BIOLOGICALLY-INSPIRED OBJECT RECOGNITION SYSTEM WITH FEATURES FROM COMPLEX WAVELETS

Tao Hong, Nick Kingsbury, Michael D. Furman

Signal Processing and Communications Laboratory
Dept. of Engineering, University of Cambridge, UK
{th315,ngk10,mdf32}@cam.ac.uk

ABSTRACT

In this paper, a novel cortex-inspired feed-forward hierarchical object recognition system based on complex wavelets is proposed and tested. Complex wavelets contain three key properties for object representation: shift invariance, which enables the extraction of stable local features; good directional selectivity, which simplifies the determination of image orientations; and limited redundancy, which allows for efficient signal analysis using the multi-resolution decomposition offered by complex wavelets. In this paper, we propose a complete cortex-inspired object recognition system based on complex wavelets. We find that the implementation of the HMAX model for object recognition in [1, 2] is rather over-complete and includes too much redundant information and processing. We have optimized the structure of the model to make it more efficient. Specifically, we have used the Caltech5 standard dataset to compare with Serre's model in [2] (which employs Gabor filter bands). Results demonstrate that the complex wavelet model achieves a speed improvement of about 4 times over the Serre model and gives comparable recognition performance.

Index Terms— Complex Wavelets, Visual Cortex, Object Recognition, Visual Hierarchial Model

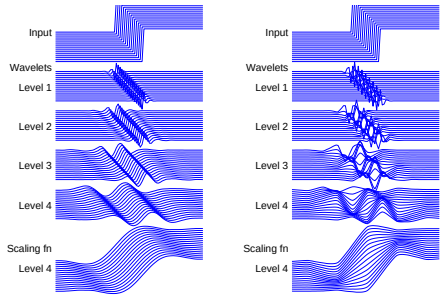
1. INTRODUCTION

Much early work in computer vision was inspired by the pioneering biologically based studies of David Marr [3]. More recently after many researchers have studied the field, the scale-invariant feature transform (SIFT) [4] was developed and shown to be suitable for recognition of specific objects (i.e. different views of the same object). However it still exhibits some problems for performing generic object recognition (i.e. recognizing different objects of the same general class). At the same time wavelet transforms [5] have been developed, and extended more recently to the dual-tree complex wavelet transform [6]. The quantitative model of the visual cortex system is based on Hubel and Wiesel's work [7] on models of simple and complex cells. Recently, M. Riesenhuber and T. Poggio developed the HMAX model [1] to explain how the visual processing in cortex could work.

The discrete wavelet transform (DWT) has the ability to localize functions in space, scale and orientation, and for natural images produces sparse representations. Since Mallat [8] first demonstrated wavelets as the foundation of multi-resolution theory for signal processing and analysis in 1987, the DWT has been widely and successfully used in many areas of image processing, e.g. denoising, enhancement, deconvolution and compression. For object recognition, however, the DWT has a critical shortcoming, its lack of shift invariance, which becomes apparent when we observe that the distribution of energy between coefficients at different scales varies sharply with shifts in the input signal.

In 1998-9 the dual-tree complex wavelet transform (DT CWT) [6] was introduced. It overcomes the disadvantages of the DWT by using Hilbert-pairs of wavelets to introduce limited redundancy and makes it possible to extract stable local features efficiently.

Fig.1 shows this in detail. At the top of the figure, the input signal is a unit step which is shifted to get 16 adjacent sampling instances. The following rows of the figure show the output signal components (after an inverse DT CWT or DWT) when wavelet and scaling function coefficients at levels 1 to 4 are retained, just one level at a time. From the right column (b), we see that the DWT is far from shift invariant, making it very difficult to capture signatures of the signal under shift, and hence the recognition of signals with wavelet signatures is not robust or stable. In the left column (a), the DT CWT produces an almost shift-invariant set of responses, which make it possible to extract stable local features efficiently.



(a) DT CWT type B (b) Real DWT

Fig. 1. Wavelet and scaling function components at levels 1 to 4 of 16 shifted step responses of the DT CWT (a) and real DWT (b).

There are two key aspects for computer vision algorithms: accuracy (recognition performance) and computational efficiency. In this paper, we evaluate a biologically-inspired object recognition system from the viewpoints of both signal processing and computer vision. A feedforward cortex-inspired object-recognition system, based on complex wavelets, is proposed and developed. Test results using the public Caltech5 dataset show that our system is significantly faster than Serre's Gabor-based system [2] while providing comparable recognition performance.

There are two specific innovations in this paper: the implementation of the early stages of the visual cortex with complex wavelets (DT CWT); and the demonstration that the Gabor-based model is too over-complete. We have optimized the front end of the model by reducing its existing redundancies. The complex wavelets make this

possible.

This paper is arranged as follows: in section 2 we describe the implementation of the system in detail; in section 3 we show how we tested the systems on the Caltech5 dataset; and in section 4 we discuss the results.

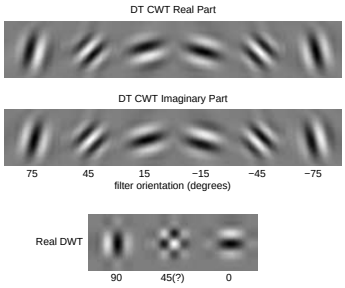


Fig. 2. DT CWT and DWT filter impulse responses.

2. IMPLEMENTATION

2.1. Model based on Complex Wavelets

A (1D) discrete wavelet transform can be described as the vector inner product of an input signal $x(n\tau)$ with $\psi^*_{a,b}(n\tau)$ for a range of shifts a and scales b [5], where $\psi^*_{a,b}$ is the complex conjugate of the wavelet $\psi_{a,b}$, related to the mother wavelet $\psi(x)$ by

$$\psi_{a,b}(n\tau) = \frac{1}{\sqrt{b}}\psi(\frac{n\tau - a}{b}) \quad (1)$$

In the normal DWT, ψ is real, whereas in the DT CWT it is complex and its real and imaginary parts are computed by two separate DWT filter trees (hence the name 'dual tree'). This 2:1 redundancy largely eliminates aliasing, leading to the translation-invariant results shown in fig. 1a. The DT CWT has already been used in several areas of object recognition [9, 10].

We have built a feedforward hierarchical model based on complex wavelets, which can be divided into four processing layers, S1, C1, S2, and C2, as in [2]. In order to compare our model with the Gabor-bases model of [2], we retain as much similarity with Serre's system as is feasible, including his style of diagrams.

S1 Layer:

The first stage of the hierarchical model is achieved by filtering the original input image with the DT CWT. (The classic simple cells in the cortex model V1 [7] correspond to this layer.) Note that in fig. 2 the DT CWT has 6 orientations ($15°$, $45°$, $75°$, $105°$, $135°$ and $165°$). We let $q^l_s(x, y)$ be the response of a simple cell in the first layer S1, where (x, y) is its position, s is its scale and l is its direction; the responses $q^l_s(x, y)$ can be computed by convolving the original image I with the DT CWT subband responses, W^l_s from fig. 2, as in equation (2). Here we take the magnitude of the complex coefficients to help to achieve translation invariance:

$$q^l_s(x, y) = |W^l_s * I| \quad (2)$$

Since the standard DT CWT has only 1 scale per octave and the scale difference between octaves is 2:1, intermediate scales may be generated by resizing the original image by the desired scale change followed by a DT CWT to filter the resized images. By interleaving the filtered images, the desired S1 layer responses $q^l_s(x, y)$ may be achieved. This includes octaves and the intermediate scales within each octave. We find that 2 scales per octave are needed, so the original image I_1 (1024×768 pixels, say) is resized into a second image, I_2, which is scaled by $2^{-\frac{1}{2}}$. I_2 is now 724×544 pixels. Fig.4 shows the S1 layer responses $q^l_s(x, y)$ for this. If there are 4 levels (octaves) of the DT CWT with 2 scales per octave, then the S1 layer has a 4D structure having the same 3D pyramid shape for each of the 6 directions.

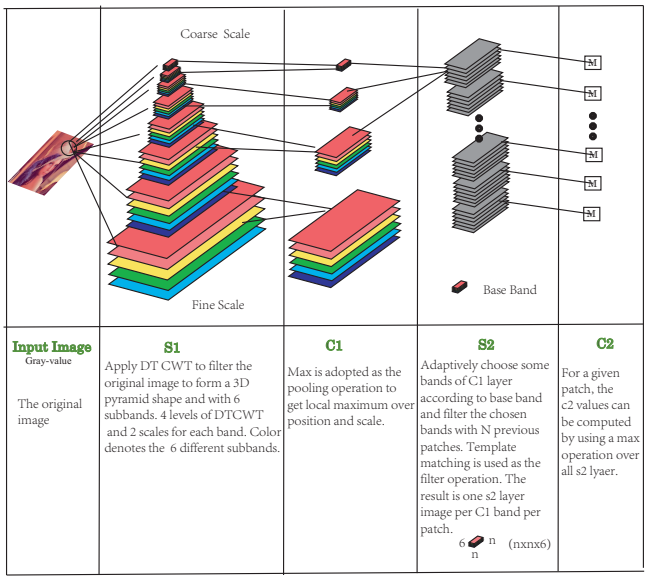


Fig. 3. Model of the algorithm including C1, S1, C2, and S2 layers. First, the input image is filtered by a DT CWT to form a pyramid shape with 6 directional subbands per scale. Next, with a local MAX pooling operation, the S1 layer is sub-sampled to form the C1 layer. In the S2 units, some bands of the C1 layer are adaptively chosen according to the base band. The chosen bands are filtered with N previous patches with a template matching operation. Finally, the S2 layer is pooled with a max operation to obtain the C2 layer.

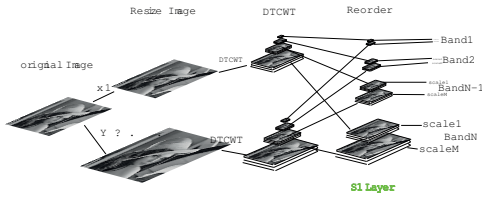


Fig. 4. S1 layer of the Model. The original image is resized to get 2 scales per octave. The original and resized images are filtered by the DT CWT and then reordered to form the S1 layer.

C1 Layer:

The second stage of the hierarchical system is the C1 layer, which corresponds to the complex cells in the V1 cortex. The C1 pooling operation reduces the resolution of the data and saves computation time, although some information is lost. In [11] several methods for the pooling operations are proposed, including Winner Take Most (WTM), Max and Square. In our system, as in others [1, 2], we adopt Max as the pooling operation due to its speed and relative

robustness.

S2 Layer:
The S2 layer corresponds to the simple cells in the V2 or V4 area of the the visual cortex, like the view-tuned-units (VTU) in [1]. The process of the $S2$ layer is shown in fig.3, where K bands, $k1 \dots k2$ of the C1 layer are chosen to form the S2 layer (e.g. bands 2, 3 and 4, according to the scale of the image). The chosen patches $X_{k1} \dots X_{k2}$ at a given location are filtered with N previously seen patches $(P_1, ..., P_N)$ from a reference band $k0$ using the template matching formula (equ.(4) in [2]):

$$r_{k,i} = \exp(-\beta||X_k - P_i||^2) \quad (3)$$

The result is one S2 layer matrix of $r_{k,i}$ values per C1 band per patch.

C2 Layer:
A global max operation is performed in the C2 layer to acquire the shift and scale invariance of the C2 response. In this layer, the S2 responses are pooled in all positions and scales of the S2 image. The value of the best match for each prototype feature P_i is kept as M_i, which leaves a vector M for the N prototype features $(P_1, ..., P_N)$.

The Learning stage:
The purpose of this stage is to select the N prototype features for the S2 layer. A random sampling function is used to extract the prototypes from a large pool of prototypes of various sizes and at random positions.

The Classification Stage:
The last stage of the model is the Classification Stage. In this stage as in [2], we adopted linear SVM classifiers to process the C1 and C2 model features.

2.2. The differences between two models

The primary differences between our proposed model and Serre's model are: 1. our use of complex wavelets instead of Gabor filters in S1; and 2. the optimization of the structure of the model by reducing the over-complete redundancies in S1 and S2, arising from the near-critical subsampling of the DT CWT. In order to obtain a fair comparison, we keep layers C1 and C2, the learning stage, and the classification stage all the same as in Serre's model. We now focus on the differences in the S1 and S2 layers.

S1 Layer:
Our model uses the DT CWT for the S1 layer while Serre's model uses Gabor filters. Due to decimation in the complex wavelet transform, the S1 layer is now a pyramid structure which does not require other operations and is properly subsampled for multi-scale signal processing. For the Gabor system, the ratio of effective width σ to wavelength λ (see [2]) is always about 0.8 (with the filter length being proportional to σ) as the scale changes. The Gabor σ determines the normalized scale. In the DT CWT system the image is filtered with the same length filter (14 taps) in every band due to the optimal subsampling, which leads to good filter characteristics with high computational efficiency.

S2 Layer:
In the S2 Layer, we have optimized the choice of scale bands, $k1 \dots k2$, from the C1 layer to form the S2 layer. Our aim was to speed up the model while providing acceptable scale tolerance. Hence we needed to optimize the number of bands $M = k2 - k1 + 1$ from the C1 layer which form the S2 layer, the number of levels L of DT CWT, and the scale difference S_d in the S1 layer. Meanwhile we needed to ensure that the scale tolerance is larger than a minimum requirement ε.

Now Serre's Gabor-based system uses 8 bands with a range of wavelengths λ from 3.5 to 22.8 (6.5:1). Our wavelet system with two $\sqrt{2}$-interleaved octave decompositions can achieve a similar scale ratio using just 3 scale bands (2, 3 and 4) covering sampling intervals from 4 to $16\sqrt{2}$ samples (5.66:1). Hence if the learning stage is performed using scale band $k_0 = 3$, we only need to set $k_1 = 2$ and $k_2 = 4$ to obtain a similar amount of scale tolerance during the S2 matching process, to the 8 bands of the Gabor system. This, together with the optimal subsampling of every band, contributes to the high efficiency of the DTCWT-based system.

For similar reasons we can use patch sizes of just 2×2, 4×4, 8×8 and 12×12 samples, instead of 4×4, 8×8 and 12×12 and 16×16.

3. EMPIRICAL EVALUATION

3.1. Model Optimization and Parameter Optimization for Object Recognition

This process tends to be time consuming due to the large test datasets. Many experiments are needed to select the optimal parameters, such as: number of levels in DT CWT to build the S1 layer; number of scales in each band in the S1 layer; the pooling size in the C1 layer; and the patch size in the learning stage along with the S2 matching stage. Because our system is computationally efficient, the parameter optimizations for each dataset are relatively convenient compared to a Gabor-based system. Our experiments show that the optimized structure and parameter values for the DT CWT model should be as follows. In the learning stage, band 3 is used with different patch sizes [2 4 8 12] and 6 directional subbands to extract the prototype features. In the S1 layer, 4 levels of the DT CWT are used with a scale difference of $\sqrt{2}$ and this gives 4 bands with 2 scales in each band as shown in Fig.4. In the C1 layer, the pyramid pooling width is [5 2 1 1] samples in the 4 bands. We set $M = 3$, so in the S2 layer, bands 2, 3 and 4 of the C1 layer are used to match the prototype features from the learning stage. A linear SVM classifier is adopted to classify the test images based on the C2 model features.The computer used in our tests was an Intel Core2 6600 2.4GHz CPU, with 3.24GB memory and 32bit Matlab 2007a.

3.2. Results for the Caltech5 dataset

Caltech5[1] contains 5 main types of object: front face, motorcycle, rear-car, airplane and a background dataset. It has been widely used for object recognition tasks [12, 13, 2]. In order to form a fair comparison, the same pre-processing is adopted as in [2], the same fixed splits are used whenever possible, and random splits are used otherwise. We resized all images to 140-pixels in height with conversion to gray-scale.

In [2], Serre et al. compared the performance of C2 $SMFs$ (standard model features) with existing object recognition systems on the Caltech5 dataset and they found that C2 $SMFs$ features performed best. Here, we compare our object recognition system based on complex wavelets with Serre's Gabor-based system, because both are biologically-inspired and based on the mammalian visual cortex. We selected the performance measure of [2] as the metric for comparison, which is the accuracy at the ROC equilibrium point, where the false positive rate equals the miss rate. A linear SVM classifier is used in both systems and the number of features is 1000. Table

[1] http://www.vision.caltech.edu/html-files/archive.html; http://www.robots.ox.ac.uk/ vgg/data/data-cats.html

1 shows the results. From the table we can see that the DTCWT-based system achieves similar object recognition performance to the Gabor-based system, but with a speed improvement of about 4 times. In addition, the DT CWT based system always performed better than Serre's benchmark system.

Datasets	System	Recognition Accuracy	Time(s)
Leaves	Gabor	**0.9786**	79067.7
	DT CWT	0.9652	**20179.1**
	Benchmark[12]	0.840	-
Cars	Gabor	**0.9931**	141826.1
	DT CWT	0.9763	**37364.7**
	Benchmark[13]	0.848	-
Faces	Gabor	**0.9832**	94820.1
	DT CWT	0.9651	**26065.2**
	Benchmark[13]	0.964	-
Airplane	Gabor	**0.9668**	122709.0
	DT CWT	0.9622	**31627.2**
	Benchmark[13]	0.940	-
Motorcycles	Gabor	**0.9918**	122423.0
	DT CWT	0.9835	**32230.7**
	Benchmark[13]	0.950	-

Table 1. Results Obtained with 1,000 C_2 features. SVM classifiers were used.

The Caltech5 dataset contains the following numbers of images: 186 leaf, 1155 car, 450 face, 1074 airplane, 826 motorcycle, 900 background, and 1370 car backgrounds. Table2 gives the average time taken per image for the S1, C1, S2 and C2 stages. The DT CWT system is much more efficient than the Gabor-based system and takes less than 2 seconds per image using Matlab code. Fig.5 shows the results graphically.

System	Leaves	Cars	Faces	Airplanes	Motorcycles
DT CWT	**1.85s**	**1.47s**	**1.93s**	**1.60s**	**1.86s**
Gabor	7.28s	5.61s	7.02s	6.21s	7.09s

Table 2. The average time per image for the complete algorithms.

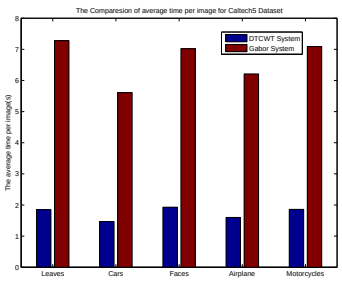


Fig. 5. The average time per image for the complete algorithm, tested on Caltech5 dataset, which include leaves, cars, faces, airplanes, motorcycles, and background data classes

4. DISCUSSION AND CONCLUSION

In this paper, a cortex-inspired object recognition system with complex wavelets is proposed and tested. There are mainly two contributions: firstly, a novel biological-inspired object recognition system is developed, based on complex wavelets; secondly, the Gabor-based HMAX model of [1, 2] is shown to be rather too over-complete. Optimal values for the DTCWT-based system are selected and compared with the Gabor-based system of [2] using the standard Caltech5 dataset. The results demonstrate that the DT CWT system is significantly more efficient than the Gabor-based system while achieving comparable recognition performance.

5. REFERENCES

[1] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 1999.

[2] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411, 2007.

[3] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 204, no. 1156, pp. 301–328, 1979.

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

[5] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999.

[6] N. Kingsbury, "Shift invariant properties of the dual-tree complex wavelettransform," *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on*, vol. 3, 1999.

[7] DH Hubel and TN Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106, 1962.

[8] S. Mallat, "A compact multiresolution Representation: The Wavelet Model," *Proc.IEEE Computer Society Workshop on Computer Vision*, pp. 2–7, 1987.

[9] J. Fauqueur, N. Kingsbury, and R. Anderson, "Multiscale keypoint detection using the dual-tree complex wavelet transform," in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2007, pp. 1625–1628.

[10] T. Hong and N. Kingsbury, "Estimation of the fundamental matrix based on complex wavelets," in *Networking and Information Technology (ICNIT), 2010 International Conference on*, 2010, pp. 350–354.

[11] H. Wersing and E. Korner, "Learning optimized features for hierarchical models of invariant object recognition," *Neural computation*, vol. 15, no. 7, pp. 1559–1588, 2003.

[12] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," *Computer Vision-ECCV 2000*, pp. 18–32, 2000.

[13] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, 2003, vol. 2.