Exploration of application of Machine Learning in Fine Arts analysis.

Aratrika Sarkar, Ashish Singh
College of Information and Computer Sciences
University of Massachusetts Amherst

1 Introduction

Machine learning applications have been suggested for many tasks. The suitability of applying machine learning to the problem of fine arts, is a new and promising field. Research on automated art identification is very sparse, and only recently has there been an increased interest in applying machine learning to the context of paintings. Through this project, we want to explore the possibility of using machine learning algorithms to efficiently classify fine art works i.e.,given the proper dataset, can machine learning techniques accurately segregate and predict given paintings based on different parameters like genres, painters and era. Researchers are marshaling advances in digital image processing, machine learning, and computer vision to solve problems of the attribution and interpretation of fine-art paintings.These existing researches focuses on painter identification and therefore stresses high degrees of accuracy on small target datasets. As a result of this focus, the problem of the broad classification of style in painting receives relatively little attention.

Artist identification is a complex process, and experts employ a wide variety of techniques such as ultraviolet fluorescence, x-radiography and paint sampling. Each artist has his own unique signature, which can be the style of brush strokes, 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.

In this paper, we look into the problem of computational categorization of digitized paintings into artistic genres (or art movements) and artists. In contrast to other directions of image classification, such as scene or object recognition, where large databases and evaluation protocols do exist, such an aspect is less emphasized for digitized paintings. Typically, the evaluation of a new method is carried on a small database with few paintings belonging to few genres. Given the latest advances of machine learning, two aspects should be noted: (1) deep networks with the many parameters easily overfit on small databases and (2) to have progress, we need larger databases.

The given problem can be categorized as multiclass classification and clustering problem, where we aim at analyzing how effective are machine learning algorithms in classifying and grouping paintings of same classes for different parameters such as genres, artists, and era. We have created our own dataset for our experimentation and we have considered 3 Genres namely Cubism, Impressionism, and Romanticism. Cubism is an early 20th-century style and movement in art, especially painting, in which perspective with a single viewpoint was abandoned and use was made of simple geometric shapes and interlocking planes. Impressionism is a style of painting associated mainly with French artists of the late nineteenth century, which seeks to re-create the artist's or viewer's general impression of a scene. Artistic and intellectual movement that originated in the late 18th century and stressed strong emotion, imagination, freedom from classical art forms, and rebellion against social conventions. So, these genres are a perfect combination which share both different and common traits at the same time. The study reports the results of a review of features currently applied to this domain and supplements the review with commonly used features in image retrieval. The study considers several features of an image which has proved to differentiate paintings such as to represent texture and shape information, a transformation of raw data input is implemented as part of feature learning and for extracting color information of paintings, Hue saturation and brightness values of the images were computed. These are then used to build a histogram of these values, which is then used as our final color descriptor. The survey of color features revealed that preserving frequency and spatial information of the color content of a painting did not improve classification accuracy. The features with the best performance were tested against a standard test database composed of images from the Web obtained by Web Scraping. Several supervised and unsupervised techniques were used for classification, visualization, and evaluation including k-Nearest Neighbor, Support Vector machine, Random Forest, K means Clustering. Training, testing accuracy analysis and Receiver operating characteristic(ROC) are proposed as evaluation techniques for classification results.

In the next section we describe our approach to the problem. In particular, we outline how we obtained our data, what features in the paintings we considered and which classification algorithms we employed. We conclude by presenting and analyzing the results obtained.

2 Related Work

According to Artyfactory [1], art movements are “collective titles that are given to artworks which share the same artistic ideals, style, technical approach or timeframe”.While the actual characteristics place a work in some art movement, its painter, for personal reasons, refused to be categorized in such a way, giving birth to disputes.Here, we deal with two major problems : Style/Genre classification and Artist Classification.Previous solutions proposed are often style-specific addressing only particular kinds of art or even the work of particular Artists[2].Here,the author selected 7 different Artists and classification of paintings based upon Artist were carried out using Machine Learning Techniques.But as feature selection of the images plays a crucial role in this task, the main features are selected here by a Color Palette Algorithm which represent the color bins of an image and does not take into consideration any other feature of an image which might be successful in classifying an Artist’s work from the other.Moreover, in the classification task , only KNN is explored while other Learning methods could also yield better results. Some of the previous works have tried to classify paintings based upon the "Texture" feature of an image but ignoring the "color structure" of the images [3].This is mainly used in the applications of Art authentication by an Artist . Study shows that previous work has also been done for Artist Identification using techniques of Image Processing only and exploring the paintings of a single Artist.[4].Here, the authors have focused on Vincent Van Gogh’s paintings and Brush Stroke analysis and after feature extraction,Stochastic model based comparison was carried out,using the Hidden Markov Model. Later, some works have been done using Machine Learning techniques to classify paintings based upon the Artists[5].Here, the authors have used mainly "color features" of an image as the features to represent the images of a painting.The authors have focused on 7 different artists and the main objective was to figure out the Artists with similarities in their paintings rather than the differences.The feature of an image considered here, also mainly concentrated on the color features of an image and for classification of an image, mainly SVM was used as a classifier.Another work in the same domain[6] attempted to classify paintings based upon the artists.Here,the authors have used the works of 5 different artists and classified them using Machine Learning Techniques.The main classifier used here was SVM with various kernels, Naive Bayes and 2D Hidden Markov Models.The features used here are mainly the color features of an image for instance the color histograms.
From our study of related work, we comprehended that the data selection must be carried out in such a way that the paintings must be similar as well was different in order to enhance proper classification of genre and artists.Secondly, the features of the images of paintings must be selected such that it includes the texture,color and brush strokes of an image.Thirdly,during selection of classification methods, all supervised and unsupervised learning method must be explored with various pre processing,cross validation and hyper parameter tuning in order to get the desired results.

3 Data Set

For our project, we created our own data set of paintings and art works. For this, we have utilized Google Images and online collections/exhibitions of art museums to find and scrape relevant art works based on different classifying parameters like genres, artists, era and styles. We have selected 3 genres - Cubism, Impressionism, Romanticism. Cubism is an early 20th-century style and movement in art, especially painting, in which perspective with a single viewpoint was abandoned and use was made

Table 1: Structure of Dataset for Genre and Era classification

GENRE	NUMBER OF PAINTINGS	HISTORICAL PERIOD
Cubism	920	1900 – present
Impressionism	984	1860 – 1925
Romanticism	874	1770 – 1850

Table 2: Structure of Data set for Artist Classification

GENRE/ARTIST	NUMBER OF PAINTINGS
Cubism/ Pablo Ruiz Picasso	58
Realism / Gustave Courbet	58
Romanticism	58

of simple geometric shapes and interlocking planes. Impressionism is a style of painting associated mainly with French artists of the late nineteenth century, which seeks to re-create the artist's or viewer's general impression of a scene. Romanticism was an artistic and intellectual movement that originated in the late 18th century and stressed strong emotion, imagination, freedom from classical art forms, and rebellion against social conventions. These genres share several common as well as different traits which help us to evaluate the various machine learning techniques.

We focused on a set of paintings from three different artists: Pablo Ruiz Picasso, Gustave Courbet, Victor Brauner. The main reason for choosing these artists were due to their prolificacy as we anticipated to require a large number of training examples for successful classification. Moreover, we tried to include pairs of artists from genres which share similar traits (Cubism and Realism) and genres which are completely different (Cubism, Surrealism and Realism, Surrealism) so that we could evaluate the performance of thee classifiers efficiently.

While considering the era, we have picked data from the genres of Cubism, Impressionism, and Romanticism. The Historical period to which they belong are : Cubism 920 1900 – present, Impressionism 984 1860 – 1925 , Romanticism 874 1770 – 1850 .So, we see that these 3 genres belong to different era, Romanticism being the oldest followed by Impressionism and Cubism which is still in use. But one detail not to be missed here is that during the time period 1900 – 1925, both Cubism and Impressionism were in use and hence it was a challenge for the classifier to segregate these genres.

The precise database structure is shown in Table 1 and Table 2 and some examples representative for the art movements are in Figure 1. Figure 2.

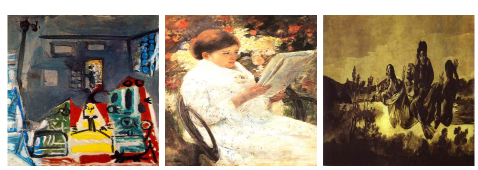

Figure 1: Sample Images from 3 genres (a) Cubism ,(b) Impressionism and (c) Romanticism for Genre Classification

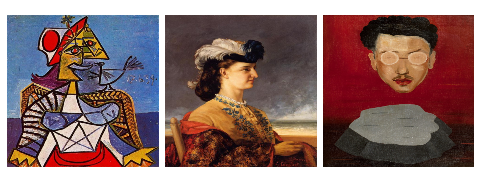

Figure 2: Sample Images by artists (a)Publo -Cubism (b)Courbet - Realism (c)Brauner - Surrealism for Painter Classification

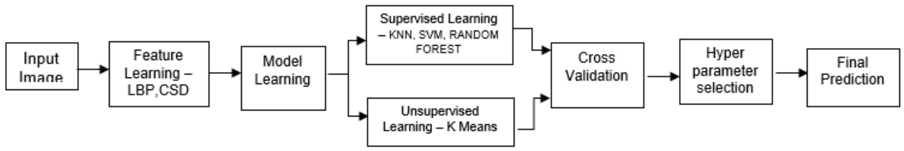

Figure 3: Flow chart showing the steps followed in our experimentation

4 Proposed Solution

To effectively evaluate application of Machine learning techniques for analyzing art works and paintings following framework is proposed, which has four major components:Firstly, our collected dataset is segregated into training and testing sets, followed by learning appropriate representation of pixels, which is then used as attributes of the data samples and are used for classification of paintings using appropriate learning models. Multiple alternatives for each components are used to formulate a comparative study. A system level block diagram of our proposed pipeline is given in Figure 3.

4.1 Training and Testing

To organize our dataset into appropriate training and testing set, our collected images are randomly sub-sampled into two equal parts for every class.This leads in creation of unbiased and balanced training and testing set. For the purpose of cross-validation, the training set is divided into k-folds ($k = 10$ for genre and $k = 8$ for artist) and accordingly used for hyper-parameter tuning.

4.2 Pre-processing

For an image, pixel intensities are the basic building blocks which uniquely represent the particular image. However, the observations from prior works [7] in literature shows that direct usage of pixel intensities fails to represent inherent structures present n these art works in terms of textures and shape information. For instance, to differentiate between impressionism and Cubism is the brush stroke thus the texture. Cubism has a darker tone with respect to Romanticism while the later depicts exoticism or extraordinary things" .In Computer terms, detecting the differences is something quite challenging.So, to better represent texture and shape information, a transformation of raw data input is implemented as part of feature learning.
For extracting color information of paintings, Hue saturation and brightness values of the images were computed. These are then used to build a histogram of these values, which is then used as our final **color descriptor**. To incorporate spatial information, the images are divided into five regions and color histogram of region is calculated and concatenated to give the final color descriptor representing the image.
To represent texture information of the paintings, pixel intensity values are converted to **Local binary patterns**, which is a histogram of local patterns in an image, quantized into binary format, computed over a $3x3$ window size.
For final representation of the images, the above two learned features are concatenated and are used as input for classification by different learning models.

4.3 Model Selection

For holistic evaluation of our problem statement, we use different classifiers with different properties like learning rate, prediction time, complexity etc. Details for the classifers employed are as follows:

- **K nearest neighbors(KNN)** It is a non-parametric classifier that classifies every new given instance by finding the nearest match to the known training data through computing a majority vote over the set K neighbors with minimum distance to the given instance. Mathematically, the model can be represented as, $f_{KNN}(x) = argmax_{y \subset Y} \sum_{i \subset N_k(x)} I[y_i = y]$ In our implementation, the value of k is being determined by k-fold cross validation and grid search over the range $[3, 4, 5, 6, 7, 8, 9, 10]$.

- **Multiclass Support Vector Machine** - It is another variant Support vector machine being used for multiclass problem. Support Vector Machines are a type of supervised learning algorithm, that given a set of training data, produces a hyperplane seperating the two class of data and maximizes the margin from this hyperplane. For multiclass problem,it is used in a one versus all framework.
  For our framework, SVM with 'Linear'kernel is used, with optimal value of hyper parameter 'C' being selected using k-fold cross validation and grid search over the values $[0.1, 0.5, 1.0, 5.0, 10.0, 50.0, 100.0]$
- **Random Forest** - It is a type of an ensemble classifier, formed by combination of 'Bagged' decision trees i.e. k different decision trees generated from k different randomly sub-sampled data (with replacements) from the same dataset.In Random Forests, there is a further de-correlation, by only taking a random subset of the given instances when deciding the variable to split on at each node of the tree.
  In our implementation, the optimal hyperparameter is the number of trees being used .Optimal number of trees in the random forest is selected sing k-fold cross validation and grid search over the values $[10, 20, 50, 100, 150, 200, 350, 500]$
- **K Means Clustering** - It is a type of unsupervised learning algorithm, where initially,i k random points are selected as centroids in the space defined by training points. All the data points are grouped into different classes based on closest centroid defining that class, wherein the following objective function is minimized (sum of squared distance between data points and cluster centre):$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - c_j\|^2$. Centroid of each class is recalculated after assignment of each datapoint to groups. The above steps are repeated till the centroid becomes stationary.
  For our work, the number of cluster center is selected as 3, keeping in mind that we have three classes in our dataset for artist and art genres respectively.

4.4 Hyperparameter Selection Selection

For Hyperparameter selection, Grid search method with K-Fold cross validation was employed In this, for a range of hyperparameter values, classifier modesl are trained and evaluated using k fold cross validation for each of the hyperparameter value. The model that performs the best is taken as the final model, and the corresponding hyperparameter value as the optimum value. In K-fold cross validation, the data set is divided into k subsets.Out of the k subsets, one subset , known as the validation set, is taken as the test set and the other other k-1 subsets are combined to form the training set. The classifier is trained on the new training set and tested on the validation set. This process is repeated k times, with all k subsets being used as a validation set once. The accuracy rate obtained from each of the k cycles are stored and the corresponding average accuracy rate is computed, which acts as the final evaluation metric of the model

5 Experiments and Results

5.1 Genre Classification

We first report the results for Genre based classification. This dataset is divided into 3 classes : Cubism,Impressionism and Romanticism.The results achieved for various combinations of features and classifiers, are listed in Table for Figure 4. The table shows the accuracies achieved for Supervised Learning whereas the Table in Figure 5 shows the accuracies achieved for Unsupervised Learning.For Random forest, the hyper parameter is the number of trees and the range of hyper parameter is [10,50,100,150,200,250,350,500].The optimal hyper parameter obtained was 100. For KNN, the hyper parameter is the number of nearest neighbors considered and the range searched over was [2,3,4,5,6,7,8,9,10] and the optimal value of hyper parameter was 4.For SVM, 'C' is used as a hyper parameter where C means The C parameter tells the SVM optimization how much you want to avoid misclassifying each training. So the range of C is [0.1,0.5,1.0,5,10,50,100] and the optimal value of C obtained is 1.0.So,from Fig 4,we find that Random Forest with Cross Validation yields the best result. Fig 7 (a,b,c) shows the Hyper parameter vs Accuracy curve for KNN,SVM and Random Forest classifiers. Next we use a ROC Curve(receiver operating characteristic) which is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.Figure 8(a,b,c)shows the ROC Curve for KNN,SVM and Random Forest. For the curve,Class 0 represents Cubism,class 1 represents Impressionism and Class 2 represents Romanticism. From the obtained ROC curves, we can see that for the task of Genre classification, all the three supervised classification methods yields good ratio of True positive over False positive. Area Under Curve (AUC) comes maximum for Random forest, and minimum for KNN.

FEATURE	KNN without CV	KNN with CV	SVM without CV	SVM with CV	Random Forest without CV	Random Forest with CV
LBP	0.233	0.300	0.263	0.296	0.321	0.328
CSD	0.298	0.340	0.280	0.312	0.357	0.400
LBP+CSD	0.421	0.499	0.521	0.577	0.656	**0.666**

Figure 4: Table for accuracies when various combinations of features and classifiers are used for genre classification

FEATURE	K MEANS CLUSTERING
LBP	0.115
CSD	0.187
LBP+CSD	0. 235

Figure 5: Table for accuracies when K MEANS Clustering is used for genre classification

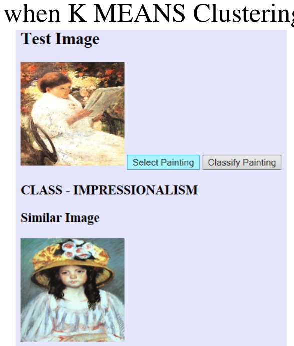

Figure 6: Interactive Classificationas an application for this task

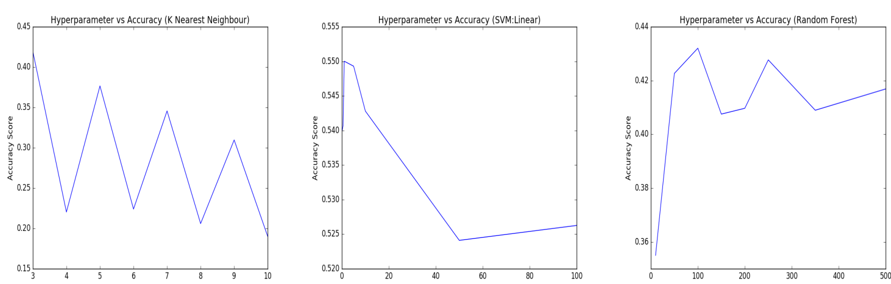

Figure 7: (a,b,c)Hyper parameter vs Accuracy for KNN,SVM and Random Forest

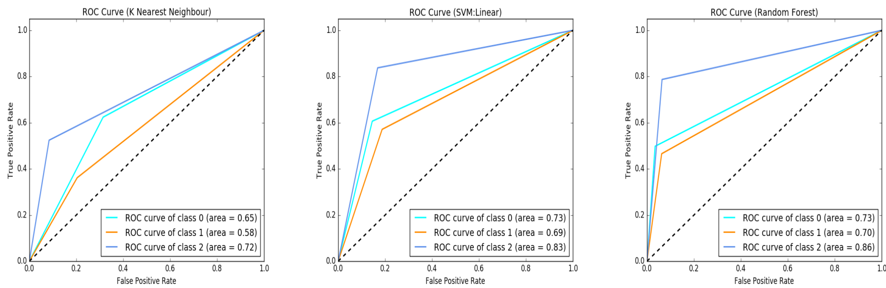

Figure 8: (a,b,c)ROC Curve for KNN,SVM and Random Forest

5.2 Artist Classification

We now report the results for Artist based classification.Here we consider 3 artists belonging to 3 classes as stated above- Publo(Cubism),Courbet(Realism) and Brauner(Surrealism).The results achieved for various combinations of features and classifiers used, are listed in Table of Figure 10. The table shows the accuracies achieved for Supervised Learning whereas the table in Figure 11 shows the accuracies achieved for Unsupervised Learning.For Random forest, the hyper parameter is the number of estimators and the range of hyper parameter is [10,50,100,150,200,250,350,500].The optimal hyper parameter obtained was 150. For KNN, the hyper parameter is the number of nearest neighbors considered and the range searched over was [3,4,5,6,7,8,9,10] and the optimal value of hyper parameter was 3.For SVM, 'C' is used as a hyper parameter and range of C is [0.1,0.5,1.0,5,10,50,100] and the optimal value of C obtained is 10.0.From Fig 8, we find that the maximum accuracy obtained is for Random Forest with Cross Validation.Fig 12 (a,b,c) shows the Hyper parameter vs Accuracy curve for KNN,SVM and Random Forest classifiers.Next we plot ROC Curve(receiver operating characteristic). Figure 13(a,b,c)shows the ROC Curve for KNN,SVM and Random Forest. For the given plot, class 0 represents art work by artist 'Brauner', class 1 by artist 'Courbet' and class 2 by artist 'Picaso'. From the obtained ROC plots, we can see that, KNN performs ver poorly, with ratio of True positives over False positives less than or equal to 0.5. For SVM the result improves, but for one class, the ratio is still around 0.5.However, Random Forest shows promising results, with best AUC amongst all the methods.

Figure 9: Interactive Classification as an application for this task

FEATURE	KNN without CV	KNN with CV	SVM without CV	SVM with CV	Random Forest without CV	Random Forest with CV
LBP	0.101	0.154	0.167	0.218	0.398	0.209
CSD	0.168	0.265	0.254	0.365	0.471	0.317
LBP+CSD	0.290	0.314	0.386	0.436	0.689	**0.609**

Figure 10: Table for accuracies when various combinations of features and classifiers are used for Artist classification

FEATURE	K MEANS CLUSTERING
LBP	0.119
CSD	0.187
LBP+CSD	0. 219

Figure 11: Table for accuracies when K MEANS Clustering is used for genre classification

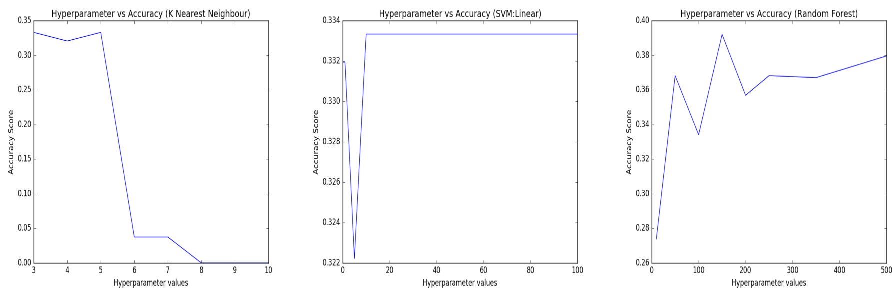

Figure 12: (a,b,c)Hyper parameter vs Accuracy for KNN,SVM and Random Forest

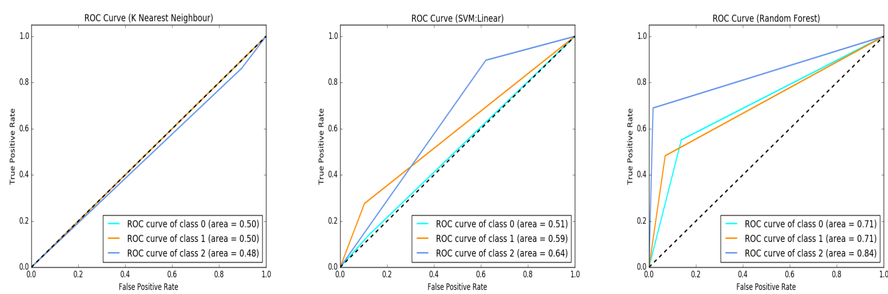

Figure 13: (a,b,c)ROC Curve for KNN,SVM and Random Forest

6 Discussion and Conclusion

For the given collected dataset, for different genre and artist, we find that Random Forest tends to produce better results as compared to other two methods for supervised classification. We find that in particular, methods like K-nearest neighbors are not suitable for such tasks, due to lack of internal complexity.This is backed by our results, wherein KNN gave the lower most accuracies among all three classifiers.
In terms of comparison between supervised and unsupervised methods, we find that supervised methods perform much better than unsupervised methods(K-means clustering). This might be attributed to very minute differences in each class of image, which the unsupervised methods are not able to learn and there is a requirement for learning on each of the exemplar instances to be able to better classify test images based on learnt differences through features of these exemplar instances

In conclusion, this project aimed at evaluating viability and application of using machine learning tools for better analyses of paintings and art works, by classification of these into sub-categories based on Genre and artist. A framework was formulated, wherein features representing color and texture information is learned from raw pixel intensities.Several machine learning techniques like KNN, SVM, Random Forest and K-means clustering are explored for analyzing and classifying paintings and their relations. For evaluating each model, we are computing model accuracy, False positive and True positive rates which are analyzed through respective ROC curves.

Our test results suggests that, given proper representation of the input art works that best defines them, it is possible to develop general classifiers that inspite of not domain specific, are able to classify art works on the basis of different genres and artists with better odds than random guessing.These classifiers can then be used in other different applications like image retrieval tasks, as shown by our interactive classification module.

7 References

[1]$http://www.artyfactory.com/art_appreciation/art_movements/art_movements.htm$
[2]$https://www.researchgate.net/publication/31873213_The_classification_of_style_in_fine-art_painting$
[3]$http://www.pnas.org/content/101/49/17006.long$
[4]$http://infolab.stanford.edu/\ wangz/project/imsearch/ART/SP08/sp_vangogh.pdf$
[5]$http://cs229.stanford.edu/proj2010/BlessingWen-UsingMachineLearningForIdentificationOfArtPaintings.pdf$
[6]$https://pdfs.semanticscholar.org/399c/6be15a90e94cc3610121566147e29520c9ea.pdf$
[7]$FahadShahbazKhan, ShidaBeigpour, JoostvandeWeijer, andMichaelFelsberg.Painting-91: alargescaledatabaseforcomputationalpaintingcategorization.Mach.Vis.App., 25(6): 1385\breve{}1397, 2014.$