Artist Identification for Renaissance Paintings

Jonathan Jou, Sandeep Agrawal

{jj,sandeep}@cs.duke.edu

Computer Science, Duke University

Abstract

Current work in author identification is primarily directed towards music classification. Identification of artists by analyzing features of their work has recently gained interest. As a multiclass classification problem, potentially applicable machine learning approaches to the problem are numerous.

We propose to extend present work in this area, which uses Naïve Bayes classifiers and multi-class SVMs, by picking a more unique set of paintings across prolific artists. We initially use a histogram of colors as our features, and then we analyze more advanced features like the histogram of gradient orientations (HOG). Due to the large number of features as compared to the number of paintings, we use PCA to condition features based on the highest variance. We apply several multi-class classification techniques like Naïve Bayes, Linear Discriminant Analysis, Logistic Regression, K-Means and SVMs to our problem and achieve a maximum classification accuracy of 65% for an unknown painting across 5 artists.

1. Introduction

Machine learning has been widely used in various vision related applications like pedestrian identification\(^1\), human detection\(^2\) and object detection. The human mind excels at classification, and current machine learning techniques are advanced enough to be suitable for a wide variety of classification tasks. Artist identification is a complex process, and experts employ a wide variety of techniques such as ultraviolet fluorescence, x-radiography and paint sampling\(^3\). Each artist has his own unique signature, which can be the style of brushstrokes, preferred color choices or preferred
landscape/portrait types. As such, machine learning tools can greatly augment judgment on the identification of the artist of an unknown painting based on a given set of features and prior knowledge.

Recently, the Van Gogh and Kröller-Müller Museums in The Netherlands released a data set of 101 high-resolution gray-scale scans of paintings within their collections to groups of image processing researchers from several different universities\(^4\). Of the 101 paintings, 82 have consistently been attributed to Van Gogh, six have always been known to be non-Van Gogh, and 13 have been or are currently questioned by experts. Therefore, there is a large potential for machine learning researchers to come up with ways to firmly establish the identity of the authors of these unknown paintings. Present work in this area is relatively sparse, with usage of Naïve Bayes\(^5\)\(^6\)\(^7\) for local features and 2D HMMs for brush stroke detection\(^3\).

In this work, we explore several popular classification schemes to classify a set of unique paintings for a group of 5 artists and use cross-validation to test a set of unknown paintings based on model parameters generated from known paintings. We use both color histogram and HOG feature sets in our models, and take a different approach to existing schemes, which use SVMs for a diverse set of features. Using just 2 varieties of features, basic and advanced, we attempt to characterize the separability of these feature spaces by trying to find a classification scheme that can reliably identify each artist.

Section 2 describes our features and methodology. Section 3 covers the various classification approaches used and experimental results. Section 4 describes our use of SVMs with various kernels and results. Section 5 describes future work and discussion and we finally conclude in Section 6.

2. The Features and Methodology

We pick a set of five artists, Braque, Matisse, Monet, Picasso and Van Gogh, as they are more prolific and unique paintings for each artist were readily available. Using a python script, we downloaded 150
unique paintings for each of these artists from collections on the web (http://www.georgesbraque.org/, http://www.claude-monet.com/, http://www.henrimatisse.org/, http://www.pablopicasso.org/ and http://www.vangoghart.org/). Figure 1 shows a small sample of our data set. In order to normalize across different paintings, each painting is scaled to a size of 100x100 pixels. We use the raw pixels themselves as one type of feature set. For a total of just 750 observations, a large number of features results in an ill-conditioned design matrix. To reduce the number of features, we shrink our color space into two alternatives: 12 bits per pixel, 4 bits per channel, and 18 bits per pixel, 6 bits per channel. We initially use a color histogram for our 12-bit color space, and later use more advanced features like HOG\(^2\). HOG is one of the more popular descriptors for human detection, and relies on counting occurrences of different gradient orientations in overlapping cells across the image. For HOG we use a cell size of 30 pixels, and a block size of 3 cells and 9 gradient bins, giving a feature vector dimensionality of 81. For estimating the HOG descriptors we use the function provided by Leo et al.\(^8\). We perform 10-fold cross validation for each of our experiments.

**Figure 1** A sample from our data set. Works by Braque, Matisse, Monet, Picasso and Van Gogh.
3. Experiments

3.1 Naïve Bayes classifier

The Naïve Bayes classifier assumes that the features are independent given the class variable. It allows us to quickly estimate the baseline classification accuracy for our feature sets. We use a simple bag of words model for the histogram features and a normal distribution for the raw pixel and HOG features. We use the built in MATLAB *NaiveBayes* classifier.

3.2 Linear Discriminant Analysis

We model the classes of known paintings as Gaussians with a shared covariance matrix and compute the posterior probabilities of each class for an unknown painting. As the number of dimensions for the histograms is much larger than the number of observations, the training matrix is ill-conditioned and we apply Principal Components Analysis (PCA) to reduce the dimensionality of our feature space. Figure 2 shows the variance of the PCA features for the various feature sets. We can see that variance is quite evenly distributed among the features. As our training set in cross validation comprises of 135 images

![Figure 2](image.png)

*Figure 2* Variance captured with increasing dimensionality for different feature sets.
per artist, we need to reduce the number of features to 135, which for the histograms evaluates to just 75% and 58% of the total variance. We use the built in MATLAB `classify` and `princomp` functions to perform this analysis.

3.3 Logistic Regression

We next try to fit a logistic classification model to our feature set. We apply PCA to condition the design matrix and do a “one vs. all” classification. Logistic regression makes only one assumption about the data—that there exists a decision boundary separating the classes. This potentially proves untenable in one vs. all classification, where the data becomes highly inseparable.

3.3 K-Means clustering

We also try unsupervised classification techniques and analyze the separability of our data based on an isotropic Gaussian model. In order to get the best case classification accuracy, we supply K-Means with the means of our feature set as starting centroids.

Figure 3 shows the classification accuracy for the above described approaches for our various feature sets.

![Figure 3](image-url)
sets. We can see that Naïve Bayes gives the maximum classification accuracy for the histogram features of 65%. The high misclassification rates of LDA, LR and K-Means show that the data set distribution is not separable. To gain further insight into this, we analyzed the distributions of each class and its relative spread in the feature space. The distance between means for any two classes varies between 377 and 616 in the 12-bit histogram space, and is smaller than the largest standard deviation among features of a class, which varies from 363 to 900. This produces a distribution which is cannot be easily separated, and it is our belief that selection of a more sophisticated feature set is the most straightforward and likeliest way to classify more effectively.

4. Support Vector Machines

In addition to MATLAB’s inbuilt SVM API, we use LIBSVM\(^9\) for support vector classification. The results indicate that the data is not linearly separable: no one vs. all classifier classified correctly, and the resulting multiclass algorithm achieved a maximum of 61% classification accuracy with a quadratic kernel and 12-bit histogram data. Kernels of degree one, two, three, and five were attempted, along

![Figure 4](image-url)  
*Figure 4* Classification accuracy for different SVM kernels for various feature sets.
with a radial basis function kernel, with cross-validation classification performing worse past the quadratic kernel. The radial basis function achieved between a maximum of 58% classification accuracy for the 18-bit histogram features, and we tried the HOG feature selection, but none of these performed better than Naïve Bayes. The decreasing trend in performance as we increased the degree of our kernel is consistent with overfitting, and did not indicate that higher-degree kernels would perform any better.

In order to gain insight into these high misclassification accuracies, we analyzed the results of one vs. one SVM classifiers for each pair of artists. Table 1 shows our results for a linear SVM in the 12-bit histogram space, which supports our hypothesis regarding the separability in our chosen feature spaces. Most of the classifiers performed at less than 80% accuracy, demonstrating the strong overlap across artist distributions.

<table>
<thead>
<tr>
<th></th>
<th>Braque</th>
<th>Matisse</th>
<th>Monet</th>
<th>Picasso</th>
<th>Van Gogh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Braque</td>
<td>100.00</td>
<td>89.67</td>
<td>69.00</td>
<td>75.67</td>
<td>64.00</td>
</tr>
<tr>
<td>Matisse</td>
<td>89.67</td>
<td>100.00</td>
<td>82.00</td>
<td>91.33</td>
<td>85.33</td>
</tr>
<tr>
<td>Monet</td>
<td>69.00</td>
<td>82.00</td>
<td>100.00</td>
<td>78.00</td>
<td>68.67</td>
</tr>
<tr>
<td>Picasso</td>
<td>75.67</td>
<td>91.33</td>
<td>78.00</td>
<td>100.00</td>
<td>80.67</td>
</tr>
<tr>
<td>Van Gogh</td>
<td>64.00</td>
<td>85.33</td>
<td>68.67</td>
<td>80.67</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table 1 Classification accuracies for a one vs. one linear kernel.

5. Future Work and Discussion

It is possible that an HMM-based brushstroke detection protocol for feature selection as used by Johnson et al. would improve accuracy on our data set, but the results suggest that our painting collection produces a distribution in the feature space which is evenly populated by most classes. It may be worthwhile to characterize each distribution with a Gaussian, to measure probabilistic distance metrics between each class.

There is a large disparity between the results in literature and our own; given that our methods are comparable, it is worthwhile to consider the representative nature of the paintings we used, and the paintings classified by others. While our own data set proved to be largely inseparable, it remains to be seen if the vast body of existing art by all artists of all eras suffers from these same issues. Differences
across eras, cultures, and schools of painting are distinct and unmistakable, and feature extraction techniques exploiting these differences have the potential to perform as well as or better than a trained human eye.

6. Conclusion

Artist identification provides ample opportunity for machine learning techniques, and continues to be a challenge even for experts. In this work we analyze a group of 5 prolific artists, and attempt to classify an unknown painting based on a model trained from known paintings. Using several classification techniques, we achieve a maximum accuracy of 65% with a Naïve Bayes classifier for simple histogram features. This amply motivates the need for more sophisticated features for reliable classification, and leaves open the question if there exist more suitable classification techniques.

7. References