Statistical Modeling in Wavelet domain for Bayesian Texture Classification and Retrieval

Seyyedeh Leila Nejadhashemi
School of Electrical engineering, Faculty of Engineering
University of Guilan, Rasht, Iran

Babak Nasersharif, Asadollah Shahbahrami
School of Computer engineering, Faculty of Engineering
University of Guilan, Rasht, Iran

Abstract— Feature extraction and similarity measure in feature space are two basic steps in a texture based image retrieval system. In this paper, we propose to extract statistical model-based features in wavelet domain where we use dual tree complex wavelet transform (DT-CWT). To this end, we employ generalized Gaussian density (GGD) to describe the statistical characteristics of DT-CWT coefficients. On the other hand, we utilize Bayesian classifier for measuring similarity and so texture classification. In addition, for improving the classification rate and computational complexity we project the features onto a low dimensional space using three methods: linear discriminate analysis (LDA), locality preserving projections (LPP) and kernel LDA (KLDA). Our experiments are conducted on two different texture databases, i.e. VisTex and Brodatz. We achieve the classification rates up to 97.5% and 96.54% for these two databases respectively which validate the robustness of the proposed method.

Keywords—generalized Gaussian density; Bayesian classification; dual tree complex wavelet transform

I. INTRODUCTION

Recent years have seen an increasing interest in developing efficient approaches for browsing through digital image databases utilized in different areas, such as industry, medicine, crime detection, digital libraries etc. Since the sizes of these databases have been growing, traditional text-based image retrieval methods in which images are manually annotated by text seem to be inefficient due to their subjectivity, inaccuracy and high costs.

To prevent such drawbacks, content-based image retrieval in which images are indexed based on their visual content is introduced. The main goal in CBIR is to develop approaches to achieve the highest possible retrieval precision in the shortest possible length of time. A recent extensive review on the subject can be found in [1].

There are two main processes in a CBIR system: feature extraction and similarity measurement. In the first process a set of features, called image signatures, is generated to accurately represent the content of each image in the database. The set has to be much smaller in size than the original image while capturing as much of the image information as possible. These features are usually extracted from shape [2], color [3] or texture [4] of images among which texture features play an important role in characterizing the image content. The second process requires a similarity measure criterion including different distance and correlation metrics and also classification algorithms to determine how similar each image in the database is to a query image. In the present work, we focus on the texture classification as one of the most important parts in a CBIR system.

Some of the most popular texture extraction methods are based on filtering or transform domain methods among which two dimensional discrete wavelet transform (2-D DWT) has been widely used for this application for years [4]-[6]. The basic assumption of these approaches is that the energy distribution in the frequency domain identifies a texture. According to this assumption which is supported by physiological studies of the visual cortex [7], traditional approaches computed energies of wavelet subbands as texture features. Do et al. [8] proposed an extension of the energy method by modeling a texture using the marginal densities of wavelet subbands coefficients. They used generalized Gaussian density (GGD) modeling of wavelet coefficients and Kullback–Leibler distance metric to perform texture retrieval. Some recent work applied this statistical modeling based image signature in wavelet domain [9]. However, the effectiveness of DWT in complex texture analysis is seriously undermined due to a lack of directionality and shift sensitivity. In other words, DWT only gives the edge information in the horizontal, vertical and diagonal directions. Additionally, since small shifts in the input signal can result in large differences of DWT coefficients at different scales, same two patterns with small spatial shifts will produce widely different feature vectors. Consequently, Gabor wavelet transform (GWT) which is both directionally selective and shift invariant can be applied as an alternative to traditional DWT [10]. However, non-orthogonality of GWT’s basis function leads to high redundancy in transformed images which increases memory requirement and limits retrieval and classification speed.

All afore-mentioned problems can be overcome using dual tree complex wavelet transform (DT-CWT) which gives texture information oriented in six different directions at each scale with approximately shift invariance property and limited redundancy [11]. Some recent work has shown that by using DT-CWT retrieval performance is improved in
comparison to real DWT [12]-[14]. Celik and Tjahjadi [12] used multi scale feature vectors consisting of the mean and standard deviation of DT-CWT subbands magnitude together with simple distance measure to perform the final classification. They improved the performance of their previous work in [13] by using both magnitude and phase of DT-CWT complex subbands and a Bayesian classifier. Additionally, the dimension of multiscale feature vectors was reduced using principal component analysis (PCA), a classical linear technique that projects the data along the directions of maximal variance, to cope with the curse of dimensionality and benefit from computational efficiency.

Dimensionality reduction is the transformation of high dimensional data into a meaningful representation of reduced dimensionality. Ideally, the reduced representation should have a dimensionality that corresponds to the intrinsic dimensionality of the data. They can be divided into two general groups of linear and nonlinear methods. Well known approaches like PCA and linear discriminant analysis (LDA) [15] belongs to the first group. Nonlinear methods are comprised of kernel based algorithms such as kernel PCA (KPCA), kernel LDA (KLDA) [16] and manifold based techniques like Isomap [17], Locally Linear Embedding (LLE) [18] and Laplacian Eigenmaps [19]. Locality preserving projections (LPP) [20] is an alternative to PCA that shares many of the data representation properties of nonlinear techniques while it is still linear and can be defined everywhere in ambient space rather than just on the training data points.

In this paper, first, we utilize GGD to model marginal distribution of both real and imaginary parts of coefficients obtained from DT-CWT. Then, we apply Bayesian inference to feature vectors consisting of GGD parameters estimated by maximum likelihood estimator to perform final texture classification. Additionally, we explore the effect of applying LDA, LPP and KLDA to feature vectors on classification results and speed.

The rest of this paper is organized as follows: In Section 2, DT-CWT is described briefly. The ideas of LDA, LPP and KLDA methods are given in Section 3. Section 4 presents the proposed approach including texture features and Bayesian classification. Section 5 includes our experimental results. Finally, the conclusion is given in Section 6.

II. DUAL-TREE COMPLEX WAVELET TRANSFORM (DT-CWT)

DT-CWT is a recent enhancement to the DWT with important additional properties: it is nearly shift-invariant and directionality selective in two and higher dimensions.

The idea behind DT-CWT, as shown in Fig. 1, is employing two real DWTs. The first DWT gives the real part of the transform while the second one gives the imaginary part [11].

2-D DT-CWT decomposes an image $f(x,y)$ using dilation and translation of a complex scaling function and six complex wavelet functions $\Psi^\theta$, i.e.:

$$
 f(x,y) = \sum_{l \in \mathbb{Z}^2} S_{j_0 l} \phi_{j_0 l}(x,y) + \sum_{b \in \mathbb{Z}^2} \sum_{j \geq j_0} \sum_{l \in \mathbb{Z}^2} C_{j,l}^0 \psi_{j,l}^0(x,y), 
\tag{1}
$$

Where $S_{j_0 l}$ is scaling coefficient, $C_{j,l}^0$ is complex wavelets coefficient, $j$ denotes the decomposition level and $\theta = \{\pm 15^\circ, \pm 45^\circ, \pm 75^\circ\}$ refers to the directionality of complex wavelet subbands [11]. The impulse responses of the six complex wavelets are illustrated in Fig. 2.

III. DIMENSIONALITY REDUCTION TECHNIQUES

A. Linear Discriminant Analysis (LDA)

LDA is a well-known supervised method which discriminates between features and reduces the feature vector dimensionality linearly. In other words, the LDA method makes data as linearly discriminative as possible by maximizing the proportion of between-class covariance, given in (2), to within-class covariance defined in (3).

![Fig. 1. Implementation of the two levels 1D DT-CWT using two filter banks operating as two parallel trees on the same data [13].]

![Fig. 2. Typical wavelets associated with 2-D DT-CWT, (a) the real part, (b) the imaginary part [11].]
Where ‘N’ is the total number of samples, ‘N_i’ is the sample size of i-th class, ‘µ_i’ and ‘µ’ are the mean vector of i-th class and whole samples in dataset respectively, ‘I’ is the total number of classes and ‘X_{ni}’ is the n-th sample of i-th class [15].

B. Kernel LDA

The reason why nonlinear methods are developed is the poor performance of linear methods for samples which cannot be discriminated linearly. As a nonlinear extension of LDA, KLDA defines a nonlinear mapping from the original space to a high dimensional space to obtain a linearly separable distribution in the new space. In kernel based methods, nonlinear mapping which is not available is estimated using kernel trick, i.e., inner product of mapped data and an uncertain nonlinear function.

The between and within class covariance matrices in mapped space is computed by (5) and (6), respectively. In addition, the objective function maximized based on (7).

\[
S_B = \frac{1}{N} \sum_{i=1}^{I} N_i (\mu_i^w - \mu^w)(\mu_i^w - \mu^w)^T
\]  
(5)

\[
S_w = \frac{1}{N} \sum_{i=1}^{I} \sum_{n=1}^{N_i} (X_{ni} - \mu_i)(X_{ni} - \mu_i)^T
\]  
(6)

\[
J(W) = \frac{W^T S_B^w W}{W^T S_w^w W}
\]  
(7)

Where all parameters are the same as in (2) to (4) and ‘φ’ denotes the quantities in the mapped space [16].

C. Locality Preserving Projections (LPP)

While PCA aims to preserve the global structure of data, and LDA aims to preserve the discriminating information, the goal of LPP method is to preserve the local structure of samples. The procedure of LPP can be simply summarized as follows [20]:

1) Constructing the adjacency graph. The graph G can be constructed in three different manners:
   - K-nearest neighbors (KNN): An edge will be put between two nodes if and only if they are among the k nearest neighbors of each other.
   - ε–neighborhoods: Nodes i and j will be connected by an edge if \( \|x_i - x_j\| \leq \varepsilon \).
   - Supervised manner: An edge will be put between two nodes if and only if they belong to the same class.

2) Choosing the weights: The weights evaluate the local structure of the data space. W is a sparse and symmetric n×n matrix. In the simplest case the weight wij is 1 if there is an edge joining the nodes i and j and 0 if there is no connection between them.

3) Eigenmaps: Consider the problem of mapping the weighted graph G to a line so that connected points stay as close together as possible. Let \( y = (y_1, y_2, \ldots, y_m)^T \) be such a map. The optimal projection preserving the locality can be solved by minimizing the following objective function based on the standard spectral graph theory [21].

\[
\sum_{ij} (y_i - y_j)^2 w_{ij}
\]  
(8)

This minimization problem is to ensure that if \( x_i \) and \( x_j \) are close, then \( y_i \) and \( y_j \) are close as well. The objective function can be modified as

\[
\frac{1}{2} \sum_{ij} (y_i - y_j)^2 w_{ij} = \frac{1}{2} \sum_{ij} (a^TX_i - a^TX_j) w_{ij}
\]

\[
= \sum_{i} a^TX_i d_{ii} x_i^T a - \sum_{ij} a^TX_i w_{ij} x_j^T a
\]

\[
= a^T X(D - W) X^T a = a^T XLX^T a.
\]

\[
D_{ii} = \sum_{j} w_{ij}
\]

\[ L = D - S \]  
(11)

Where \( D \) is a diagonal matrix, \( L \) is the Laplacian matrix and \( a \) denotes a transformation vector, that is, \( y = a^TX \), where the i-th column vector of X is \( x_i \). The larger the value of \( d_{ii} \) (corresponding to \( y_i \)), the more “important” \( y_i \) is. Therefore, a constraint is imposed

\[
y^T D y = 1 \Rightarrow a^T XDX^T a = 1
\]  
(12)
Consequently, the minimization problem reduces to find

\[
\arg \min_{a} a^T XLX^T a/a^T XDX^T a = 1. 
\]  

(13)

It is equivalent to find the solution of the generalized eigenvalue and eigenvector problem.

\[
XLX^T a = \lambda XDX a
\]

(14)

Where \(a_0,a_1,\ldots,a_{L-1}\) is the solutions of (14), ordered according to their eigenvalues \(\lambda_0 < \lambda_1 < \cdots < \lambda_{L-1}\). Thus, the embedding of each data point is represented by

\[
x_i \rightarrow y_i = A^T x_i, \quad A = (a_0,a_1,\ldots,a_{L-1})
\]

(15)

IV. PROPOSED METHOD

A. Texture Features

In order to characterize textures, generalized Gaussian density, defined in (16), is used to model the distribution of real and imaginary parts of DT-CWT coefficients.

\[
p(x;\alpha,\beta) = \frac{\beta}{2\alpha \Gamma(\frac{1}{\alpha})} e^{\frac{|x|^\alpha}{\beta^\alpha}}
\]

(16)

where \(\Gamma(.)\) is the Gamma function, \(\alpha\) is the scale parameter modeling the width of the PDF peak and \(\beta\) is the shape parameter which is inversely proportional to the decreasing rate of the peak.

Some typical examples in Fig. 3 show acceptable fits of GGD to the actual distribution of real and imaginary parts of DT-CWT coefficients in different scales.

We use two parameters of GGD \((\alpha,\beta)\) estimated by maximum likelihood estimator for each of real and imaginary parts of DT-CWT coefficients in each decomposition scale. Therefore, image decomposition on \('S' scales with each scale consisting of 6 complex subbands leads to an feature vector \((F1)\) of length equal to 24S:

\[
F1 = \{a_s^1, b_s^1, \ldots, a_s^6, b_s^6\}, s = 1,2,\ldots,S.
\]

(18)

In addition to the proposed feature vector, we build two others to make a comparison between them. The first one includes the GGD parameters estimated from DWT coefficients distribution \((F2)\). The mean and standard deviation of DT-CWT subbands are used to create the other image signature \((F3)\) which consists of features extracted from real and imaginary parts of complex subbands.

B. Texture Classification

Bayesian classification and decision making are based on probability theory and the principle of choosing the most probable or the lowest risk option. Let \(x = \{x_1, x_2, \ldots, x_D\}\) be a feature vector of dimensionality \(D\). The probability (or posteriori probability) that a feature vector \(x\) belongs to texture class \(w_k\) is \(p(w_k|x)\). The posterior probabilities can be computed using the a priori probabilities in the framework of Bayes formula, i.e.:

\[
p(w_k|x) = \frac{p(x|w_k)p(w_k)}{p(x)},
\]

(19)

where \(p(x|w_k)\) and \(p(w_k)\) are the probability density function and a priori probability of class \(w_k\), respectively. The factor in the denominator is a normalization factor used to ensure that the weighted sum of \(p(x|w_k)\)’s for a training database is one, i.e.:

\[
p(x) = \sum_{k=1}^{K} p(x|w_k)p(w_k),
\]

(20)

where \(K\) is the total number of different texture classes.

The feature vector \(x\) is assigned to the texture class \(w_k\) with the highest a posteriori probability which produces the minimum error probability, i.e.,

\[p(w_k|x) > p(w_i|x), \quad \forall i \in \{1,2,\ldots,K\}, i \neq k\]

which can be reformulated using (19) as
The classification problem reduces to the estimation of $p(w_k | x)$ and $p(w_k)$ for each texture class $w_k$ [11]. The a priori probabilities are estimated from the relative frequencies of the classes in the training set. Since the number of samples belonging to each class is equal in our implementation, the a priori probabilities of the texture classes will be equal.

V. EXPERIMENTAL RESULTS

The efficiency of proposed texture classification approach is evaluated on two commonly-used databases. D1 database consists of the same 40 textures used in [8] from MIT Vision Texture (VisTex) database [22] and D2 database includes 110 monochrome textures from Brodatz album [23].

Each of $512 \times 512$ images is divided into 16 non-overlapping $128 \times 128$ subimages among which 4 ones are chosen randomly for testing and the twelve others are used for training. Consequently, there are a total number of 480 images in training set and 160 images in test set generated from D1 database. Similarly, we obtain 1320 images for training and 440 images for testing from D2 database.

In implementation we set the number of decomposition scales to 3 because the transformed images obtained in next scales are too small to have reliable model estimation.

We conduct different sets of experiments using two different databases and different texture features to evaluate the proposed approach performance.

In the first set of experiments, D1 database is used to compare the Bayesian classification (BC) of F1, F2 and F3 feature vectors and the approach proposed by Do et al. [8] that used F2 texture features and kullbach-leibler (K-L) distance metric.

As it is shown in Table I, our proposed approach has the best classification gain (96.87%) in comparison with three other methods. Additionally, using Bayesian classifier causes a 3.75% increase in classification rate of F2 compared to when K-L distance is applied.

Table II shows the results obtained from the second set of experiments in which D2 database is used to compare the Bayesian classification of F1, F2 and F3 feature vectors. According to the results, the proposed approach still outperforms others with an average classification rate up to 92.5%.

Finally, we explore how dimensionality reduction techniques affect the classification performance. In this experiment, we apply two linear techniques, i.e. LDA and LPP and a nonlinear technique, i.e. KLDA only to our proposed statistical model-based features which result in the best classification rates in two previous experiments. We use the supervised method to construct the adjacency graph in LPP and Gaussian kernel function in KLDA. The target dimensionalities of features is set to be 20 and 40 for D1 and D2 databases respectively as they give the best performance empirically. The classification results of their projected feature vectors are compared in Table III and Table IV. Clearly, the linear techniques outperform KLDA in all cases. While LDA and LPP projected features have comparable classification rates, they lead to a considerable improvement over non-projected ones. Additionally, low dimensional feature vectors have advantages over original ones from computational point of view. According to Table V, the average classification time of one image in both D1 and D2 databases is reduced by using dimensionality reduction techniques.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1&amp;BC</th>
<th>F2&amp;BC</th>
<th>F3&amp;BC</th>
<th>F2&amp;K-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Rate</td>
<td>96.87</td>
<td>81.25</td>
<td>74.36</td>
<td>77.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>F1&amp;BC</th>
<th>F2&amp;BC</th>
<th>F3&amp;BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Rate</td>
<td>92.5</td>
<td>78.86</td>
<td>75.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>F1&amp;LDA</th>
<th>F1&amp;LPP</th>
<th>F1&amp;KLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Rate</td>
<td>97.5</td>
<td>97.5</td>
<td>90.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>F1&amp;LDA</th>
<th>F1&amp;LPP</th>
<th>F1&amp;KLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Rate</td>
<td>96.54</td>
<td>96.14</td>
<td>88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimensionality Reduction Technique</th>
<th>None</th>
<th>LDA</th>
<th>LPP</th>
<th>KLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>0.4688</td>
<td>0.2383</td>
<td>0.3854</td>
<td>0.4794</td>
</tr>
<tr>
<td>D2</td>
<td>0.5817</td>
<td>0.3175</td>
<td>0.3913</td>
<td>0.5502</td>
</tr>
</tbody>
</table>
VI. CONCLUSION

In this paper, we propose a new efficient method for Bayesian based texture classification. The GGD which has been proved to be a proper model of DWT subbands histogram is used to model the complex subbands of DT-CWT, a recently-used multiresolution transform which has important advantages over the traditional DWT. It is shown that GGD has a good capability to model the distribution of both real and imaginary parts of coefficients obtained from DT-CWT. Furthermore, we apply three effective dimension reduction techniques including LDA, LPP and KLDA to our model based feature vectors. Our experiments show that LDA and LPP projected features give both of high speed and high texture classification rates.

VII. REFERENCES