SCATTERING CONVOLUTIONAL HIDDEN MARKOV TREE

J.B. REGLI, J. D. B. NELSON

UCL, Department of Statistical Science

ABSTRACT

We here combine the rich, overcomplete signal representation afforded by the scattering transform together with a probabilistic graphical model which captures hierarchical dependencies between coefficients at different layers. The wavelet scattering network component results in a high-dimensional representation which is translation invariant and stable to deformations whilst preserving informative content. Such properties are achieved by cascading wavelet transform convolutions with non-linear modulus and averaging operators. The network structure and its distributions are described using a Hidden Markov Tree. This yields a generative model for high-dimensional inference and offers a means to perform various inference tasks such as prediction. Our proposed scattering convolutional hidden Markov tree displays promising results on classification tasks of complex images in the challenging case where the number of training examples is extremely small.

***Index Terms*—** Scattering network, Hidden Markov Model, Classification, Deep network

1 Introduction

The standard approach to classify high dimensional signals can be expressed as a two step process. First the data are projected in a feature space where the task at hand is simplified. Then prediction is done using a simple predictor in this new representational space. The mapping can either be hand-built —e.g. Fourier transform, wavelet transform— or learned. In the last decade methods for learning the projection have drastically improved under the impulsion of the so called deep learning. Deep neural networks (often enriched by convolutional architecture) have been able to learn very effective representations for a given dataset and a given task [1–3]. Such method have achieved state of the art on many standard problems [4, 5] as well as real world applications [6].

However deep learning methods are only efficient when we have access to a vast quantity of training examples [7]. But in some cases, such as in medical or defence applications for example, datapoints are rare or using an expert for hand-labelling them is time-consuming, costly or subjective. Hence in situations where training examples are expensive to collect, learning has to be performed on smaller datasets. In that case using a fixed, hand crafted set of filters seems to be one of the best solution [8]. Recently Mallat introduced the scattering transform [9]— a fixed bank of wavelet filters used to generate data representation in a convolutional neural networks like architecture. This representational approach was used together with a support vector machine classifier and achieved close to state of the art performance on a number of standard datasets [10]. Moreover, it has been shown that this method performs very well on a relatively smaller numbers of training examples [11] —i.e. around 1000 training samples.

When only a very small number of training samples are available one-shot learning [12] generative classification methods achieve significantly better results than their discriminative models [13], however they require pre-training. Generative probabilistic graphical models have been successfully constructed for various wavelet transforms ; in particular, Hidden Markov trees have been used to model the dependencies between the wavelet coefficients [14–16].

In a similar fashion, we propose to model Mallat's scattering convolutional network [10] using hidden Markov trees. This combines a recently proposed deterministic, analytically tractable transformation inspired by deep convolutional with a probabilistic graphical model. It creates a potentially powerful probabilistic tool to handle high-dimensional prediction problems. Unlike previous work on hidden Markov wavelet trees, the use of scattering transforms allow us to exploit their full range of invariances. However, it also compels us to adapt the HMT model to non-homogeneous, non-regular trees. In contrast to simply passing the raw scattering coefficients into a classifier, our proposed framework captures dependencies between different layers in a generative probabilistic model. Moreover, unlike standard classification, once trained our model can tackle not only prediction problems but also other inference tasks such as generation, sensitivity analysis, etc and can also outperform SVMs when only a very small number of training examples are available.

The remainder of this paper introduces the scattering convolutional hidden Markov tree and is organised as follow. In section 2 we review the scattering transform and some of its properties. Section 3 introduces the proposed Scattering Hidden Markov Tree (SCHMT). In Section 4 we perform several classification experiments on MNIST [17] restricted to only a few training samples. We draw conclusions in Section 5.

J.-B. Regli is funded by a Dstl/UCL Impact studentship. J. D. B. Nelson is partially supported by grants from the Dstl and Innovate UK/EPSRC.

2 Scattering networks

Scattering convolutional networks (SCNs) [18] are Convolutional Neural Networks (CNNs) [3] that use a fixed filter bank of wavelets. Filters can be hand-crafted to yield descriptors with various desired invariances [9]. For image classification tasks, one is interested in descriptors that are at least stable to deformations and invariant to translations. Although SCNs can produce a more complex set of invariances, we here focus attention to deformations and translations only.

2.1 Scattering transform

Wavelets are localized functions stable to deformations and can be adapted to construct descriptors that are translation invariant. A two-dimensional spatial wavelet transform W is obtained by scaling by 2^j and rotating by r_θ a mother wavelet ψ :

$$\psi_\lambda(u) = \psi_{j,\theta}(u) = 2^{-2j}\psi(2^{-j}r_\theta p) \quad (1)$$

In the remainder of this paper we restrict attention to Morlet wavelets defined on $\Lambda = G \times [\![0, J]\!]$ where G is a finite group of rotations of cardinality L and where the wavelet is taken at scale J, namely

$$W_J\mathbf{x} = \{\mathbf{x} * \phi_J(u); \mathbf{x} * \psi_\lambda(u)\}_{p\in\mathbb{R}^2,\lambda\in\Lambda} \quad (2)$$

Whilst the averaging part ϕ_J of the wavelet transform is invariant to translations, the high frequency part ψ_λ is covariant to them [9]. Invariance within a limited range inferior tp 2^J can be achieved by averaging the positive envelope with a smooth window,

$$S_J[\lambda]\mathbf{x}(u) = |\mathbf{x} * \psi_\lambda| * \phi_J(u) \quad (3)$$

Such non-linearised averaged wavelet coefficients are used in various form in computer vision (SIFT [19], DAISY [20]), but the scattering transform proposes a new non-linearity as well as a layer based architecture.

2.2 Scattering convolutional network

While providing local translation invariance, the averaging convolution introduced in (3) also removes the spatial variability of the wavelet transform. SCNs cascade this wavelet modulus operator to recover the lost information and compute progressively more invariant descriptors. The wavelet transform and modulus operations are combined into a single wavelet modulus operator, thus

$$\mathcal{U}_J\mathbf{x} = \{S[\emptyset]\mathbf{x}; U[\lambda]\mathbf{x}\}_{\lambda\in\Lambda_J} = \{\mathbf{x}*\phi_J; |\mathbf{x} * \psi_\lambda|\}_{\lambda\in\Lambda_J}, \quad (4)$$

A scattering transform can be interpreted as a CNN [21] illustrated in Figure 1 which propagates a signal $\mathbf{x}$ across multiple layers of the network and which outputs at each layer m the scattering invariant coefficients $S[p_m]\mathbf{x}$ where $p_m = (\lambda_1 \dots \lambda_m)$ is a path of m orientations and scales.

The scattering energy is mainly concentrated along frequency decreasing paths, i.e. for which $|\lambda_{k+1}| \leq |\lambda_k|$ [9]. The energy contained in the other paths is negligible and thus for

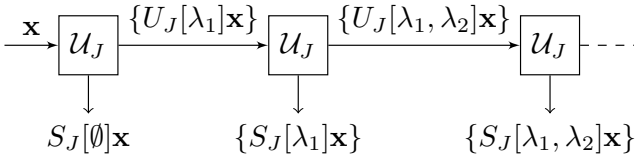

Fig. 1. Scattering networks can be seen as neural networks iterating over wavelet modulus operators $\mathcal{U}_j$. Each layer m outputs the averaged invariants $S[p_m]\mathbf{x}$ and covariant coefficients $U[p_{m+1}]\mathbf{x}$.

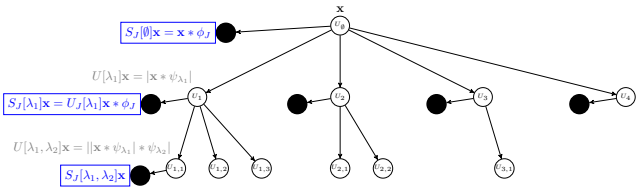

Fig. 2. Frequency decreasing scattering convolution network with $J = 4$, $L = 1$ and $M = 2$. A node i at scale j_i generates $(j_i - 1) \times L$ nodes.

applications only frequency decreasing paths are considered. Moreover there exist a path length $M > 0$ after which all longer paths can be neglected. For signal processing applications, this decay appears to be exponential. And for classification applications, paths of length $M = 3$, i.e. two convolutions, provide the most interesting results [10, 22].

This restrictions yield a more convenient parametrization of a scattering network. Indeed its now completely defined by the mother wavelet ϕ, the maximum path length considered M, the finest scale level considered J, and the number of orientation considered L.

Hence for a given set of parameters (ψ, M, J, L), let $ST_{(\psi,M,J,L)}(\mathbf{x})$ denote the unique frequency decreasing windowed scattering convolutional network with those parameters evaluated for signal $\mathbf{x}$. Each node i of this network generates a, possibly empty, set of of nodes of size $(j_i - 1) \times L$ where j_i is the scale of node i and L is the number of orientations considered and it has the architecture displayed by Figure 2.

2.3 Scattering convolutional classifier

In the original framework [18], the scattering network $ST_{(\psi,M,J,L)}(.)$ is used for a classification task using a SVM classifier on the outputs of the network. Performance can be slightly improved by adding a feature selection step to perform PCA on the scattering coefficients and keep only the most informative ones. This classification framework provides results comparable with the state of the art on several datasets [10].

3 The Scattering hidden Markov tree

State of the art performance can be achieved using SCNs associated with SVMs. However this approach is not adapted

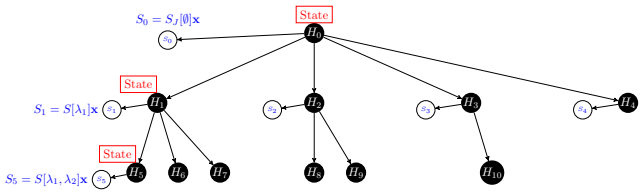

Fig. 3. Scattering convolutional hidden Markov tree.

to very small training sets ; nor does it deliver a probabilistic output. To overcome these limitations we propose an adaptation of the Crouse [14] and Durand [16] wavelet hidden Markov trees to the non-regular, non-homogeneous tree structure of SCNs.

3.1 Hidden Markov tree model

The HMT models the marginal distribution of each real ST coefficient S_i as a Gaussian mixture. To each S_i, we associate a discrete hidden state H_i that takes on values in $[\![1, K]\!]$ with probability mass function (PMF) $P(H_i)$. Conditioned on $H_i = k$, S_i is Gaussian with mean $\mu_{i,k}$ and variance $\sigma_{i,k}$. Thus, its overall marginal PDF is given by $P(w_i) = \sum_{k=1}^{K} P(H_i = m)P(S_i|H_i = k)$ with $P(S_i|H_i = k) \sim \mathcal{N}(\mu_{i,k}, \sigma_{i,k})$. While each scattering coefficient S_i is conditionally Gaussian given its state variable H_i, overall it has a non-Gaussian density. Finally the probability for the hidden node H_i to be in a state k given its father's state g is characterized by a transition probability such that $\epsilon_i^{(gk)} = P(H_i = k|H_{\rho(i)} = g)$. This yields $P(H_i = k) = \sum_{g=1}^{K} \epsilon_i^{(gk)} P(H_{\rho(i)} = g)$.

Such a model is pictured in Figure 3 and for a given scattering architecture —i.e. fixed M, J and L— the SCHMT model is fully parametrized by,

$$\Theta = \left(\pi_0, \{\epsilon_i, \{\theta_{k,i}\}_{k\in[\![1,K]\!]}\}_{i\in\mathcal{T}}\right). \tag{5}$$

This model implies two assumptions on the scattering transform. We firstly posit K-populations ; i.e. that the scattering coefficients of a given signal can be described by K clusters. This is a common assumption for standard wavelets [15] and it can hence be extended to the scattering transform. The SCHMT also assumes persistence in that the informative character of a coefficient is propagated across layers. This assumption is justified by the fact that scattering coefficients are highly correlated [21].

3.2 Learning the tree parameters

The proposed SCHMT is trained using the smoothed version of the Expectation-Maximization (EM) algorithm for hidden Markov trees proposed by [23] and adapted to non-homogeneous and non-binary trees.

Let $\bar{\mathcal{S}}_i = \bar{s}_i$ be the observed sub-tree rooted at node i. By convention $\bar{\mathcal{S}}_0$ denotes the entire observed tree. The smoothed version of the E-step requires the computation of the conditional probability distributions $\xi_i(k) = P(H_i = k|\bar{\mathcal{S}}_i = \bar{s}_i)$ (smoothed probability) and $P(H_i = k, H_{\rho(i)} = g|\bar{\mathcal{S}}_i = \bar{s}_i)$ for each node $i \in \mathcal{T}$ and states k and g. This can be achieved through an upward-downward recursion displayed in Algorithm 1 and 2. The output from the downward step are used in the M-step as shown in Algorithm 3.

// Initialization :
for *All the nodes i of the tree $\mathcal{T}$* **do**
  $P_{\theta_{k,i}}(s_i) = \mathcal{N}(s_i|\mu_{k,i}, \sigma_{k,i})$
end
for *All the leaves i of the tree $\mathcal{T}$* **do**
  $\beta_i(k) = \frac{P_{\theta_{k,i}}(s_i)P(H_i=k)}{\sum_{g=1}^{K} P_{\theta_{g,i}}(s_i)P(H_i=g)}$
  $\beta_{i,\rho(i)}(k) = \sum_{g=1}^{K} \frac{\beta_i(g)\epsilon_i^{(kg)}}{P(H_i=g)}.P(H_{\rho(i)} = k)$
  $l_i = 0$
end
// Induction :
for *All non-leaf nodes i of the tree $\mathcal{T}$ (Bottom-up)* **do**
  $M_i = \sum_{k=1}^{K} P_{\theta_{k,i}}(s_i) \prod_{j\in c(i)} \frac{\beta_{j,i}(k)}{P(H_i=k)^{n_i-1}}$
  $l_i = \log(M_i) + \sum_{j\in c(i)} l_j$
  $\beta_i(k) = \frac{P_{\theta_{k,i}(s_i)} \prod_{j\in c(i)}(\beta_{j,i}(k))}{P(H_i=k)^{n_i-1}M_i}$
  for *All the children nodes j of node i* **do**
    $\beta_{i\setminus c(i)}(k) = \frac{\beta_i(k)}{\beta_{i,j}(k)}$
  end
  $\beta_{i,\rho(i)}(k) = \sum_{g=1}^{K} \frac{\beta_i(g)\epsilon_i^{(kg)}}{P(H_i=g)}.P(H_{\rho(i)} = k)$
end

Algorithm 1: Smoothed upward algorithm.

// Initialization :
$\alpha_0(k) = 1$
// Induction :
for *All nodes i of the tree $\mathcal{T}\setminus\{0\}$ (Top-Down)* **do**
  $\alpha_i(k) = \frac{1}{P(H_i=k)} \sum_{g=1}^{K} \alpha_{\rho(i)}(g)\epsilon_i^{(gk)}\beta_{\rho(i)\setminus i}(g)P(H_{\rho(i)} = g)$
end

Algorithm 2: Smoothed downward algorithm.

// Initialization :
$\pi_0(k) = \frac{1}{N}\sum_{n=1}^{N} P(H_0^n = m|s_0^n, \Theta^l)$
// Induction :
for *All nodes i of the tree $\mathcal{T}\setminus\{0\}$* **do**
  $P(H_i = k) = \frac{1}{N}\sum_{n=1}^{N} P(H_i^n = k|\bar{s}_0^n, \Theta^l)$,
  $\epsilon_i^{gk} = \frac{\sum_{n=1}^{N} P(H_i^n=k, H_{\rho(i)}^n=g|\bar{s}_0^n,\Theta^l)}{NP(H_{\rho(i)}=k)}$,
  $\mu_{k,i} = \frac{\sum_{n=1}^{N} s_i^n P(H_i^n=k|\bar{s}_0^n,\Theta^l)}{NP(H_i=k)}$,
  $\sigma_{k,i}^2 = \frac{\sum_{n=1}^{N}(s_i^n-\mu_{k,i})^2 P(H_i^n=k|\bar{s}_0^n,\Theta^l)}{NP(H_i=k)}$.
end

Algorithm 3: M-step of the EM algorithm.

3.3 MAP classification

Let Θ_c now be a set of parameters for an SCHMT $\mathcal{T}$ learned on a training set $\{\bar{S}_{0,c}^n\}_{n\in[\![1,N]\!]} = \{ST_{(\psi,J,M,L)}(\mathbf{x}_c^n)\}_{n\in[\![1,N]\!]}$ composed of the scattering representations of N realizations of a signal of class c . Let also $\mathbf{x}^{new}$ be another realization of this signal, not used for training and $\mathcal{T}^{new}$ be the instance of the SCHMT generated by this realization.

In this context the MAP algorithm [23] aims to find the optimal hidden tree $\hat{h}_0^{new} = (\hat{h}_0^{new}, \ldots, \hat{h}_{I-1}^{new})$ that maximizes the probability of this sequence given the model's parameters $P(\bar{\mathcal{H}}_0 = \hat{h}_0^{new}|\mathcal{T}^{new}, \Theta_c)$. The MAP framework also provides $\hat{P}$ the value of this maximum.

1 vs All class		MNIST	
		SCHMT	SCN+SVM
9	(Acc)	$\mathbf{97.2}\%$	90.0%
	(PTR)	$\mathbf{3.8}$	0
6	(Acc)	$\mathbf{94.1}\%$	90.0%
	(PTR)	$\mathbf{30.7}$	0
Mean	(Acc)	$\mathbf{93.9}\%$	90.0%
	(PTR)	$\mathbf{75.5}$	0

Fig. 4. Accuracy and positive likelihood ratio —the higher the better— on 1000 samples of MNIST trained using only a 5 of training points per class and tested on "1 vs All".

4 Experiments

We compare the performance of SCHMT to those of a SCN combined with an SVM (SCN+SVM) restricted to a small number of training examples by performing two experiments on the digit classification dataset MNIST [17]. It contains 60000 training and 10000 test images of 10 handwritten digits (zero to nine), with 28×28 pixels. For both experiments we use a scattering transform with $M = 3$ orders, $J = 3$ scales, $L = 3$ orientations and a Morlet mother wavelet. The hidden Markov tree has $K = 2$ states and uses a mixture of Gaussian to describe the relationship between the scattering coefficients and the hidden states. For the SVM, the best parameters are selected by cross-validation.

4.1 MNIST - "1 vs All"

In a similar fashion to [24], we first test SCHMT on a "one vs all" basis. However we propose this experiment with a more challenging setup. Indeed they pretrain their model with 100 samples from each class but one. They then provide only limited amount of training examples, say N, for this last class. Instead we propose a framework where all the classes have the same limited amount of training points N. We then test the models on 1000 examples.

Table 4 displays the accuracy and positive likelihood ratio (PLR) of belonging to one class versus all the others for both SCHMTs and SCN+SVM. With this amount of training example, SVM is not able to discriminate the digit of interest and simply classify everything as "All". This yields a PLR of 0 and an uninformative test. SCHMT, however, is able to correctly discriminate the digit of interest and provide a very informative test — high PLR.

4.2 MNIST - Full

SCHMT and SCN+SVM are tested on the more complex problem of mutliclass labelling. SCHMT and SCN+SVM are both trained on a limited number of training examples per class and tested on 20 test samples per class and on the full testset.

The best results for each models are displayed in Table 5. SCHMT displays better generalization and prediction properties than SCN+SVM when trained on a very limited number of training points. For $N = 2$ training points per class, SVM hardly beats random —i.e. 10%— while SCHMT provides a three folds improvement. With only 5 training examples per class, SCHMT does about five times better than a random selector. As expected, when the number of training samples grow large enough —i.e. 10 and more— for the SVMs, SCN+SVM reaches better maximum classification score.

Training samples per class	MNIST	
	SCHMT	SCN+SVM
$N = 2$	$\mathbf{28.6}\%$	18.7%
$N = 5$	$\mathbf{48.0}\%$	43.2%
$N = 10$	45.2%	$\mathbf{49.9}\%$

Fig. 5. Classification score on the complete testset of MNIST (10000 samples) trained using only a limited number N of training points per class

The drop in performance of SCHMT for $N = 10$ training examples is explained by the fact that the EM algorithm performances are undermined by convergence to local minima issues [25] yielding sometimes to poor learning quality for SCHMTs. However when a good minima is found, SCHMTs has acceptable generalization performance.

While confirming the superiority of a our model in terms of generalizability for limited number of training points, this experiment also highlights a potential weakness of it in that sometimes convergence problems occur. However, in the main, SCHMT provides good classification score for such a low number of training examples.

5 Conclusion

A SCHMT framework has been proposed which comprises a scattering transform and a hidden Markov tree model. The scattering transform projects the data into a representational space of even higher dimensionality but of reduced volume along the invariants in the data. Then a probabilistic graphical model —hidden Markov tree— was used to fit a generative model to the distribution of the representation of the data. As such, the proposed model takes advantage of the way in which the scattering transform introduces invariances into the representation but also the manner in which hidden Markov models capture dependencies between coefficients. Experiments have demonstrated that the modelled distribution can be used to perform efficient classification tasks even with small training sizes. Even though we only here consider classification, a generative model is much more versatile than a simple —yet efficient— discriminative one. Because they model the full distribution of the data they can express more complex relationships between the observed and the unknown variables than simple discrimination.

To enhance SCHMT and especially the chance of converging toward a good minima during the EM learning, further work will include development of variational methodology learn the model parameters [26].

6 References

[1] R. Salakhutdinov and G.E. Hinton, "Deep boltzmann machines," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 448–455.

[2] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.A. Manzagol, "Stacked denoising autoencoders : Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[3] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, 1995.

[4] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv :1207.0580*, 2012.

[6] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, R. Cheng-Yue, F. Mujica, A. Coates, et al., "An empirical evaluation of deep learning on highway driving," *arXiv preprint arXiv :1504.01716*, 2015.

[7] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[8] S.T. Hsiang and J.W. Woods, "Embedded video coding using invertible motion compensated 3-d subband/wavelet filter bank," *Signal Processing : Image Communication*, vol. 16, no. 8, pp. 705–724, 2001.

[9] S. Mallat, "Group invariant scattering," *Communications in Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, Oct. 2012.

[10] J. Bruna and S. Mallat, "Classification with scattering operators," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1561–1566.

[11] L. Sifre and S. Mallat, "Rotation, scaling and deformation invariant scattering for texture discrimination," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. 2013, pp. 1233–1240, IEEE.

[12] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 594–611, 2006.

[13] A. Jordan, "On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, pp. 841, 2002.

[14] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *Signal Processing, IEEE Transactions on*, vol. 46, no. 4, pp. 886–902, 1998.

[15] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and computational harmonic analysis*, vol. 10, no. 3, pp. 234–253, 2001.

[16] J.B. Durand, P. Goncalves, and Y. Guédon, "Computational methods for hidden markov tree models-an application to wavelet trees," *Signal Processing, IEEE Transactions on*, vol. 52, no. 9, pp. 2551–2560, 2004.

[17] Y. LeCun, "Personal webpage : Mnist," `http://yann.lecun.com/exdb/mnist/`, 2016, Accessed : 18-01-2016.

[18] J. Bruna, *Scattering representations for recognition*, Ph.D. thesis, Ecole Polytechnique X, 2013.

[19] M. Grabner, H. Grabner, and H. Bischof, "Fast approximated sift," in *Computer Vision–ACCV 2006*, pp. 918–927. Springer, 2006.

[20] S. Winder, G. Hua, and M. Brown, "Picking the best daisy," in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*. 2009, pp. 178–185, IEEE.

[21] E. Oyallon and S. Mallat, "Deep roto-translation scattering for object classification," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. 2015, pp. 2865–2873, IEEE.

[22] J. Andén and S. Mallat, "Multiscale scattering for audio classification.," in *ISMIR*, 2011, pp. 657–662.

[23] J.B. Durand and P. Gonçalves, *Statistical inference for hidden Markov tree models and application to wavelet trees*, Ph.D. thesis, INRIA, 2001.

[24] R. Salakhutdinov, J. Tenenbaum, and A. Torralba, "One-shot learning with a hierarchical nonparametric bayesian model," 2010.

[25] T.K. Moon, "The expectation-maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[26] M.J. Wainwright and M.I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.