th**	**26th**
All 54 features	92.5%	✓	✓	✓	✓	×	×	✓
Frequency		0.95	0.83	0.56	0.76	0.15	0.4	0.63
Misclassification		6(R),	13(N)					

R" means this painting is Raphael's while "N" means non-Raphael's

"Frequency" is the frequency of a particular painting classified as Raphael's in 100 repeated experiments

When we look into the features subsets we choose in Distance Discriminant Analysis, we notice that no standard deviation is chosen under the hypothesis. However, they may still be useful in a more complicated and delicate model. If we rule out all the 18 standard deviations and construct the neural network with the remaining 36 features, accuracy drops down to about 80%. This provides another motivation for our Combined Model.

Combined Model

As we have mentioned above, the normalized Distance Discriminant Analysis model or Scale Model gives a pretty good True Positive Rate (TPR) while a bad False Negative Rate (FNR). In this section, I try to re-predict those paintings that are first considered as non-Raphael's using the same Neural Network above, with all 54 features used.

And the classification rule will be:

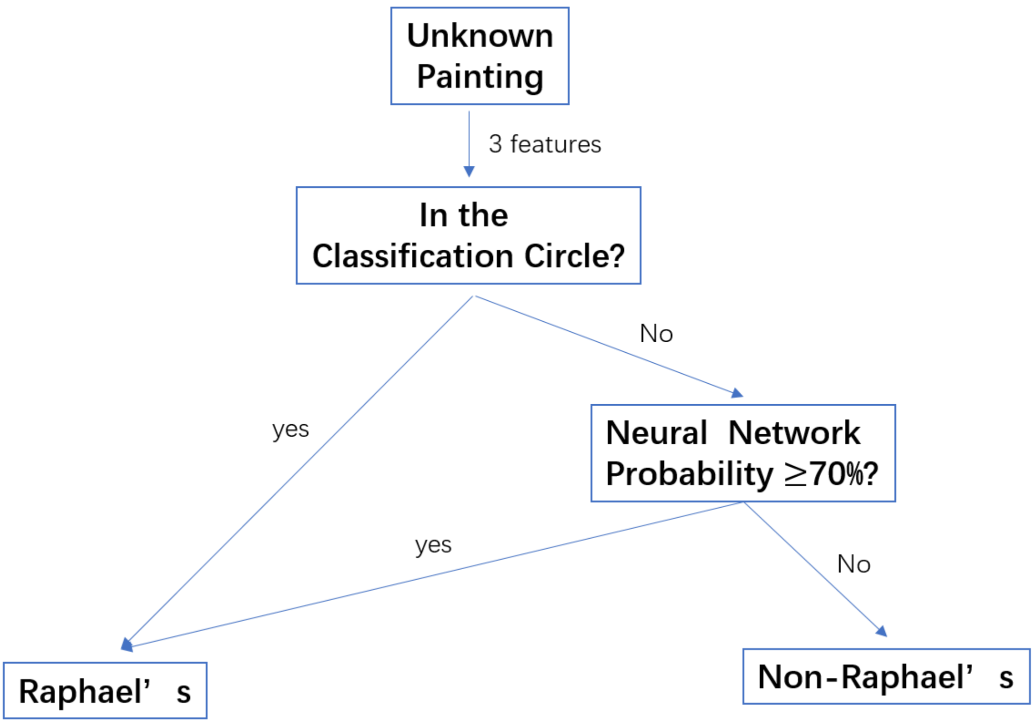
(1) If a painting is considered as Raphael's in Distance Discriminant Analysis, then we accept it as Raphael's;
(2) If a painting is considered as non-Raphael in Distance Discriminant Analysis but has a probability larger than 70% to be Raphael's in Neural Network, we accept is as Raphael's;
(3) All the other cases will be rejected to be Raphael's.

Main result of Combined Model

Accuracy	**1st**	**7th**	**10th**	**20th**	**23th**	**25th**	**26th**
96.5%	✓	✓	✓	×	×	✓	✓
Frequency	0.5333	0.7	1	0.2667	0	1	1

"Accuracy" is the accuracy of the leave-one-out cross validation test.

"Frequency" is the frequency of a particular painting classified as Raphael's in 30 repeated experiments.



Basic Procedure of Combined Model

Analysis

In the first model, Distance Discriminant Analysis, we select 3 features under which Raphael' s paintings will concentrated around a specify circle while non-Raphael' s ones will spread as outliers.

However, this hypothesis seems too strong that we may misclassified some Raphael' s paintings as Non-Raphael' s. This case is very likely to happen since we only choose such a small number of features: it is difficult even for Raphael to perform consistently on a small set of particular features. A re-classification process seems reasonable since even a painting lies as outlier under several features, it is still possible to be Raphael' s if other features resemble those of Raphael' s ones.

In Combined Model, we first classify paintings using the circle hypothesis, then we re-predict the outliers based on a non-circle model, trying to capture as more information as we can, like the standard deviation information that we ignore in circle or ellipse model.

Conclusion

From the result of the two classification methods we discuss above, we are confident that the 7/10/25/26th are Raphael' s , 20/23rd are non-Raphael' s while the 1st is still under uncertainty.

1st (Uncertain)

Link:

[1] Haixia Liu, Raymond H. Chan, Yuan Yao, Geometric tight frame based stylometry for art authentication of van Gogh paintings :https://doi.org/10.1016/j.acha.2015.11.005

[2] ZhicongLiang Github: https://github.com/ZhicongLiang/Identification-of-Raphael-Paintings