Invariant Scattering Convolution Networks

Joan Bruna, *Member*, *IEEE*, and Stéphane Mallat, *Fellow*, *IEEE*

Abstract—A wavelet scattering network computes a translation invariant image representation which is stable to deformations and preserves high-frequency information for classification. It cascades wavelet transform convolutions with nonlinear modulus and averaging operators. The first network layer outputs SIFT-type descriptors, whereas the next layers provide complementary invariant information that improves classification. The mathematical analysis of wavelet scattering networks explains important properties of deep convolution networks for classification. A scattering representation of stationary processes incorporates higher order moments and can thus discriminate textures having the same Fourier power spectrum. State-of-the-art classification results are obtained for handwritten digits and texture discrimination, with a Gaussian kernel SVM and a generative PCA classifier.

Index Terms—Classification, convolution networks, deformations, invariants, wavelets

1 Introduction

A major difficulty of image classification comes from the considerable variability within image classes and the inability of euclidean distances to measure image similarities. Part of this variability is due to rigid translations, rotations, or scaling. This variability is often uninformative for classification and should thus be eliminated. In the framework of kernel classifiers [32], the distance between two signals x and x' is defined as a euclidean distance $\|\Phi x - \Phi x'\|$ applied to a representation Φx of each x. Variability due to rigid transformations is removed if Φ is invariant to these transformations.

Nonrigid deformations also induce important variability within object classes [17], [3]. For instance, in handwritten digit recognition, one must take into account digit deformations due to different writing styles [3]. However, a full deformation invariance would reduce discrimination since a digit can be deformed into a different digit, for example, a one into a seven. The representation must therefore not be deformation invariant. It should linearize small deformations, to handle them effectively with linear classifiers. Linearization means that the representation is Lipschitz continuous to deformations. When an image x is slightly deformed into x', then $\|\Phi x - \Phi x'\|$ must be bounded by the size of the deformation, as defined in Section 2.

Translation invariant representations can be constructed with registration algorithms [33], autocorrelations, or with the Fourier transform modulus. However, Section 2.1 explains that these invariants are not stable to deformations and hence not adapted to image classification. Trying to avoid Fourier transform instabilities suggests replacing sinusoidal waves by localized waveforms such as wavelets. However, wavelet transforms are not invariant but covariant to translations. Building invariant representations from wavelet coefficients requires introducing nonlinear operators, which leads to a convolution network architecture.

Deep convolutional networks have the ability to build large-scale invariants which seem to be stable to deformations [20]. They have been applied to a wide range of image classification tasks. Despite the successes of this neural network architecture, the properties and optimal configurations of these networks are not well understood because of cascaded nonlinearities. Why use multiple layers? How many layers? How do we optimize filters and pooling nonlinearities? How many internal and output neurons? These questions are mostly answered through numerical experimentations that require significant expertise.

We address these questions from a mathematical and algorithmic perspective by concentrating on a particular class of deep convolutional networks, defined by the scattering transforms introduced in [24] and [25]. A scattering transform computes a translation invariant representation by cascading wavelet transforms and modulus pooling operators, which average the amplitude of iterated wavelet coefficients. It is Lipschitz continuous to deformations, while preserving the signal energy [25]. Scattering networks are described in Section 2 and their properties are explained in Section 3. These properties guide the optimization of the network architecture to retain important information while avoiding useless computations.

An expected scattering representation of stationary processes is introduced for texture discrimination. As opposed to the Fourier power spectrum, it gives information on higher order moments and can thus discriminate non-Gaussian textures having the same power spectrum. Scattering coefficients provide consistent estimators of expected scattering representations.

Classification applications are studied in Section 4. Classifiers are implemented with a Gaussian kernel SVM and a generative classifier which selects affine space models computed with a PCA. State-of-the-art results are obtained for handwritten digit recognition on MNIST and USPS

- *J. Bruna is with Courant Institute, New York University, 715 Broadway, New York, NY 10003. E-mail: joan.bruna@gmail.com.*
- *S. Mallat is with the Ecole Normale Supérieure, 45 rue d'Ulm, Paris 75005, France. E-mail: mallat@di.ens.fr.*

Manuscript received 7 Mar. 2012; revised 28 July 2012; accepted 5 Sept. 2012; published online 12 Oct. 2012.
Recommended for acceptance by S. Bengio, L. Deng, H. Larochelle, H. Lee, and R. Salakhutdinov.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMISI-2012-03-0168.
Digital Object Identifier no. 10.1109/TPAMI.2012.230.

databases, and for texture discrimination. These are problems where translation invariance, stationarity, and deformation stability play a crucial role. Software is available at www.di.ens.fr/data/scattering.

2 Toward a Convolution Network

Small deformations are nearly linearized by a representation if the representation is Lipschitz continuous to the action of deformations. Section 2.1 explains why high frequencies are sources of instabilities, which prevent standard invariants to be Lipschitz continuous. Section 2.2 introduces a wavelet-based scattering transform, which is translation invariant and Lipschitz relative to deformations. Section 2.3 describes its convolution network architecture.

2.1 Fourier and Registration Invariants

A representation Φx is invariant to global translations $x_c(u) = x(u-c)$ by $c = (c_1, c_2) \in \mathbb{R}^2$ if

$$\Phi x_c = \Phi x. \tag{1}$$

A canonical invariant [17], [33] $\Phi x(u) = x(u - a(x))$ registers x with an anchor point $a(x)$, which is translated when x is translated: $a(x_c) = a(x) + c$. It is therefore invariant: $\Phi x_c = \Phi x$. For example, the anchor point may be a filtered maximum $a(x) = \arg\max_u |x \star h(u)|$ for some filter $h(u)$.

The Fourier transform modulus is another example of translation invariant representation. Let $\hat{x}(\omega)$ be the Fourier transform of $x(u)$. Since $\widehat{x_c}(\omega) = e^{-ic.\omega}\,\hat{x}(\omega)$, it results that $|\widehat{x_c}| = |\hat{x}|$ does not depend upon c. The autocorrelation $Rx(v) = \int x(u)x(u-v)du$ is also translation invariant: $Rx = Rx_c$.

To be stable to additive noise $x'(u) = x(u) + \epsilon(u)$, we need a Lipschitz continuity condition which supposes that there exists $C > 0$ such that for all x and x':

$$\|\Phi x' - \Phi x\| \leq C\, \|x' - x\|,$$

where $\|x\|^2 = \int |x(u)|^2\, du$. The Plancherel formula proves that the Fourier modulus $\Phi x = |\hat{x}|$ satisfies this property with $C = 2\pi$.

To be stable to deformation variabilities, Φ must also be *Lipschitz continuous to deformations*. A small deformation of x can be written $x_\tau(u) = x(u - \tau(u))$, where $\tau(u)$ is a nonconstant displacement field that deforms the image. The deformation gradient tensor $\nabla\tau(u)$ is a matrix whose norm $|\nabla\tau(u)|$ measures the deformation amplitude at u and $\sup_u |\nabla\tau(u)|$ is the global deformation amplitude. A small deformation is invertible if $|\nabla\tau(u)| < 1$ [1]. Lipschitz continuity relative to deformations is obtained if there exists $C > 0$ such that for all τ and x:

$$\|\Phi x_\tau - \Phi x\| \leq C\, \|x\| \sup_u |\nabla\tau(u)|. \tag{2}$$

This property implies global translation invariance because if $\tau(u) = c$, then $\nabla\tau(u) = 0$, but it is much stronger.

If Φ is Lipschitz continuous to deformations τ, then the Radon-Nykodým property proves that the map that transforms τ into Φx_τ is almost everywhere differentiable in the sense of Gâteaux [22]. It means that for small deformations, $\Phi x - \Phi x_\tau$ is closely approximated by a bounded linear operator of τ, which is the Gâteaux derivative. Deformations are thus linearized by Φ, which enables linear classifiers to effectively handle deformation variabilities in the representation space.

A Fourier modulus is translation invariant and stable to additive noise but unstable to small deformations at high frequencies. Indeed, $|\,|\hat{x}(\omega)| - |\widehat{x_\tau}(\omega)|\,|$ can be arbitrarily large at a high frequency ω, even for small deformations, and in particular for a small dilation $\tau(u) = \epsilon u$. As a result, $\Phi x = |\hat{x}|$ does not satisfy the deformation continuity condition (2) [25]. The autocorrelation $\Phi x = Rx$ satisfies $\widehat{Rx}(\omega) = |\hat{x}(\omega)|^2$. The Plancherel formula thus proves that it has the same instabilities as a Fourier transform:

$$\|Rx - Rx_\tau\| = (2\pi)^{-1} \|\,|\hat{x}|^2 - |\hat{x_\tau}|^2\|.$$

Besides deformation instabilities, the Fourier modulus and the autocorrelation lose too much information. For example, a Dirac $\delta(u)$ and a linear chirp e^{iu^2} are two signals having Fourier transforms whose moduli are equal and constant. Very different signals may not be discriminated from their Fourier modulus.

A registration invariant $\Phi x(u) = x(u - a(x))$ carries more information than a Fourier modulus, and characterizes x up to a global absolute position information [33]. However, it has the same high-frequency instability as a Fourier transform. Indeed, for any choice of anchor point $a(x)$, applying the Plancherel formula proves that

$$\|x(u - a(x)) - x'(u - a(x'))\| \geq (2\pi)^{-1}\, \|\,|\hat{x}(\omega)| - |\hat{x}'(\omega)|\,\|.$$

If $x' = x_\tau$, the Fourier transform instability at high frequencies implies that $\Phi x(u) = x(u - a(x))$ is also unstable with respect to deformations.

2.2 Scattering Wavelets

A wavelet transform computes convolutions with dilated and rotated wavelets. Wavelets are localized waveforms and are thus stable to deformations, as opposed to Fourier sinusoidal waves. However, convolutions are translation covariant, not invariant. A scattering transform builds nonlinear invariants from wavelet coefficients, with modulus and averaging pooling functions.

Let G be a group of rotations r of angles $2k\pi/K$ for $0 \leq k < K$. Two-dimensional directional wavelets are obtained by rotating a single band-pass filter ψ by $r \in G$ and dilating it by 2^j for $j \in Z$:

$$\psi_\lambda(u) = 2^{-2j}\psi(2^{-j}r^{-1}u) \text{ with } \lambda = 2^{-j}r. \tag{3}$$

If the Fourier transform $\hat{\psi}(\omega)$ is centered at a frequency η, then $\hat{\psi}_{2^{-j}r}(\omega) = \hat{\psi}(2^j r^{-1}\omega)$ has a support centered at $2^{-j}r\eta$ and a bandwidth proportional to 2^{-j}. The index $\lambda = 2^{-j}r$ gives the frequency location of ψ_λ and its amplitude is $|\lambda| = 2^{-j}$.

The wavelet transform of x is $\{x \star \psi_\lambda(u)\}_\lambda$. It is a redundant transform with no orthogonality property. Section 3.1 explains that it is stable and invertible if the wavelet filters $\hat{\psi}_\lambda(\omega)$ cover the whole frequency plane. On discrete images, to avoid aliasing, we only capture frequencies in the circle $|\omega| \leq \pi$ inscribed in the image frequency square. Most camera images have negligible energy outside this frequency circle.

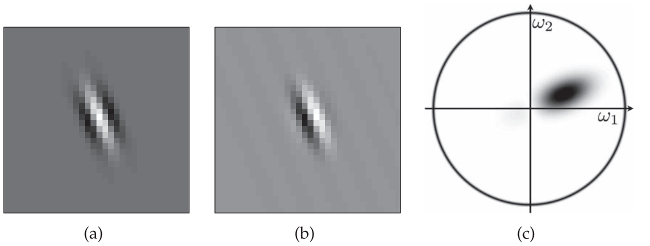

Fig. 1. Complex Morlet wavelet. (a) Real part of $\psi(u)$. (b) Imaginary part of $\psi(u)$. (c) Fourier modulus $|\hat{\psi}(\omega)|$.

Let $u.u'$ and $|u|$ denote the inner product and norm in $\mathbb{R}^2$. A Morlet wavelet ψ is an example of complex wavelet given by

$$\psi(u) = \alpha\, (e^{iu.\xi} - \beta)\, e^{-|u|^2/(2\sigma^2)},$$

where $\beta \ll 1$ is adjusted so that $\int \psi(u)\, du = 0$. Its real and imaginary parts are nearly quadrature phase filters. Fig. 1 shows the Morlet wavelet with $\sigma = 0.85$ and $\xi = 3\pi/4$, used in all classification experiments.

A wavelet transform commutes with translations, and is therefore not translation invariant. To build a translation invariant representation, it is necessary to introduce a nonlinearity. If Q is a linear or nonlinear operator that commutes with translations, then $\int Qx(u)\, du$ is translation invariant. Applying this to $Qx = x \star \psi_\lambda$ gives a trivial invariant $\int x \star \psi_\lambda(u)\, du = 0$ for all x because $\int \psi_\lambda(u)\, du = 0$. If $Qx = M(x \star \psi_\lambda)$ and M is linear and commutes with translations, then the integral still vanishes. This shows that computing invariants requires a nonlinear pooling operator M, but which one?

To guarantee that $\int M(x \star \psi_\lambda)(u)\, du$ is stable to deformations, we want M to commute with the action of any diffeomorphism. Additionally, to preserve stability to additive noise we also want M to be nonexpansive: $\|My - Mz\| \leq \|y - z\|$. If M is a nonexpansive operator that commutes with the action of diffeomorphisms, then one can prove [7] that M is necessarily a pointwise operator. It means that $My(u)$ is a function of the value $y(u)$ only. If, moreover, we want invariants which also preserve the signal energy, we shall choose a modulus operator over complex signals $y = y_r + i\, y_i$:

$$My(u) = |y(u)| = (|y_r(u)|^2 + |y_i(u)|^2)^{1/2}. \tag{4}$$

The resulting translation invariant coefficients are then $\mathbf{L}^1(\mathbb{R}^2)$ norms:

$$\|x \star \psi_\lambda\|_1 = \int |x \star \psi_\lambda(u)|\, du.$$

The $\mathbf{L}^1(\mathbb{R}^2)$ norms $\{\|x \star \psi_\lambda\|_1\}_\lambda$ form a crude signal representation which measures the sparsity of wavelet coefficients. The loss of information does not come from the modulus that removes the complex phase of $x \star \psi_\lambda(u)$. Indeed, one can prove [37] that x can be reconstructed from the modulus of its wavelet coefficients $\{|x \star \psi_\lambda(u)|\}_\lambda$, up to a multiplicative constant. The information loss comes from the integration of $|x \star \psi_\lambda(u)|$, which removes all nonzero frequencies. These nonzero frequencies are recovered by calculating the wavelet coefficients $\{|x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(u)\}_{\lambda_2}$ of $|x \star \psi_{\lambda_1}|$. Their $\mathbf{L}^1(\mathbb{R}^2)$ norms define a much larger family of invariants, for all λ_1 and λ_2:

$$\||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1 = \int \||x \star \psi_{\lambda_1}(u)| \star \psi_{\lambda_2}|\, du.$$

More translation invariant coefficients can be computed by further iterating on the wavelet transform and modulus operators. Let $U[\lambda]x = |x \star \psi_\lambda|$. Any sequence $p = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ defines a *path*, along which is computed an ordered product of nonlinear and noncommuting operators:

$$\begin{aligned} U[p]x &= U[\lambda_m]\ \cdots\ U[\lambda_2]\ U[\lambda_1]x \\ &= |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|\ \cdots\ | \star \psi_{\lambda_m}|, \end{aligned}$$

with $U[\emptyset]x = x$. A scattering transform along the path p is defined as an integral, normalized by the response of a Dirac:

$$\overline{S}x(p) = \mu_p^{-1} \int U[p]x(u)\, du \text{ with } \mu_p = \int U[p]\delta(u)\, du.$$

Each scattering coefficient $\overline{S}x(p)$ is invariant to a translation of x. We shall see that this transform has many similarities with the Fourier transform modulus, which is also translation invariant. However, a scattering is Lipschitz continuous to deformations, as opposed to the Fourier transform modulus.

For classification, it is often better to compute localized descriptors that are invariant to translations smaller than a predefined scale 2^J while keeping the spatial variability at scales larger than 2^J. This is obtained by localizing the scattering integral with a scaled spatial window $\phi_{2^J}(u) = 2^{-2J}\phi(2^{-J}u)$. It defines a windowed scattering transform in the neighborhood of u:

$$S[p]x(u) = U[p]x \star \phi_{2^J}(u) = \int U[p]x(v)\phi_{2^J}(u - v)\, dv,$$

and hence

$$S[p]x(u) = |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|\ \cdots\ | \star \psi_{\lambda_m}| \star \phi_{2^J}(u),$$

with $S[\emptyset]x = x \star \phi_{2^J}$. For each path p, $S[p]x(u)$ is a function of the window position u, which can be subsampled at intervals proportional to the window size 2^J. The averaging by ϕ_{2^J} implies that if $x_c(u) = x(u - c)$ with $|c| \ll 2^J$, then the windowed scattering is nearly translation invariant: $S[p]x \approx S[p]x_c$. Stability relatively to deformations is reviewed in Section 3.1.

2.3 Scattering Convolution Network

If $p = (\lambda_1, \ldots, \lambda_m)$ is a path of length m, then $S[p]x(u)$ is called a windowed scattering coefficient of order m. It is computed at the layer m of a convolution network that is specified. For large scale invariants, several layers are necessary to avoid losing crucial information.

For appropriate wavelets, first-order coefficients $S[\lambda_1]x$ are equivalent to SIFT coefficients [23]. Indeed, SIFT computes the local sum of image gradient amplitudes among image gradients having nearly the same direction in a histogram having eight different direction bins. The DAISY approximation [34] shows that these coefficients are well approximated by $S[\lambda_1]x = |x \star \psi_{\lambda_1}| \star \phi_{2^J}(u)$, where ψ_{λ_1} are partial derivatives of a Gaussian computed at the finest

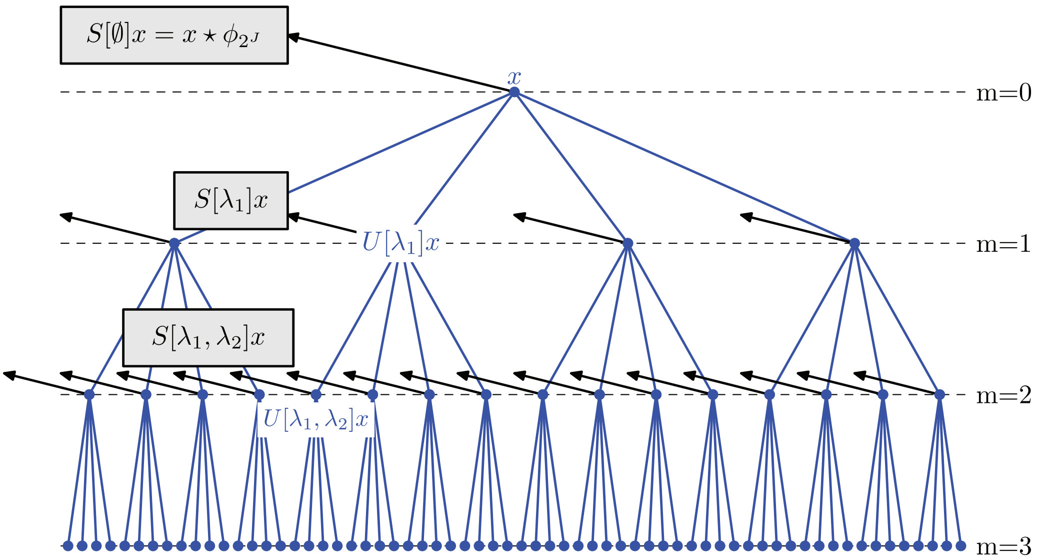
$S[\emptyset]x = x \star \phi_{2^J}$
x
m=0
$S[\lambda_1]x$
$U[\lambda_1]x$
m=1
$S[\lambda_1,\lambda_2]x$
m=2
$U[\lambda_1,\lambda_2]x$
m=3

Fig. 2. A scattering propagator $\widetilde{W}$ applied to x computes the first layer of wavelet coefficients modulus $U[\lambda_1]x = |x \star \psi_{\lambda_1}|$ and outputs its local average $S[\emptyset]x = x \star \phi_{2^J}$ (black arrow). Applying $\widetilde{W}$ to the first layer signals $U[\lambda_1]x$ outputs first-order scattering coefficients $S[\lambda_1] = U[\lambda_1] \star \phi_{2^J}$ (black arrows) and computes the propagated signal $U[\lambda_1,\lambda_2]x$ of the second layer. Applying $\widetilde{W}$ to each propagated signal $U[p]x$ outputs $S[p]x = U[p]x \star \phi_{2^J}$ (black arrows) and computes the next layer of propagated signals.

image scale, along eight different rotations. The averaging filter ϕ_{2^J} is a scaled Gaussian.

Partial derivative wavelets are well adapted to detecting edges or sharp transitions but do not have enough frequency and directional resolution to discriminate complex directional structures. For texture analysis, many researchers [21], [31] have been using averaged wavelet coefficient amplitudes $|x \star \psi_\lambda| \star \phi_{2^J}(u)$, calculated with a complex wavelet ψ having a better frequency and directional resolution.

A scattering transform computes higher order coefficients by further iterating on wavelet transforms and modulus operators. Wavelet coefficients are computed up to a maximum scale 2^J and the lower frequencies are filtered by $\phi_{2^J}(u) = 2^{-2J}\phi(2^{-J}u)$. For a Morlet wavelet ψ, the averaging filter ϕ is chosen to be a Gaussian. Since images are real-valued signals, it is sufficient to consider "positive" rotations $r \in G^+$ with angles in $[0,\pi)$:

$$Wx(u) = \{x \star \phi_{2^J}(u), x \star \psi_\lambda(u)\}_{\lambda \in \mathcal{P}}, \tag{5}$$

with an index set $\mathcal{P} = \{\lambda = 2^{-j}r : r \in G^+, j \leq J\}$. Let us emphasize that 2^J and 2^j are spatial scale variables, whereas $\lambda = 2^{-j}r$ is a frequency index giving the location of the frequency support of $\hat{\psi}_\lambda(\omega)$.

A wavelet modulus propagator keeps the low-frequency averaging and computes the modulus of complex wavelet coefficients:

$$\widetilde{W}x(u) = \{x \star \phi_{2^J}(u), |x \star \psi_\lambda(u)|\}_{\lambda \in \mathcal{P}}. \tag{6}$$

Iterating on $\widetilde{W}$ defines a convolution network illustrated in Fig. 2.

The network nodes of the layer m correspond to the set $\mathcal{P}^m$ of all paths $p = (\lambda_1, \ldots, \lambda_m)$ of length m. This mth layer stores the propagated signals $\{U[p]x\}_{p \in \mathcal{P}^m}$ and outputs the scattering coefficients $\{S[p]x\}_{p \in \mathcal{P}^m}$. For any $p = (\lambda_1, \ldots, \lambda_m)$, we denote $p + \lambda = (\lambda_1, \ldots, \lambda_m, \lambda)$. Since $S[p]x = U[p]x \star \phi_{2^J}$ and $U[p+\lambda]x = |U[p]x \star \psi_\lambda|$, it results that

$$\widetilde{W}\, U[p]x = \{S[p]x, U[p+\lambda]x\}_{\lambda \in \mathcal{P}}.$$

Applying $\widetilde{W}$ to all propagated signals $U[p]x$ of the mth layer $\mathcal{P}^m$ outputs all scattering signals $S[p]x$ and computes all propagated signals $U[p+\lambda]$ on the next layer $\mathcal{P}^{m+1}$. All output scattering signals $S[p]x$ along paths of length $m \leq \overline{m}$ are thus obtained by first calculating $\widetilde{W}x = \{S[\emptyset]\,x, U[\lambda]x\}_{\lambda \in \mathcal{P}}$ and then iteratively applying $\widetilde{W}$ to each layer of propagated signals for increasing $m \leq \overline{m}$.

The translation invariance of $S[p]x$ is due to the averaging of $U[p]x$ by ϕ_{2^J}. It has been argued [8] that an average pooling loses information, which has motivated the use of other operators such as hierarchical maxima [9]. A scattering avoids this information loss by recovering wavelet coefficients at the next layer, which explains the importance of a multilayer network structure.

A scattering is implemented by a deep convolution network [20] having a very specific architecture. As opposed to standard convolution networks, output scattering coefficients are produced by each layer as opposed to the last layer [20]. Filters are not learned from data but are predefined wavelets. Indeed, they build invariants relative to the action of the translation group, which does not need to be learned. Building invariants to other known groups such as rotations or scaling is similarly obtained with predefined wavelets which perform convolutions along rotation or scale variables [25], [26].

Different complex quadrature phase wavelets may be chosen, but separating signal variations at different scales is fundamental for deformation stability [25]. Using a modulus (4) to pull together quadrature phase filters is also important to remove the high-frequency oscillations of wavelet coefficients. The next section explains that it guarantees a fast energy decay of propagated signals $U[p]x$ across layers so that we can limit the network depth.

For a fixed position u, windowed scattering coefficients $S[p]x(u)$ of order $m = 1, 2$ are displayed as piecewise

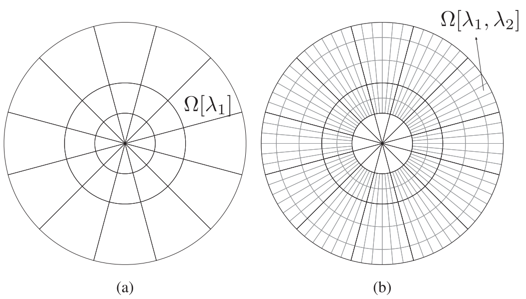

Fig. 3. To display scattering coefficients, the disk covering the image frequency support is partitioned into sectors $\Omega[p]$, which depend upon the path p. (a) For $m = 1$, each $\Omega[\lambda_1]$ is a sector rotated by r_1 that approximates the frequency support of $\hat{\psi}_{\lambda_1}$. (b) For $m = 2$, all $\Omega[\lambda_1, \lambda_2]$ are obtained by subdividing each $\Omega[\lambda_1]$.

constant images over a disk representing the Fourier support of the image x. This frequency disk is partitioned into sectors $\{\Omega[p]\}_{p \in \mathcal{P}^m}$ indexed by the path p. The image value is $S[p]x(u)$ on the frequency sectors $\Omega[p]$, shown in Fig. 3.

For $m = 1$, a scattering coefficient $S[\lambda_1]x(u)$ depends upon the local Fourier transform energy of x over the support of $\hat{\psi}_{\lambda_1}$. Its value is displayed over a sector $\Omega[\lambda_1]$ that approximates the frequency support of $\hat{\psi}_{\lambda_1}$. For $\lambda_1 = 2^{-j_1} r_1$, there are K rotated sectors located in an annulus, corresponding to each $r_1 \in G$, as shown by Fig. 3a. Their areas are proportional to $\|\psi_{\lambda_1}\|^2 \sim K^{-1}\, 2^{-j_1}$.

Second-order scattering coefficients $S[\lambda_1, \lambda_2]x(u)$ are computed with a second wavelet transform that performs a second frequency subdivision. These coefficients are displayed over frequency sectors $\Omega[\lambda_1, \lambda_2]$ that subdivide the sectors $\Omega[\lambda_1]$ of the first wavelets $\hat{\psi}_{\lambda_1}$, as illustrated in Fig. 3b. For $\lambda_2 = 2^{-j_2} r_2$, the scale 2^{j_2} divides the radial axis, and the resulting sectors are subdivided into K angular sectors corresponding to the different r_2. The scale and angular subdivisions are adjusted so that the area of each $\Omega[\lambda_1, \lambda_2]$ is proportional to $\||\psi_{\lambda_1}| \star \psi_{\lambda_2}\|^2$.

Fig. 4 shows the Fourier transform of two images and the amplitude of their scattering coefficients. In this case, the scale 2^J is equal to the image size. The top and bottom images are very different, but they have the same first-order scattering coefficients. The second-order coefficients clearly discriminate these images. Section 3.1 shows that the second-order scattering coefficients of the top image have a larger amplitude because the image wavelet coefficients are more sparse. Higher order coefficients are not displayed because they have a negligible energy, as explained in Section 3.

3 Scattering Properties

A convolution network is highly nonlinear, which makes it difficult to understand how the coefficient values relate to the signal properties. For a scattering network, Section 3.1 analyzes the coefficient properties and optimizes the network architecture. Section 3.2 describes the resulting computational algorithm. For texture analysis, the scattering transform of stationary processes is studied in Section 3.3. Section 3.4 shows that a cosine transform further reduces the size of a scattering representation.

3.1 Energy Propagation and Deformation Stability

A windowed scattering S is computed with a cascade of wavelet modulus operators $\widetilde{W}$, and its properties, thus, depend upon the wavelet transform properties. Conditions are given on wavelets to define a scattering transform that is nonexpansive and preserves the signal norm. This analysis shows that $\|S[p]x\|$ decreases quickly as the length of p increases, and is nonnegligible only over a particular subset of frequency-decreasing paths. Reducing computations to these paths defines a convolution network with much fewer internal and output coefficients.

The norm and distance on a transform $Tx = \{x_n\}_n$ which output a family of signals will be defined by

$$\|Tx - Tx'\|^2 = \sum_n \|x_n - x'_n\|^2.$$

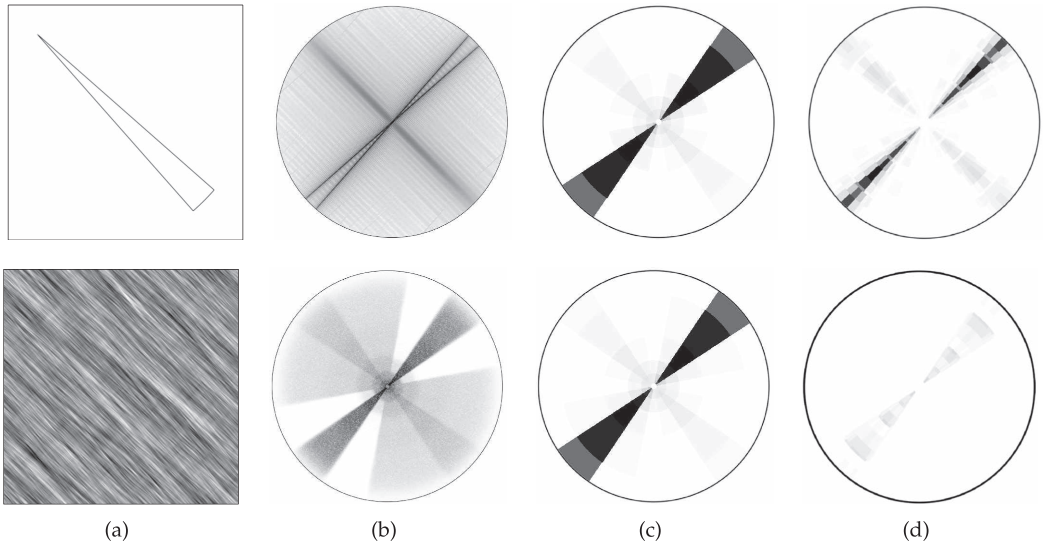

Fig. 4. (a) Two images $x(u)$. (b) Fourier modulus $|\hat{x}(\omega)|$. (c) First-order scattering coefficients $Sx[\lambda_1]$ displayed over the frequency sectors of Fig. 3a. They are the same for both images. (d) Second-order scattering coefficients $Sx[\lambda_1, \lambda_2]$ over the frequency sectors of Fig. 3b. They are different for each image.

If there exists $\epsilon > 0$ such that, for all $\omega \in \mathbb{R}^2$,

$$1 - \epsilon \leq |\hat{\phi}(\omega)|^2 + \frac{1}{2}\sum_{j=0}^{\infty}\sum_{r\in G} |\hat{\psi}(2^j r\omega)|^2 \leq 1, \tag{7}$$

then applying the Plancherel formula proves that if x is real, then $Wx = \{x \star \phi_{2^J}\ ,\ x \star \psi_\lambda\}_{\lambda\in\mathcal{P}}$ satisfies

$$(1 - \epsilon)\ \|x\|^2 \leq \|Wx\|^2 \leq \|x\|^2, \tag{8}$$

with

$$\|Wx\|^2 = \|x \star \phi_{2^J}\|^2 + \sum_{\lambda\in\mathcal{P}} \|x \star \psi_\lambda\|^2.$$

In the following, we suppose that $\epsilon < 1$ and, hence, that the wavelet transform is a nonexpansive and invertible operator with a stable inverse. If $\epsilon = 0$, then W is unitary. The Morlet wavelet ψ shown in Fig. 1 together with $\phi(u) = \exp(-|u|^2/(2\sigma^2))/(2\pi\sigma^2)$ for $\sigma = 0.7$ satisfy (7) with $\epsilon = 0.25$. These functions are used in all classification applications. Rotated and dilated cubic spline wavelets are constructed in [25] to satisfy (7) with $\epsilon = 0$.

The modulus is nonexpansive in the sense that $||a| - |b|| \leq |a - b|$ for all $(a, b) \in \mathbb{C}^2$. Since $\widetilde{W} = \{x \star \phi_{2^J}\ ,\ |x \star \psi_\lambda|\}_{\lambda\in\mathcal{P}}$ is obtained with a wavelet transform W followed by a modulus, which are both nonexpansive, it is also nonexpansive

$$\|\widetilde{W}x - \widetilde{W}y\| \leq \|x - y\|.$$

Let $\mathcal{P}_\infty = \cup_{m\in\mathbb{N}}\mathcal{P}^m$ be the set of all paths for any length $m \in \mathbb{N}$. The norm of $Sx = \{S[p]x\}_{p\in\mathcal{P}_\infty}$ is

$$\|Sx\|^2 = \sum_{p\in\mathcal{P}_\infty} \|S[p]x\|^2.$$

Since S iteratively applies $\widetilde{W}$, which is nonexpansive, it is also nonexpansive:

$$\|Sx - Sy\| \leq \|x - y\|.$$

It is thus stable to additive noise.

If W is unitary, then $\widetilde{W}$ also preserves the signal norm $\|\widetilde{W}x\|^2 = \|x\|^2$. The convolution network is built layer by layer by iterating on $\widetilde{W}$. If $\widetilde{W}$ preserves the signal norm, then the signal energy is equal to the sum of the scattering energy of each layer plus the energy of the last propagated layer:

$$\|x\|^2 = \sum_{m=0}^{\overline{m}}\sum_{p\in\mathcal{P}^m} \|S[p]x\|^2 + \sum_{p\in\mathcal{P}^{\overline{m}+1}} \|U[p]\|^2. \tag{9}$$

For appropriate wavelets, it is proven in [25] that the energy of the mth layer $\sum_{p\in\mathcal{P}^m} \|U[p]\|^2$ converges to zero when m increases, as well as the energy of all scattering coefficients of order $\geq m$. This result is important for numerical applications because it explains why the network depth can be limited with a negligible loss of signal energy. By letting the network depth $\overline{m}$ go to infinity in (9), it results that the scattering transform preserves the signal energy:

$$\|x\|^2 = \sum_{p\in\mathcal{P}_\infty} \|S[p]x\|^2 = \|Sx\|^2. \tag{10}$$

TABLE 1
Percentage of Energy $\sum_{p\in\mathcal{P}_\downarrow^m} \|S[p]x\|^2/\|x\|^2$ of Scattering Coefficients on Frequency-Decreasing Paths of Length m, Depending upon J

J	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m\leq 3$
1	95.1	4.86	-	-	-	99.96
2	87.56	11.97	0.35	-	-	99.89
3	76.29	21.92	1.54	0.02	-	99.78
4	61.52	33.87	4.05	0.16	0	99.61
5	44.6	45.26	8.9	0.61	0.01	99.37
6	26.15	57.02	14.4	1.54	0.07	99.1
7	0	73.37	21.98	3.56	0.25	98.91

These average values are computed on the Caltech-101 database, with zero mean and unit variance images.

This scattering energy conservation also proves that the more sparse the wavelet coefficients, the more energy propagates to deeper layers. Indeed, when 2^J increases, one can verify that at the first layer, $S[\lambda_1]x = |x \star \psi_{\lambda_1}| \star \phi_{2^J}$ converges to $\|\phi\|^2\ \|x \star \psi_\lambda\|_1^2$. The more sparse $x \star \psi_\lambda$, the smaller $\|x \star \psi_\lambda\|_1$ and, hence, the more energy is propagated to deeper layers to satisfy the global energy conservation (10).

Fig. 4 shows two images having the same first-order scattering coefficients, but the top image is piecewise regular and, hence, has wavelet coefficients that are much more sparse than the uniform texture at the bottom. As a result, the top image has second-order scattering coefficients of larger amplitude than at the bottom. For typical images, as in the CalTech101 dataset [12], Table 1 shows that the scattering energy has an exponential decay as a function of the path length m. Scattering coefficients are computed with cubic spline wavelets which define a unitary wavelet transform and satisfy the scattering energy conservation (10). As expected, the energy of scattering coefficients converges to 0 as m increases, and it is already below 1 percent for $m \geq 3$.

The propagated energy $\|U[p]x\|^2$ decays because $U[p]x$ is a progressively lower frequency signal as the path length increases. Indeed, each modulus computes a regular envelope of oscillating wavelet coefficients. The modulus can thus be interpreted as a nonlinear "demodulator" that pushes the wavelet coefficient energy toward lower frequencies. As a result, an important portion of the energy of $U[p]x$ is then captured by the low-pass filter ϕ_{2^J} that outputs $S[p]x = U[p]x \star \phi_{2^J}$. Hence, less energy is propagated to the next layer.

Another consequence is that the scattering energy propagates only along a subset of frequency decreasing paths. Since the envelope $|x \star \psi_\lambda|$ is more regular than $x \star \psi_\lambda$, it results that $|x \star \psi_\lambda(u)| \star \psi_{\lambda'}$ is nonnegligible only if $\psi_{\lambda'}$ is located at lower frequencies than ψ_λ, and, hence, if $|\lambda'| < |\lambda|$. Iterating on wavelet modulus operators thus propagates the scattering energy along frequency-decreasing paths $p = (\lambda_1, \ldots, \lambda_m)$, where $|\lambda_k| < |\lambda_{k-1}|$ for $1 \leq k < m$. We denote by $\mathcal{P}_\downarrow^m$ the set of frequency decreasing paths of length m. Scattering coefficients along other paths have a negligible energy. This is verified by Table 1 that shows not only that the scattering energy is concentrated on low-order paths, but also that more than 99 percent of the energy is absorbed by frequency-decreasing paths of length $m \leq 3$. Numerically, it is therefore sufficient to compute the scattering transform along frequency-decreasing paths. It defines a much smaller convolution network. Section 3.2 shows that the resulting coefficients are computed with $O(N \log N)$ operations.

Preserving energy does not imply that the signal information is preserved. Since a scattering transform is calculated by iteratively applying $\widetilde{W}$, one needs to invert $\widetilde{W}$ to invert S. The wavelet transform W is a linear invertible operator, so inverting $\widetilde{W}z = \{z \star \phi_{2^J},\ |z \star \psi_\lambda|\}_{\lambda \in \mathcal{P}}$ amounts to recovering the complex phases of wavelet coefficients removed by the modulus. The phase of Fourier coefficients cannot be recovered from their modulus, but wavelet coefficients are redundant, as opposed to Fourier coefficients. For particular wavelets, it has been proven that the phase of wavelet coefficients can be recovered from their modulus, and that $\widetilde{W}$ has a continuous inverse, and the phase can be recovered with a convex optimization [37].

Still, one cannot exactly invert S because we discard information when computing the scattering coefficients $S[p]x = U[p] \star \phi_{2^J}$ of the last layer $\mathcal{P}^{\overline{m}}$. Indeed, the propagated coefficients $|U[p]x \star \psi_\lambda|$ of the next layer are eliminated because they are not invariant and have a negligible total energy. The number of such coefficients is larger than the total number of scattering coefficients kept at previous layers. Initializing the inversion by considering that these small coefficients are zero produces an error. This error is further amplified as the inversion of $\widetilde{W}$ progresses across layers from $\overline{m}$ to 0. Numerical experiments conducted over one-dimensional audio signals [2], [7] indicate that reconstructed signals have good audio quality with $\overline{m} = 2$ as long as the number of scattering coefficients is comparable to the number of signal samples. Audio examples in www.di.ens.fr/data/scattering show that reconstructions from first-order scattering coefficients are typically of much lower quality because there are much fewer first-order than second-order coefficients. When the invariant scale 2^J becomes too large, the number of second-order coefficients also becomes too small for accurate reconstructions. Although individual signals cannot be recovered, reconstructions of equivalent stationary textures are possible with arbitrarily large scale scattering invariants [7].

For classification applications, besides computing a rich set of invariants, the most important property of a scattering transform is its Lipschitz continuity to deformations. Indeed, wavelets are stable to deformations and the modulus commutes with deformations. Let $x_\tau(u) = x(u - \tau(u))$ be an image deformed by the displacement field τ. Let $\|\tau\|_\infty = \sup_u |\tau(u)|$ and $\|\nabla\tau\|_\infty = \sup_u |\nabla\tau(u)| < 1$. If Sx is computed on paths of length $m \leq \overline{m}$, then it is proven in [25] that for signals x of compact support:

$$\|Sx_\tau - Sx\| \leq C\, \overline{m}\, \|x\|\, (2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty), \tag{11}$$

with a second-order Hessian term which is part of the metric definition on $\mathbf{C}^2$ deformations, but which is negligible if $\tau(u)$ is regular. If $2^J \geq \|\tau\|_\infty / \|\nabla\tau\|_\infty$, then the translation term can be neglected and the transform is Lipschitz continuous to deformations:

$$\|Sx_\tau - Sx\| \leq C\, \overline{m}\, \|x\|\, \|\nabla\tau\|_\infty. \tag{12}$$

If $\overline{m}$ goes to ∞, then $C\,\overline{m}$ can be replaced by a more complex expression [25] which is numerically converging for natural images.

3.2 Fast Scattering Computations

We describe a fast scattering implementation over frequency decreasing paths where most of the scattering energy is concentrated. A frequency decreasing path $p = (2^{-j_1}r_1, \ldots, 2^{-j_m}r_m)$ satisfies $0 < j_k \leq j_{k+1} \leq J$. If the wavelet transform is computed over K rotation angles, then the total number of frequency-decreasing paths of length m is $K^m\binom{J}{m}$. Let N be the number of pixels of the image x. Since ϕ_{2^J} is a low-pass filter scaled by 2^J, $S[p]x(u) = U[p]x \star \phi_{2^J}(u)$ is uniformly sampled at intervals $\alpha 2^J$, with $\alpha = 1$ or $\alpha = 1/2$. Each $S[p]x$ is an image with $\alpha^{-2}2^{-2J}N$ coefficients. The total number of coefficients in a scattering network of maximum depth $\overline{m}$ is thus

$$P = N\,\alpha^{-2}\,2^{-2J}\sum_{m=0}^{\overline{m}} K^m\binom{J}{m}. \tag{13}$$

If $\overline{m} = 2$, then $P \simeq \alpha^{-2}\,N2^{-2J}K^2J^2/2$. It decreases exponentially when the scale 2^J increases.

Algorithm 1 describes the computations of scattering coefficients on sets $\mathcal{P}_\downarrow^m$ of frequency decreasing paths of length $m \leq \overline{m}$. The initial set $\mathcal{P}_\downarrow^0 = \{\emptyset\}$ corresponds to the original image $U[\emptyset]x = x$. Let $p + \lambda$ be the path that begins by p and ends with $\lambda \in \mathcal{P}$. If $\lambda = 2^{-j}r$, then $U[p+\lambda]x(u) = |U[p]x \star \psi_\lambda(u)|$ has energy at frequencies mostly below $2^{-j}\pi$. To reduce computations, we can thus subsample this convolution at intervals $\alpha 2^j$, with $\alpha = 1$ or $\alpha = 1/2$, to avoid aliasing.

Algorithm 1. Fast Scattering Transform

```
for $m = 1$ to $\overline{m}$ do
    for all $p \in \mathcal{P}_\downarrow^{m-1}$ do
        Output $S[p]x(\alpha 2^J n) = U[p]x \star \phi_{2^J}(\alpha 2^J n)$
    end for
    for all $p + \lambda_m \in \mathcal{P}_\downarrow^m$ with $\lambda_m = 2^{-j_m}r_m$ do
        Compute
            $U[p+\lambda_m]x(\alpha 2^{j_m} n) = |U[p]x \star \psi_{\lambda_m}(\alpha 2^{j_m} n)|$
    end for
end for
for all $p \in \mathcal{P}_\downarrow^{\overline{m}}$ do
    Output $S[p]x(\alpha 2^J n) = U[p]x \star \phi_{2^J}(\alpha 2^J n)$
end for
```

At the layer m there are $K^m\binom{J}{m}$ propagated signals $U[p]x$ with $p \in \mathcal{P}_\downarrow^m$. They are sampled at intervals $\alpha 2^{j_m}$ which depend on p. One can verify by induction on m that layer m has a total number of samples equal to $\alpha^{-2}\,(K/3)^m\,N$. There are also $K^m\binom{J}{m}$ scattering signals $S[p]x$, but they are subsampled by 2^J and thus have much fewer coefficients. The number of operations to compute each layer is therefore driven by the $O((K/3)^m\,N\,\log N)$ operations needed to compute the internal propagated coefficients with FFTs. For $K > 3$, the overall computational complexity is thus $O((K/3)^{\overline{m}}\,N\,\log N)$.

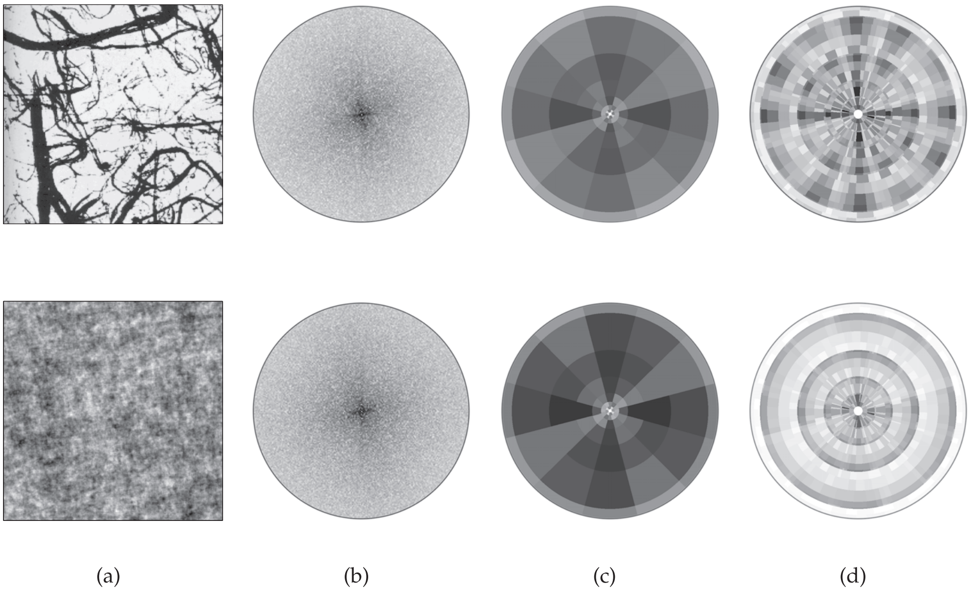

Fig. 5. (a) Realizations of two stationary processes $X(u)$. Top: Brodatz texture. Bottom: Gaussian process. (b) The power spectrum estimated from each realization is nearly the same. (c) First-order scattering coefficients $S[p]X$ are nearly the same for 2^J equal to the image width. (d) Second-order scattering coefficients $S[p]X$ are clearly different.

3.3 Scattering Stationary Processes

Image textures can be modeled as realizations of stationary processes $X(u)$. We denote the expected value of X by $E(X)$, which does not depend upon u. Despite the importance of spectral methods, the power spectrum is often not sufficient to discriminate image textures because it only depends upon second-order moments. Fig. 5 shows very different textures having the same power spectrum. A scattering representation of stationary processes depends upon second-order and higher order moments, and can thus discriminate such textures. Moreover, it does not suffer from the large variance curse of high-order moments estimators [36] because it is computed with a nonexpansive operator.

If $X(u)$ is stationary, then $U[p]X(u)$ remains stationary because it is computed with a cascade of convolutions and modulus which preserve stationarity. Its expected value thus does not depend upon u and defines the expected scattering transform:

$$\overline{S}X(p) = E(U[p]X).$$

A windowed scattering gives an estimator of $\overline{S}X(p)$, calculated from a single realization of X, by averaging $U[p]X$ with ϕ_{2^J}:

$$S[p]X(u) = U[p]X \star \phi_{2^J}(u).$$

Since $\int \phi_{2^J}(u)\, du = 1$, this estimator is unbiased:

$$E(S[p]X) = E(U[p]X) = \overline{S}X(p).$$

For appropriate wavelets, it is proven in [25] that a windowed scattering transform conserves the second moment of stationary processes:

$$\sum_{p \in \mathcal{P}_\infty} E(|S[p]X|^2) = E(|X|^2). \tag{14}$$

The second-order moments of all wavelet coefficients, which are useful for texture discrimination, can also be recovered from scattering coefficients. Indeed, for $p = (\lambda_1, \ldots, \lambda_m)$, if we write $\lambda + p = (\lambda, \lambda_1, \ldots, \lambda_m)$, then

$$S[p]|X \star \psi_\lambda| = S[p]U[\lambda]X = S[\lambda + p]X,$$

and replacing X by $|X \star \psi_\lambda|$ in (14) gives

$$\sum_{p \in \mathcal{P}_\infty} E(|S[\lambda + p]X|^2) = E(|X \star \psi_\lambda|^2). \tag{15}$$

However, if p has a length m because of the m successive modulus nonlinearities, one can show [25] that $\overline{S}X(p)$ also depends upon normalized high-order moments of X, mainly of order up to 2^m. Scattering coefficients can thus discriminate textures having the same second-order moments but different higher order moments. This is illustrated by the two textures in Fig. 5, which have same power spectrum and hence the same second order moments. Scattering coefficients $S[p]X$ are shown for $m = 1$ and $m = 2$, with the frequency tiling illustrated in Fig. 3. The squared distance between the order 1 scattering coefficients of these two textures is of the order their variance. Indeed, order 1 scattering coefficients mostly depend upon second-order moments and are thus nearly equal for both textures. On the contrary, scattering coefficients of order 2 are different because they depend on moments up to 4. Their squared distance is more than five times bigger than their variance.

High-order moments are difficult to use in signal processing because their estimators have a large variance [36], which can introduce important errors. This large variance comes from the blow up of large coefficient outliers produced by X^q for $q \geq 2$. On the contrary, a scattering is computed with a nonexpansive operator and thus has much lower variance estimators. The estimation of $\overline{S}X(p) = E(U[p]X)$ by $S[p]X = U[p]X \star \phi_{2^J}$ has a variance which is reduced when the averaging scale 2^J increases. For

TABLE 2
Normalized Scattering Variance $\sum_{p\in\mathcal{P}_\infty} E(|SX-\overline{S}X(p)|^2)/E(|X|^2)$, as a Function of J, Computed on Zero-Mean and Unit Variance Images of the Brodatz Dataset, with Cubic Spline Wavelets

$J=1$	$J=2$	$J=3$	$J=4$	$J=5$	$J=6$	$J=7$
0.85	0.65	0.45	0.26	0.14	0.07	0.0025

TABLE 3
Percentage of Energy $\sum_{p\in\mathcal{P}^m_\downarrow} |\overline{S}X(p)|^2/E(|X|^2)$ Along Frequency Decreasing Paths of Length m, Computed on the Normalized Brodatz Dataset, with Cubic Spline Wavelets

$m=0$	$m=1$	$m=2$	$m=3$	$m=4$
0	74	19	3	0.3

all image textures, it is numerically observed that the scattering variance $\sum_{p\in\mathcal{P}_\infty} E(|S[p]X-\overline{S}X(p)|^2)$ decreases to zero when 2^J increases. Table 2 gives the decay of this scattering variance, computed, on average, over the Brodatz texture dataset. Expected scattering coefficients of stationary textures are thus better estimated from windowed scattering transforms at the largest possible scale 2^J, equal to the image size.

Let $\overline{\mathcal{P}}_\infty$ be the set of all paths $p=(\lambda_1,\ldots,\lambda_m)$ for all $\lambda_k=2^{j_k}r_k\in 2^{\mathbb{Z}}\times G^+$ and all length m. The conservation (14) together with the scattering variance decay also implies that the second moment is equal to the energy of expected scattering coefficients in $\overline{\mathcal{P}}_\infty$:

$$\|\overline{S}X\|^2=\sum_{p\in\overline{\mathcal{P}}_\infty}|\overline{S}X(p)|^2=E(|X|^2). \tag{16}$$

Indeed, $E(S[p]X)=\overline{S}X(p)$, so

$$E(|S[p]X|^2)=\overline{S}X(p)^2+E(|S[p]X-E(S[p]X)|^2).$$

Summing over p and letting J go to ∞ gives (16).

Table 3 gives the ratio between the average energy along frequency decreasing paths of length m and second moments for textures in the Brodatz dataset. Most of this energy is concentrated over paths of length $m\leq 3$.

3.4 Cosine Scattering Transform

Natural images have scattering coefficients $S[p]X(u)$ that are correlated across paths $p=(\lambda_1,\ldots,\lambda_m)$, at any given position u. The strongest correlation is between coefficients of a same layer. For each m, scattering coefficients are decorrelated in a Karhunen-Loève basis that diagonalizes their covariance matrix. Fig. 6 compares the decay of the sorted variances $E(|S[p]X-E(S[p]X)|^2)$ and the variance decay in the Karhunen-Loève basis computed over half of the Caltech image dataset, for the first and second layer of scattering coefficients. Scattering coefficients are calculated with a Morlet wavelet. The variance decay (computed on the second half of the data) is much faster in the Karhunen-Loève basis, which shows that there is a strong correlation between scattering coefficients of the same layers.

A change of variables proves that a rotation and scaling $X_{2^lr}(u)=X(2^{-l}ru)$ produces a rotation and inverse scaling on the path variable:

$$\overline{S}X_{2^lr}(p)=\overline{S}X(2^lrp)\ \text{ where }\ 2^lrp=\big(2^lr\lambda_1,\ldots,2^lr\lambda_m\big),$$

and $2^lr\lambda_k=2^{l-j_k}rr_k$. If natural images can be considered as randomly rotated and scaled [29], then the path p is randomly rotated and scaled. In this case, the scattering transform has stationary variations along the scale and rotation variables. This suggests approximating the Karhunen-Loève basis by a cosine basis along these variables. Let us parameterize each rotation r by its angle $\theta\in[0,2\pi]$. A path $p=(2^{-j_1}r_1,\ldots,2^{-j_k}r_k)$ is then parameterized by $((j_1,\theta_1),\ldots,\ (j_m,\theta_m))$.

Since scattering coefficients are computed along frequency decreasing paths for which $0<j_k<j_{k+1}\leq J$, to reduce boundary effects a separable cosine transform is computed along the variables $l_1=j_1,\ l_2=j_2-j_1,\ldots,\ l_m=j_m-j_{m-1}$, and along each angle variable $\theta_1,\ \theta_2,\ldots,\theta_m$. Cosine scattering coefficients are computed by applying this separable discrete cosine transform along the scale and angle variables of $S[p]X(u)$ for each u and each path length m. Fig. 6 shows that the cosine scattering coefficients have variances for $m=1$ and $m=2$ which decay nearly as fast as the variances in the Karhunen-Loève basis. It shows that a DCT across scales and orientations is nearly optimal to decorrelate scattering coefficients. Lower frequency DCT coefficients absorb most of the scattering energy. On natural images, more than 99.5 percent of the scattering energy is absorbed by the 1/2 lowest frequency cosine scattering coefficients.

We saw in (13) that without oversampling $\alpha=1$, when $\overline{m}=2$, an image of size N is represented by $P=N\,2^{-2J}\,(KJ+K^2J(J-1)/2)$ scattering coefficients. Numerical computations are performed with $K=6$ rotation angles and the DCT reduces at least by 2 the number

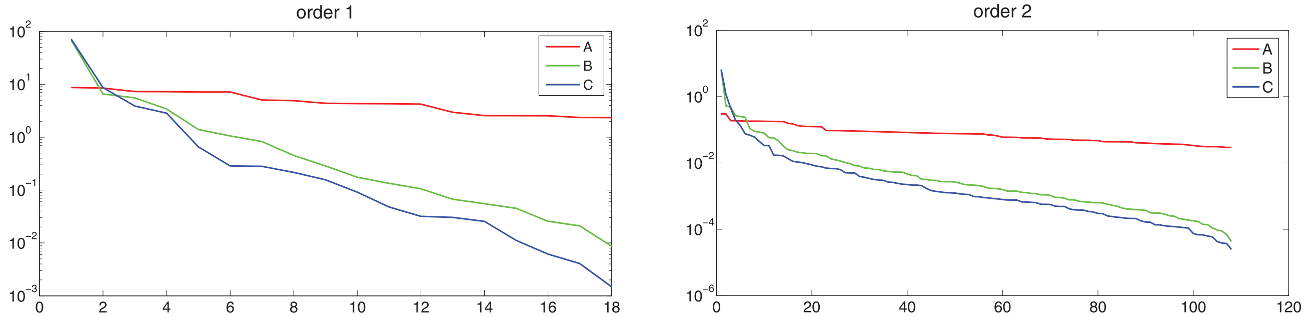

Fig. 6. A: Sorted variances of scattering coefficients of order 1 (left) and order 2 (right), computed on the CalTech101 database. B: Sorted variances of cosine transform scattering coefficients. C: Sorted variances in a Karhunen-Loève basis calculated for each layer of scattering coefficients.

of coefficients. At a small invariant scale $J = 2$, the resulting cosine scattering representation has $P = 3N/2$ coefficients. As a matter of comparison, SIFT represents small blocks of 4^2 pixels with eight coefficients, and a dense SIFT representation thus has $N/2$ coefficients. When J increases, the size of a cosine scattering representation decreases like 2^{-2J}, with $P = N$ for $J = 3$ and $P \approx N/40$ for $J = 7$.

4 Classification

A scattering transform eliminates the image variability due to translations and linearizes small deformations. Classification is studied with linear generative models computed with a PCA, and with discriminant SVM classifiers. State-of-the-art results are obtained for hand-written digit recognition and for texture discrimination. Scattering representations are computed with a Morlet wavelet.

4.1 PCA Affine Space Selection

Although discriminant classifiers such as SVM have better asymptotic properties than generative classifiers [28], the situation can be inverted for small training sets. We introduce a simple robust generative classifier based on affine space models computed with a PCA. Applying a DCT on scattering coefficients has no effect on any linear classifier because it is a linear orthogonal transform. Keeping the 50 percent lower frequency cosine scattering coefficients reduces computations and has a negligible effect on classification results. The classification algorithm is described directly on scattering coefficients to simplify explanations. Each signal class is represented by a random vector X_k whose realizations are images of N pixels in the class.

Each scattering vector SX_k has P coefficients. Let $E(SX_k)$ be the expected vector over the signal class k. The difference $SX_k - E(SX_k)$ is approximated by its projection in a linear space of low dimension $d \ll P$. The covariance matrix of SX_k has P^2 coefficients. Let $\mathbf{V}_k$ be the linear space generated by the d PCA eigenvectors of this covariance matrix having the largest eigenvalues. Among all linear spaces of dimension d, it is the space that approximates $SX_k - E(SX_k)$ with the smallest expected quadratic error. This is equivalent to approximating SX_k by its projection on an affine approximation space:

$$\mathbf{A}_k = E\{SX_k\} + \mathbf{V}_k.$$

The classifier associates to each signal x the class index $\hat{k}$ of the best approximation space:

$$\hat{k}(x) = \underset{k \le C}{\operatorname{argmin}} \|Sx - P_{\mathbf{A}_k}(Sx)\|. \tag{17}$$

The minimization of this distance has similarities with the minimization of a tangential distance [14] in the sense that we remove the principal scattering directions of variability to evaluate the distance. However, it is much simpler since it does not evaluate a tangential space that depends upon Sx. Let $\mathbf{V}_k^{\perp}$ be the orthogonal complement of $\mathbf{V}_k$ corresponding to directions of lower variability. This distance is also equal to the norm of the difference between Sx and the average class "template" $E(SX_k)$, projected in $\mathbf{V}_k^{\perp}$:

$$\|Sx - P_{\mathbf{A}_k}(Sx)\| = \|P_{\mathbf{V}_k^{\perp}}(Sx - E(SX_k))\|. \tag{18}$$

Minimizing the affine space approximation error is thus equivalent to finding the class centroid $E(SX_k)$ which is the closest to Sx, without taking into account the first d principal variability directions. The d principal directions of the space $\mathbf{V}_k$ result from deformations and from structural variability. The projection $P_{\mathbf{A}_k}(Sx)$ is the optimum linear prediction of Sx from these d principal modes. The selected class has the smallest prediction error.

This affine space selection is effective if $SX_k - E(SX_k)$ is well approximated by a projection in a low-dimensional space. This is the case if realizations of X_k are translations and limited deformations of a single template. Indeed, the Lipschitz continuity implies that small deformations are linearized by the scattering transform. Hand-written digit recognition is an example. This is also valid for stationary textures where SX_k has a small variance, which can be interpreted as structural variability.

The dimension d must be adjusted so that SX_k has a better approximation in the affine space $\mathbf{A}_k$ than in affine spaces $\mathbf{A}_l$ of other classes $l \neq k$. This is a model selection problem, which requires an optimization of the dimension d to avoid overfitting [5].

The invariance scale 2^J must also be optimized. When the scale 2^J increases, translation invariance increases, but it comes with a partial loss of information, which brings the representations of different signals closer. One can prove [25] that the scattering distance $\|Sx - Sx'\|$ decreases when 2^J increases, and it converges to a nonzero value when 2^J goes to ∞. To classify deformed templates such as hand-written digits, the optimal 2^J is of the order of the maximum pixel displacements due to translations and deformations. In a stochastic framework where x and x' are realizations of stationary processes, Sx and Sx' converge to the expected scattering transforms $\overline{S}x$ and $\overline{S}x'$. To classify stationary processes such as textures, the optimal scale is the maximum scale equal to the image width because it minimizes the variance of the windowed scattering estimator.

A cross-validation procedure is used to find the dimension d and the scale 2^J which yield the smallest classification error. This error is computed on a subset of the training images, which is not used to estimate the covariance matrix for the PCA calculations.

As in the case of SVM, the performance of the affine PCA classifier is improved by equalizing the descriptor space. Table 1 shows that scattering vectors have unequal energy distribution along its path variables, in particular as the order varies. A robust equalization is obtained by dividing each $S[p]X(u)$ by

$$\gamma(p) = \max_{x_i} \Big(\sum_u |S[p]x_i(u)|^2 \Big)^{1/2}, \tag{19}$$

where the maximum is computed over all training signals x_i. To simplify notations, we still write SX the vector of normalized scattering coefficients $S[p]X(u)/\gamma(p)$.

TABLE 4
Percentage of Errors of MNIST Classifiers, Depending on the Training Size

Training	x		Wind. Four.		Scat. $\overline{m}=1$		Scat. $\overline{m}=2$		Conv.
size	PCA	SVM	PCA	SVM	PCA	SVM	PCA	SVM	Net.
300	14.5	15.4	7.35	7.4	5.7	8	**4.7**	5.6	7.18
1000	7.2	8.2	3.74	3.74	2.35	4	**2.3**	2.6	3.21
2000	5.8	6.5	2.99	2.9	1.7	2.6	**1.3**	1.8	2.53
5000	4.9	4	2.34	2.2	1.6	1.6	**1.03**	1.4	1.52
10000	4.55	3.11	2.24	1.65	1.5	1.23	0.88	1	**0.85**
20000	4.25	2.2	1.92	1.15	1.4	0.96	0.79	**0.58**	0.76
40000	4.1	1.7	1.85	0.9	1.36	0.75	0.74	**0.53**	0.65
60000	4.3	1.4	1.80	0.8	1.34	0.62	0.7	**0.43**	0.53

Affine space scattering models can be interpreted as generative models computed independently for each class. As opposed to discriminative classifiers such as SVM, we do not estimate cross-correlation interactions between classes, besides optimizing the model dimension d. Such estimators are particularly effective for a small number of training samples per class. Indeed, if there are few training samples per class, then variance terms dominate bias errors when estimating off-diagonal covariance coefficients between classes [4].

An affine space approximation classifier can also be interpreted as a robust quadratic discriminant classifier obtained by coarsely quantizing the eigenvalues of the inverse covariance matrix. For each class, the eigenvalues of the inverse covariance are set to 0 in $\mathbf{V}_k$ and to 1 in $\mathbf{V}_k^{\perp}$, where d is adjusted by cross validation. This coarse quantization is justified by the poor estimation of covariance eigenvalues from few training samples. These affine space models are robust when applied to distributions of scattering vectors having non-Gaussian distributions, where a Gaussian Fisher discriminant can lead to significant errors.

4.2 Handwritten Digit Recognition

The MNIST database of hand-written digits is an example of structured pattern classification where most of the intraclass variability is due to local translations and deformations. It is comprised of at most 60,000 training samples and 10,000 test samples. If the training dataset is not augmented with deformations, the state of the art was achieved by deep-learning convolution networks [30], deformation models [17], [3], and dictionary learning [27]. These results are improved by a scattering classifier.

All computations are performed on the reduced cosine scattering representation described in Section 3.4, which keeps the lower frequency half of the coefficients. Table 4 computes classification errors on a fixed set of test images, depending upon the size of the training set, for different representations and classifiers. The affine space selection of Section 4.1 is compared with an SVM classifier using RBF kernels, which are computed using Libsvm [10], and whose variance is adjusted using standard cross validation over a subset of the training set. The SVM classifier is trained with a renormalization which maps all coefficients to $[-1, 1]$. The PCA classifier is trained with the renormalization factors (19). The first two columns of Table 4 show that classification errors are much smaller with an SVM than with the PCA algorithm if applied directly on the image. The third and fourth columns give the classification error obtained with a PCA or an SVM classification applied to the modulus of a windowed Fourier transform. The spatial size 2^J of the window is optimized with a cross validation that yields a minimum error for $2^J = 8$. It corresponds to the largest pixel displacements due to translations or deformations in each class. Removing the complex phase of the windowed Fourier transform yields a locally invariant representation but whose high frequencies are unstable to deformations, as explained in Section 2.1. Suppressing this local translation variability improves the classification rate by a factor 3 for a PCA and by almost 2 for an SVM. The comparison between PCA and SVM confirms the fact that generative classifiers can outperform discriminative classifiers when training samples are scarce [28]. As the training set size increases, the bias-variance tradeoff turns in favor of the richer SVM classifiers, independently of the descriptor.

Columns 6 and 8 give the PCA classification result applied to a windowed scattering representation for $\overline{m} = 1$ and $\overline{m} = 2$. The cross validation also chooses $2^J = 8$. Fig. 7 displays the arrays of normalized windowed scattering coefficients of a digit "3." The first- and second-order coefficients of $S[p]X(u)$ are displayed as energy distributions over frequency disks described in Section 2.3. The spatial parameter u is sampled at intervals 2^J so each image of N pixels is represented by $N2^{-2J} = 4^2$ translated disks, both for order 1 and order 2 coefficients.

Increasing the scattering order from $\overline{m} = 1$ to $\overline{m} = 2$ reduces errors by about 30 percent, which shows that second-order coefficients carry important information even at a relatively small scale $2^J = 8$. However, third-order coefficients have a negligible energy and including them brings marginal classification improvements while increasing computations by an important factor. As the learning set increases in size, the classification improvement of a scattering transform increases relative to windowed Fourier transform because the classification is able to incorporate more high-frequency structures, which have deformation instabilities in the Fourier domain as opposed to the scattering domain.

Table 4 shows that below 5,000 training samples, the scattering PCA classifier improves results of a deep-learning convolution network, which learns all filter coefficients with a back-propagation algorithm [20]. As more training samples are available, the flexibility of the SVM classifier brings an improvement over the more rigid affine classifier, yielding a 0.43 percent error rate on the

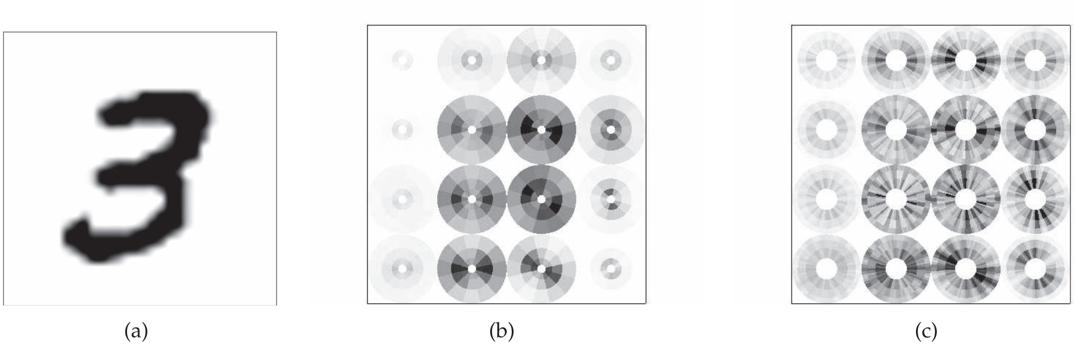

Fig. 7. (a) Image $X(u)$ of a digit "3." (b) Arrays of windowed scattering coefficients $S[p]X(u)$ of order $m = 1$, with u sampled at intervals of $2^J = 8$ pixels. (c) Windowed scattering coefficients $S[p]X(u)$ of order $m = 2$.

original dataset, thus improving upon previous state-of-the-art methods.

To evaluate the precision of affine space models, we compute an average normalized approximation error of scattering vectors projected on the affine space of their own class, over all classes k:

$$\sigma_d^2 = C^{-1} \sum_{k=1}^{C} \frac{E(\|SX_k - P_{\mathbf{A_k}}(SX_k)\|^2)}{E(\|SX_k\|^2)}. \tag{20}$$

An average separation factor measures the ratio between the approximation error in the affine space $\mathbf{A_k}$ of the signal class and the minimum approximation error in another affine model $\mathbf{A}_l$ with $l \neq k$, for all classes k:

$$\rho_d^2 = C^{-1} \sum_{k=1}^{C} \frac{E(\min_{l \neq k}\|SX_k - P_{\mathbf{A}_l}(SX_k)\|^2)}{E(\|SX_k - P_{\mathbf{A_k}}(SX_k)\|^2)}. \tag{21}$$

For a scattering representation with $\overline{m} = 2$, Table 5 gives the dimension d of affine approximation spaces optimized with a cross validation. It varies considerably, ranging from 5 to 140 when the number of training examples goes from 300 to 40,000. Indeed, many training samples are needed to reliably estimate the eigenvectors of the covariance matrix and thus to compute reliable affine space models for each class. The average approximation error σ_d^2 of affine space models is progressively reduced while the separation ratio ρ_d^2 increases. It explains the reduction of the classification error rate observed in Table 4 as the training size increases.

TABLE 5
For Each MNIST Training Size, the Table Gives the Cross-Validated Dimension d of Affine Approximation Spaces, Together with the Average Approximation Error σ_d^2 and Separation Ratio ρ_d^2 of These Spaces

Training	d	σ_d^2	ρ_d^2
300	5	$3 \cdot 10^{-1}$	2
5000	100	$4 \cdot 10^{-2}$	3
40000	140	$2 \cdot 10^{-2}$	4

The US-Postal Service is another handwritten digit dataset, with 7,291 training samples and 2,007 test images of 16×16 pixels. The state of the art is obtained with tangent distance kernels [14]. Table 6 gives results obtained with a scattering transform with the PCA classifier for $\overline{m} = 1, 2$. The cross validation sets the scattering scale to $2^J = 8$. As in the MNIST case, the error is reduced when going from $\overline{m} = 1$ to $\overline{m} = 2$ but remains stable for $\overline{m} = 3$. Different renormalization strategies can bring marginal improvements on this dataset. If the renormalization is performed by equalizing using the standard deviation of each component, the classification error is 2.3 percent, whereas it is 2.6 percent if the supremum is normalized.

The scattering transform is stable but not invariant to rotations. Stability to rotations is demonstrated over the MNIST database in the setting defined in [18]. A database with 12,000 training samples and 50,000 test images is constructed with random rotations of MNIST digits. The PCA affine space selection takes into account the rotation variability by increasing the dimension d of the affine approximation space. This is equivalent to projecting the distance to the class centroid on a smaller orthogonal space by removing more principal components. The error rate in Table 7 is much smaller with a scattering PCA than with a convolution network [18]. Much better results are obtained for a scattering with $\overline{m} = 2$ than with $\overline{m} = 1$ because second-order coefficients maintain enough discriminability despite the removal of a larger number d of principal directions. In this case, $\overline{m} = 3$ marginally reduces the error.

Scaling and rotation invariance is studied by introducing a random scaling factor uniformly distributed between $1/\sqrt{2}$ and $\sqrt{2}$, and a random rotation by a uniform angle. In this case, the digit "9" is removed from the database so as to avoid any indetermination with the digit "6" when rotated. The training set has 9,000 samples (1,000 samples per class). Table 8 gives the error rate on the original MNIST database when transforming the training and testing samples with either random rotations or scalings, or with both. Scalings

TABLE 6
Percentage of Errors for the Whole USPS Database

Tang. Kern.	Scat. $\overline{m} = 2$ SVM	Scat. $\overline{m} = 1$ PCA	Scat. $\overline{m} = 2$ PCA
2.4	2.7	3.24	2.6 / **2.3**

TABLE 7
Percentage of Errors on an MNIST Rotated Dataset [18]

Scat. $\overline{m} = 1$ PCA	Scat. $\overline{m} = 2$ PCA	Conv. Net.
8	**4.4**	8.8

TABLE 8
Percentage of Errors on Scaled/Rotated MNIST Digits

Transformations on MNIST images	Scat. $\overline{m} = 1$ PCA	Scat. $\overline{m} = 2$ PCA
None	1.6	0.8
Rotations	6.7	3.3
Scalings	2	1
Rot. + Scal.	12	5.5

have a smaller impact on the error rate than rotations because scaled scattering vectors span an invariant linear space of lower dimension. Second-order scattering outperforms first-order scattering, and the difference becomes more significant when rotation and scaling are combined. Second-order coefficients are highly discriminative in the presence of scaling and rotation variability.

4.3 Texture Discrimination

Visual texture discrimination remains an outstanding image processing problem because textures are realizations of non-Gaussian stationary processes, which cannot be discriminated using the power spectrum. The affine PCA space classifier removes most of the variability of $SX - E(SX)$ within each class. This variability is due to the residual stochastic variability, which decays as J increases, and to variability due to illumination, rotation, scaling, or perspective deformations when textures are mapped on surfaces.

Texture classification is tested on the CUReT texture database [21], [35], which includes 61 classes of image textures of $N = 200^2$ pixels. Each texture class gives images of the same material with different pose and illumination conditions. Specularities, shadowing, and surface normal variations make classification challenging. Pose variation requires global rotation and illumination invariance. Fig. 8 illustrates the large intraclass variability after a normalization of the mean and variance of each textured image.

Table 9 compares error rates obtained with different image representations. The database is randomly split into a training and a testing set, with 46 training images for each class as in [35]. Results are averaged over 10 different splits. A PCA affine space classifier applied directly on the image pixels yields a large classification error of 17 percent. The lowest published classification errors obtained on this dataset are 2 percent for Markov random fields [35], 1.53 percent for a dictionary of textons [15], 1.4 percent for basic image features [11], and 1 percent for histograms of image variations [6]. A PCA classifier applied to a Fourier power spectrum estimator also reaches 1 percent error. The power spectrum is estimated with windowed Fourier transforms calculated over half-overlapping windows whose squared modulus are averaged over the whole image to reduce the estimator variance. A cross-validation optimizes the window size to $2^J = 32$ pixels.

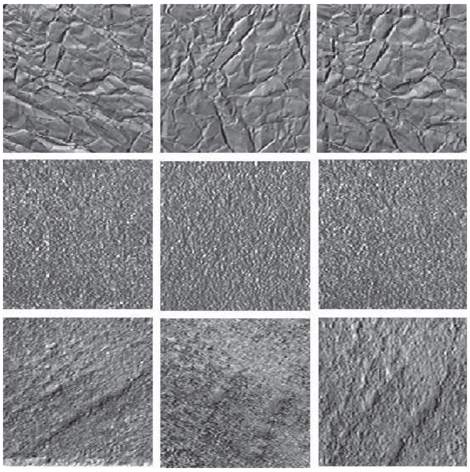

Fig. 8. Examples of textures from the CUReT database with normalized mean and variance. Each row corresponds to a different class, showing intraclass variability in the form of stochastic variability and changes in pose and illumination.

For the scattering PCA classifier, the cross validation chooses an optimal scale 2^J equal to the image width to reduce the scattering estimation variance. Indeed, contrarily to a power spectrum estimation, the variance of the scattering vector decreases when 2^J increases. Fig. 9 displays the scattering coefficients $S[p]X$ of order $m = 1$ and $m = 2$ of a CureT textured image X. A PCA classification with only first-order coefficients ($\overline{m} = 1$) yields an error 0.5 percent, although first-order scattering coefficients are strongly correlated with second-order moments whose values depend on the Fourier spectrum. The classification error is improved relative to a power spectrum estimator

TABLE 9
Percentage of Classification Errors of Different Algorithms on CUReT

Training size	X PCA	MRF [35]	Textons [15]	BIF [11]	Histo. [6]	Four. Spectr. PCA	Scat. $\overline{m} = 1$ PCA	Scat. $\overline{m} = 2$ PCA
46	17	2	1.5	1.4	1	1	0.5	**0.2**

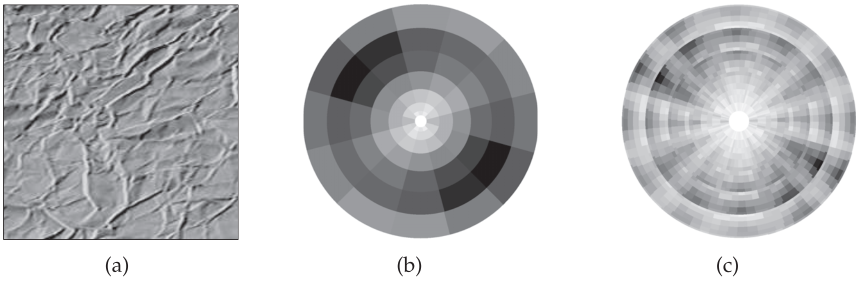

Fig. 9. (a) Example of CureT texture $X(u)$. (b) First-order scattering coefficients $S[p]X$, for 2^J equal to the image width. (c) Second-order scattering coefficients $S[p]X(u)$.

because $SX[\lambda_1]X = |X \star \psi_{\lambda_1}| \star \phi_{2^J}$ is an estimator of a first-order moment $\overline{S}[\lambda_1]X = E(|X \star \psi_{\lambda_1}|)$ and thus has a lower variance than second-order moment estimators. A PCA classification with first- and second-order scattering coefficients ($\overline{m} = 2$) reduces the error to 0.2 percent. Indeed, scattering coefficients of order $m = 2$ depend upon moments of order 4, which are necessary to differentiate textures having same second-order moments, as in Fig. 5. Moreover, the estimation of $\overline{S}[\lambda_1, \lambda_2]X = E(\|X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|)$ has a low variance because X is transformed by a nonexpansive operator as opposed to X^q for high-order moments $q \geq 2$. For $\overline{m} = 2$, the cross validation chooses affine space models of small dimension $d = 16$. However, they still produce a small average approximation error (20) $\sigma_d^2 = 2.5 \cdot 10^{-1}$ and the separation ratio (21) is $\rho_d^2 = 3$.

The PCA classifier provides a partial rotation invariance by removing principal components. It mostly averages the scattering coefficients along rotated paths. The rotation of $p = (2^{-j_1}r_1, \ldots, 2^{-j_m}r_m)$ by r is defined by $rp = (2^{-j_1}rr_1, \ldots, 2^{-j_m}rr_m)$. This rotation invariance obtained by averaging comes at the cost of a reduced representation discriminability. As in the translation case, a multilayer scattering along rotations recovers the information lost by this averaging with wavelet convolutions along rotation angles [26]. It preserves discriminability by producing a larger number of invariant coefficients to translations and rotations, which improves rotation invariant texture discrimination [26]. This combined translation and rotation scattering yields a translation and rotation invariant representation which remains stable to deformations [25].

5 Conclusion

A scattering transform is implemented by a deep convolution network. It computes a translation invariant representation which is Lipschitz continuous to deformations, with wavelet filters and a modulus pooling nonlinearity. Averaged scattering coefficients are provided by each layer. The first layer gives SIFT-type descriptors, which are not sufficiently informative for large-scale invariance, whereas the second layer brings additional stable and discriminative coefficients.

The deformation stability gives state-of-the-art classification results for handwritten digit recognition and texture discrimination, with SVM and PCA classifiers. If the dataset has other sources of variability due to the action of another Lie group such as rotations, then this variability can also be eliminated with an invariant scattering computed on this group [25], [26].

In complex image databases such as CalTech256 or Pascal, important sources of image variability do not result from the action of a known group. Unsupervised learning is then necessary to take into account this unknown variability. For deep convolution networks, it involves learning filters from data [20]. A wavelet scattering transform can then provide the first two layers of such networks. It eliminates translation or rotation variability, which can help in learning the next layers. Similarly, scattering coefficients can replace SIFT vectors for bag-of-feature clustering algorithms [8]. Indeed, we showed that second layer scattering coefficients provide important complementary information, with a small computational and memory cost.

Acknowledgments

This work is funded by the French ANR grant BLAN 0126 01 and the ERC grant InvariantClass 320959.

References

[1] S. Allassonniere, Y. Amit, and A. Trouve, “Toward a Coherent Statistical Framework for Dense Deformable Template Estimation,” *J. Royal Statistical Soc.*, vol. 69, pp. 3-29, 2007.

[2] J. Anden and S. Mallat, “Scattering Audio Representations,” *IEEE Trans. Signal Processing*, to be published.

[3] Y. Amit and A. Trouve, “POP. Patchwork of Parts Models for Object Recognition,” *Int'l J. Computer Vision*, vol 75, pp. 267-282, 2007.

[4] P.J. Bickel and E. Levina, “Covariance Regularization by Thresholding,” *Annals of Statistics*, 2008.

[5] L. Birge and P. Massart, “From Model Selection to Adaptive Estimation,” *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pp. 55-88, 1997.

[6] R.E. Broadhurst, “Statistical Estimation of Histogram Variation for Texture Classification,” *Proc. Workshop Texture Analysis and Synthesis*, 2005.

[7] J. Bruna, “Scattering Representations for Pattern and Texture Recognition,” PhD thesis, CMAP, Ecole Polytechnique, 2012.

[8] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, “Learning Mid-Level Features For Recognition,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[9] J. Bouvrie, L. Rosasco, and T. Poggio, “On Invariance in Hierarchical Models,” *Proc. Advances in Neural Information Processing Systems Conf.*, 2009.

[10] C. Chang and C. Lin, “LIBSVM: A Library for Support Vector Machines,” *ACM Trans. Intelligent Systems and Technology*, vol. 2, pp. 27:1-27:27, 2011.

[11] M. Crosier and L. Griffin, “Using Basic Image Features for Texture Classification,” *Int'l J. Computer Vision*, pp. 447-460, 2010.

[12] L. Fei-Fei, R. Fergus, and P. Perona, “Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.

[13] Z. Guo, L. Zhang, and D. Zhang, “Rotation Invariant Texture Classification Using LBP Variance (LBPV) with Global Matching,” *J. Pattern Recognition*, vol. 43, pp. 706-719, Aug. 2010.

[14] B. Haasdonk and D. Keysers, “Tangent Distance Kernels for Support Vector Machines,” *Proc. 16th Int'l Conf. Pattern Recognition*, 2002.

[15] E. Hayman, B. Caputo, M. Fritz, and J.O. Eklundh, “On the Significance of Real-World Conditions for Material Classification,” *Proc. European Conf. Computer Vision*, 2004.

[16] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, “What Is the Best Multi-Stage Architecture for Object Recognition?” *Proc. 12th IEEE Int'l Conf. Computer Vision*, 2009.

[17] D. Keysers, T. Deselaers, C. Gollan, and H. Ney, “Deformation Models for Image Recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1422-1435, Aug. 2007.

[18] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, “Exploring Strategies for Training Deep Neural Networks,” *J. Machine Learning Research*, vol. 10, pp. 1-40, Jan. 2009.

[19] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.

[20] Y. LeCun, K. Kavukvuoglu, and C. Farabet, “Convolutional Networks and Applications in Vision,” *Proc. IEEE Int'l Symp. Circuits and Systems*, 2010.

[21] T. Leung and J. Malik, “Representing and Recognizing the Visual Appearance of Materials Using Three-Dimensional Textons,” *Int'l J. Computer Vision*, vol. 43, no. 1, pp. 29-44, 2001.

[22] J. Lindenstrauss, D. Preiss, and J. Tise, *Fréchet Differentiability of Lipschitz Functions and Porous Sets in Banach Spaces.* Princeton Univ. Press, 2012.

[23] D.G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[24] S. Mallat, “Recursive Interferometric Representation,” *Proc. European Signal Processing Conf.*, Aug. 2010.

[25] S. Mallat, "Group Invariant Scattering," *Comm. Pure and Applied Math.*, vol. 65, no. 10, pp. 1331-1398, Oct. 2012.

[26] L. Sifre and S. Mallat, "Combined Scattering for Rotation Invariant Texture Analysis," *Proc. European Symp. Artificial Neural Networks*, Apr. 2012.

[27] J. Mairal, F. Bach, and J. Ponce, "Task-Driven Dictionary Learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791-804, Apr. 2012.

[28] A.Y. Ng and M.I. Jordan, "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," *Proc. Advances in Neural Information Processing Systems Conf.*, 2002.

[29] L. Perrinet, "Role of Homeostasis in Learning Sparse Representations," *Neural Computation J.*, vol. 22, pp. 1812-1836, 2010.

[30] M. Ranzato, F. Huang, Y. Boreau, and Y. LeCun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[31] C. Sagiv, N.A. Sochen, and Y.Y. Zeevi, "Gabor Feature Space Diffusion via the Minimal Weighted Area Method" *Proc. Third Int'l Workshop Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 621-635, 2001.

[32] B. Scholkopf and A.J. Smola, *Learning with Kernels.* MIT Press, 2002.

[33] S. Soatto, "Actionable Information in Vision," *Proc. 12th IEEE Int'l Conf. Computer Vision*, 2009.

[34] E. Tola, V. Lepetit, and P. Fua, "DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 815-830, May 2010.

[35] M. Varma and A. Zisserman, "Texture Classification: Are Filter Banks Necessary?" *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003.

[36] M. Welling, "Robust Higher Order Statistics," *AISTATS*, 2005.

[37] I. Waldspurger, A. D'Aspremont, and S. Mallat, "Phase Recovery, Maxcut and Complex Semidefinite Programming," *ArXiv: 1206.0102*, June 2012.

Joan Bruna received the graduate degree from the Universitat Politecnica de Catalunya in both mathematics and electrical engineering in 2002 and 2004, respectively, and the MSc degree in applied mathematics from ENS Cachan, France, in 2005. He is currently working toward the PhD degree in applied mathematics at the Ecole Polytechnique, Palaiseau, France. From 2005 to 2010, he was a research engineer in an image processing startup, developing real-time video processing algorithms. His research interests include invariant signal representations, stochastic processes, and functional analysis. He is a member of the IEEE.

Stéphane Mallat received the engineering degree from the Ecole Polytechnique, Paris, the PhD degree in electrical engineering from the University of Pennsylvania, Philadelphia, in 1988, and the habilitation in applied mathematics from the Université Paris-Dauphine, France. In 1988, he joined the Computer Science Department of the Courant Institute of Mathematical Sciences where he became an associate professor in 1994 and a professor in 1996. From 1995 to 2012, he was a full professor in the Applied Mathematics Department at the Ecole Polytechnique, Paris. From 2001 to 2008, he was a cofounder and CEO of a start-up company. Since 2012, he has been with the Computer Science Department of the Ecole Normale Supérieure in Paris. His research interests include computer vision, signal processing, and harmonic analysis. He received the 1990 IEEE Signal Processing Society's paper award, the 1993 Alfred Sloan fellowship in mathematics, the 1997 Outstanding Achievement Award from the SPIE Optical Engineering Society, the 1997 Blaise Pascal Prize in applied mathematics from the French Academy of Sciences, the 2004 European IST Grand prize, the 2004 INIST-CNRS prize for most cited French researcher in engineering and computer science, and the 2007 EADS prize of the French Academy of Sciences. He is a fellow of the IEEE and EURASIP.