



# SCATTERING HIDDEN MARKOV TREES

## IMAGE REPRESENTATION AND SCATTERING TRANSFORM MODELING

Jean-Baptiste Regli

2013-2014

RESEARCH REPORT

Academic supervisor : James Nelson
Sponsor : Dstl/UCL Impact studentship

UCL
Department of Statistical Science
London

# Contents :

# List of figures :

# 1 Introduction :

## 1.1 Image representation :

The first part of this report will be dedicated to explaining the motivations as well as the construction process of the scattering transform (ST). To understand those motivations we will first provide an intuition of what a "good" representation of signal -for classification- is. We will then provide more formal mathematical definitions for our intuition and how this can be achieved.

### 1.1.1 Intuition of a "good" image representation :

One way to develop an intuition on what properties a "good" representation for classification have is to look at the human visual function and what he is able to tell apart.

Based on that, we think our representation should be :

— ***Informative*** enough to permit classification.

— ***Invariant to translations***. Indeed to a human eye there is no difference in the information carried by a signal if it is shifted.

— ***Stable to deformations***. Once again to a human eye, it is still possible to recognize a signal if it has undergone -small- deformations. Yet if the deformations is too important the informational content of the signal is lost.

— ***To a certain degree invariant to rotations***. Rotations cannot be handled as easily as translation. Indeed here one is after a local rotational invariance rather than a global one. Solutions exist to develop a scattering transform with such behaviour ***citation*** but this will not be addressed in this review.

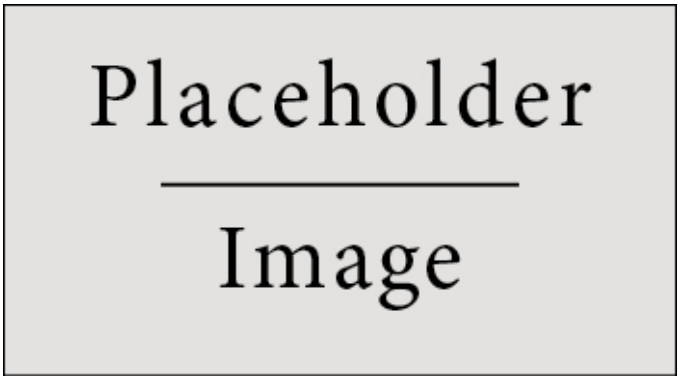


FIGURE 1.1 – A human can easily tell that those two images are from the same class.

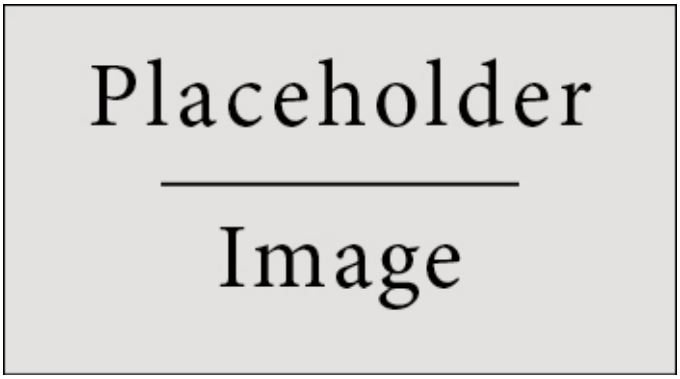


FIGURE 1.2 – A human can easily tell that (a) and (b) are from the same class. (c) can still be recognized even though it is slightly more challenging.

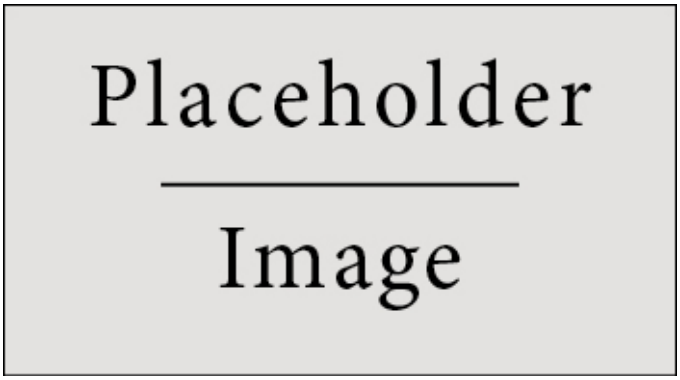


FIGURE 1.3 – A human can easily tell that (a) and (b) are from the same class. (c) could be a '6' slightly rotated or a '9' heavily rotated.




### 1.1.2 Formalization of a "good" image representation :

Let us now try to provide formal mathematical framework to the properties listed above.

In this document we define a signal as,

**Definition 1.1.1. Signal**
A signal $f$ is a square-integrable $d$ dimensional real function.

$$f \in \mathcal{L}^2(\mathbb{R}^d).$$

We will call $L_{(.)}$ the translation operator for the function in $\mathcal{L}^2(\mathbb{R}^d)$, **i.e.** for $f \in \mathcal{L}^2(\mathbb{R}^d)$ and $(x, c) \in (\mathbb{R}^d)^2 \quad L_c f(x) = f(x - c)$. An operator is translation invariant -resp : canonical translation invariant- if,

**Proposition 1.1.1.** ***Translation invariant***
*An operator $\Phi : \mathcal{L}^2(\mathbb{R}^d) \to \mathcal{H}$ where $\mathcal{H}$ is an Hilbert space is translation invariant if :*

$$\forall c \in \mathbb{R}^d \text{ and } \forall f \in \mathcal{L}^2(\mathbb{R}^d) \ \Phi(L_c f) = \Phi(f).$$

**Proposition 1.1.2.** ***Canonical translation invariant***
*An operator $\Phi : \mathcal{L}^2(\mathbb{R}^d) \to \mathcal{H}$ where $\mathcal{H}$ is an Hilbert space is canonical translation invariant if :*

$$\forall f \in \mathcal{L}^2(\mathbb{R}^d) \ \Phi(L_a f) = \Phi(f) \text{ where } a \in \mathbb{R}^d \text{ is function of } f.$$

For the usual representation operators instabilities to deformations are known to appear -especially at high frequencies. To prevent this, one would like the representation to be non-expansive,

**Definition 1.1.2. Non-expensive representation** A representation $\Phi$ is non-expensive if,

$$\forall (f, h) \in (\mathcal{L}^2(\mathbb{R}^d))^2 \quad \|\Phi(f) - \Phi(h)\| \leq \|f - h\|. \tag{1.1}$$

The stability to deformations of a non-expansive operator can be expressed as its Lipschitz continuity to the action of deformations close to translations ***cite mallat GIS***. Such a diffeormorphism transform can be expressed as,

$$\begin{aligned} L_\tau : \mathcal{L}^2(\mathbb{R}^d) &\to \mathcal{L}^2(\mathbb{R}^d) \\ f \quad &\to f(\mathbb{1} - \tau) \end{aligned}$$

where $\tau(x) \in \mathbb{R}^d$ is a displacement field.

**Proposition 1.1.3.** ***Lipschitz continuous*** *A translation invariant operator $\Phi$ is said to be Lipschitz continuous to the action of mathcalC$^2$ diffeomorphisms if for any compact $\Omega \in \mathbb{R}^d$ there exists $C$ such that for all $f \in \mathcal{L}^2(\mathbb{R}^d)$ supported in $\Omega$ and all $\tau \in$ mathcalC$^2(\mathbb{R}^d)$,*

$$\|\Phi(f) - \Phi(L_\tau f)\|_{\mathcal{H}} \leq C \|f\| \left( \sup_{x \in \mathbb{R}^d} |\nabla \tau(x)| + \sup_{x \in \mathbb{R}^d} |H\tau(x)| \right) \tag{1.2}$$

where $|\nabla \tau(x)|$ and $|H\tau(x)|$ are respectively the sup-norm and the sup-norm of the Hessian tensor of the matrix $\tau(x)$.

Hence a Lipschitz continuous operator $\Phi$ is almost invariant to "local" translations by $\tau(x)$, up to the fist and second order deformations terms. The equation 1.2 also implies that $\Phi$ is invariant to global translations.

### 1.1.3 State of the art in image representation :

Now that we have listed the properties we would like our representation to have, let us have a look at the usual signal representation tools and see if they which of them they fulfil.

The first representation method one can think of is the modulus of the ***Fourier transform***. This operator is informational enough to allow -to a certain extent- discrimination different type of signal ***find a citation for clf with fourier transform***. It is also translation invariant ***find a citation***. However it is well known that those operators present instabilities to deformation at high frequencies ***cite 10 from mallat*** and thus are not Lipschitz continuous to the action of diffeomorphisms.

***Wavelet transform*** is another popular representation method. Again they provide a "good enough" representation to allow classification of different signals ***find citations***. Plus by grouping high frequencies into dyadic packet in $\mathbb{R}^d$, wavelet operators are stable to deformations ***citation mallat's book***. However wavelets are known to be non-invariant to translations.

Another signal representation method popular at the moment are the ***convolutional neural networks cite LeCun***. As opposed to the two previously mentioned representation methods, those operators are not fixed but learned from the data ***cite learning method from CNN***. Over the past few year they have provided state of the art results on many standard classification task, such as MNIST ***cite***, CIFAR ***cite***, ImageNet ***cite*** or ***find a example in speech processing***. Those good results are used to advocate that those networks are learning "good" representations. However it seems that in certain cases they learn representation of the data that are -for example- not invariant to deformations ***cite Bruna and Al strange pties of NN***.

In the next section we will focus on the construction of a wavelet based operator with a structure somehow similar to a convolutional neural network which will be fulfilling all the properties of what we have defined as a "good" signal representation for classification.
Another representation method popular at the moment is the convolutional neural networks.

## 1.2 Scattering transform :

? ? ? - not sure yet

## 1.3 Probabilistic graphical model :

? ? ? - not sure yet

## 1.4 Outline of the report :

# 2 The Scattering transform :

In this section we describe the contruction process of a mathematical operator - the scattering transform (ST)- designed to generate what we have considered to be an interesting representation of our data (see 1.1). Therefore a scattering transform builds ***invariant***, ***stable*** and ***informative representation*** of signals through a ***non-linear***, ***unitary transform***. It is an operator delocalizing signal informational content into scattering decomposition path, computed by ***cascading wavelet/modulus operators***. This architecture is similar to a ***convolutional neural network*** (CNN) where the synaptic weights would be given by a wavelet operator instead of learned.

In 2.1, we will quickly introduce the standard architecture of a CNN, then we will explain how are built the scattering operators (see 2.2) and review some of their important properties (see ***ref : correct section***). Finally, in 2.3, we will describe how the scattering transform is usually used for classification tasks.

## 2.1 Convolutional Neural Network :

A convolutional neural network is a multilayer architecture cascading... ***TBD***

## 2.2 Scattering Convolution Network :

In this section we first introduce a wavelet-based scattering transform built to have interesting properties for classification tasks, meaning being translation invariant and stable to $\mathcal{L}^2$ deformations. Then we describe its convolutional architecture.

### 2.2.1 Scattering wavelets :

A two-dimensional directional wavelet is obtained by scaling and rotating a single band-pass filter $\psi$. If we let $G$ be a discrete, finite rotation group of $\mathbb{R}^2$, multi-scale directional wavelet filters are defined for any scale $j \in \mathbb{Z}$ and rotation $r \in G$ by

$$\psi_{2^j r}(u) = 2^{2j}\psi(2^j r^{-1}u). \tag{2.1}$$

To simplify the notations, we will now denote $\lambda = \lambda(j, r) \overset{d}{=} 2^j r \in \Lambda \overset{d}{=} G \times \mathbb{Z}$.

A wavelet transform filters the signal $x$ using a family of wavelets $\{x * \psi_\lambda(u)\}_\lambda$. This is computed from a filter bank of dilated and rotated wavelets having no orthogonality property and it creates a multi-scale and orientation representation of the input.

If $u.u'$ and $\|u\|$ define respectively the inner product and the norm in $\mathbb{R}^2$, the Morlet wavelet $\psi$ is an example of wavelet given by

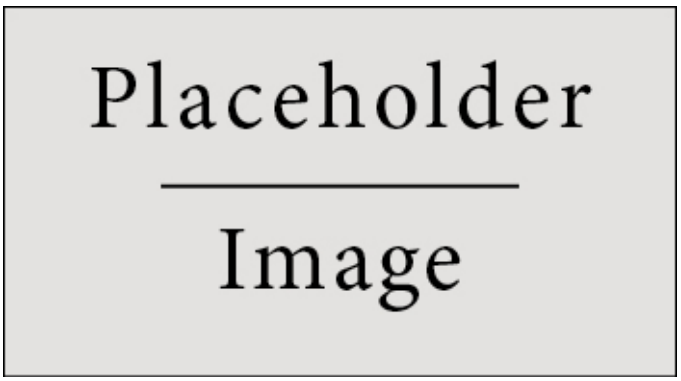


FIGURE 2.1 – Complex Morlet wavelet.

$$\psi(u) = C_1(e^{iu.\xi} - C_2)e^{\|u\|^2/(2\sigma^2)},$$

where $C_1$, $\xi$ and $\sigma$ are meta-parameters of the wavelet and $C_2$ is adjusted so that $\int \psi(u)du = 0$. Figure 2.1 shows a Morlet wavelet for $\xi = 3\pi/4$ and $\sigma = 0.85$.

As opposed to the Fourier sinusoidal waves, wavelets are operators stable to $\mathcal{L}^2$ deformations as they can be expressed as localized waveforms (***citation***). However, wavelet transforms compute convolutions with wavelets, hence they are translation covariant operators (***citation***).

To ensure a translation invariant behavior to an operator commuting with them, one has to introduce a non-linearity. For example if $R$ is a linear or non-linear operator commuting with translations $L_c$, **i.e.** $R(L_c x) = L_c R(x)$), then the integral $\int R(x(u))du$ is translation invariant. One can apply this to $R(x) = x * \psi_\lambda$ and gets the trivial invariant,

$$\int x * \psi_\lambda(u)du = 0,$$

for all $x$ as $\int \psi_\lambda(u)du = 0$. However to preserve the informative character of the scattering operator, one has to ensure that the integral does not vanish. To do so an operator $M$ such that $R(x) = M(x * \psi_\lambda)$ is introduced. If $M$ is a linear transformation commuting with translation then the integral still vanishes. Hence one has to choose $M$ to be a non-linear.

Keeping in mind that the scattering transform has to be stable to deformations and taking advantages of the wavelet transform stability to small deformations in the input space, we also impose that $M$ commutes with deformations

$$\forall \tau(u)\ ,\ ML_\tau = L_\tau M.$$

If a weak differentiability condition is added, one can prove (***ref ISCN 6***) that $M$ must necessarily be a point-wise operator, **i.e.** $Mx(u)$ only depends on the value of $x(u)$. Finally, by adding an $\mathcal{L}^2(\mathbb{R}^2)$ stability constraint,

$$\forall (x,y) \in \mathcal{L}^2(\mathbb{R}^2)^2\ ,\ \|Mx\| = \|x\| \text{ and } \|Mx - My\| \leq \|x - y\|\,,$$

one can show (***ref ISCN 6***) that necessarily $Mx = e^{i\alpha}\,|x|$. For the scattering transform $\alpha$ is set to 0 and therefore the resulting coefficients are the $\mathcal{L}^1(\mathbb{R}^2)$ norms :

$$\|x * \psi_\lambda\|_1 = \int |x * \psi_\lambda|\, du$$

The family of $\mathcal{L}^1(\mathbb{R}^2)$ normed wavelet $\{\|x * \psi_\lambda\|_1\}_\lambda$ generate a crude signal representation which measures the sparsity of the wavelet coefficients. One can prove (***ref ISCN 36***) that $x$ can be reconstructed from $\{|x * \psi_\lambda(u)|\}_\lambda$ up to a multiplicative constant. Which means that the information loss in $\{\|x * \psi_{\lambda_1}\|\}_\lambda$ comes from the integration of the absolute value $|x * \psi_\lambda(u)|$ which removes all non-zero frequencies. However those components can be recovered by calculating the wavelet coefficients $|x * \psi_{\lambda_1}| * \psi_{\lambda_2}(u)$. By doing so their $\mathcal{L}^1(\mathbb{R}^2)$ norms define a much larger family of invariants :

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

By further iterating on the wavelet/modulus operators more translation invariant coefficients can be computed. Let us define :

**Definition 2.2.1. Scattering Propagator**
The scattering operator $U$ for a scale and an orientation $\lambda \in G \times \mathbb{Z}$ is defined as the absolute value of the input convolved with the wavelet operator at this scale and orientation.

$$U[\lambda](x) \overset{d}{=} |x * \psi_\lambda| \tag{2.2}$$

**Definition 2.2.2. Path Ordered Scattering Propagators**
Any sequence $p = (\lambda_1, \lambda_2, ..., \lambda_m)$ where $\forall i \in [\![1, m]\!] \lambda_i \in G \times \mathbb{Z}$ defines a ***path*** of length $m$, **i.e.** the ordered product of non-linear and non-commuting operators

$$\begin{aligned} U[p]x &\overset{d}{=} U[\lambda_m]...U[\lambda_2]U[\lambda_1](x) \\ &= ||| |x * \psi_{\lambda_1}| * \psi_{\lambda_2}| ...| * \psi_{\lambda_m}| \end{aligned} \tag{2.3}$$

With the convention : $U[\emptyset]x = x$

From there one can provide a first formal definition of the scattering transform :

**Definition 2.2.3. Scattering Coefficient**
A scattering coefficient along the path $p$ is defined as an integral of the $p$ ordered scattering propagators, normalized by the response of a Dirac :

$$\bar{S}[p](x) \overset{d}{=} \mu_p^{-1} \int U[p]x(u)du \tag{2.4}$$

with,

$$\mu_p \overset{d}{=} \int U[p]\delta(u)du$$

We shall see later (***reference***) that each scattering coefficient $\bar{S}[p](x)$ is -as desired - invariant to translation of the input $x$ and Lipschitz continuous to deformations.

For classification tasks, one might want to compute localized descriptors only invariant to translations smaller than a predefined scale $2^J$, while keeping the spatial variability at scales larger than $2^J$. One can achieved this by localizing the scattering integral with a scaled spatial window $\phi_{2^J}(u) = 2^{-2J}\phi(2^{-2J}u)$. This yield to the definition of the windowed scattering transform :

**Definition 2.2.4. -Windowed- Scattering Coefficient Of Order $m$**
If $p$ is a path of length $m \in \mathbb{N}$, the -windowed- scattering coefficient of order m at scale $2^J$ is defined as :

$$\begin{aligned} S_J[p](x) &\stackrel{d}{=} U[p]x * \phi_{2^J}(u) \\ &= \int U[p]x(v)\phi_{2^J}(u-v)dv \\ &= ||||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| ...| * \psi_{\lambda_m}| * \phi_{2^J}(u) \end{aligned} \tag{2.5}$$

With the convention : $S_J[\emptyset]x = x * \phi_{2^J}$

### 2.2.2 Scattering Convolution Network :

**The first order scattering coefficient :**

A scattering transform computes higher-order coefficients by further iterating on the wavelet transform/modulus operators. At a maximum scale $2^J$, wavelet coefficients are computed at frequencies $2^j \geq 2^{-J}$, and lower frequencies are filtered by $\phi_{2^J} = 2^{-2J}\phi(2^{-J}u)$.

In the image processing case, as images are real-valued signal, one can only consider the "positive" rotations $r \in G^+$ with angles $[0, \pi)$ :

$$W_J x(u) \tag{2.6}$$

### 2.2.3 Spatial Wavelet transform :

Introduction of the translation invariance.

### 2.2.4 Roto-translation wavelet transform :

Introduction of the pseudo rotation invariance.

## 2.3 Application to classification :

Examples of Mallat's work.

# 3 Probabilistic graphical models :

In this chapter, we will introduce the two mains classes of probabilistic graphical models. (1) The Bayesian networks. (2) The Hidden Markov Models.

## 3.1 Bayesian Network :

Quick overview over the main methods for BNs.

### 3.1.1 Architecture :

TBD

### 3.1.2 Learning :

TBD

### 3.1.3 Inference :

TBD

## 3.2 Hidden Markov Models :

Quick overview over the main methods for HMMs.

### 3.2.1 Architecture :

TBD

### 3.2.2 Learning :

TBD

### 3.2.3 Inference :

TBD

# 4 Hidden Markov trees :

More at length description of the specific properties of HMTs.

## 4.1 The tree structure :

TBD

## 4.2 Learning :

### 4.2.1 Expectation maximization :

EM by Crouse and EM by Durand

### 4.2.2 Variational methods :

Variational like Crouse and variational like Durand

## 4.3 Generation : Vitterbi algorithm :

# 5 Scattering hidden Markov tree :

## 5.1 Related work :

Wavelet hidden markov trees

## 5.2 Hypothesis :

(1) 2 populations. (2) Persistence.

## 5.3 Persistence :

Hopefully some kind of proof there.

# 6 Experimental results :

TBD

# 7 Conclusion :

**Scattering hidden Markov tree :** TBD

**Next steps :** Variational methods General graphical models Bayesian neural networks

# 8 Acknowledgements :

# Bibliographie