DUAL-TREE WAVELET SCATTERING NETWORK WITH PARAMETRIC LOG TRANSFORMATION FOR OBJECT CLASSIFICATION

Amarjot Singh and Nick Kingsbury

Signal Processing Group, Department of Engineering, University of Cambridge, U.K.

ABSTRACT

We introduce a ScatterNet that uses a parametric log transformation with Dual-Tree complex wavelets to extract translation invariant representations from a multi-resolution image. The parametric transformation aids the OLS pruning algorithm by converting the skewed distributions into relatively mean-symmetric distributions while the Dual-Tree wavelets improve the computational efficiency of the network. The proposed network is shown to outperform Mallat's Scatter-Net [1] on two image datasets, both for classification accuracy and computational efficiency. The advantages of the proposed network over other supervised and some unsupervised methods are also presented using experiments performed on different training dataset sizes.

Index Terms— DTCWT, Scattering network, Convolutional neural network, Orthogonal least squares, CIFAR.

1. INTRODUCTION

Object classification is a difficult problem due to the translation, rotation and scale variability of objects within the images as well as external variabilities such as noise and illumination. Hand-engineered features such as SIFT [2] and HOG [3] modeled the geometric properties of the objects to achieve decent classification accuracy. However, these features have been recently replaced by trained networks [4], [5], [6], especially, Convolutional Neural Networks (CNNs) [6] that have achieved state-of-the-art accuracy by learning invariant and discriminative class-specific image representations. Despite the success of CNNs, design and optimal configuration of these networks is not well understood which makes it difficult to develop these networks.

Mallat [7], [8], [9], [1] has shown that ScatterNets incorporate geometric knowledge of images to produce discriminative and invariant (translation and rotation) representations which can give performance comparable to that of trained networks. The invariants at the first layer of the network are obtained by filtering the image with multi-scale and multi-directional complex Morlet wavelets followed by a point-wise nonlinearity and local smoothing. The high frequencies lost due to smoothing are recovered at the later layers using cascaded wavelet transformations, justifying the need for a multilayer network. A log transformation may be applied to de-correlate the multiplicative low-frequency components from the concatenated invariants obtained at all layers [1]. Next, orthogonal least squares (OLS) selects the subset of object class-specific dimensions across the training data, similar to that of the fully connected layers in CNNs [1].The presence of outliers in the extracted features or unwanted features extracted from the background clutter, noise, and illumination can hinder feature selection due to their effect on the least squares parameter estimates. Hence, it is important to introduce approximate symmetry in the extracted features to suppress the effect of these outliers.

We propose an improved computationally efficient ScatterNet that extracts relatively symmetric translation invariant representations from a multi-resolution image using the *dual-tree complex wavelet transform* (DTCWT) [10] and the proposed parametric log transformation layer. Here, we only introduce translation invariance, as the orientation of an object in the image plane is often well-known as a strong prior (e.g. side-view images). The OLS layer next selects a subset of object specific components without undesired bias from outliers due to the introduced symmetry. The selected features are finally used by a Gaussian-kernel support vector machine (G-SVM) to perform object classification on CIFAR-10 and CIFAR-100 datasets.

The contributions of the paper are as follows:

- *Multi-resolution Input Image*: The input image is transformed into multi-resolution images of 2 or more different sizes such that the dual-tree wavelet decompositions produce more densely spaced feature maps over scale. These allow the OLS algorithm to learn additional discriminatory features which can aid the classification.

- *Parametric Log transformation*: Log transformation reduces the effect of outliers by introducing approximate symmetry in representations with parameters learnt from the data. The transformation also de-correlates the multiplicative low-frequency components (illumination) while simultaneously creating a form of contrast normalization which enhances weaker features.

- *Computational Efficiency*: Dual-tree wavelets are used as opposed to Morlet [7] because of their discrete form, short support, perfect reconstruction, and limited redundancy [10]. They provide similar rich features to
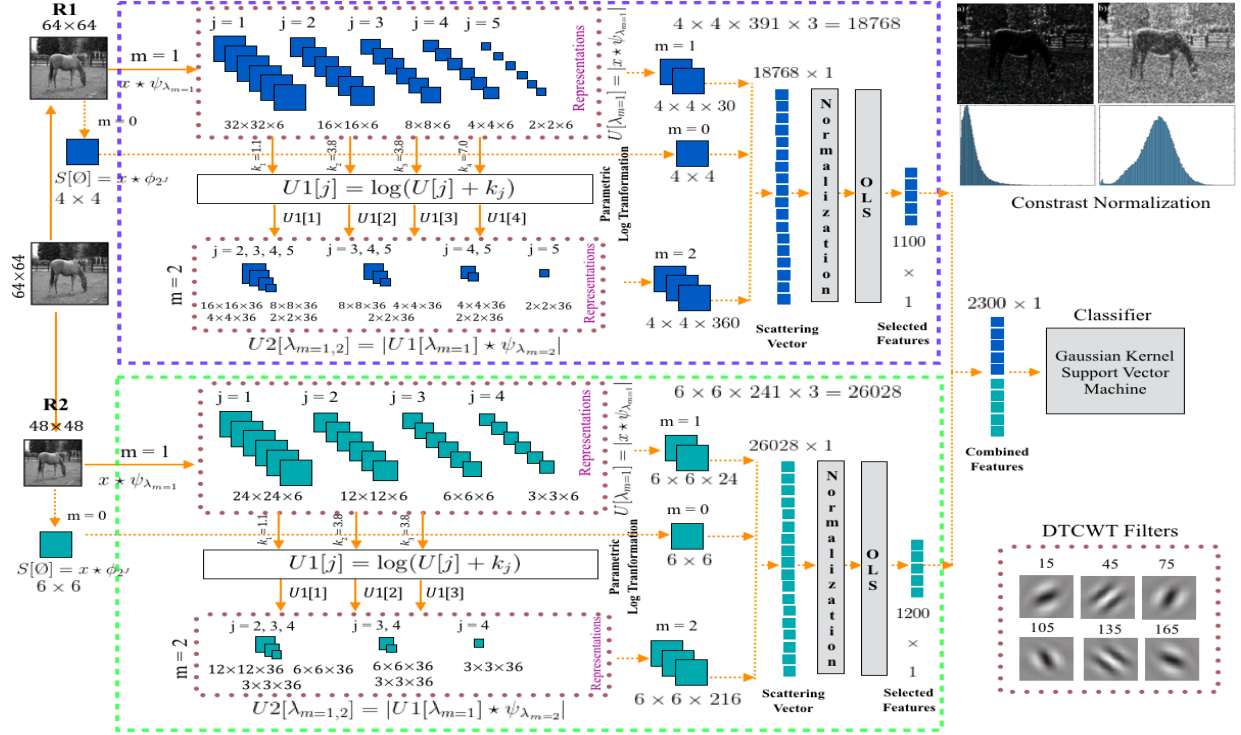
Fig. 1. Illustration shows the input image (x) of size 64×64 resized to images of resolution, R1 (64×64 (x)) and R2 (48×48 (x_1)) respectively. Image representations at m = 1 are obtained using DTCWT filters at 5 scales for R1, 4 scales for R2 and 6 orientations ($x \star \psi_{\lambda_{m=1}}$). Next, L2 non-linearity (complex modulus) is applied on the representations to obtain the regular envelope $|x \star \psi_{\lambda_{m=1}}|$. Log transformation $U1[j] = \log(U[j] + k_j)$ with parameters k_j is applied on the envelope for all scales j except the coarsest scale. Next, local smoothing is applied to extract the translation invariant coefficients $U1[\lambda_{m=1}] \star \phi_{2^J}$. The information lost due to smoothing are recovered by cascaded wavelet filtering at the second layer $|U1[\lambda_{m=1}] \star \psi_{\lambda_{m=2}}|$. Translation invariance is introduced in the recovered frequencies using L2 non-linearity and local smoothing $U2[\lambda_{m=1}, \lambda_{m=2}] \star \phi_{2^J}$. The contrast normalization effect of the parametric log transformation is shown in the top right while the DTCWT filters at six fixed orientations are shown in the bottom.

Morlet wavelets but with less computation and somewhat lower redundancy in the output vectors. In addition, dual-tree wavelets can be efficiently implemented in the spatial domain, rather than requiring the complexities and constraints of Fourier domain filtering.

The proposed network improves on Mallat's ScatterNet on classification accuracy and computational efficiency on two datasets. Multiple experiments on different training dataset sizes are performed to highlight the advantages of the proposed network against supervised and unsupervised methods.

The paper is divided into the following sections. Section 2 briefly presents our proposed DTCWT scattering network with parametric log transformation. Section 3 presents the experimental results while Section 4 draws conclusions.

2. DTCWT SCATTERNET

The proposed ScatterNet uses dual-tree wavelets to decompose the multi-resolution input image into multi-scale and multi-directional representations at multiple layers, because

the DT-CWT is lossless and has high computational efficiency in the spatial domain. The parametric log transformation is applied to the outputs of the first scatternet layer to introduce relative symmetry to the distributions of coefficient magnitudes and thus aid OLS feature selection. Subsequent scatternet layers then apply local smoothing and bandpass filtering with wavelet-modulus operations to gradually increase translation invariance while preserving information about the higher frequency components of the image [7]. Below we present the formulation of the proposed ScatterNet for a single input image which may then be applied to each of the multi-resolution images.

An input image x is filtered using dual-tree complex wavelets $x \star \psi_{\lambda_1}$ where $\lambda_1 = (j, r)$. At the first layer, the real and imaginary parts of the complex coefficients are combined from real filters in the dual tree using:

$$x \star \psi_{\lambda_1} = x \star \psi_{\lambda_1}^a + \iota x \star \psi_{\lambda_1}^b \tag{1}$$

where ψ^a is the real and ψ^b the imaginary part of the wavelet. The six orientations (r) in the transform are pre-defined to be: $15°, 45°, 75°, 105°, 135°$ and $165°$.

The wavelet filtering signal commutes with translations, and is therefore not translation invariant. To build a more translation invariant representation, a point-wise L_2 non-linearity is applied to the filtered signal, as described below:

$$U[\lambda_{m=1}] = |x \star \psi_{\lambda_1}| = \sqrt{|x \star \psi_{\lambda_1}^a|^2 + |x \star \psi_{\lambda_1}^b|^2} \quad (2)$$

This step produces the regular envelope of the filtered signal and reduces the redundancy of each representation to 2:1. L_2 is a good non-linearity as it is stable to deformations and additive noise [7]. However, the representations may also contain outliers that can hinder the performance of the orthogonal least squares based feature selection layer (explained in Section. 1). Hence, the parametric log transformation layer is applied to all the oriented representations ($U[j]$) extracted at a particular scale j with a parameter k_j, to reduce the effect of outliers by introducing relative symmetry, as shown below:

$$U1[j] = \log(U[j] + k_j), \quad U[j] = |x \star \psi_j|, \quad (3)$$

Good symmetry is achieved for the distribution of oriented representations obtained by selecting the parameter k_j that minimizes the difference between the mean and median of the distribution. The parametric log transformation also decorrelates the low-frequency multiplicative components arising due to illumination variation and noise [1] as well as *normalizing the contrast* of the representations by elevating the weak features and suppressing the stronger ones as shown top right corner in Fig. 1.

Next, a local average is computed on the envelope $|U1[\lambda_{m=1}]|$ that aggregates the coefficients to build the desired translation-invariant representation:

$$S[\lambda_{m=1}] = |U1[\lambda_{m=1}]| \star \phi_{2^J} \quad (4)$$

The high frequency components lost due to smoothing are retrieved by cascaded wavelet filtering performed at the second layer. The retrieved components are again not translation invariant. Translation invariance is achieved by first applying the L2 non-linearity of eq(2) to obtain the regular envelope:

$$U2[\lambda_{m=1}, \lambda_{m=2}] = |U1[\lambda_{m=1}] \star \psi_{\lambda_{m=2}}| \quad (5)$$

A local-smoothing operator is then applied to the regular envelope ($U2[\lambda_{m=1}, \lambda_{m=2}]$) to extract the desired second layer ($m = 2$) translation invariant coefficients:

$$S[\lambda_{m=1}, \lambda_{m=2}] = U2[\lambda_{m=1}, \lambda_{m=2}] \star \phi_{2^J} \quad (6)$$

The scattering coefficients obtained at each layer are:

$$S = \begin{pmatrix} x \star \phi_{2^J} \\ U1[\lambda_{m=1}] \star \phi_{2^J} \\ U2[\lambda_{m=1}, \lambda_{m=2}] \star \phi_{2^J} \end{pmatrix}_{j=(2,3,4,5...)} \quad (7)$$

The coefficients extracted from each layer are concatenated to generate a feature vector for each of the images in the training dataset as shown in Fig. 1. The scattering feature vectors are then normalized across each dimension and given as input to the feature selection layer.

The feature selection layer is implemented using a supervised orthogonal least square (OLS) regression [11] that greedily selects discriminative features specific to class C with a one-versus-all linear regression using the following indicator function:

$$f_C(x) = \begin{cases} 1 & \text{if x belongs to class C} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The regression is applied to a training set of scattering feature vectors where each vector of N dimensions is reduced to N' selected dimensions ($N' << N$) that belong to a specific class C. Let $(\Phi_t^{M \times N})_C$ be the dictionary at the tth iteration for a specific class C. The tth feature x is selected such that the linear regression of $f_C(x)$ has a minimum mean-squared error, computed on the training set corresponding to class C. The reduced training feature dataset is given as input to the G-SVM that learns the weights that best discriminate the classes in the dataset. Feature selection makes training and applying a classifier more efficient due to the decreased vector size. It also tends to improve performance by eliminating unnecessary components of the input and their associated noise.

3. OVERVIEW OF RESULTS

The performance of the proposed network is evaluated on CIFAR-10 and CIFAR-100 datasets with 10 and 100 classes respectively. Each dataset contains a total of 50000 training and 10000 test images of size 32×32 equally divided between the classes. The evaluation is performed on the classification accuracy, computational efficiency and feature richness. A comparison with Mallat's ScatterNet [1], unsupervised [12], [4] and supervised methods [6] is also performed.

In order to extract the scattering representations, every 32×32 image is first upsampled into two images of resolution 64×64 (R1) and 48×48 (R2). The upsampled image is then transformed into two images of resolution 64×64 (R1) and 48×48 (R2). R1 and R2 are decomposed using DTCWT filters with 6 fixed orientations at 5 and 4 scales respectively, followed by L2 non-linearity, as shown in Fig. 1. Next, the log transformation is applied to the representations (except the ones obtained at the coarsest scale) obtained from both R1 and R2 pipeline with parameters $k_1 = 1.1$, $k_2 = 3.8$, $k_3 = 3.8$ and $k_4 = 7$ chosen for scale j = 1, 2, 3 and 4 respectively. An smoothing operator is then applied to introduce translation invariance in the representations. The classification accuracy for representations obtained at various scales (J), with and without the use of parametric log transformation and the concatenated coefficients at m=1 with G-SVM, are shown for both R1 and R2 pipelines in Table 1. The G-SVM parameter (c) is selected as 14 while gamma parameter is set to 0.00002 using 5-fold cross validation on the training feature set. We see that the parametric log transformation results in a small improvement in classification accuracy.

The information lost due to smoothing at the first layer is retrieved at the next layer using cascaded filtering as shown in Fig. 1. The retrieved information is made translation invariant by local smoothing. Representations for the three color channels at m = 0, 1, 2 are concatenated to produce a 18768 (6256×3) dimensional vector for R1 image and a vector of length 26028 (8676×3) for R2 as shown in Fig. 1. OLS is then applied on the training dataset (50000×18768) to select 108 dimensions per class resulting into a total of 1080 discriminative dimensions for every R1 image (50000×1080). Similarly, 1200 dimensions per image are chosen for the R2 image. This reduced feature dataset results in a classification accuracy of 81.6% (80.7% without log transformation) for R1 images while an accuracy of 81.8% (80.9% without log transformation) is recorded for R2 images, using the above-mentioned SVM for the CIFAR-10 datasets as shown in Table. 1. A classification accuracy of 82.4% is obtained by concatenating the selected dimensions of R1 and R2. A decrease in classification accuracy is recorded on selecting more than the above-mentioned feature dimensions.

Table 1. Accuracy (%) on CIFAR-10 for both R1 and R2 for each scale (J) and coefficients at m = 1, with and without applying log transformation. The accuracy for features selected from the final scattering vector at $m_{1,2}$ using OLS is presented in the last column.

	$J=1$	$J=2$	$J=3$	$J=4$	m_1	$m_{1,2}$
R1: No-log	62.7	66.9	69.0	70.2	70.4	80.7
R1: log	**65.6**	**69.9**	**71.5**	**72.4**	**72.5**	**81.6**
R2: No-log	65.9	70.0	71.2	–	71.7	80.9
R2: log	**68.0**	**71.5**	**72.6**	–	**73.4**	**81.8**

Next, scattering coefficients extracted using DTCWT ScatterNet with the above-mentioned parameters result in a classification accuracy of 56.7% for the CIFAR-100 dataset, as shown in Table. 2. The translation invariant coefficients extracted using the proposed network outperform the translation as well as Roto-translation invariant features of Mallat's ScatterNet [1], on both datasets. The network also outperformed state-of-the-art unsupervised methods [12], [4] but underperformed by nearly 10% against supervised deep learning models [6], as shown in Table. 2.

Table 2. Accuracy (%) and comparison on both datasets. Pro.: Proposed, Sup: Supervised and Unsup: Unsupervised, learning.

Dataset	Pro.	ScatNet [1]	Unsup	Sup
CIFAR-10	**82.4**	81.6	82.2 [12]	89.6 [6]
CIFAR-100	**56.7**	55.8	54.2 [4]	64.3 [6]

The proposed network can be an attractive choice over Mallat's ScatterNet due to its computational efficiency and gain in classification accuracy. The proposed network extracts the coefficients from both R1 and R2 images in almost three-quarters (0.78 (s)) of the time needed by Mallat's network (0.98 (s)) to decompose only the R1 image, as shown in Table. 3. This marginal difference is significant for large image datasets such as CIFAR. In addition, since the scattering vector produced by the proposed network is smaller (44796)

as compared to Mallat's network (113712) (three-layer network) [1], the OLS layer can select the desired feature dimensions (1080 and 1200) in almost 3/4 of the time (2.21(h) vs 3.22(h)). The selected dimensions with OLS from the scattering vector are more for the proposed network (1080 + 1200 = 2300) as compared to Mallat's network (1080). This suggests that the features extracted by the proposed network are significantly richer in information as compared to Mallat's network as feature richness is defined as the number of dimensions selected with OLS divided by the total feature dimensions. The simulations are computed on a server with 32 Gb RAM per node in uniform conditions.

Table 3. Arc.: Architectures, Pro.: Proposed, R1, R2: Resolution - 1,2 pipeline, FVL: Feature vector length, SD: Selected dimensions using OLS, FR: Feature richness (%), TS (s): Scattering time an image in seconds, T-OLS: Feature selection time using OLS in hours.

Arch.	FVL	SD	FR (%)	TS (s)	T-OLS (h)
ScatNet [1]	113712	2000	1.75	0.98	3.22
R1	18762	1100	**5.86**	**0.46**	**1.07**
R2	26028	1200	**4.61**	**0.32**	**1.14**
Pro. (R1+R2)	44796	2300	**5.13**	**0.78**	**2.21**

However, supervised models require large training datasets to learn which may not exist for most application. Table. 4 shows that DTCWT ScatterNet outperformed LeNet [5] and Network in Network (NIN) [6] supervised learning networks on the CIFAR-10 datasets with less than 10k images. The experiments were performed by dividing the training dataset of 50000 images into 8 datasets of different sizes. The images for each dataset are obtained by randomly selecting the required number of images from the full 50000 training dataset. It is made sure that an equal number of images per object class are sampled from the training dataset. The full test set of 10000 images is used for all the experiments. Deeper models like NIN [6] result in low classification accuracy due to their inability to train on the small training dataset.

Table 4. Comparison of Proposed (Pro.) network on accuracy (%) with two supervised learning methods (LeNet [5] and NIN: Network in Network [6]) against different training dataset sizes on CIFAR-10.

Arch.	300	500	1K	2K	5K	10K	20K	50K
Pro.	**39.3**	**48.8**	**55.9**	**61.8**	**67.0**	**72.9**	76.8	82.4
LN	34.9	44.7	53.1	57.9	63.0	69.0	74.0	77.6
NIN	10.1	10.3	10.9	40.4	63.4	72.0	**83.1**	**89.6**

4. CONCLUSION

The paper proposes an improved version of Mallat's ScatterNet using dual-tree wavelets and parametric log non-linearity. The DTCWT ScatterNet gives enhanced performance on classification accuracy and computational efficiency as compared to Mallat's ScatterNet on two datasets. The network has also shown to outperform unsupervised learning methods while evidence of the advantage of DTCWT ScatterNet over supervised learning (CNNs) methods is presented for applications with small training datasets.

5. REFERENCES

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

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[4] Y. Jia, C. Huang, and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[5] C. Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply- supervised nets," *arXiv preprint*, 2014.

[6] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv:1312.4400*, 2013.

[7] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1872 –1886, 2013.

[8] L. Sifre and S. Mallat, "Rotation, scaling and deformation invariant scattering for texture discrimination," *IEEE conference on Computer Vision and Pattern Recognition*, , no. 1233 - 1240, 2013.

[9] L. Sifre and S. Mallat, "Rigid-motion scattering for texture classification," *arXiv preprint*, 2014.

[10] N.G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Applied and computational harmonic analysis*, vol. 10, pp. 234–253, 2001.

[11] T. Blumensath and M. E. Davies, "On the difference between orthogonal matching pursuit and orthogonal least squares," 2007.

[12] K. Sohn and H. Lee, "Learning invariant representations with local transformations," *International Conference on Machine Learning*, 2012.