KERNELIZED LEARNING IN DEEP SCATTERING CONVOLUTION NETWORKS

Yuehan Xiong, Can Xu, Hongkai Xiong

Department of Electronic Engineering, Shanghai Jiao Tong University, China
{xiongyuehan, jamiexucan, xionghongkai}@sjtu.edu.cn

ABSTRACT

This paper addresses the problem of automatic scattering feature selection for signal classification. While features derived from group invariant scattering networks are quite effective for signal classification. We argue that scattering networks are not always the appropriate choice as they are not learned for the objective at hand. In this paper, we explore jointly learning a deep scattering convolution network with a support vector machine by casting the problem as a multiple kernel learning problem. The convolution paths of the network are kernelized respectively to be selected in a large-margin context. We deduce scattering paths from the corresponding kernels after solving the kernel learning problem. Experiments on several datasets demonstrate the effectiveness of the proposed method over state-of-the-art techniques.

Index Terms— wavelet filter, multiple kernel learning, scattering transform, digit analysis, texture recognition

1. INTRODUCTION

Many signal classification problems are solved by feature extraction and regression. The ability for features to characterize similarities within a class and discrepancies between classes is called feature discrimination, which obviously affects the classification accuracy. Thus, data-driven learning of feature extractors has been investigated in a variety of contexts.

In [1], an evolution strategy was evolved to optimize a wavelet packet-based feature representation for signal classification. In [2], a filter bank is learned in union with a hidden Markov models-based classifier through an evolutionary algorithm. [3] explored a joint learning of a filter bank layer and a deep neural network. [4] proposed to jointly learn a combination of wavelet coefficients and a classifier in a kernelized large-margin context. More recently, Sangnier et al. recommended to jointly learn the filters of a filter bank and a Support Vector Machine (SVM) by casting the problem as a multiple kernel learning problem [5][6]. However, these works can be understood to optimize a single layer of filter bank.

Recently, the scattering convolution network has been a promising approach for feature extraction of signal classification, as shown in audio classification [7], image textures [8] and object recognition [9]. It is a local descriptor which attains high-order statistical information by cascading wavelet convolution with non-linear modulus and averaging operators. It is thus translation invariant and linearizes small deformations. However, the number of convolution paths and the corresponding wavelet filters are fixed given the scale and rotation levels. The canonical scattering representations may not be appropriate as it is not tuned to the task at hand.

This paper proposes a new approach to jointly learn a deep wavelet scattering network and a Support Vector Machine classifier, which aims at finding an optimal cascade of wavelet convolution layers, instead of a single convolution layer. It ensures that the scattering network is learned for the underlying task. By associating each path coefficient resulting from a parameterized scattering transform to a kernel, the problem can be formulated as a Multiple Kernel Learning (MKL) problem with infinitely many kernels (Infinite Kernel Learning, IKL) [10], which can be solved by active constraint methods. The combined kernel is the product of base kernels which corresponds to the tensor product of their feature spaces [11]. It is worth mentioning that the multiple kernel learning hereinafter is not a linear combination of base kernels which focuses on the concatenation of individual kernel feature spaces. This Generalized MKL (GMKL) leads to a much higher-dimensional feature representation in comparison to feature concatenation. In turn, the wavelet convolution paths can be deduced from the selected kernels. The general framework of proposed method is illustrated in Fig. 1.

The remainder of the paper is organized as follows. Section II presents the overall framework and details of proposed method. Section III provides experimental results on various datasets. Finally, Section IV concludes the paper with some future challenges.

2. PROPOSED METHOD

2.1. Formulation

Suppose that we have a training set of instance-label pairs $\{(x_i, y_i)\}_{i=1}^{N}$ consisting of C classes, $x_i \in R^n$ is the ith train-

The work was supported in part by the NSFC, under grants 61425011, U1201255, 61271218, 61529101, and Shu Guang project (13SG13).

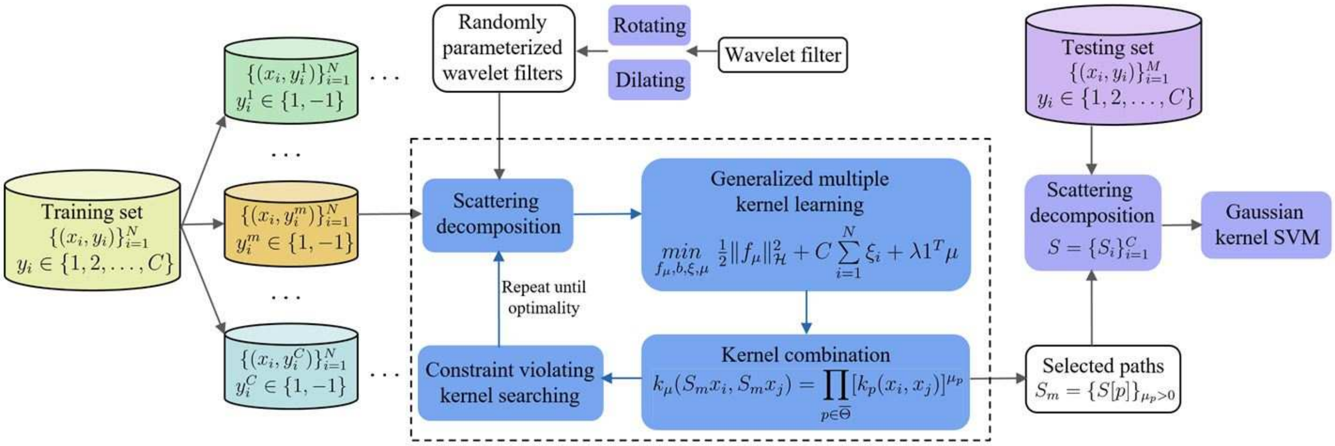
Randomly parameterized wavelet filters
Rotating
Dilating
Wavelet filter
$\{(x_i, y_i^1)\}_{i=1}^N$ $y_i^1 \in \{1, -1\}$
$\{(x_i, y_i^m)\}_{i=1}^N$ $y_i^m \in \{1, -1\}$
$\{(x_i, y_i^C)\}_{i=1}^N$ $y_i^C \in \{1, -1\}$
Training set $\{(x_i, y_i)\}_{i=1}^N$ $y_i \in \{1, 2, \ldots, C\}$
Scattering decomposition
Generalized multiple kernel learning
$\min_{f_\mu, b, \xi, \mu} \frac{1}{2}\|f_\mu\|_{\mathcal{H}}^2 + C\sum_{i=1}^N \xi_i + \lambda 1^T \mu$
Repeat until optimality
Constraint violating kernel searching
Kernel combination
$k_\mu(S_m x_i, S_m x_j) = \prod_{p \in \overline{\Theta}} [k_p(x_i, x_j)]^{\mu_p}$
Testing set $\{(x_i, y_i)\}_{i=1}^M$ $y_i \in \{1, 2, \ldots, C\}$
Scattering decomposition $S = \{S_i\}_{i=1}^C$
Gaussian kernel SVM
Selected paths $S_m = \{S[p]\}_{\mu_p > 0}$

Fig. 1. The proposed scattering network learning framework.

ing instance and $y_i \in \{1, 2, \ldots, C\}$ is its label. Our goal is to find a group of scattering operators $\{S_1, S_2, \ldots, S_C\}$ as feature extractors for an SVM, where S_j is learned by promoting maximum margin between the jth class and other classes in a binary classifier training process. This can be achieved by jointly optimizing the parameters of each scattering transform S_j and an SVM classifier, which can be cast as an MKL problem in which kernel weights are learned jointly with model parameters. According to GMKL, the problem can be formed as:

$$\begin{aligned} \min_{f_\mu, b, \xi, \mu} \quad & \frac{1}{2}\|f_\mu\|_{\mathcal{H}}^2 + C\sum_{i=1}^N \xi_i + \eta \mathbb{1}^T \mu \\ s.t. \quad & y_i f_\mu(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \\ & \mu \succeq 0 \end{aligned} \tag{1}$$

where η is the coefficient of the sparsity penalty for kernel weight vector μ, $\mathcal{H}$ corresponds to the feature space that implicitly constructs the combined kernel function $k_\mu(\cdot, \cdot)$. The decision function $\hat{f}_\mu$ is:

$$\hat{f}_\mu(x) = \sum_{i=1}^N \hat{\alpha}_i y_i k_\mu(x, x_i) + \hat{b} \tag{2}$$

2.2. Parameterized scattering transform

Scattering transform computes cascades of wavelet transforms and nonlinear modulus operators. Let ψ be a complex directional wavelet whose imaginary and real parts are orthogonal, ψ_λ is computed by rotating and dilating ψ respectively by θ and 2^j:

$$\psi_\lambda(u) = 2^{-2j}\psi(2^{-j}\theta^{-1}u) \tag{3}$$

where $\lambda = 2^{-j}\theta$, $u \in \mathbb{R}^2$. The wavelet transform of an input signal $x \in \mathbb{R}^2$ is $x \star \psi_\lambda(u)$. To maintain the good localization property of wavelets and construct invariability, the modulus nonlinearity and spatial averaging are introduced:

$$U[\lambda]x = |x \star \psi_\lambda| \tag{4}$$

$$S[\lambda]x = U[\lambda]x \star \phi_{2^J} \tag{5}$$

where the spatial window is $\phi_{2^J}(u) = 2^{-2J}\phi(2^{-J}u)$.

Define a frequency-decreasing path $p = (\lambda_1, \lambda_2, ..., \lambda_m)$, $|\lambda_i| < |\lambda_{i+1}|$, $i = 1, 2, ..., m-1$. The corresponding scattering propagator is:

$$\begin{aligned} U[p]x &= U[\lambda_m] \cdots U[\lambda_2]U[\lambda_1]x \\ &= |||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| \cdots | * \psi_{\lambda_m}| \end{aligned} \tag{6}$$

The output of the path p is computed by performing localized integration:

$$S[p]x = U[p]x \star \phi_{2^J} \tag{7}$$

Averaging by the scaled spatial window ϕ_{2^J} guarantees translation invariance in a range smaller than 2^J. For each path p, the scattering coefficients $S[p]x$ are down-sampled at intervals proportional to 2^J. As depicted in Figure 2, scattering transform decomposes signals into finer scales and different orientations iteratively. The paths of the mth layer correspond to the set $\mathcal{P}^m$ of all paths $p = (\lambda_1, \lambda_2, ..., \lambda_m)$ of length m. Typically, Scattering transforms decompose signals to the second layer, i.e., $m \in \{0, 1, 2\}$, which is the case we consider in this work. Concatenating coefficients of each path gives the scattering feature representation. Considering the strong correlation between scattering paths and their different impact on classification, this work focuses on learning a data-dependent wavelet scattering network within a large-margin setting.

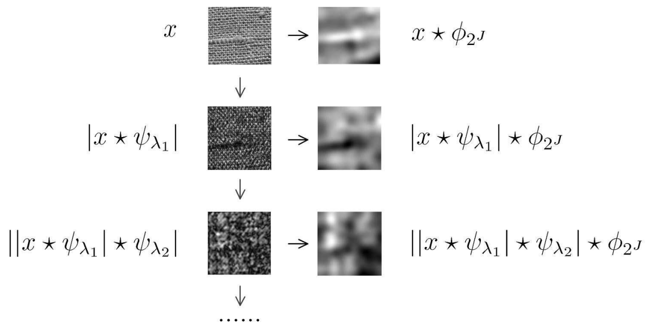
x
$x \star \phi_{2^J}$
$|x \star \psi_{\lambda_1}|$
$|x \star \psi_{\lambda_1}| \star \phi_{2^J}$
$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$
$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J}$
......

Fig. 2. A scattering transform is a cascade of wavelet decompositions.

2.3. Kernel extraction

The proposed method automatically selects optimal wavelet scattering paths for a given classification problem. This can be achieved by associating each path of the scattering transform with a base kernel and employing an MKL approach. We use Morlet wavelets to scatter signal information into multiple paths:

$$\psi(x,y) = \frac{\zeta^2}{2\pi\sigma_\psi^2} e^{-\frac{x'^2+\zeta^2 y'^2}{2\sigma_\psi^2}} (e^{i\xi x'} - \beta) \tag{8}$$

Let $\theta \in [0,\pi]$ denote the orientation of the filter, $x' = xcos\theta + ysin\theta$, $y' = -xsin\theta + ycos\theta$. $\beta \ll 1$ ensures that $\int \psi(x,y)dxdy = 0$. ζ and $\sigma_\psi = 2^j * \sigma_{\psi_0}$ are the eccentricity and the standard deviation of the elliptic envelope, respectively.

The spatial window ϕ is defined as:

$$\phi(x,y) = \frac{\zeta^2}{2\pi\sigma_\phi^2} e^{-\frac{x'^2+y'^2}{2\sigma_\phi^2}} \tag{9}$$

where $\sigma_\phi = 2^{J-1} * \sigma_{\phi_0}$ is the standard deviation of the Gaussian envelope.

Given a training set for binary classification which is $\{(x_i,y_i)\}_{i=1}^N$, $x_i \in R^n$, $y_i \in \{1,-1\}$. Instead of predefining dilation and rotation parameters of wavelets as $(j,\theta) \in \{0,1,\dots,J-1\} \times \{0, \frac{1}{L}\pi, \dots, \frac{L-1}{L}\pi\}$ as in a standard scattering transform setting, we define the continuous set of all possible wavelet parameter vectors (j,θ) as $\mathcal{P} = [0,J-1) \times [0,\pi)$. In this case, for a wavelet scattering net of M layers, the set of path parameters $\Theta = \bigcup_{i=1}^M \mathcal{P}^i$ is infinite. We randomly extract a finite set $\Theta_0 \subset \Theta$, $|\Theta_0| = d$ for initialization of scattering paths. This is achieved by first generating a group of wavelets defined in (8) where dilation and orientation parameters $\{j_i\}_{i=1}^J$ and $\{\theta_i\}_{i=1}^L$ are randomly sampled from $[0, J-1)$ and $[0,\pi)$ and sorted in ascending order, respectively. Then the finite set of paths are derived by constructing a randomly parameterized scattering network with these wavelets in a frequency-decreasing way.

We consider Gaussian kernels in this paper : $k(\mathbb{R}^c, \mathbb{R}^c) \to \mathbb{R}$. For any scattering path $p = (\lambda_1, \dots, \lambda_i, \dots, \lambda_m)$, $\lambda_i = 2^{-j_i}\theta_i$, the corresponding kernel can be generated as following:

$$\begin{aligned} k_p &= k(S[p]x_i, S[p]x_j) \\ &= exp(-\frac{\| S[p]x_i - S[p]x_j \|_F^2}{2\sigma^2}) \end{aligned} \tag{10}$$

where σ is the bandwidth of kernels. The Gaussian kernel expresses the similarity between two training instances mapped in an infinite-dimensional space. The finite set of symmetric positive definite kernels $\{k_p\}_{p\in\Theta_0}$ are called spanning kernels. The multiple kernel k_μ is the weighted product of spanning kernels in GMKL setting:

$$\begin{aligned} &\forall (x_i,x_j), \quad x_i,x_j \in R^n \,, \\ &k_\mu(Sx_i, Sx_j) \\ &= k([\sqrt{\mu_1}S[p_1]x_i, \dots, \sqrt{\mu_d}S[p_d]x_i], \\ &\qquad [\sqrt{\mu_1}S[p_1]x_j, \dots, \sqrt{\mu_d}S[p_d]x_j]) \\ &= \prod_{m=1}^d k(\sqrt{\mu_m}S[p_m]x_i, \sqrt{\mu_m}S[p_m]x_j) \qquad (11) \\ &= \prod_{m=1}^d k(S[p_m]x_i, S[p_m]x_j)^{\mu_m} \\ &= \prod_{p\in\Theta_0} (k_p)^{\mu_p} \end{aligned}$$

where $\mu \succeq 0$ is the kernel weight vector. As the weighted product of similarities between each scattering path p, the combined kernel k_μ measures the proximity between scattering representations of two training instances.

2.4. Generalized IKL for scattering kernel selection

We employ the Generalized IKL algorithm to learn a combination of scattering kernels. The GIKL approach is a variant of the IKL algorithm that can handle MKL problems with a large number of kernels. As shown in equation (11), for combination of multiple base kernels, GIKL uses a weighted product instead of a convex combination of kernels, which enables the usage of much richer feature representations compared to the original IKL. In order to handle an endless number of kernls, GIKL applies an active constraint approach [4] which starts from a guess on the active kernel set called spanning kernels, then uses an optimality condition to iteratively add kernels to this set until optimality. This active set principle defines active kernels as kernels with positive weights, other kernels have no impact on the solution as their weights are null.

Reformulate equation 1 as the following optimization

Algorithm 1: Scattering kernel learning algorithm

```
Input: Training set $(x_i, y_i)_{1\leq i\leq N}$
Output: Scattering path set $\Theta_t$ and classifier $\hat{f}$
$(k_p)_{p\in\Theta_0} \leftarrow$ initialize kernels with randomly parameterized scattering paths;
$\mu \leftarrow \frac{1}{|\Theta_0|}\mathbb{1}$;
$t \leftarrow 0$;
while not suboptimal do
    $\mu \leftarrow$ GMKL solution with $\Theta_t$;
    $\Theta_t \leftarrow \{p \in \Theta_t, \mu_p > 0\}$;
    $\overline{\Theta} \leftarrow$ random sample from $\Theta$;
    $\hat{p} \leftarrow \underset{p\in\overline{\Theta}}{argmax}\, T(p)$;
    if $T(\hat{p}) > 0$ then
        $\Theta_{t+1} = \Theta_t \bigcup \{\hat{p}\}$;
        $t = t + 1$;
    else
        Suboptimality reached;
```

problem:

$$\underset{\mu}{min} \quad \omega(\mu) = \begin{cases} \underset{f_\mu, b, \xi}{min} & \frac{1}{2}\|f_\mu\|^2_{\mathcal{H}} + C\sum\limits_{i=1}^{N}\xi_i + \eta\mathbb{1}^T\mu \\ s.t. & y_i f_\mu(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{cases}$$
$$s.t. \quad \mu \in \mathbb{R}^{\Theta}_{+} \tag{12}$$

where η balances the sparsity of μ and the SVM structural risk. The problem is non-convex so the point μ is suboptimal when $\omega(\mu)$ achieves a local minimum, then the KKT suboptimality conditions are satisfied:

$$\exists\gamma \in \mathbb{R}^d_+ \ , \ \nabla\omega(\mu) + \eta\mathbb{1} - \gamma = 0 \tag{13}$$
$$s.t. \quad \mu \succeq 0\,,\, \gamma_i\mu_i = 0$$

Then, in any case,

$$\nabla\omega(\mu) + \eta\mathbb{1} \succeq 0 \tag{14}$$

The combined kernel matrix can be formulated as $K = exp(-\sum_{p\in\Theta}\mu_p D_p/2\sigma^2)$ according to (11), where $(D_p)_{p\in\Theta}$ are the distance matrices in the scattering domain. The suboptimality condition can then be formulated with the optimal dual variable $\hat{\alpha}$ of equation (1):

$$\begin{aligned} & T \leq 0 \ \ with \\ & \forall p \in \Theta\,, \\ & T(p) = -\frac{\partial\omega}{\partial\mu_p}(\mu) - \eta \\ & \quad = -\frac{1}{4\sigma^2}\hat{\alpha}^T Y(D_p \circ K)Y\hat{\alpha} - \eta \end{aligned} \tag{15}$$

where $Y = diag(y)$, $\circ$ is the Hadamard product. The problem boils down to solving the following non-convex subproblem:

$$\hat{p} = \underset{p\in\Theta}{argmax}\, T(p) \tag{16}$$

To solve this problem, as is shown in Algorithm 1, we start with a few parameters $\Theta_0 \subset \Theta$, and iterates between GMKL algorithm and the search for the constraint violating kernel. A random search $\overline{\Theta} \subset \Theta$ is applied at each iteration to find a violating kernel $k_{\hat{p}}$. Given the number $n_{\hat{p}}$ of it among $\overline{\Theta}$, since it can be computed as:

$$n_{\hat{p}} = \begin{cases} g_1(j_{\hat{p}}, \theta_{\hat{p}}) & m_{\hat{p}} = 1 \\ g_2(j_{\hat{p}}, \theta_{\hat{p}}) & m_{\hat{p}} = 2 \end{cases} \tag{17}$$

$$\begin{aligned} g_1(j_{\hat{p}}, \theta_{\hat{p}}) &= (j^1_{\hat{p}} - 1)L + \theta^1_{\hat{p}} \\ g_2(j_{\hat{p}}, \theta_{\hat{p}}) &= \frac{1}{2}L^2(2J - j^1_{\hat{p}} - 2)(j^1_{\hat{p}} - 1) + (j^2_{\hat{p}} - j^1_{\hat{p}} - 1)L \\ &\quad + (\theta^1_{\hat{p}} - 1)(J - j^1_{\hat{p}})L + \theta^2_{\hat{p}} + JL \end{aligned}$$

where $m_{\hat{p}}$ denotes the layer of scattering path $\hat{p}$. Let $l \in \{1, \ldots, m_{\hat{p}}\}$ denote the layer of a wavelet in $\hat{p}$, $j^l_{\hat{p}}$ and $\theta^l_{\hat{p}}$ are the positions of wavelet paramters among sorted dilation and orientation factors $\{j_i\}^J_{i=1}$ and $\{\theta_i\}^L_{i=1}$, respectively. Thus the corresponding scattering path parameters $(j_{\hat{p}}, \theta_{\hat{p}}) \in \mathcal{P}^{m_{\hat{p}}}$ can be easily deduced from $n_{\hat{p}}$.

At each iteration of the Algorithm 1, non-active kernels are removed after solving the GMKL subproblem (1), then a new kernel is added with a null weight. These new weights form a feasible point of the new GMKL subproblem with an objective equals to the optimal objective of the previous GMKL problem. As the new kernel violates the KKT conditions, the current point is not optimal. The objective can then be improved by solving the GMKL problem again. The collection of path coefficients gives the scattering output $Sx = \{S[p]x\}_{p\in\Theta_t}$ once suboptimality is reached.

3. EXPERIMENTAL ANALYSIS

In this section, we present our experimental analysis on three datasets, i.e., MNIST digits, KTH-TIPS texture, and CIFAR-10 images. We apply a logarithm non-linearity to scattering coefficients in training and testing process to seperate multiplicative low-frequency components owing to illumination variations. For all the experiments, we learn translation invariant scattering networks computed to the second order, and initialize Morlet wavelets by randomly sampling in the parameter set for each iteration. One-versus-all Gaussian kernel SVMs are used for classification of data-dependent scattering features. As we will see, our algorithm learns scattering kernels properly for each classification task.

3.1. Digit Recognition

For hand-written digit recognition, we apply our method on the large scale MNIST database. MNIST is a data basis of

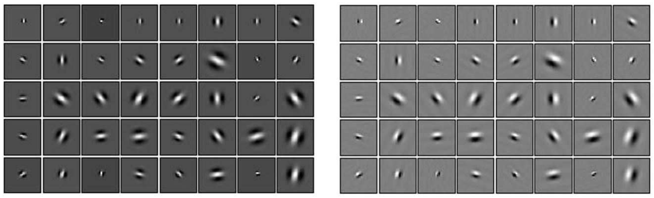

Fig. 3. Quadrature phase complex Morlet wavelets learned by promoting maximum margin between digit 4 and all other digits. Their real and imaginary parts are displayed respectively on the left and on the right. Each pair of adjacent wavelets are of the first and the second layer of a convolution path, respectively.

hand-written digits with at most 6×10^4 training samples and 1×10^4 test samples. The digits are size-normalized and centred in fixed-size images. We randomly select 250 images among each class from the training samples for scattering network learning process and initialize randomly parameterized wavelets with $J = 4$ dilations, $L = 4$ orientations. The classification result is averaged over 10 runs.

Table 1. Classification accuracy (%) on MNIST Digits, with different algorithms.

Method	Accuracy
Trans scattering [12]	99.54
Haar scattering [13]	99.41
ConvNet [14]	99.47
Proposed	**99.56**

Morlet wavelets learned while promoting maximum margin between training samples of digit 4 and others are shown in Figure 3. All scattering paths selected are of the second layer, which suggests that path coefficients of deeper layers carry more discriminative information. During our experiments, we notice that out method selects only frequency-decreasing paths if we also include other paths. Results of several algorithms without preprocessing and distortion are given in Table 1, which shows that our method outperforms the state of the art.

3.2. Texture Analysis

For texture analysis, we use the KTH-TIPS texture dataset for evaluation. The database is composed of 10 classes each with 81 grey scale images of size 200×200. Controlled scaling, shearing and illumination changes exist within each class. We initialize wavelets with dilation and orientation parameters $J = 5$, $L = 4$ for each iteration and use the mean of each path as the output coefficient. SVM and MKL parameters C, η and σ are tuned according to a five-fold cross validation with resamplings among the training set.

Table 2. Classification accuracy on (KTH-TIPS, 2004) database, obtained by different algorithms. Columns correspond to different training sizes for each class.

Training size	5	20	40
BIF [15]	-	-	98.5
SRP [16]	-	-	**99.3**
COX [17]	80.2±2.2	92.4±1.1	95.7±0.5
Trans scat [12]	69.1±3.5	94.8±1.3	98.0±0.8
Roto-trans scat [8]	69.5±3.6	94.9±1.4	98.3±0.9
Proposed	**80.5**±**3.4**	**95.3**±**1.2**	98.5±0.7

Table 2 gives the mean classification rate and standard deviation over 200 random splits of training and testing sets for different training sizes. ”Trans scat” corresponds to a translation invariant scattering as in [12]. ”Roto-trans scat” denotes a joint translation and rotation invariant scattering in [8]. As we can see, with a flexible model capturing more discriminative information, our approach improves to 80.5% for training size $T = 5$ compared to the standard translation invariant scattering. Due to the lack of orientation changes within each class in the database, our method has an advantage compared to state of the art algorithms when the training size is small, and achieves performance close to the best results for larger training sizes.

3.3. CIFAR-10 Images

CIFAR-10 color images contains at most 5×10^4 training samples and 1×10^4 testing samples of 32×32 pixels, which are much more complex than MNIST digits. It is composed of 10 classes such as ”airplanes”, ”birds”, ”ships”. The 3 color bands are represented with Y, U, V channels and we learn scattering networks independently in each channel. Setting $J = 4$, $L = 4$, we again learn scattering transforms with different training sizes.

Table 3. Classification accuracy (%) comparison on CIFAR-10 images using scattering network learned with GMKL algorithm, first-order data-dependent scattering network, and proposed second-order scattering network.

Algorithm	GMKL	First-order	Proposed
T = 500	56.87	52.68	**60.82**
T = 1000	65.50	56.24	**68.72**
T = 2000	69.13	60.43	**73.17**

To evaluate the quality of our learned scattering network, we compare with ”GMKL” and ”First-order”. In which

”GMKL” denotes learning a second-order scattering network with a finite number of kernels using GMKL algorithm, and ”First-order” denotes learning a first-order scattering transform where $J = 4$, $L = 8$ with GIKL algorithm. Table 3 lists the mean recognition accuracy for each method over 10 runs where T denotes the training size. The results demonstrate that our data-dependent second-order scattering transform outperforms the other two significantly for all training sizes, implying that second-order scattering coefficients carry more discriminative information compared to first-order ones, and that the active kernel approach succeeds in improving the discrimination of the multiple kernel by handling a large number of base kernels.

4. CONCLUSION

In this paper, we propose a novel approach to learn a translation invariant scattering network jointly with an SVM by casting the problem as an MKL problem. By choosing a specified family of wavelet filters which can be defined by a few parameters, The wavelet scattering paths can be deduced from the learned sparse multiple kernels. One advantage of our approach is that the number of paths and the corresponding wavelets are automatically learned from training data instead of predetermined. Numerical experiments on three datasets, including handwritten digits, images textures, and tiny color images show promising results of the proposed algorithm. For future works, we will investigate learning wavelet convolution networks with other types of classifiers, e.g., graphical models, for better discriminative analysis.

5. REFERENCES

[1] L. D. Vignolo, D. H. Milone, H. L. Rufiner, et al., “Parallel implementation for wavelet dictionary optimization applied to pattern recognition,” in *Proceedings of the 7th Argentine Symposium on Computing Technology*. 2006.

[2] L. D. Vignolo, H. L. Rufiner, D. H. Milone, et al., “Evolutionary splines for cepstral filterbank optimization in phoneme classification,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 8, 2011.

[3] T. N. Sainath, B. Kingsbury, A. Mohamed, et al., “Learning filter banks within a deep neural network framework,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 297–302. 2013.

[4] F. Yger and A. Rakotomamonjy, “Wavelet kernel learning,” *Pattern Recognition*, vol. 44, no. 10, pp. 2614–2629, 2011.

[5] M. Sangnier, J. Gauthier, and A. Rakotomamonjy, “Filter bank kernel learning for nonstationary signal classification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3183–3187. 2013.

[6] M. Sangnier, J. Gauthier, and A. Rakotomamonjy, “Filter bank learning for signal classification,” *Signal Processing*, vol. 113, pp. 124–137, 2015.

[7] J. Andén and S. Mallat, “Multiscale scattering for audio classification,” in *Proceedings of the ISMIR 2011 conference*, pp. 657–662. 2011.

[8] L. Sifre and S. Mallat, “Rotation, scaling and deformation invariant scattering for texture discrimination,” in *Computer Vision and Pattern Recognition*, vol. II, pp. 1233–1240. 2013.

[9] E. Oyallon and S. Mallat, “Deep roto-translation scattering for object classification,” in *Computer Vision and Pattern Recognition*. 2015.

[10] P. Gehler and S. Nowozin, “Infinite kernel learning,” in *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*. 2008.

[11] M. Varma and B. R. Babu, “More generality in efficient multiple kernel learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1065–1072. 2009.

[12] J. Bruna and S. Mallat, “Invariant scattering convolution networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[13] X. Chen, X. Cheng, and S. Mallat, “Unsupervised deep haar scattering on graphs,” in *Advances in Neural Information Processing Systems*, pp. 1709–1717. 2014.

[14] Y. LeCun, K. Kavukcuoglu, and C. Farabet, “Convolutional networks and applications in vision,” in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 253–256. 2010.

[15] M. Crosier and L. D. Griffin, “Texture classification with a dictionary of basic image features,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7. 2008.

[16] L. Liu, P. Fieguth, G. Kuang, et al., “Sorted random projections for robust texture classification,” in *IEEE International Conference on Computer Vision*, pp. 391–398. 2011.

[17] H. Nguyen, R. Fablet, and J. Boucher, “Visual textures as realizations of multivariate log-gaussian cox processes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2945–2952. 2011.