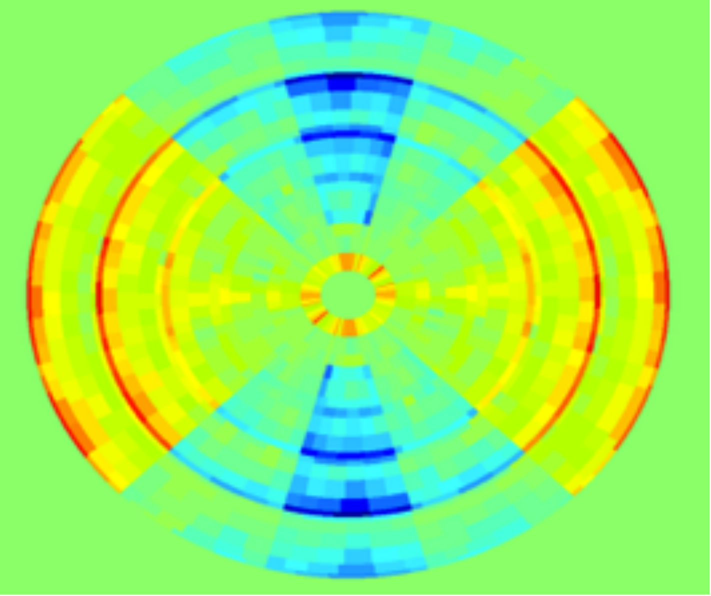
On the road to Affine Scattering Transform

Edouard Oyallon

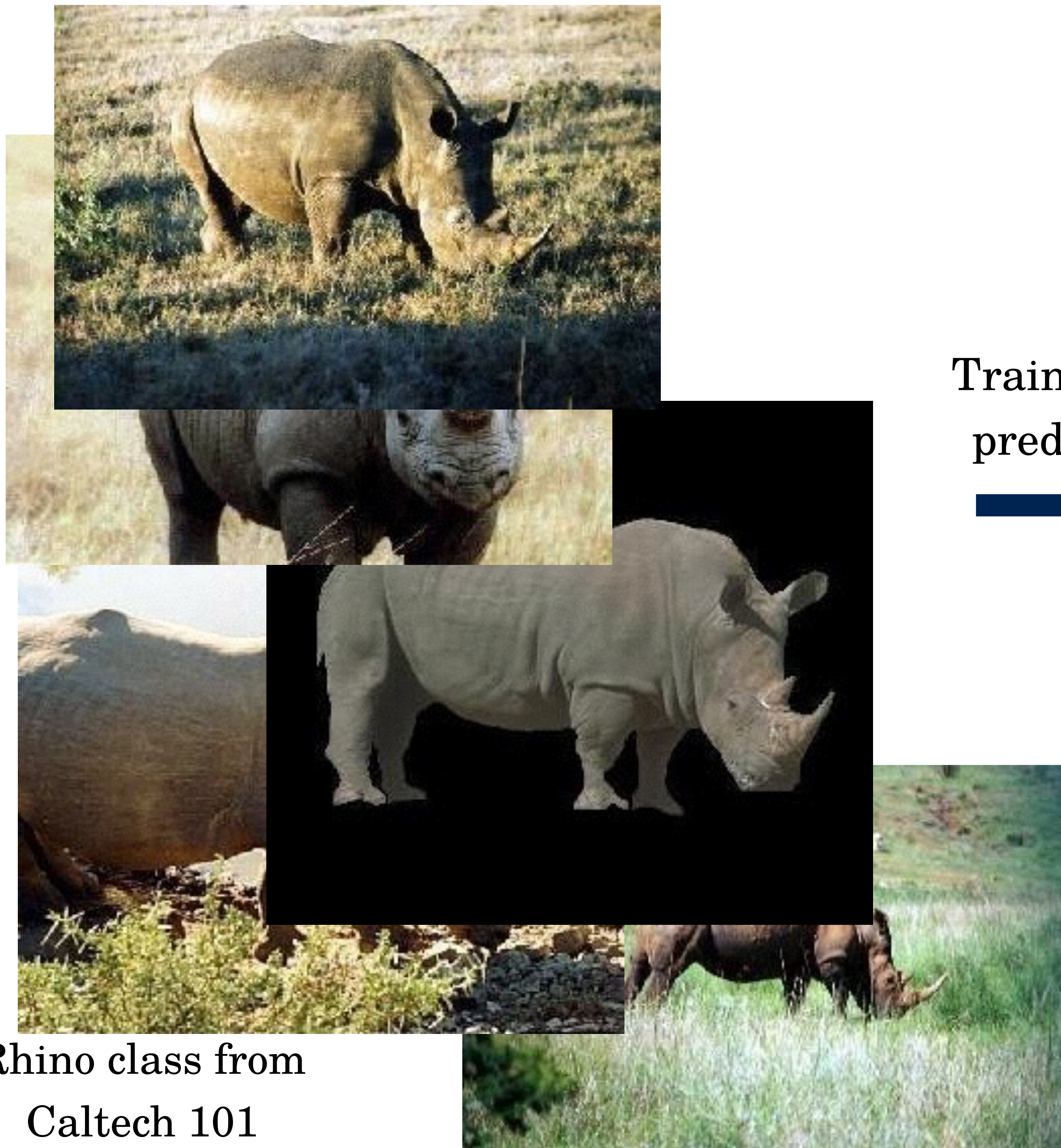


with Stéphane Mallat

following the work of Laurent Sifre, Joan Bruna, …

High Dimensional classification

$$(x_i, y_i) \in \mathbb{R}^{512^2} \times \{1, ..., 100\}, i = 1...10^4 \quad \longrightarrow \hat{y}(x)?$$



Rhino class from Caltech 101

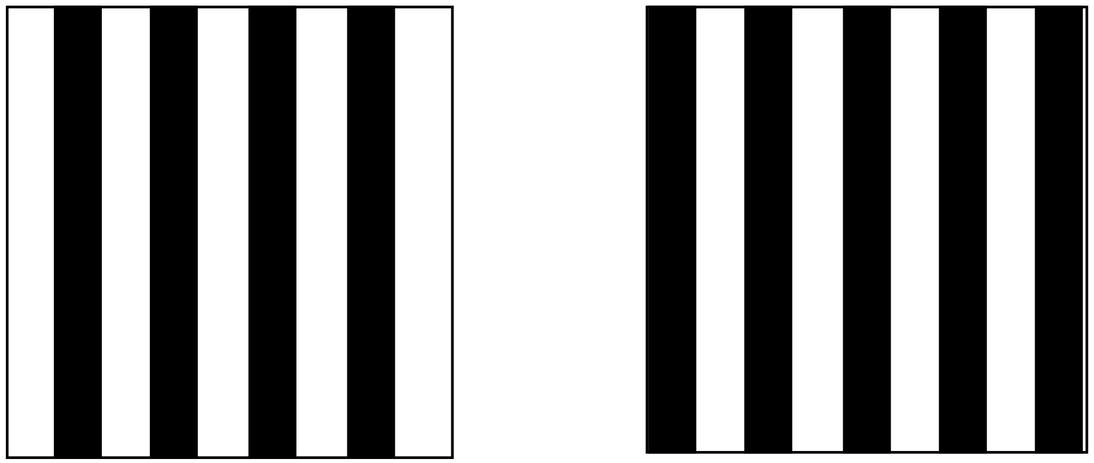
Training set to predict labels

Rhino

Not a rhino

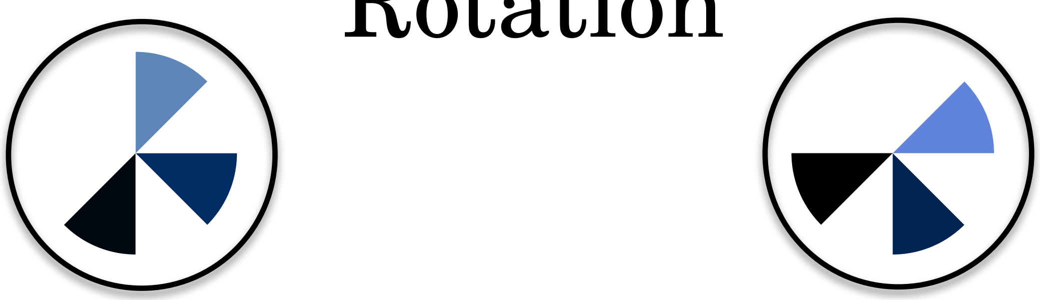
DATA

Translation



x y

Rotation



x y

$$\|x - y\|_2 = 2$$

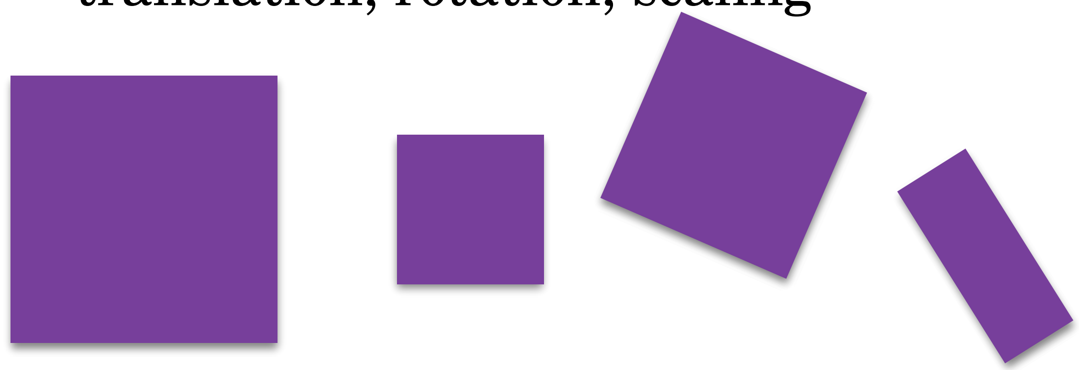
Averaging is the key to get invariants

High dimensionality issues

Fighting the curse of dimensionality

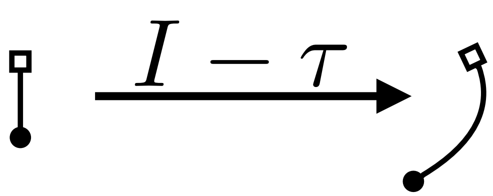
Geometric variability

Groups acting on images:
translation, rotation, scaling



Other sources : luminosity, occlusion, small deformations

$$x_\tau(u) = x(u - \tau(u)), \tau \in \mathcal{C}^\infty$$



Can be carefully handled

Class variability

Intraclass variability
<u>Not informative</u>

Extraclass variability

Needs to be learned

$x \rightarrow \Phi x \rightarrow L\Phi x \rightarrow$ Classifier

L reduces class variability

Existing approach

Ref.: Discovering objects and their location in images . Sivic et al.
High-dimensional Signature Compression for Large-Scale Image Classification, Perronnin et al.

- Unsupervised learning: Bag of Words, Fisher Vector,… $\{x_1, ..., x_N\} \longrightarrow \Phi$

- Supervised learning: Deep Learning,…

$$\{(x_1, y_1), ..., (x_N, y_N)\} \longrightarrow L\Phi$$

- Non learned: HMax, Scattering Transform.

$$\{G_1, ...\} \longrightarrow \Phi$$

Ref.: Robust Object Recognition with Cortex-Like Mechanisms. Serre et al.

DeepNet?

- It is a cascade based on **a lot of** linear operators followed by non linearities.

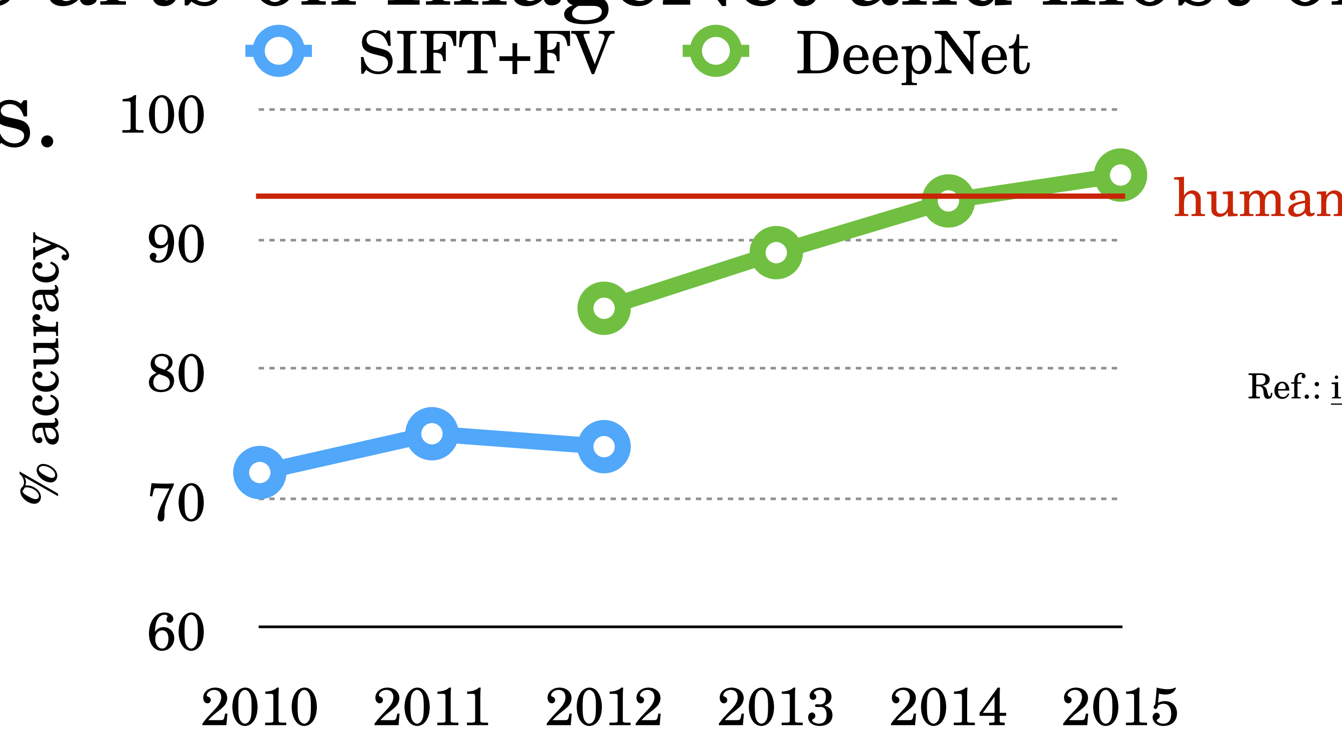
Ref.: Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick et al.

Convolutional network and applications in vision. Y. LeCun et al.

- Each operator is **supervisedly** learned

- Sort of “Super SIFT”

- State-of-the arts on ImageNet and most of the benchmarks.



Ref.: image-net.org

Complexity of the architecture

- Requires a **huge** amount of data

- Need **many engineering** to select the hyper parameters and to optimise it

- **Interpreting** the learned operators is hard when the network is deep(i.e. more than 3 layers)

Ref.: Intriguing properties of neural networks, C. Szegedy et al.

- Few theoretical results, yet outstanding numerical results.

Operators of a Deep architecture

- Linear operators are often convolutional whose kernels are **small** filters.

$$x_j(u,\lambda_2) = f(\sum_{\lambda_1} x_{j-1}(.,\lambda_1) \star h_{j,\lambda_1}(u)) \to \boxed{x_j = f(F_j x_{j-1})}$$

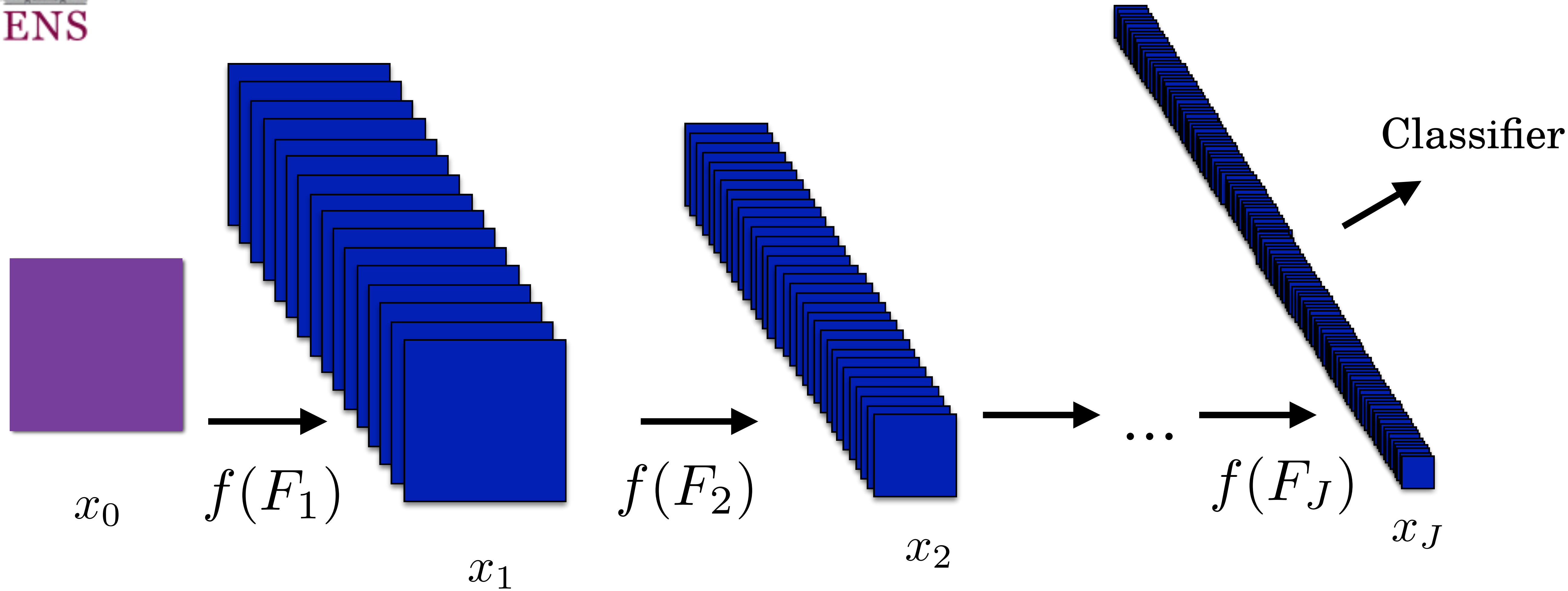
- **Deep.**

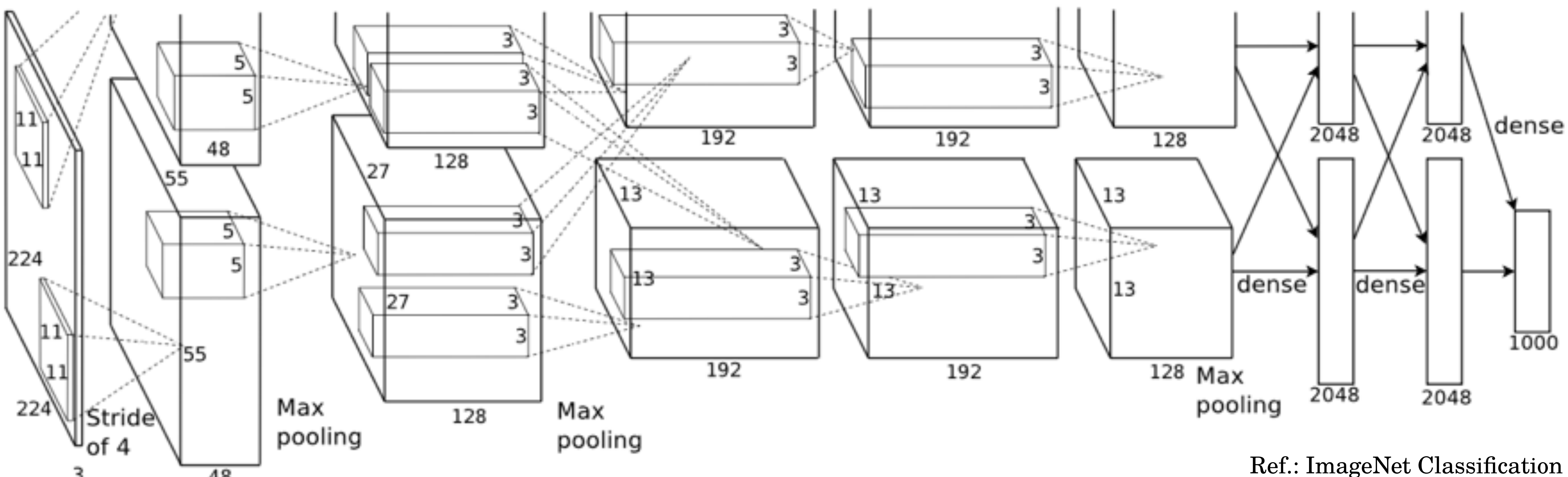
Contains a phase information

- f is often a ReLu: $x \to \max(0,x)$

- Sometimes "pooling" which leads to a down sampling.

DATA





DeepNetwork

Ref.: ImageNet Classification with Deep Convolutional Neural Networks. A Krizhevsky et al.

Does everything need to be learned?

- A **Scattering Network** is a deep architecture, where all the filters are **predefined**.

Ref.: Invariant Convolutional Scattering Network, J. Bruna et Mallat S

- We challenge the **necessity** to learn the weights of every filters of a deep architecture.

- Scattering gets state-of-the-art results on **unsupervised learning** for some complex datasets.

Ref.: Deep Rototranslation Scattering Network for complex Image Recognition, EO, S Mallat

- We highlight the similarity of the Scattering Network with the DeepNet architectures.

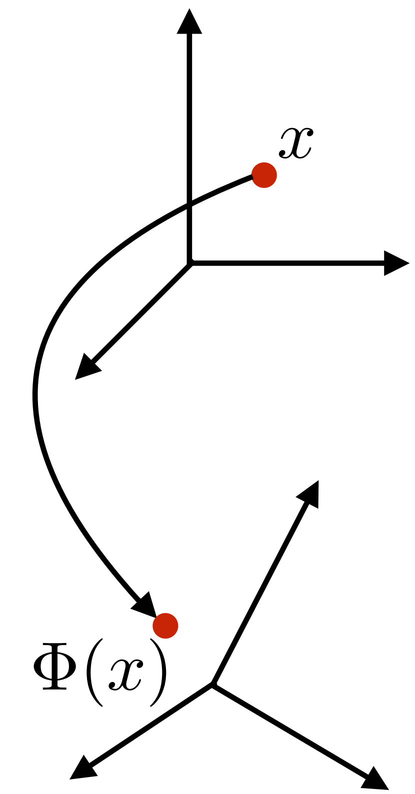
Desirable properties of a representation

- **Invariance** to group G of transformation

$$\forall x, \forall g \in G, \Phi(g.x) = \Phi(x)$$

- **Stability** to noise

$$\forall x, y, \|\Phi(x) - \Phi(y)\|_2 \leq \|x - y\|_2$$



- **Reconstruction** properties

$$y = \Phi(x) \Longleftrightarrow x = \Phi^{-1}(y)$$

- **Linear separation** of the different classes

$$\forall i \neq j, \|E(\Phi(X_i)) - E(\Phi(X_j))\|_2 \gg 1$$

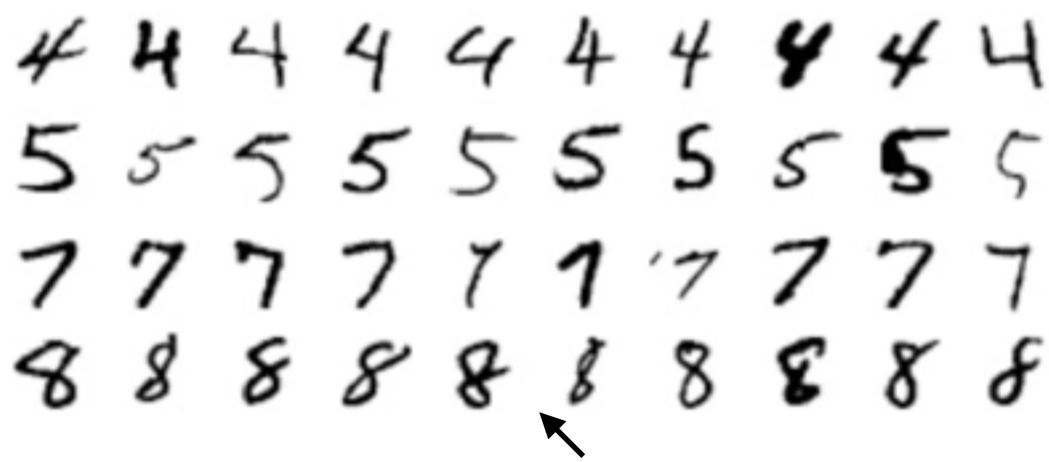
$$\forall i, \sigma(\Phi(X_i)) \ll 1$$

Success story
Scattering for Textures&Digits

- Non-learned representation have been successively used on:

Ref.: Invariant Convolutional Scattering Network, J. Bruna and S Mallat

- Digits (patterns):



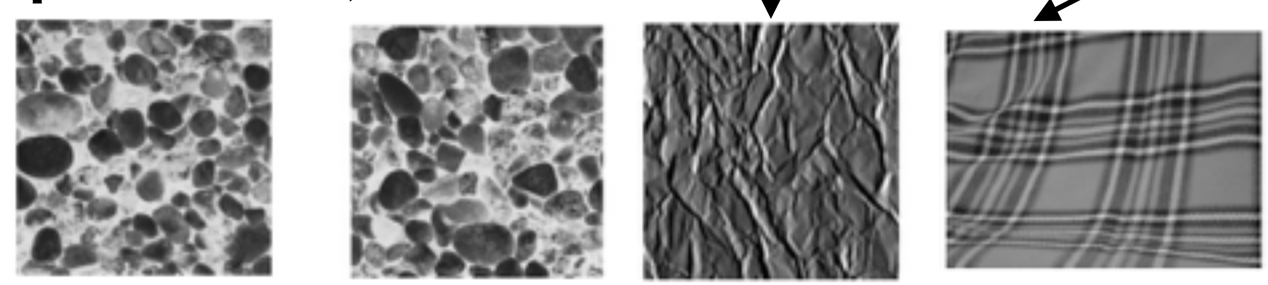
Small deformations

- Textures (stationary process):

Ref.: Rotation, Scaling and Deformation Invariant Scattering for texture discrimination, Sifre L and Mallat S.

+Translation

Rotation+Scale



- However all the variabilities(groups) here are **perfectly** understood.

Wavelets

- Wavelets help to describe signal structures. ψ is a wavelet iff

$$\psi \in \mathcal{L}^2(\mathbb{R}^2, \mathbb{C}) \text{ and } \int_{\mathbb{R}^2} \psi(u)du = 0$$

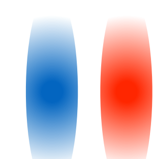
- Learned in the first layers of a DeepNet.

Ref.: ImageNet Classification with Deep Convolutional Neural Networks. A Krizhevsky et al.

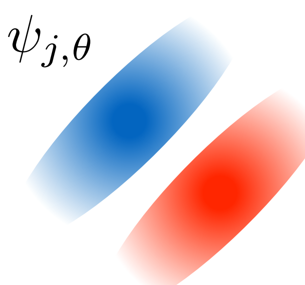
- Wavelets can be dilated in order to be a **multi-scale** representation of signals, **rotated** to describe rotations.

$$\psi_{j,\theta} = \frac{1}{2^{2j}}\psi(\frac{r_\theta(u)}{2^j})$$

ψ



$\psi_{j,\theta}$



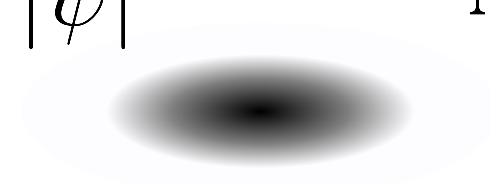
- Design wavelets selective to an **informative** variability.

Isotropic

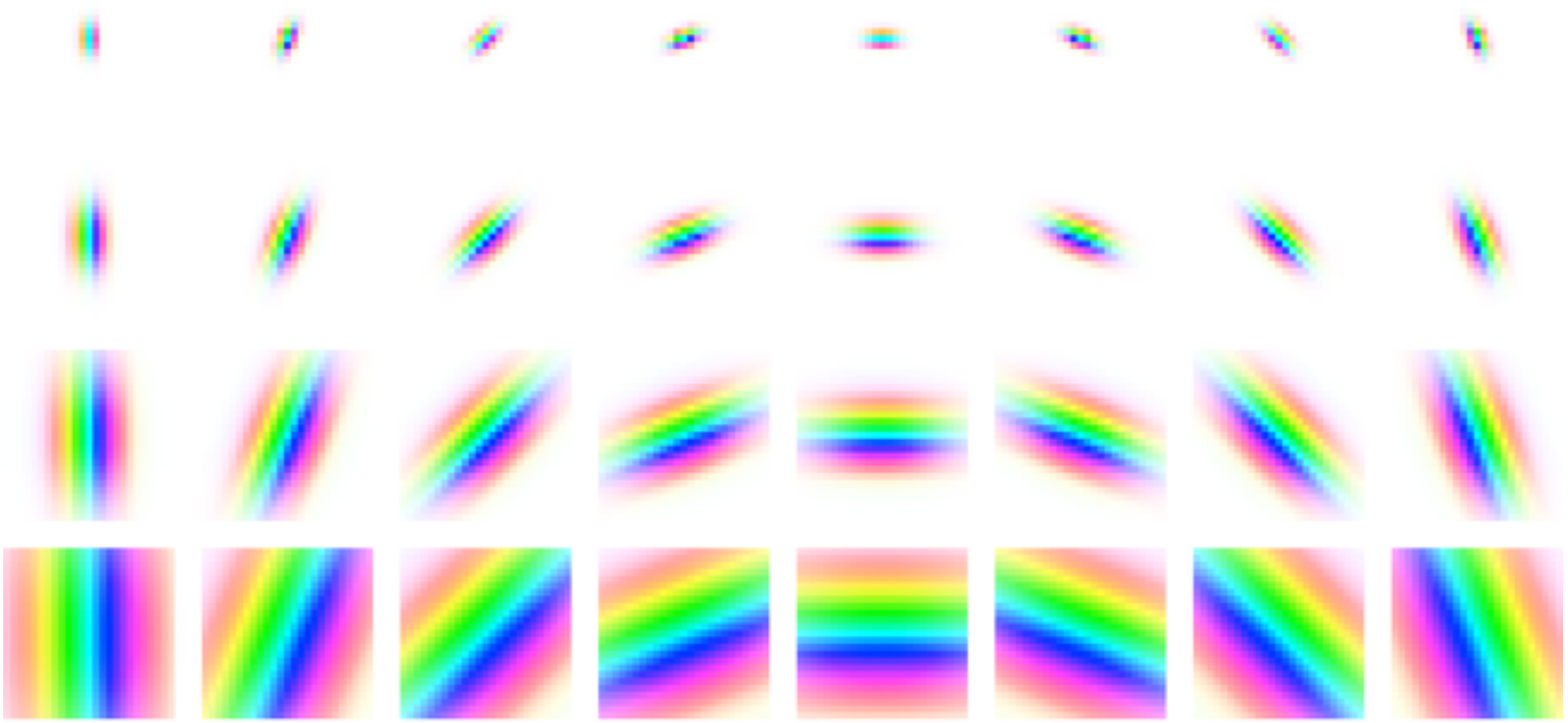
$|\hat{\psi}|$

VS

$|\hat{\psi}|$



Non-Isotropic



$$\psi(u) = \frac{1}{2\pi\sigma} e^{-\frac{\|u\|^2}{2\sigma}} \left(e^{i\xi . u} - \kappa\right)$$

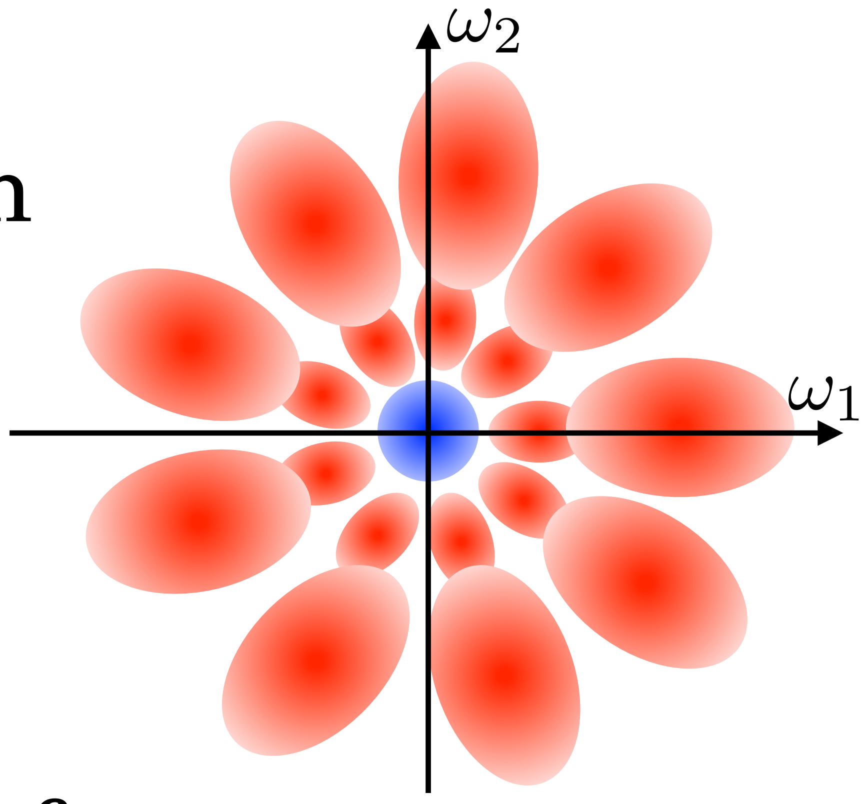
$$\phi(u) = \frac{1}{2\pi\sigma} e^{-\frac{\|u\|^2}{2\sigma}}$$

(for sake of simplicity, formula are given in the isotropic case)

The Gabor wavelet

Wavelet Transform

- Wavelet transform : $Wx = \{x \star \psi_{j,\theta}, x \star \phi_J\}_{\theta, j \leq J}$



- Isometric and linear operator, with

$$\|Wx\|^2 = \sum_{\theta, j \leq J} \int |x \star \psi_{j,\theta}|^2 + \int x \star \phi_J^2$$

- Covariant with translation

$$W(x_{\tau=c}) = (Wx)_{\tau=c}$$

- Nearly commutes with the action of diffeomorphism

$$\|[W, ._\tau]\| \leq C \|\nabla \tau\|$$

Ref.: Group Invariant Scattering, Mallat S

- Why wavelets are not enough? Invariance…

Filter bank implementation of a Fast WT

Ref.: Fast WT, Mallat S, 89

- Assume it is possible to find h and g such that

$$\hat{\psi}_\theta(\omega) = \frac{1}{\sqrt{2}}\hat{g}_\theta(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2}) \quad \text{and} \quad \hat{\phi}(\omega) = \frac{1}{\sqrt{2}}\hat{h}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$$

- Set:

$$x_j(u,0) = x \star \phi_j(u) = h \star (x \star \phi_{j-1})(2u) \quad \text{and}$$

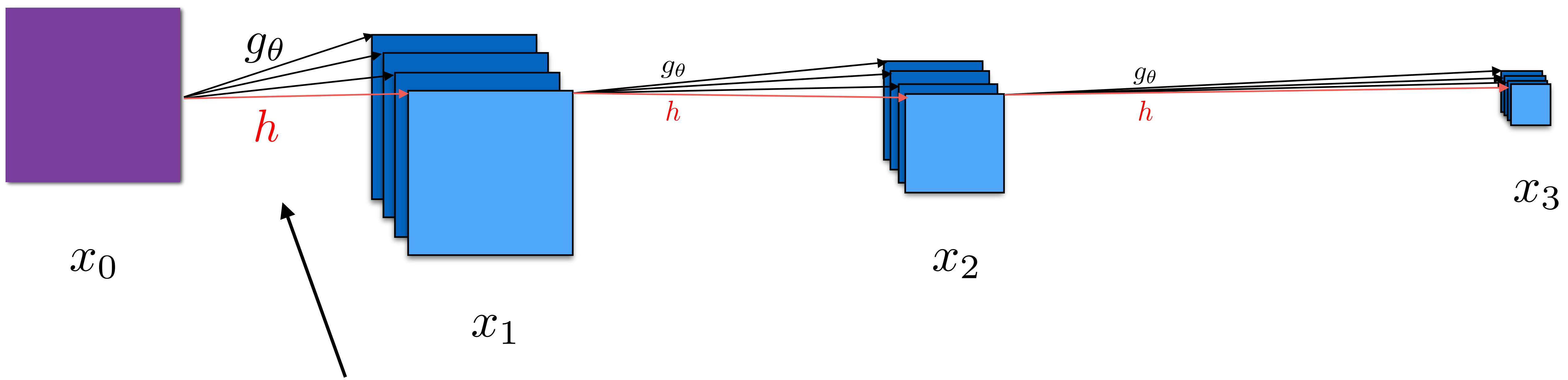
$$x_j(u,\theta) = x \star \psi_{j,\theta}(u) = g_\theta \star (x \star \phi_{j-1})(2u)$$

- The WT is then given by $Wx = \{x_j(.,\theta), x_J(.,0)\}_{j\leq J,\theta}$

- A WT can be interpreted as a **deep cascade** of linear operator, which is approximatively verified for the Gabor Wavelets.

$$\hat{\phi}_j = \frac{1}{\sqrt{2}}\hat{h}(\frac{\cdot}{2})\hat{\phi}_{j-1}$$

$$\hat{\psi}_{j,\theta} = \frac{1}{\sqrt{2}}\hat{g}_\theta(\frac{\cdot}{2})\hat{\phi}_{j-1}$$



There is an oversampling

$h \geq 0$

Deep implementation of a WT

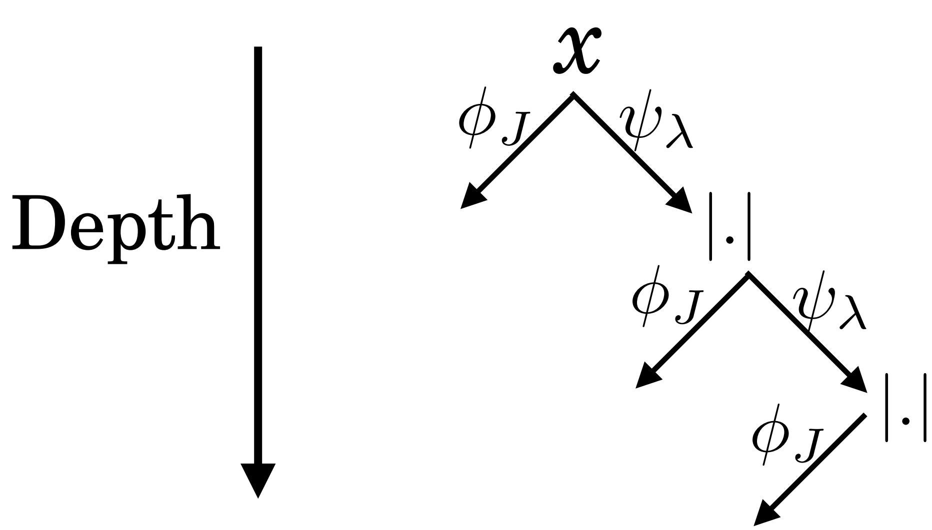
Scattering Transform

- Scattering transform at scale J is the cascading of complex WT with modulus non-linearity, followed by a low pass-filtering:

Ref.: Group Invariant Scattering, Mallat S

$$S_J x = \{x \star \phi_J, \quad |x \star \psi_{\lambda_1}| \star \phi_J, \quad ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J\}$$

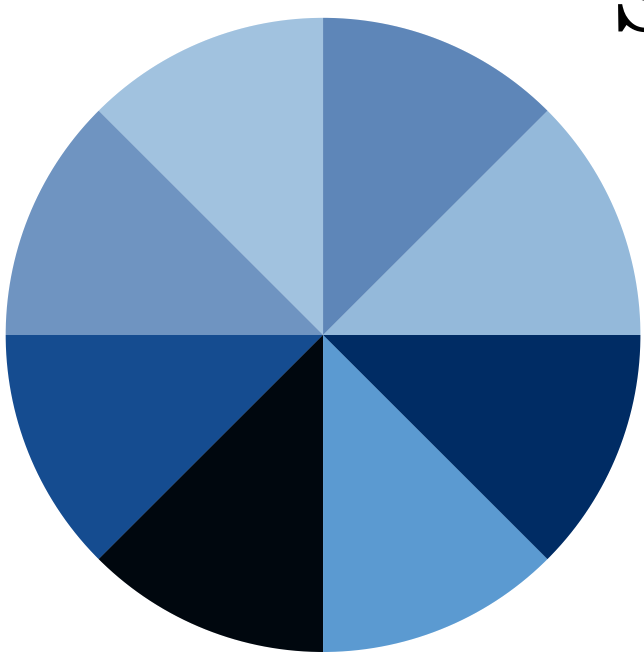
with $\lambda_i = \{j_i, \theta_i\}, j_i \leq J$



- **Mathematically** well defined for a large class of wavelets.

For people into computer vision:

SIFT performs a histogram of gradient



$$h(\theta) = \sum_{\angle g \in [\theta, \theta+\eta]} \|g\|$$

Gradient

$$= \sum_{g} \|1_{\angle g \in [\theta, \theta+\eta]} g\|$$

Quantification...

$$\approx |x \star \psi_\theta| \star \phi$$

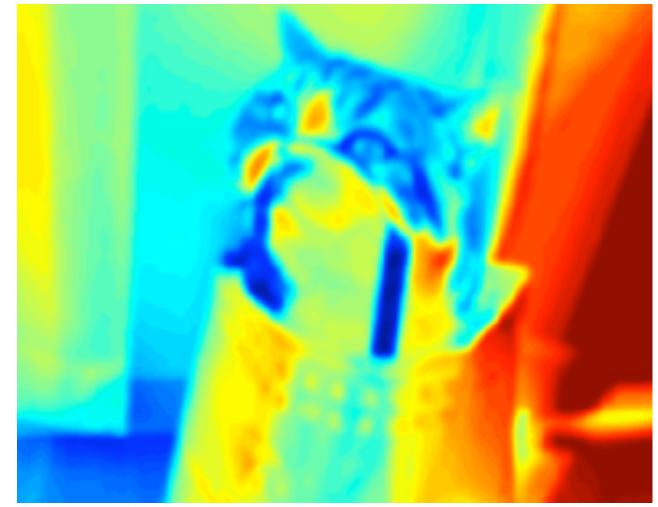
The averaging leads to a loss of information...

Relations with SIFT

Feature map

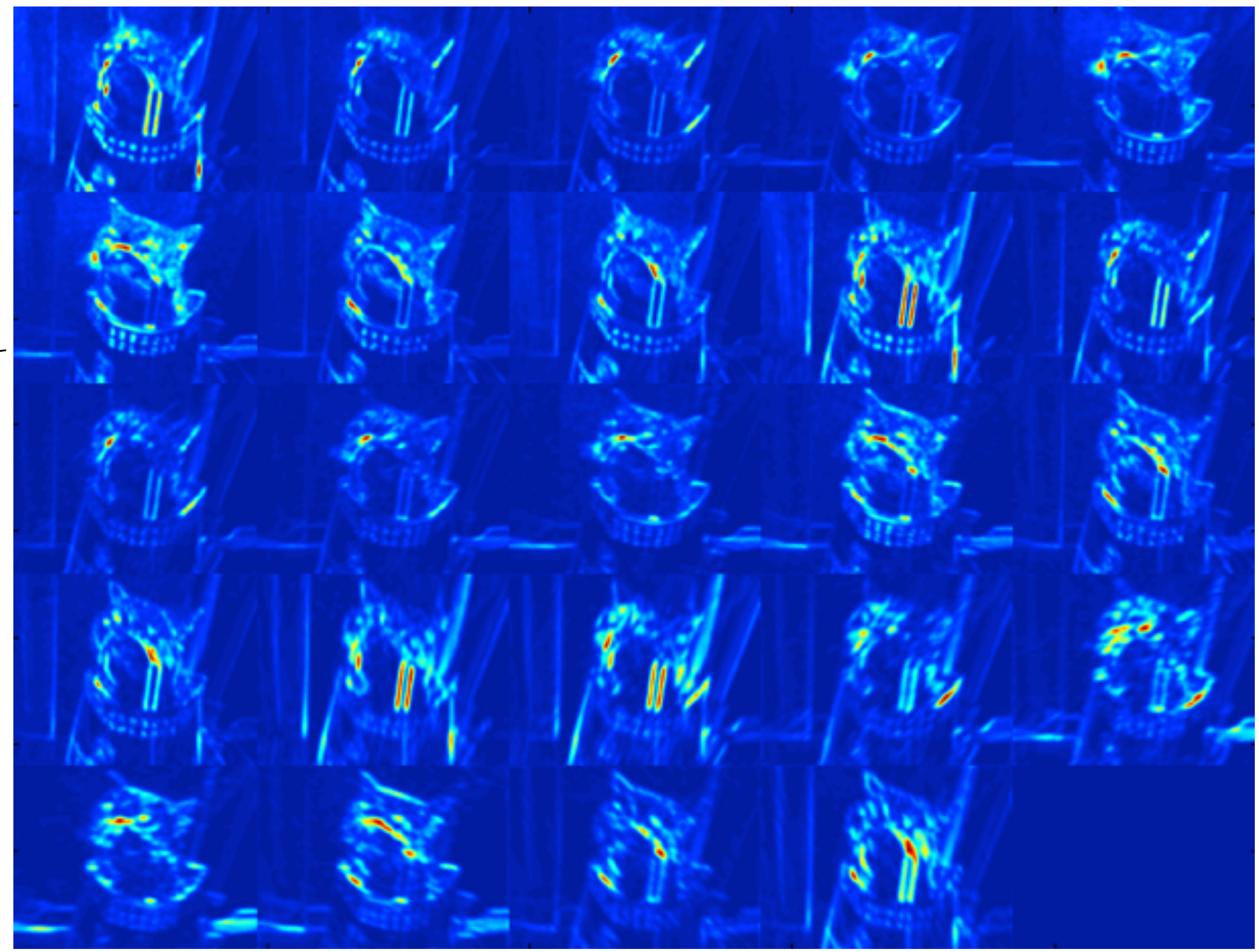
x

$x \star \phi$



$|x \star \psi| \star \phi$

1st order coefficients



Example of Scattering coefficients

Properties

Ref.: Group Invariant Scattering, Mallat S

- Non-linear
- Isometric

  $$\|S_J x\| = \|x\|$$
- Stable to noise

  $$\|S_J x - S_J y\| \leq \|x - y\|$$
- Covariant with translation

  $$S_J(x_{\tau=c}) = S_J(x)_{\tau=c}$$
- Invariant to small translation

  $$|c| \leq 2^J \Rightarrow S_J(x_{\tau=c}) \approx S_J(x)$$
- Sensitive to the action of rotation

  $$S_J(r_\theta x) \neq S_J(x)$$
- Linearize the action of small deformation

  $$\|S_J x_\tau - S_J x\| \leq C\|\nabla \tau\|$$
- Reconstruction properties

allow a linear classifier to build class invariant on informative variability

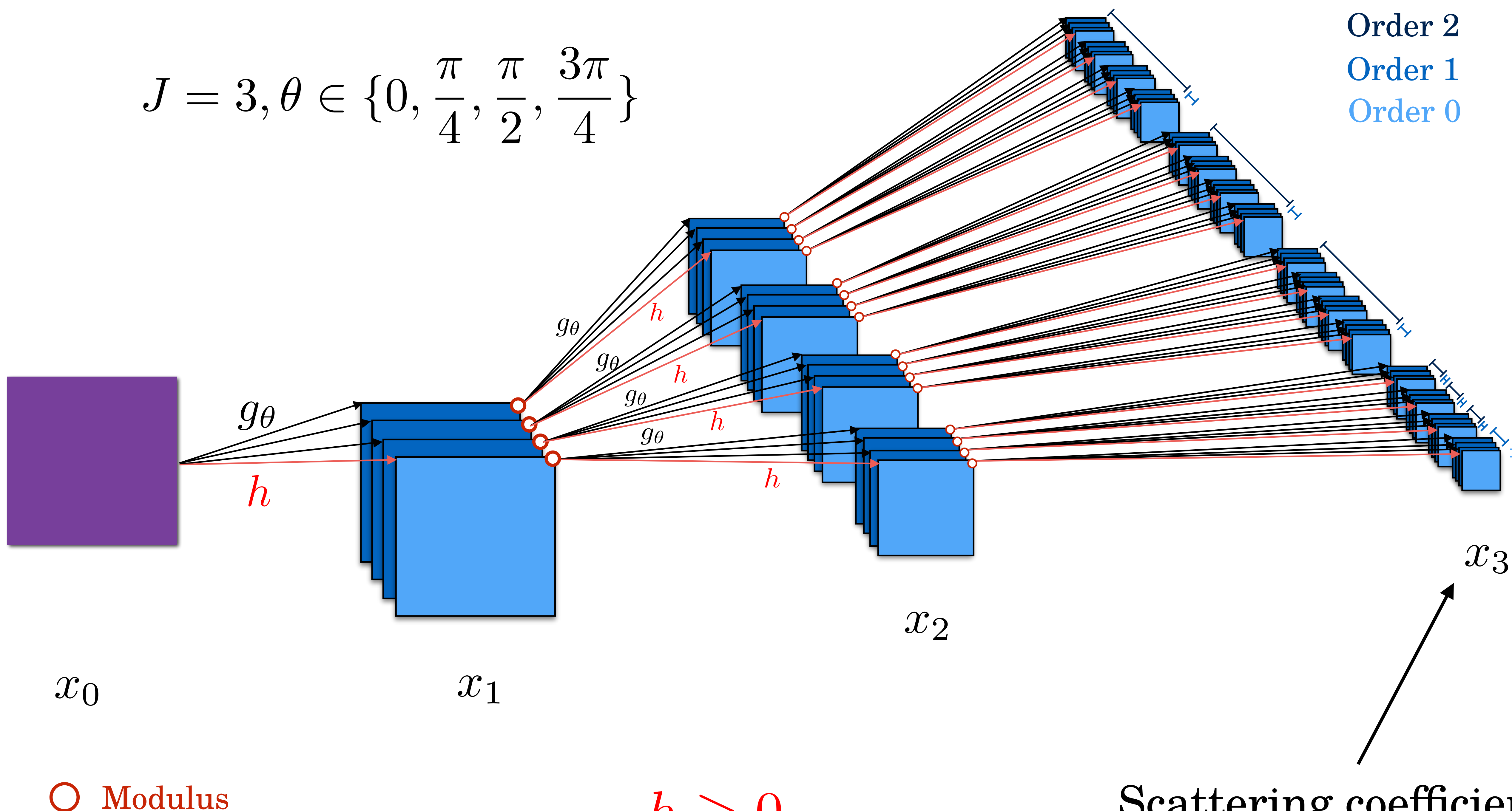
Ref.: Reconstruction of images scattering coefficient. Bruna J

Parameters&dimensionality of the Scattering

- For 256x256 images, $\Phi x \in \mathbb{R}^{10^5}$
- Yet it is possible to reduce it up to $L\Phi x \in \mathbb{R}^{2000}$
- It depends on a few parameters L, J , the shape of the mother wavelet…
- Identical parameters can be chosen for natural images on different datasets.

It is a generic representation

$$J = 3, \theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$$



$$h \geq 0$$

Scattering coefficients are only at the output

2nd order Translation Scattering

Affine Scattering Transform

- Observe that $|W|$ is covariant with the affine group Aff(E):

$$|W|[\lambda](g.x) = |(g.x) \star \psi_\lambda| = |W|[g.\lambda]x$$

Ref.: PhD, L Sifre

- See $|W|$ as a signal parametrised by some elements of the affine group:

$$|x \star \psi_{j,\theta}(u)| = \tilde{x}(g), g = (u, r_\theta, j)$$

- We can define a WT on any compact Lie group(even not commutative) via:

Ref.: Topics in harmonic analysis related to the Littlewood-Paley theory Stein EM

$$x \star^G \psi(g) = \int_G \psi(g')x(g'^-1.g)dg'$$

- The same previous properties hold for this WT/Scattering.

Ref.: Group Invariant Scattering, Mallat S

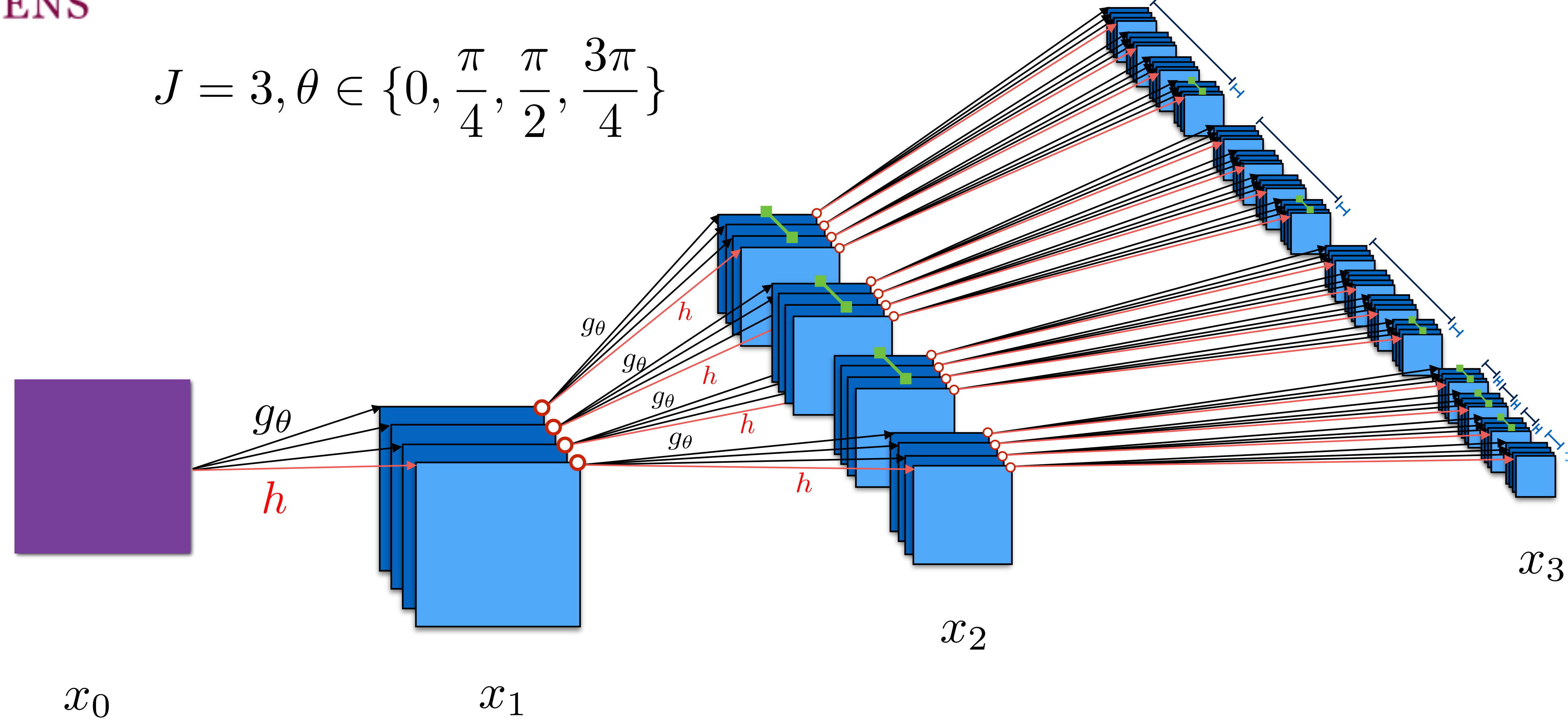
Separable Roto-Translation Scattering

Ref.: Deep Roto-Translation Scattering for Object Classification. EO and S Mallat

- Roto-translation group is not separable yet we used a **separable wavelet transform** on it.
- No averaging along angle (sensitivity increased)
- Separable (simple to implement and fast)
- Equal (slightly better) results as with non separable

DATA

$$J = 3, \theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$$



Ref.: Deep Roto-Translation Scattering for Object Classification. EO and S Mallat

Separable convolution that recombines angles on 2nd order

Separable Roto-Translation scattering

Linearization of the rotation

Work in progress

- For a deformation:

$$u - \tau(u) \approx v - \tau(v) + (I - \nabla\tau)(v)(u - v)$$

- In fact the the affine group acts also on the deformation diffeomorphism:

$$g.(I - \tau)(u) \approx g.(I - \tau)(v) + g.(I - \nabla\tau)(v)(u - v)$$

- Decompose it on the affine group: $(I - \nabla\tau)(v) = r_{\tau_\theta(v)} K(v)$

- A way to linearize the action of the rotation?

$$\|Sx - Sx_\tau\| \leq C(\|\nabla\tau_\theta\| + ...)$$

Classification pipeline

$$x \longrightarrow \Phi x \longrightarrow L\Phi x \longrightarrow \langle L\Phi x, w \rangle \geq 0$$

Scattering Transform on YUV +Log

Supervised dimensionality Reduction

Gaussian SVM

- We learn L(**select features**) and w(**select samples**) from the data

- Getting L is the most costly part of the algorithm…(Orthogonal Least Square)

OLS?

Ref.: On the difference between orthogonal matching pursuit and orthogonal least squares. T. Blumensath and M. E. Davies.

- Supervised forward selection of features. The selection is done class per class. (similar to OMP)

- Principle: given a dictionary of feature $\{\phi_k\}_k$

  - Set $y_i = 1$ if i is in the class, 0 otherwise

    - Find the most correlated feature ϕ_k with y

    - Pull it from the dictionary, orthogonalize the dictionary and **normalize the dictionary**.

    - Select it. Iterate.

DATA

Dataset	Type	Paper	Accuracy
Caltech101	Scattering		79.9
	Unsupervised	Ask the locals	77.3
		RFL	75.3
		M-HMP	82.5
	Supervised	DeepNet	***91.4***
Caltech256	Scattering		43.6
	Unsupervised	Ask the Locals	41.7
		M-HMP	50.7
	Supervised	DeepNet	***70.6***
CIFAR10	Scattering		82.3
	Unsupervised	RFL	83.1
	Supervised	DeepNet	***91.8***
CIFAR100	Scattering		56.8
	Unsupervised	RFL	54.2
	Supervised	DeepNet	***65.4***

Identical Representation

Numerical results

DATA

Method	Caltech101	CIFAR10
Translation Scattering order 1	59.8	72.6
Translation Scattering order 2	70.0	80.3
Translation Scattering order 2+ OLS	75.4	81.6
Roto-translation Scattering order 2	74.5	81.5
Roto-translation Scattering order 2+ OLS	***79.9***	***82.3***

Improvement layer wise

How to fill in the Gap?

- Adding more supervision in the pipeline

- **Building a DeepNet on top** of it? Initialising a DeepNet with wavelets filters? Question is opened.

- A more complex classifier could help to handle class variabilities. **Fisher vector** with scattering?

- Adding a layer means identifying the next complex source of variabilities. **Are they geometric? classes?**

Other application of ST

- Audio → Vincent Lostanlen
- Quantum chemistry → Matthew Hirn
- Temporal data, video
- Reconstruction of WT → Irène Waldspurger
- Unstructured data… → Mia Chen, Xu Cheng

Conclusion

- We provide:
  - Mathematical analysis and algorithms
  - Competitive numerical results
  - Software: http://www.di.ens.fr/~oyallon/ (or send me an email to get the latest!)