Scattering Transform and Sparse Linear Classifiers for Art Authentication

Roberto Leonarduzzi\textsuperscript{a,}\textsuperscript{*,} Haixia Liu\textsuperscript{a,}\textsuperscript{*,} Yang Wang\textsuperscript{a}

\textsuperscript{a}Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Abstract

Recently, a novel signal-processing tool was proposed, the scattering transform, which uses a cascade of wavelet filters and nonlinear (modulus) operations to build translation-invariant and deformation-stable representations. Despite being aimed at providing a theoretical understanding of deep neural networks, it also shows state-of-the-art performance in image classification. In this paper, we explore its performance for art authentication purposes. We analyze two databases of art objects (postimpressionist paintings and Renaissance drawings) with the goal of determining those authored by van Gogh and Raphael, respectively. To that end, we combine scattering coefficients with several linear classifiers, in particular sparse \( \ell_1 \)-regularized classifiers. Results show that these tools provide excellent performance, superior to state-of-the-art results. Further, they suggest the benefits of using sparse classifiers in combination with deep networks.

Keywords

Art authentication Stylometry Scattering transform Sparse classifiers.

1. Introduction

Art authentication. Art authentication is an important problem in both art history and art collection. Two of the main problems in art authentication are authorship attribution and forgery detection. These tasks have traditionally been performed by art experts resorting to various forms of analysis including, among others, historical knowledge, characteristics of physical substrate (canvas, paper), and visual analysis of style. Recent technological advances, enabling the cheap acquisition and storage of high-resolution images, have led to an increased interest in the use of image processing tools to help art experts perform their tasks. Thus, recent years have seen an explosion of research work applying a vast number of image processing techniques for art authentication problems: fractal \[1\] and multifractal \[2\] analyses, tight frames \[3\], wavelets \[4\] \[5\] \[6\], empirical mode decomposition \[7\], Gabor filters \[8\] \[9\], and custom linear filters \[10\], among many others. One recurrent feature shared by all these techniques is the computation of a single layer of representations, based on space, scale and/or orientation information.

Scattering coefficients. The recent eruption in applications of deep neural networks (DNN) has brought forth breakthrough performance on image recognition tasks \[10\]. Such networks build many layers of hierarchical representations by cascading data-driven linear filters with nonlinear operators, and adding a classifier in the last layer. However, they suffer from two major drawbacks: (i) they lack a strong theoretical support, and (ii) they require large amounts of training data to achieve good performance. More recently, scattering networks have been proposed to address the first issue: they share the hierarchical structure of DNNs, but replace data-driven filters by wavelets \[11\], enabling a thorough mathematical analysis \[12\] \[11\] \[13\]. Moreover, they have also been shown to deliver excellent practical performance, notably in small-sample scenarios, and in many different domains: texture and digit discrimination \[14\], audio \[15\], and heart rate analysis \[10\].

Small-sample problems. Scattering networks show a better performance than DNNs on small sample-problems (cf. e.g. \[14\] for a comparison) mainly because they remove the need of learning linear filters from data. Thus, they only need to learn the coefficients in their last (classification) layer. However, scattering networks still suffer from a major drawback: their hierarchical structure makes them produce a fairly large amount of features. This not only makes the classification problem difficult when the number of samples is small—as is typically the case in art authentication problems—but also precludes a meaningful analysis of the features: it is difficult, if not impossible, to determine exactly which features are relevant. This information turns out to be crucial for art authentication problems, where the experts need to know what details of style are driving the decisions made by automatic tools. Moreover, knowledge of which features are involved may help designing the structure of the network.

Goals, contributions and outline. In consequence, the goal of the present work is twofold. First, we explore the use of scattering networks for author attribution on two datasets consisting of paintings and drawings. Second, we explore the use of sparse classifiers on scattering networks, with the aim of performing joint classification and feature selection. To these ends, we apply the scattering transform—described in Sec. \[3.1\]—jointly with a

\textsuperscript{*}Corresponding authors

Email addresses: roberto.leonarduzzi@sens-lyon.fr (Roberto Leonarduzzi), mahxliu@ust.hk (Haixia Liu), yangwang@ust.hk (Yang Wang)

© 2018. This manuscript version is made available under the CC-BY-NC-ND 4.0 license.
collection of linear classifiers—detailed in Sec. [3.2]—to two datasets of high-resolution scans of oil paintings and drawings (Sec. [2]). Our results (Sec. [4]) indicate that scattering networks achieve excellent performance, superior to state-of-the-art results reported in the literature. Further, we show that high classification performances can be achieved using a highly-sparse subset of scattering coefficients, greatly favoring the interpretation of results.

2. Data Collection and Preprocessing

Here we briefly describe the two datasets we analyze.

2.1. Van Gogh paintings

The van Gogh dataset consists of 79 high-resolution impressionist and post-impressionist color paintings from the van Gogh and Kröller-Müller museums. The painting sizes range from $1452 \times 833$ to $5614 \times 7381$ pixels. It is now generally agreed that 64 of these paintings were painted by van Gogh himself, while the remaining 15 were either forgeries or painted by his contemporaries in a similar style. The paintings by van Gogh are mainly from the Paris, Arles, and Saint-Remy periods, but also include four from his Auvers-sur-Oise days, a few months before his death. In the following, we will denote the two groups of paintings, according to their authorship, as VG (van Gogh) and NVG (non van Gogh). Figure 1 (top row) shows sample images of VG and NVG paintings. It should be noted that six of the NVG paintings have been attributed to van Gogh at different moments in history, but are currently known not to be painted by him. Table 1 lists these six once-debatable paintings, which are regarded as difficult examples for stylistic analysis.

2.2. Raphael Drawings

The Raphael dataset is provided by a private collector. It consists of 21 drawings of which 12 are considered genuine while the remaining 9 are not. Their sizes range from $2188 \times 3312$ to $6330 \times 4288$ pixels. In the following, we will denote the two groups of paintings, according to their authorship, as RA (Raphael) and NRA (non Raphael). Figure 1 (bottom row) shows sample RA and NRA drawings.

2.3. Preprocessing

All images were converted to gray scale, with double-precision values in the range $[0, 1]$. Since many of the images in the dataset included the edges of the canvas—which provide information that is obviously not useful for art authentication [6, 3]—we have used an automated edge detection algorithm based on Prewitt’s method [17] to adaptively detect and eliminate the canvas edges. The algorithm was limited to an edge width of at most 100 pixels.

3. Methods

3.1. Scattering transform

Notation. We will follow loosely the notation in [14, 11]. Let $X \in L^2(\mathbb{R}^2)$ be the image under analysis. Let $\psi : \mathbb{R}^2 \to \mathbb{C}$ be a mother wavelet—a complex, oscillating band-pass filter. A multiscale and oriented filter bank can be constructed by scaling and rotating $\psi$. Let $\theta \in \mathcal{R}(\mathbb{R}^2)$ be a rotation matrix in the finite, discrete rotation group $\mathcal{R}(\mathbb{R}^2)$. For simplicity, let us denote with $\lambda_i = 2^{-j_i}\theta_i$, $i = 1, \ldots, m$, the joint scaling and rotation operations. Then, the multiscale and oriented wavelet filters are defined as

$$\psi_{\lambda_i}(x) = 2^{-2j_i}\psi(2^{-j_i}\lambda_i x),$$

where the normalization factor $2^{-2j_i}$ ensures proper energy conservation in the frame of $L^2(\mathbb{R}^2)$ defined by $\{\psi_{\lambda_i}\}_{\lambda_i}$. Finally, let $\phi$ denote a low-pass filter, and let $\phi_{J}(x) = 2^{-2j_J}\phi(2^{-j_J}x)$.

Scattering coefficients. Wavelet coefficients are defined by the convolutions of $X$ with the wavelet filters: $X \ast \psi_{\lambda_i}$. A translation-invariant and deformation-stable representation, which is well suited for image recognition tasks, is obtained through nonlinear (modulus) and low-pass filtering operations, giving rise to first-layer scattering coefficients [11]:

$$S_{1,J}(\lambda_1, x) = |X \ast \psi_{\lambda_1} \ast \phi_{J}(x)|,$$

for wavelet scales $j_1 \leq J$.

The low-pass filter $\phi_{J}$ provides local translation invariance at scale $2^{-j}$, but also results in the loss of high-frequency information [11]. The lost information can be
recovered by additional band-pass filtering, giving rise to second-layer scattering coefficients [11]:

\[ S_{2,J}(\lambda_1, \lambda_2), x) = |X \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \phi_J(x), \quad (3) \]

for \( j_1 < j_2 \leq J \).

Note that the parameter \( J \)—the width of the low-pass filter—determines the length of local translation invariance, as well as the number of scales available from the transform [14]. Moreover, since \( S_{1,J} \) and \( S_{2,J} \) are the outputs of low-pass filters, they can be downsampled according to the filter width \( 2^J \).

Some intuition on scattering coefficients can be gained by noting that the modulus operation demodulates the spatial evolution of wavelet coefficients, centering their Fourier spectrum at zero frequency (and causing some distortion on its shape). The energy of \(|X \ast \psi_{\lambda_1}| \) is then captured in two ways: (i) through the low-pass filter \( \phi_J \), giving rise to first-layer coefficients; (ii) through the family of band-pass filters \( \psi_{\lambda_2} \), ultimately (i.e. after further demodulation and low-pass filtering) giving rise to second-layer coefficients. Thus, we can consider \( S_{2,J}(\lambda_1, \lambda_2), x) \) as measuring the spatial evolution, at scale \( 2^J \) and orientation \( \theta_2 \), of the (modulus of) wavelet coefficients at scale \( 2^J \) and orientation \( \theta_1 \).

Higher-order scattering coefficients can be built by further cascading the same operations [11]: wavelet-filtering, modulus, and low-pass filtering, giving rise to coefficients \( S_{m,J}(\Lambda, x) \), with \( \Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m) \), \( m = 1, 2, \ldots \). However, it has been shown that the energy of scattering coefficients decreases rapidly as the layer level increases, with almost 99% of the energy contained in the first two layers [11] [14]. Thus, we will focus only on first- and second-layer coefficients.

Finally, renormalized second order coefficients are defined as [15]:

\[ \tilde{S}_{2,J}(\lambda_1, \lambda_2) = \frac{S_{2,J}(\lambda_1, \lambda_2))}{S_{1,J}(\lambda_1)} \]  \( (4) \)

and further \( \tilde{S}_{1,J}(\lambda_1) = S_{1,J}(\lambda_1) \). The renormalization decorrelates coefficients that share a parent, increasing their invariance [15].

Features. Since our goal is to apply the scattering transform to a small-sample problem, we provide further dimension reduction by taking spatial averages of scattering coefficients:

\[ \bar{S}_{m,J}(\Lambda) = \sum_x \bar{S}_{m,J}(\Lambda, x), \quad m \in \{1, 2\}. \quad (5) \]

This procedure destroys the spatial information contained in scattering coefficients, and provides only a crude approximation of their distribution which is, nonetheless, sufficient for classification purposes, as Sec. 4 will show. In this setup, the parameter \( J \) determines only the largest scales used (since \( j_1 < j_2 \leq J \)) and, thus, the total number of features.

Number of features. Let \( J \) denote the number of scales, let \( R = |R(R^2)| \) be the number of rotations, and consider the input image is of size \( N \times N \). A straightforward calculation shows that the number of values the scattering parameters \( \Lambda \) can take is \( R^2(J^2 - J)/2 + JR = O(R^2J^2) \), which is also the number of averaged scattering features \( \hat{S} \) defined in (3). If the averaging was not performed, then each of the parameters would produce a matrix of subsampled scattering coefficients with \( N^22^{-2J} \) pixels (see [2], [3]). Thus, the total number of scattering coefficients in this case would rise to \( N^22^{-2J}(R^2(J^2 - J)/2 + JR) \), consequently increasing the data requirements for the classification layer.

Computational complexity. Scattering coefficients are efficiently computed with a 2D FFT by subsampling intermediate wavelet coefficients [14]. The resulting procedure computes a scattering transform of order \( m = 2 \) in \( O(R^2J^2N^22^{-2J}\log N) \) operations [14].

3.2. Classification layer

The last layer of a scattering network combines all features in a supervised classifier to predict the class-membership of data. In this section, we briefly discuss the used classifiers.

Guidelines. The choice of classifier must be guided by two considerations: (i) the number of scattering features is usually large, and (ii) in the particular case of the art authentication problem, usually only a small number of samples (art objects) are available to train the classifiers. As discussed in [18] [14], classifier performance
depends on the size of the training sample (relative to feature size): complex classifiers with small asymptotic error might not be preferable in small-sample scenarios. Thus, we consider several simple linear classifiers, both generative and discriminative, and compare their performances.

Further, we also consider $\ell_1$-norm-regularized, sparse linear classifiers which have been shown to have good performance on high-dimensional small-sample problems (the so-called “$p \gg n$” problems) [19]. This kind of classifier reduces the number of features involved in the classification, decreasing the risks of over-fitting and making an a posteriori analysis of the relevant features easier.

**Linear classification problem.** Let $f \in \mathbb{R}^p$ denote a zero-mean random feature vector, and let $f_n, n = 1, 2, \ldots, N$, denote the actual feature vectors computed from the available image set $\{X_n\}_{n=1}^N$. Let $k(f) \in \{1, 2\}$ denote the class to which $f$ belongs.

The linear classification problem attempts to predict the classes $k(f)$ through estimates $\hat{k}(f)$ based on linear combinations:

$$\hat{k}(f) = \begin{cases} 1 & \text{if } \hat{w}^T f > t \\ 2 & \text{otherwise,} \end{cases}$$

for some weight vector $\hat{w} \in \mathbb{R}^p$ and threshold $t \in \mathbb{R}$. We will consider several algorithms that provide estimates $\hat{w}$ and $\hat{t}$.

**Linear Discriminant Analysis (LDA).** This classifier models each class $k$ with multivariate normal distributions $P(f; \mu_k, \Sigma)$ having different means $\mu_k$ but the same covariance $\Sigma$ [19]. Classification is performed according to:

$$\hat{k}(f) = \arg \max_k P(f; \mu_k, \Sigma),$$

i.e. by choosing the class with largest probability to produce $f$, where parameters $\mu_k$ and $\Sigma$ are estimated from training data. It can be shown that (in the binary case) [7] is equivalent to the linear decision problem [6] with $w = \Sigma^{-1}(\mu_2 - \mu_1)$ and $t = (\mu_2 + \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1)$ [19].

**Sparse Linear Discriminant Analysis (SLDA).** This classifier, recently proposed in [20], provides a sparse version of the LDA weight vector by recasting (7) into a linear regression, and imposing an $\ell_1$ lasso penalty to induce sparsity:

$$(\hat{w}, \hat{t}, \Theta) \in \arg \min_{w, t, \Theta} \|Y \Theta - Fw - t1\|_2^2 + \delta \|w\|_1$$

s.t. $\Theta^T Y^T Y \Theta = 1, \Theta^T Y^T Y \Theta_1 = 0, \forall l < k,$

where $Y \in \{0, 1\}^{n \times 2}$ is a matrix with dummy encodings of the class labels, $F \in \mathbb{R}^{n \times p}$ is a data matrix stacking all sample features as rows, $\Theta$ is a scoring vector that transforms the binary dummy variables into continuous ones, and $1 \in \mathbb{R}^n$ is a vector whose entries are all 1.

The regularization parameter $\delta$ controls the trade-off between sparsity and classification accuracy, and is selected by cross-validation. Note that when $\delta = 0$ the solution provided by SLDA and LDA are equivalent [20].

**Support Vector Machine (SVM).** This traditional classifier finds the vector $\hat{w}$ in (6) that provides the optimal classification on the training data, in the sense of margin maximization [19]. It can be formulated through the optimization problem:

$$(\hat{w}, \hat{t}, \gamma_{opt}) \in \arg \min_{w, t, \gamma} \sum_{n=1}^N H_{w,t}(f_n, k(f_n)) + \gamma \|w\|_2^2, \quad (9)$$

where $H_{w,t}(f, k) = \max(0, 1 - (2k - 3)(w^T f + t))$ is the hinge loss function. Problem (9) is convex, and its single solution can be easily and efficiently obtained through quadratic programming [19].

**Sparse Support Vector Machine (SSVM).** Classic SVM is known to produce a nonsparse feature vector, where most entries are nonzero. To alleviate this problem, it has been proposed to replace the $\ell_2$ norm in (9) by an $\ell_1$ norm that induces sparsity [21] [22]. This, however, leads to an optimization problem that is too difficult to solve. The difficulty is overcome by a reformulation of the penalty in terms of a square hinge loss $H_w^2$:

$$(\hat{w}, \hat{t}) \in \arg \min_{w, t} \sum_{n=1}^N H_{w,t}(f_n, k(f_n)) + \delta \|w\|_1. \quad (10)$$

This problem is solved through a block-coordinate-descent algorithm [23] [21]. The regularization parameter $\delta$ controls the trade-off between sparsity and classification accuracy, and is selected by cross-validation.

**Principal Component Analysis (PCA).** This simple classifier based on projections on principal components was proposed in [14] and shown to have a good performance when used with scattering coefficients. Let $f^{(k)}$ denote the random feature vector for images in class $k$. Let $V_{d,k}$ be the linear space generated by the first $d$ eigenvalues of the covariance matrix of $f^{(k)}$ (i.e. the first principal directions). Finally, let $P_{d,k}$ be the orthogonal projection operator onto space $A$. Then, the PCA classifier assigns features $f$ to classes $k$ according to:

$$\hat{k}(f) = \arg \min_k \|f - P_{d,k}(f)\|_2,$$

i.e. by choosing the $d$-dimensional space that better approximates the feature $f$ in an $L_2$-error sense.

**Performance comparison.** Performance comparisons between classifiers with $L_1$ and $L_2$ regularizations have been extensively made in the literature, cf. e.g. [19] [20] [24] [25] [26]. The main interest in $L_1$-regularized classifiers lies in the fact that they perform joint classification
and feature selection, thus permitting a better interpretation of classification results. Moreover, $\ell_1$-regularized classifiers typically have a larger bias, but overall smaller mean squared error due to a reduced variance [19].

4. Results

4.1. Analysis setup

Wavelet analysis was performed using the so-called Morlet mother wavelet, which is essentially a modulated gaussian (cf. [14, 27]). Also, a gaussian window was used as the low-pass filter $\phi$.

Since the paintings on both datasets had been discretized at extremely high resolutions, they were divided into small patches of size $512 \times 512$, and each patch was analyzed independently. This led to 3755 patches in the van Gogh dataset (3204 VG and 551 NVG), and to 1168 patches in the Raphael dataset (479 RA and 689 NRA). This procedure not only allowed to overcome memory limitations, but also increased the available training data. The patch sizes were similar to those used on other studies, cf. e.g. [5, 2]. The analysis was repeated for several values of $J$, ranging from 2 to 7, compatible with the chosen patch size. Individual classifier decisions for all patches of the same picture were subsequently aggregated through a majority voting to reach a single decision for each painting. The analyses analysis were repeated for patches of 1024$^2$ and 2048$^2$ pixels; results were found not to change significantly. More details on the influence of the patch size are given in Sec. 4.3.

Cross-validation was used to select the classification parameters [14]: (i) the cutoff scale for scattering coefficients was set to $J = 4$; (ii) the regularization parameters were set to $\delta = 2^{-3}$ for SSVM and $\delta = 2^{-4.5}$ for SLDA. Performance was also measured by cross validation. For both purposes, a stratified 5-fold cross validation was used, with 11 repetitions of each loop to reduce the sampling bias.

Classifier performance was measured in terms of: (i) accuracy (ACC), defined as the proportion of correctly identified instances; (ii) sensitivity (SE), defined as the proportion of correctly identified positive instances; (iii) specificity (SP), defined as the proportion of correctly classified negative instances; (iv) area under ROC curve (AUC), defined as the area under the graph $(1 - SP(t), SE(t))$, where $t \in [-\infty, +\infty]$ is the threshold in [6]. We refer the reader to, e.g. [28] for further details. We consider the original art objects authored by van Gogh or Raphael to be the positive instances, with the rest being the negative ones.

All the analyses were performed in Matlab (tm). Scattering transforms were computed using the ScatNet toolbox2 [20]. SLDA was computed using the SpaSM toolbox$^3$.

SSVM was computed using our own custom code3. SVM and LDA were computed using the Statistics and Machine Learning Toolbox, Mathworks (tm).

4.2. Scattering coefficients

In order to provide some intuition into what scattering coefficients measure, Figure 2 shows examples of such coefficients, at the first and second layers, computed from a sample van Gogh painting (top row) and Raphael drawing (bottom row). On each row it displays: (a) the gray scale image, (b) the first-layer coefficients, and (c-d) two groups of second-layer coefficients derived from (b) with different parameters.

Analysis of the van Gogh painting (first row) shows that first-layer coefficients $S_{1,j}$ (second column) mainly detect edges and sharp transitions, as expected from their bandpass nature. Interestingly, second-layer coefficients $S_{2,j}$ (third and fourth columns) show details that are not easily seen in $S_{1,j}$, despite the fact that they share parameters $j_1$ and $\theta_1$: the third column captures fine-scale details such as the lines in the floor, while the fourth column captures coarse-scale details such as the interior of the window panels.

A similar analysis holds for the Raphael drawing: details that are not evident in the first layer are brought forward with a new application of the scattering operator—compare, for instance, the three representations of the standing person in the left.

4.3. Classification performance

4.3.1. Patch classification

As a first step, we analyze classification performance for individual patches. Table 2 shows classification performance metrics for the different classifiers under consideration. Notably, identical performance trends are observed on both datasets, and are commented below. First, Table 2 shows that performance is overall very good, achieving values of AUC generally larger than 0.9 regardless of the classifier. Second, the best AUC, 0.956 for VG (resp. 0.997 for RA), is obtained by the PCA-based classifier. Third, AUC of more complex classifiers such as LDA and SVM drops slightly to around 0.88 (resp. 0.97) but remains acceptable. The a priori unreasonable fact that the best performance is achieved with the simplest classifier is in agreement with the analysis and experimental results in [14, 18] given a relatively small sample size, simple, generative classifiers have a better performance than more complex ones.

The addition of an $\ell_1$ regularization to either LDA or SVM causes a very slight increase in AUC, enough to make SSVM the second best-performing classifier. More importantly, this slight increase in performance comes together with a large reduction in the number of used

---

1Available at http://www.di.ens.fr/data/software/scatnet/
2Available at http://www2.imm.dtu.dk/projects/spasm/
3Available upon request.
features with respect to their nonsparse counterparts: half (resp. a quarter) of the features are involved in the classification. This not only allows for a more efficient training, which entails robustness to overfitting, but also permits an interpretation of the relevant features for classification. Further analysis of the selected coefficients is delayed to Sec. 4.5.

It is worth mentioning that the PCA classifier also performs a large and drastic dimension reduction, since it only uses the projections on the first 47 (resp. 26) principal components. This intrinsic dimension reduction is a large contributor to the good performance shown by the classifier. However, each of these projections is a linear combination of all the 681 (resp. 417) features. In consequence, the PCA classifier does not perform a selection of features in the same sense that SLDA or SSVM do.

### 4.3.2. Painting classification: majority vote

The classifier decisions for each patch were aggregated through a majority vote to reach a decision for the full paintings. Results are shown in Table 3, which displays the confusion matrices and accuracy for all classifiers.

**Van Gogh dataset.** Table 3 indicates that performance on this dataset is excellent. For PCA and SSVM, only three paintings are misclassified; SLDA provides the second best result with 6 errors, followed by the nonsparse classifiers with significantly worse performance. To the best of our knowledge, the accuracy reported in Table 3 for the van Gogh dataset is superior to what has been reported so far in the literature for the classification of van Gogh’s paintings, as shown in Table 3. All results in Table 3 were obtained with similar training sizes and validation methods, making their comparison meaningful.

Furthermore, it is well known that leave-one-out (LOO) provides more pessimistic estimates of the training error than 5-fold cross validation (since there is less training data available on each fold)[19]. However, if the validation of our method is repeated using LOO, an identical performance is obtained after the voting procedure.

The false positives (non-VG classified as VG) with the PCA-based classifier are: **f687 (Reaper with sickle)** and

---

4. Two datasets are analyzed in this paper, one being a subset of the other. We select results for the larger one (IP4A12) because it has almost the same size as ours, favoring a fair comparison.
s503 (Windmills near Zaandam), while the false negative (VG classified as non-VG) is f249 (Bowl with peonies and rose). Similarly, SSVM commits only one false positive error, s503, but pays the cost of two false negatives: 216a (Plaster statuette of a female torso) and f360 (Still life with plaster statuette, a rose and two novels). Note that painting s503, is misclassified in both cases, suggesting a very similar style to van Gogh. Also, painting f687 is one of those included in Table 1 and is thus so similar in style to van Gogh as to have been misclassified by experts in the past; however, the remaining five paintings in Table 1 are correctly classified in both cases, a good outcome for our automatic method. Finally, it is also worth noticing that the three false negatives are still lifes, and thus might share some stylistic features that prevent their correct classification.

Raphael dataset. Table 3 shows that classification is perfect except for LDA, which commits only one error. This result is in strong agreement with the large accuracy reported in Table 2 (bottom half) for the individual patches. These results suggest that the classification problem on this dataset is significantly easier than on the van Gogh dataset.

### 4.4. Influence of patch size and cutoff scale

In this section, we explore the influence of patch size in classification performance. Figure 3 shows classification accuracy as a function of the cutoff scale \( J \), for square patches of side 512, 1024 and 2048, on the van Gogh dataset and using the PCA classifier. First, Fig. 3 (left) considers the classification of individual patches. It can be seen that performance with patches of size 512 and 2048 is remarkably similar when \( J \geq 3 \), whereas patches of size 1024 yield a slightly higher accuracy. In all cases, the best performance is achieved for \( J = 4 \), and performance decreases rapidly as \( J \) increases. This indicates that local, fine-scale details of size 16 × 16 are preferred for the classification, and that larger patches sizes do not entail the use of more coarse-scale information.

Next, Fig. 3 (right) considers the classification of full paintings through the voting procedure. In a similar way to individual patches, the best performances are achieved for \( J = 4 \). Notably, in this case the best accuracy is achieved with the smaller 512 × 512 patches, despite their slightly worse performance for individual patches.

The results reported in this section justify our selection of \( J = 4 \) and patches of size 512 × 512, as detailed in Sec. 4.4. These parameters provide the best performance for full paintings and an acceptable performance for individual patches, while incurring a minimal computational cost.

### 4.5. Feature selection

As previously stated, sparse classifiers crucially enable a simple interpretation of the important features for the
classification task, thus bringing light into exactly which characteristics of an image are relevant. In this section we will analyze in this light the selected weights for both datasets.

**Display of weights.** The weights given by SSVM (left) and SLDA (right) to each scattering coefficient are displayed using the polar plots introduced in [14]. First-layer coefficients are shown as annular sections of surface and radial position proportional to \( j_1 \), and rotated counter-clockwise according to \( \theta_1 \). For second-layer coefficients, each first-layer annular section is further divided radially according to \( j_2 \) and angularly according to \( \theta_2 \) (note that, as \( j_1 \) increases there are fewer subdivisions due to the constraint \( j_1 < j_2 \leq J \)). We refer the reader to, e.g., [14] for more details on this type of plots.

**Van Gogh Paintings.** Figure 4 shows the classifier weights for SSVM and SLDA corresponding to the results shown in Table 2. First, they further illustrate what was already hinted in Table 2: the weights are highly sparse, with around half of them being zero. This means that half of the features produced by the scattering transform are not actually relevant for the classification task and can be safely and conveniently discarded.

Secondly, weights selected by SSVM and SLDA somewhat resemble each other: most large weights in one case correspond to an equivalent or neighboring weight in the other. The fact that similar solutions are reached by two different optimization algorithms gives confidence on their convergence and corroborates the results.

Finally, and most importantly, most first-layer coefficients are set to zero or to relatively small values. On the contrary, the largest weights are assigned to second-layer coefficients, which are also selected for many scales and orientations, and appear scattered throughout the polar plot. This suggests that classification is mainly driven by the second layer, and that there are no scales or orientations that play a dominant and preferential role; rather, a large combination of them is used.

The fact that mainly coefficients at the second layer are selected deserves a special emphasis: it means that the additional information provided by the scattering network over the “traditional” wavelet transform (i.e. the first layer) is crucial for classification. In other words, the relevant information for authentication of VG paintings does not seem to be encoded in the oscillatory behavior at a given scale or orientation—as measured by first-layer wavelet coefficients—but rather by the spatial modulation of such behavior as measured by second-layer scattering coefficients.

**Raphael Dataset.** Figure 5 shows the classifier weights for SSVM and SLDA corresponding to the results shown in Table 2. Again, and similarly to the previous case, the weights chosen by SSVM and SLDA are remarkably similar and highly sparse. However, in sharp contradiction with the previous case, it is first-layer coefficients that are mainly selected for the Raphael dataset. In fact, only a small subset of second-layer coefficients are given relatively small weights.

These results further confirm that the Raphael classification problem is much simpler, requiring only a few first-layer, oriented, band-pass filters to achieve a perfect performance. On the contrary, the much more demanding van Gogh dataset, as discussed before, requires the use of many higher-order, nonlinear features to build suitable invariants on which to base classification.

Figure 5 suggests that identification of Raphael drawings is based on a combination of high- and low-spatial frequencies at essentially diagonal orientations. Such features could be associated, for instance, with the orientation and separation of the pencil strokes.

**Remarks.** Our detailed analysis of classification weights confirms that the additional information provided by sparse classifiers is extremely valuable. Indeed, it shows that while a high-accuracy classification could be achieved by a single layer of wavelet filters on one dataset, the cascaded and nonlinear structure provided by scattering coefficients turns out to be crucial on the other. Thus, this knowledge can, for instance, provide guidelines for the design of the networks for similar classification problems.

Moreover, the physical interpretation of these quantities can provide valuable insight on the nature of the relevant features. Drawings on the Raphael dataset are characterized by the total activity at a few spatial frequencies and orientations. On the contrary, identification
in the van Gogh dataset is based on the spatial modulation of the bandpass details, rather than on their total magnitude.

5. Conclusions

This contribution proposed the combined use of the scattering transform and $\ell_1$-penalized linear sparse classifiers to perform art authentication. The performance of these techniques was illustrated on two datasets of different natures, and was shown to provide very promising results.

The scattering transform provides an excellent classification tool for application domains such as art authentication, where the available data is naturally scarce. It provides access to the complex, high-performing features introduced by deep neural networks, but at a much smaller cost: filters are fixed to be wavelets rather than learned.

Sparse classifiers tackle a remaining issue on scattering networks: their production of a large number of features. Our results show that a large part of the scattering coefficients are actually not needed for classification purposes. Further, our results show that different datasets perform better with drastically different network architectures: while the second layer is preferred for van Gogh’s paintings, the first one is chosen for Raphael’s drawings. In consequence, the feature-selection ability of sparse classifiers provides extremely relevant information for the practitioner, and should be routinely used with such networks.

We believe that these tools are useful to support art historians in their scholar tasks, by revealing relevant features deeply hidden in the data.

6. Acknowledgments

We thank Profs. Haixiang Lin (Delft University of Technology), Eric Postma (Tilburg Center for Cognition and Communication) and Raymond H. Chan (The Chinese University of Hong Kong) for providing us the 79 paintings in the van Gogh dataset. We also thank Edward Rosser for making available the 21 drawings in the Raphael dataset.

This work has been partially supported by the Hong Kong Research Grant Council grants 16306415 and 16317416.

References


