TREE-STRUCTURED WAVELET DECOMPOSITION BASED ON THE MAXIMIZATION OF FISHER'S DISTANCE

Sergio Barbarossa, Laura Parodi

INFOCOM Dpt., University of Rome 'La Sapienza'
Via Eudossiana 18, 00184 Rome, Italy
e-mail: sergio@infocom.ing.uniroma1.it

ABSTRACT

The aim of this work is to propose a method for optimizing the decomposition law of a tree-structured wavelet transform in order to maximize the capability of discriminating different textures. The optimization criterion is the maximization of the Fisher's distance. The analysis is carried out theoretically and by simulation on gaussian Markov random fields and is then applied to the classification of real Synthetic Aperture Radar images.

1. INTRODUCTION

Multiresolution processing of images by Wavelet Transform (WT) has already been proposed for texture analysis and classification. The WT is generally computed by a bank of mirror filters, whose output is undersampled in order to keep the overall number of samples constant. The WT of an image generates four images: a lowpass sub-image and three detail sub-images corresponding to the discontinuities along the horizontal, vertical and diagonal directions. The WT can then be applied recursively on the sub-images to extract informations at different scales. Conventional pyramid-type wavelet transform recursively decomposes the low-frequency channel. However, in many practical cases, most of the information useful for discriminating different classes of texture is contained in the middle frequency channels. Moreover, dealing with 2-Dimensional (2D) sequences, the spatial orientation of the discontinuities is relevant for the ensuing classification. Therefore it is useful to decompose the image in an adaptive way in order to enhance the discrimination capabilities of the classifier using the wavelet representation as a tool for the feature extraction. In [1] has been proposed an adaptive tree-structured wavelet decomposition that assumes as a criterion for the choice of the channel to be further decomposed the energy: at each iteration the WT is applied to the channel with the highest energy. In this work we propose a recursive decomposition method based on the maximization of the discrimination capabilities among different classes. The discrimination capability is assessed by the Fisher's distance. The classification is performed by evaluating a set of discriminant functions and comparing them with a set of thresholds. The discriminants are functions of the wavelet coefficients representing the image to be classified. It will be shown that the discriminants have to be nonlinear functions of the wavelet coefficients. The proposed approach is analyzed theoretically and by simulation on a gaussian Markov random field, and on a real Synthetic Aperture Radar (SAR) image.

2. CLASSIFICATION IN THE WAVELET DOMAIN

An image classification system is tipically based on two fundamental steps: feature extraction and classification. In the approach we are proposing the feature extraction is performed by computing the wavelet transform of the image and the classification is obtained by computing a set of discriminants, which are functions of the wavelet coefficients, and comparing them with a set of thresholds.

2.1 Why using wavelets ?

The image is classified according to the texture properties. The reason for using the wavelet transform lies in the fact that the texture is related to the local correlation (or spectrum) properties. The wavelet transform is a method for representing an image at different scales. In particular, it provides information on the energy content of the image over different spatial frequency bands and orientations. This information can then be exploited for discriminating regions with different texture. Moreover, by using an orthogonal WT, we have an efficient and nonredundant tool for extracting the information useful for the discrimination. The method is nonredundant because the number of samples in the representation in this case is exactly equal to the number of pixel of the original image. The method is also efficient in the sense that the two-dimensional WT can be efficiently computed by the cascade of quadrature mirror filterbanks, whose outputs are all decimated by a factor two, as shown inFig.1 [2].

By means of the wavelet decomposition, we have a tool for associating to each pixel a feature vector, instead of just one value given by the pixel intensity. For example, with reference to Fig.1, representing one stage of the wavelet transform, we can associate to each position (n, m) in the image a four-dimensional vector y(n, m) whose elements are the values of the four output sub-images, corresponding to the same position:

$$y^T(n, m) = (y_1(n, m), y_2(n, m), y_3(n, m), y_4(n, m)) \tag{1}$$

If we further decompose each sub-image by the wavelet transform, we can associate to each pixel a sixteen-elements vector and so on. It is this increase in

0-7803-2431-5/95 $4.00 © 1995 IEEE

dimensionality that provides a better capability of discriminating different texture areas.

2.2 How computing the discriminants ?

The wavelet coefficients are combined to evaluate a set of discriminant functions which are then compared with a set of suitable thresholds for the classification.

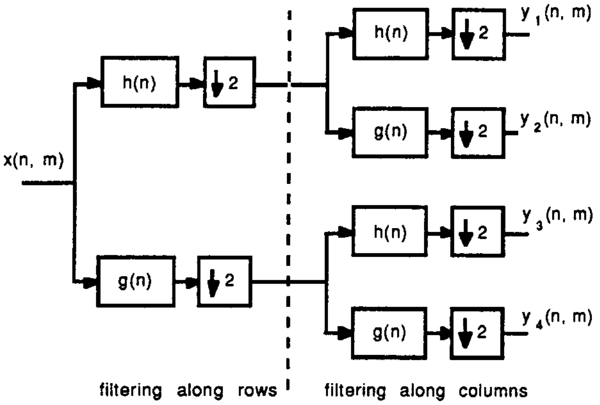

Fig.1 - One iteration of wavelet decomposition

The number of discriminant functions has to be at least equal to the number of classes minus one. Dealing with real images, it is often necessary to compute the discriminants as *nonlinear* functions of the wavelet coefficients. The reason for using nonlinear functions is the following. Let us consider one step of the WT, as shown in Fig.1. The lowpass (LP) channel (along both rows and columns) often contains most of the energy necessary for the image representation. However, since the *discrimination* problem is of course different from the *representation* problem, often the LP sub-image is not essential for the discrimination. Conversely, the highpass (HP) sub-images, even if characterized by a lower energy content, they do carry informations about the image discontinuities (their frequency and their orientation), which are often fundamental for the discrimination between different textures. Let us now consider the ideal case of an image modelled as a gaussian random field. If the samples associated to each class are characterized by different covariance matrices and mean vectors, according to the Bayes' decision rule, the optimal discriminant function turns out to be a quadratic function. The quadratic function reduces to a linear function only in the case in which the covariance matrices relative to different classes are the same, in which case the discrimination is only based on the difference between the mean vectors. Since the WT is a linear operator, the WT of a gaussian random field provides four gaussian random fields. Three of these fields (the three highpass sub-images) are zero mean gaussian random fields, being the output of highpass filters. This means that if we want to base our classification method on the highpass wavelet coefficients, we cannot exploit any difference in the mean value, but we must exploit the only possible difference which is impressed on the covariance matrices. The discriminant functions are then quadratic functions. Dealing with real images, the gaussian model may not be appropriate. However, if adjacent image pixel are weakly correlated and the filter impulse responses are long enough, the filter output can still be modelled as a gaussian random field, by virtue of the central limit theorem. This means that the previous arguments, based on the ideal gaussian model, can be often extended to the analysis of real images.

2.3 Classification criterion

The aim of the ensuing analysis is to quantify the improvement in discrimination capability of a classification method based on the wavelet decomposition. The performance of the classifier can be assessed by the probability of erroneous classification. However, since sometimes it may be difficult to find out a closed form expression for the error probability, a good performance parameter can be given by the Fisher's distance which, in a two-class problem, is defined as [3]:

$$f = \frac{(\eta_2 - \eta_1)^2}{\sigma_2^2 + \sigma_1^2} \tag{2}$$

where η_k and σ_k^2 are the expected value and variance of the discriminant function, conditioned to the fact that the input sequence belongs to the class Ω_k, with k=1, 2. The Fisher's distance will also be used to evaluate which is the tree-decomposition that maximizes the discrimination capabilities. We will examine the optimal approach based on the computation of the discriminant as a nonlinear function of the observed vector and some simpler sub-optimal approaches where the nonlinearity is applied to the output of the WT and the resulting values are then combined linearly.

2.3.1 Optimal quadratic classifier

We will now examine in detail the two-classes case where the input image is modelled as a zero mean, stationary gaussian Markov random process. Let us denote by x(n, m) the 2D input sequence. We suppose that the input random process x(n, m) may belong to two classes Ω_1 and Ω_2, in which cases, it is characterized by covariance matrices C_{X1} or C_{X2}, respectively. We will also assume that the correlation of the process is separable and has an exponential law:

$$E\{x(n, m)\, x(n+i, m+j)\} = R_k(i, j) = R_{kx}(i) \cdot R_{ky}(j) = \\ = P_k \exp(-|i|/l_{kx}) \exp(-|j|/l_{ky}) \tag{3}$$

where the subscript k refers to the class (k = 1 or 2) and l_{kx} and l_{ky} are the correlation lengths, corresponding to class k, along the rows and the columns, respectively. Since the input sequence is a zero mean gaussian process, the output sequences $y_i(n, m)$ are also zero mean gaussian processes. The goal of the classifier is to associate to each pixel x(n,m) its corresponding class. According to the Bayes approach, given a zero mean gaussian vector y that may belong to two classes Ω_1 and Ω_2, the optimal classifier is based on the computation of the quadratic discriminant function [3]:

$$z = y^T Q y \qquad (4)$$

and on its comparison with a threshold. The matrix **Q** is equal to [3]:

$$Q = C_{Y1}^{-1} - C_{Y2}^{-1} \qquad (5)$$

where C_{Y1} and C_{Y2} are the covariance matrices of **y** conditioned to the belonging of the input sequence to class Ω_1 or Ω_2, respectively. The vector **y** is formed with the outputs of the wavelet transform of the input sequence x(n,m), as in (1). The covariance matrices C_{Y1} and C_{Y2} can be expressed in terms of the input covariance matrices and the filter impulse responses, by recalling that the covariance function of the filter output is equal to the convolution between the input covariance function and the autocorrelation of the filter impulse response In the case we are examining, the conditional expected values of the discriminant function (4) are:

$$\eta_k = E\{z / \Omega_k\} = E\{y^T Q y / \Omega_k\} = \operatorname{tr}(Q C_{Yk}) \qquad (6)$$

$$\sigma_k^2 = E\{(z - \eta_k)^2 / \Omega_k\} = 2 \operatorname{tr}(Q C_{Yk} Q C_{Yk}) \qquad (7)$$

for k=1, 2 (tr(A) indicates the trace of the matrix A). These values, once substituted in Eqn.(2), allow us to evaluate the Fisher's distance. For example, with reference to the correlation model of (3), let us assume that the two input classes are characterized by the two sets of values: (P_1= 1, l_{1x}= 8, l_{1y} = 8) and (P_2= 1, l_{2x} = 16, l_{2y} = 4). In such a case, there is no way to discriminate the two classes by means of a pixel-by-pixel analysis since each pixel is a gaussian random variable (rv) with the same mean value and variance in both cases (the Fisher's distance is zero in such a case); if we compute the first order WT (we have used an 8-taps Daubechies filter) we obtain a Fisher's distance equal to 0.213. If we proceed further by decomposing each sub-image by a second WT, thus obtaining 16 sub-images, we obtain a Fisher's distance equal to 0.436. This parameter then quantifies the increase of the discriminability between the two classes gained by working with the wavelet representation.

2.3.2 Feature subset selection

As regards the representation problem, all the four sub-images obtained by the WT have to be retained if we want to invert the transform and recover the initial image. However, for what concerns the classification problem, the four sub-images carry different amount of information and some of them can be discarded without any consistent loss of discrimination capabilities. The Fisher's distance can be used again as a measure of the separability between classes. Its value can be used as a criterion for the extraction of the subset of features that still provide a good classification, within an acceptable loss, with a lower computational cost. As an example, let us consider again the classification of the 2D Markov field considered before. Table I shows the Fisher's distance corresponding to taking all the sub-images and any subset of them. The left column reports the sub-images taken into account (the numbers 1÷4 refer to the output sequences, as indicated in Fig.1); the center and right columns report the Fisher's distance relative to two cases corresponding to two pairs of correlation lenghts: (l_{1x} = 8, l_{1y} = 8; l_{2x} = 16, l_{2y} = 4) and (l_{1x}= 8, l_{1y} = 8; l_{2x} = 1, l_{2y} = 1). From both cases we can observe, for example, that the lowpass image can be discarded without any appreciable loss and that the two sub-images carrying more information are the sub-images 2 and 3.

subset	8x8/16x4	8x8/1x1
1 2 3 4	0.2126	0.6289
2 3 4	0.2124	0.6204
1 3 4	0.1101	0.5551
1 2 4	0.1027	0.5551
1 2 3	0.2121	0.6196
3 4	0.1099	0.5470
2 4	0.1025	0.5470
2 3	0.2119	0.6172
1 4	0.0001	0.4853
1 3	0.1098	0.3613
1 2	0.1025	0.3613
4	0.0016	0.0749
3	0.1027	0.3076
2	0.1104	0.3076
1	0.0090	0.4778

Table I - Fisher's distance

2.3.3 Suboptimum discriminants

The optimum discriminant is in general a nonlinear function of the observed samples. In the gaussian case, it is a quadratic function (see Eqn.(4)). In practical applications, however, it may be quite cumbersome to estimate the coefficients of the nonlinear classifier and is then desirable to devise some simpler suboptimum approach. At this purpose we propose the computation of the discriminant as a linear combination of nonlinear functions (tipically the square) of the wavelet coefficients. The combination coefficients are chosen in order to maximize the Fisher's distance. For example, with reference to Fig.1, let us denote by **z** the vector whose elements are nonlinear functions of the wavelet coefficients: $z^T(n, m) = (g(y_1(n, m)), ..., g(y_M(n, m)))$, where $y_i(n, m)$ is the generic i-th output sub-image, g(·) is a nonlinear function and M indicates the number of sub-images produced by the WT. The only important characteristics of g(·) is that it cannot be an anti-symmetric function, in order to give rise to new random variables that have an expected value different from zero. We define the new discriminants as a linear combination of the elements of **z**:

$$d(n, m) = \sum_{k=kmin}^{kmax} w_k\, g(y_k(n, m))$$

The weighting coefficients w_k can be chosen in order to maximize the Fisher's distance. Indicating by $\mathbf{m}_{Zk}$ and C_{Zk} the expected values vector and the covariance matrix of the random variables vector **z**, conditioned to the class k, with k=1, 2, the optimal weighting vector is:

$$\mathbf{w} = \left(\frac{1}{2} C_{Z1} + \frac{1}{2} C_{Z2} \right)^{-1} (\mathbf{m}_{Z2} - \mathbf{m}_{Z1})$$

If the nonlinear function g(·) is simply the square value of its argument, the expected value and the covariance matrix of **z** can be explicitely expressed in terms of the expected value and covariance matrix of the input image (in the case of a gaussian random field) and of the filter impulse responses. To improve the separability between classes, we can also average the squares of the pixel values within a moving window. The effect of this average is that, as far as the region within the window is homogeneous, the statistical parameters of **z** tend to be more separated and this increases the discrimination capabilities. However, this advantage is evidently paid in tems of resolution. To quantify the performance of the proposed approach, Table II shows the Fisher's distance F together with the classification error probability P_e, as a function of the window size and the sub-images used for the classification. In particular, the first column (Lin(4)) refers to the case in which all the four output sub-images of Fig.1 are considered, whereas the second column refers to the case in which the lowpass sub-image is discarded. As expected, the presence of the lowpass sub-image does not bring any appreciable contribution to the discrimination.

window	Lin(4)		Lin(3)	
size	F	P_e (%)	F	P_e (%)
1x1	0.198	39.15	0.197	39.15
5x5	1.604	16.52	1.603	16.53
9X9	4.032	5.43	4.032	5.43
13X13	7.439	1.02	7.433	0.99

Table II - Performance of the wavelet-based classifier

3. CLASSIFICATION OF SAR IMAGES

As an example of application to real images, we now examine the classification of Synthetic Aperture Radar (SAR) images. In such a case there is a particular need for an automatic classification method to handle the huge amount of data transmitted from the satellite carrying the radar onboard without the need for a human support. At this regard, Fig.2 shows a SAR image obtained by the German E-SAR flying over Oberpfaffenhofen. The image shows an urban area in the upper left corner, some forests in the lower left and upper right regions of the image, some cultivated areas and an airport. The imaged area was illuminated from the top, as evidenced by the shadows in the image. The pixel intensity is not compensated to correct the different attenuation due to the the different distance from the closest points to the farthest ones (this explains why the upper pixels are brighter than the lower ones). By using a linear combination of the square values of the wavelet coefficients corresponding only to the three detail images (thus discarding the lowpass image), and considering as classes of interest urban areas, forests and cultivated fields, the result of the classification is reported in Fig.3. It is worth noticing that, in spite of the variation in the pixel intensity, the method works quite well in discriminating the different areas.

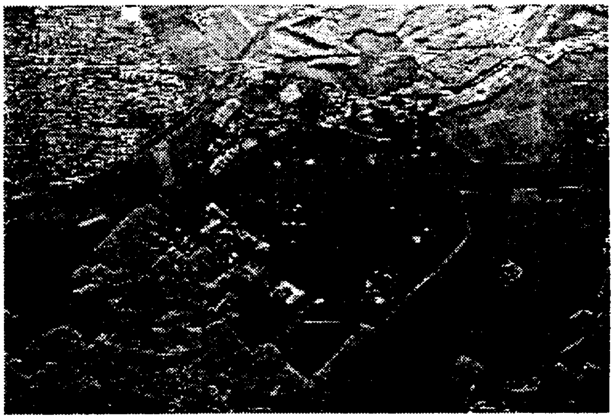

Fig.2 - SAR image of Oberpfaffenhofen

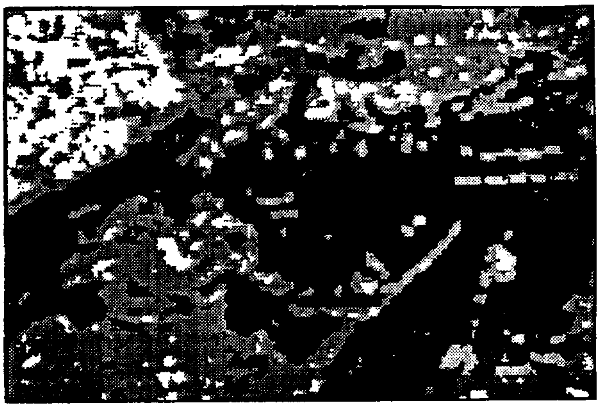

Fig.3 - Classified image (white colour indicates urban areas, gray indicates forests and black corresponds to cultivated fields)

References

[1] T.Chang, C.C.J.Kuo: "Texture Analysis and Classification with Tree-Structured Wavelet Transform", *IEEE Trans. on Image Proc.*, Vol.2, No.4, Oct.1993.

[2] S. Mallat "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.2, N.7, July 1989..

[3] K.Fukunaga: *An Introduction to Statistical Pattern Recognition*, Academic Press(2nd Ed.), San Diego, 1990.