Wavelet Feature Selection based on Support Vector Machine

MALIKA AIT AIDER[*], KAMAL HAMMOUCHE[*] and DJAMEL GACEB[**]
[*] Université Mouloud Mammeri, Département d'Automatique, Tizi-Ouzou, Algeria
[**] Laboratoire LIRIS, INSA de Lyon,
Bât. Jules Verne 20, avenue Albert Einstein, 69621, Villeurbanne Cedex – France
m_aitaider@yahoo.fr, kamal_hammouche@yahoo.fr, djamel.gaceb1@insa-lyon.fr

Abstract: - Feature selection is the process of selecting the best features among all the features because all the features are not useful in constructing the clusters: some features may be redundant or irrelevant thus not contributing to the learning process. In this paper, we proposed the combination of the discrete wavelet transform and two features selection approaches such as a principal component analysis followed by a sequential forward selection method. This strategy of combination increases the performance of a recognition system. The wavelet transform performs a local analysis to characterize the image in time and scale space. We motivate the use of this technique in selecting the optimal subset of features using the coefficients of the approximation sub-image generated by a wavelet transform. The choice of the criterion for selecting a subset features is primordial. Therefore, in this work, it is based on the correct classification achieved by the support vector machine classifier. Some well known wavelet families with their different orders (Haar, coiflet1, Daubechies 4 and symlet 4) are utilized to investigate their performance in handwriting digit recognition. Support vector machine is used again in the classification phase. Experiments conducted on a data set extracted from the USPS database show that our proposed method can increase the recognition accuracy.

Key-Words: - feature extraction, feature selection, handwritten digit recognition, wavelet transform, support vector machine.

1 Introduction

The character recognition has long been a goal of many research efforts in the optical character recognition (OCR) field. It is not only a newly developing topic due to many potential applications, such as bank check reading, postal mail sorting, automatic reading of tax forms, but it is also a benchmark for testing and verifying new pattern recognition theories and algorithms. Particularly, the recognition of handwritten digits is an important sub-problem of optical character recognition (OCR). The difficulty with handwriting recognition is large intra-class variance due to the shape variations caused by the distinct writing styles of individuals [18]. In this kind of applications, a vast amount of features which can distinguish one class of patterns from another in a more concise and meaningful way is usually needed. However, it allows increasing system complexity, processing time and on the other hand, a bad choice of some features leads to a worse rather than better performance. So, it is important and necessary to select the most relevant features to increase the performance of the method used. Feature selection is the problem of identifying features [3]. That is, this can be used to identify the important features with significant information content when the problem is poorly structured. The main goal of feature selection is to select the potential features needed to discriminate samples belonging to different classes and therefore, allowing best accuracy of a system.

A number of feature selection methods can be divided into three categories: filter method [7], wrapper method [11], and hybrid methods [6]. In a filter approach, some feature evaluation functions such as functions that measure distance, information theory and dependency are used without involving any mining algorithm. When feature extraction is carried out, a criterion function should be given for increasing classification separability. Some of these techniques such as principal component analysis (PCA), Karhunen-Loeve (K-L) transformation [8] have been successfully applied to recognize the handwritten digits. In a wrapper method, the performance of the classifier is used to evaluate the feature subsets, but it also tends to be more computationally expensive than the filter model [12]. In recent years, a number of feature selection algorithms based on wrapper techniques have been presented and tested on handwritten digit recognition. Some well-known feature selection methods include: the genetic algorithm [10], entropy based feature selection [21], independent component analysis (ICA) [4], sequential forward selection-sequential backward selection (SFS/SBS) [9], sequential forward floating selection/sequential backward floating sequential selection (SFFS-SBFS) [17] and their combinations. Finally, the hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages.

In this paper, we concentrate on improving the recognition rate by selecting efficient subset of wavelet features. To do this, we propose a hybrid model by combination of a filter method using principal component analysis (PCA) followed by a wrapper method using sequential forward feature selection (SFS) and a support vector machine is used as an evaluation criterion for searching relevant features. For classification task, support vector machine is used again as classifier.

The rest of this paper is organized as follows: general description of wavelet theory, wavelet family and feature extraction-selection techniques are given in section 2. Section 3 describes the support vector machine; section 4 is the experimental results and the conclusion is provided in section 5.

2 Wavelet features selection

2.1 Concepts of the Wavelets

The theory behind wavelets has been developed during the last twenty to thirty years. A wavelet is localized function that can be used to capture informative, efficient, and useful descriptions of a signal. If the signal is represented as a function of time, then wavelets provide efficient localization in both time and either frequency or scale. It has received significant attention recently due to their multiresolution concept [16] which is suitable for image processing tasks including image compression and texture classification.

The discrete wavelet transform (DWT) uses filter banks to perform the wavelet analysis. It decomposes the signal into wavelet coefficients from which the original signal can be reconstructed again. The wavelet coefficients represent the signal in various frequency bands. The coefficients can be processed in several ways, giving the DWT attractive properties over linear filtering. Compared to Fourier technique, wavelet transform permits much more flexibility in choosing appropriate representations for particular applications. The choice of type of wavelet is of great importance in the performance of the application. Recently, a number of wavelet families have been proposed for feature extraction to recognize the handwritten digits. Among them, Haar wavelet transform [14] which is the most commonly used wavelets because they are easy to comprehend and fast to compute, Daubechies wavelet of four order [2] and biorthogonal spline wavelet [5].

2.1.1 Discrete wavelet transforms

The wavelet transform can be viewed as a generalization and refinement of the concept of a windowed Fourier transform. With the Short Time Fourier Transform (STFT), the analysis function is a window. The window is translated in time but is not otherwise modified. The wavelet approach replaces the STFT window with a wavelet function ψ (called a mother wavelet). The wavelet function is scaled (expanded or dilated) in addition to being translated in time. A generalized wavelet family, $\psi_{a,b}$ described in the normalized form is:

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right) \qquad (1)$$

The scale parameter a indicates the level of analysis and b represents the translation variable. Small values of a provide a high frequency analysis while large values (large scale) correspond to low frequency analysis.

Typically, the scale factor between levels increases by two. Widely used a and b parameter settings that create an orthonormal bases are $a = 2^j$ and $b = 2^j k$ ($j,k \in \mathrm{Z}$). The wavelet family then becomes:

$$\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k) \qquad (2)$$

The wavelet transform calculates wavelet coefficients (details) by filtering the input signal $f(x)$ by $\psi_{j,k}(x)$. It results in a set of detail coefficients $D_{j,k}$ that represent the high-frequency signal information.

$$D_{j,k} = \langle f, \psi_{j,k} \rangle = 2^{-j/2}\int_{-\infty}^{\infty} f(x)\psi(2^{-j}x - k)dx \qquad (3)$$

These wavelet coefficients are measures of the goodness of fit between the signal and the wavelet. Large coefficients indicate a good fit.

On the other hand, the approximation coefficients at a given scale are obtained in the same way as the details coefficients, but, by using the scale function $\phi(x)$, which is orthogonal to $\psi(x)$. The wavelet approximate or scale coefficients are defined by:

$$S_j(k) = \int_{-\infty}^{\infty} \phi_{j,k}(x) f(x) dx \qquad (4)$$

Using the concept of multi-resolution, the DWT decomposed any signal into a set of discrete wavelet coefficients. Generally, the DWT uses filter banks for the analysis and synthesis of a signal. The filter banks contain wavelet (high pass filter G) and scaling filters (low pass filter H) to extract the frequency content of the signal in various sub-bands.

In one-dimensional (1-D) wavelet transform, a signal is passed through a low pass filters (scaling functions) and high pass filters (wavelet function) simultaneously. Down-sampling or decimation by a factor 2 is performed after each pass through filters. Consequently, this process constitutes one level of decomposition. As a result, the approximation coefficients and details coefficients are obtained at this level of decomposition. Multiple levels or scales of the wavelet transform are made by repeating the filtering and decimation process on the output of low pass filter only (low frequency component).

For a 2-D input signal such as images, the transform coefficients are obtained by projecting the input onto the four basis functions given in equation (5). $\phi(x,y)$ can be thought of as the 2-D scaling function; $\psi_1(x,y)$, $\psi_2(x,y)$ and $\psi_3(x,y)$ are the three 2-D wavelet functions. This results respectively in four different sub-images in the decomposition corresponding to the four types of transform coefficients; LL_1 (the image approximation), and three detail sub-images; LH_1 (contain the vertical details), HL_1 (contain the horizontal details) and HH_1 (represent the diagonal details). To obtain the next coarse level of wavelet coefficients, the sub-image LL1 alone is further decomposed. These results in 2 level wavelet decomposition as shown in Figure 2. Similarly, to obtain further decomposition, LL2 will be used. This process continues repeatedly until some final scale is reached. Figure 1 depicts the first level in a multi-resolution pyramid decomposition of an image. Figure 2 shows the decomposition result of two multi-resolution levels.

$$\begin{aligned} \phi(x,y) &= \phi(x)\phi(y) \\ \psi_1(x,y) &= \psi(x)\phi(y) \\ \psi_2(x,y) &= \phi(x)\psi(y) \\ \psi_3(x,y) &= \psi(x)\psi(y) \end{aligned} \qquad (5)$$

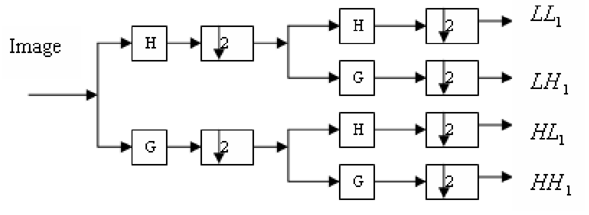

Fig.1. One level filter bank for computation of 2-D DWT

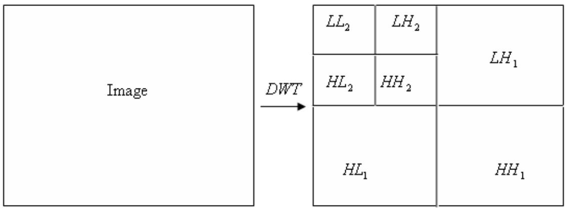

Fig.2. Two-level 2-D decomposition

2.1.2 Wavelet families

Wavelet families can be divided into two main categories, orthogonal and Biorthogonal wavelets, which have different properties of basis functions. Orthogonality decorrelates the transform coefficients by minimizing redundancy. Symmetry provides linear phase and minimize border artifacts. Other important properties of wavelet functions in image processing applications are compact support, symmetry, regularity and degree of smoothness. Haar and Daubechies wavelets are the most popular wavelets. They represent the foundations of wavelet signal processing and are used in various applications. The Haar, Daubechies, Symlets and Coiflets are compactly supported orthogonal wavelets.

2.2 Wavelet features

Feature extraction is a crucial processing step of shape recognition systems. In fact, what most distinguishes different recognition methodologies is the type of features used for representation. In this section, we describe the proposed DWT-based feature extraction method. The advantage of this transform is that it performs a local analysis to characterize the image in time and scale space and on the other hand its ability of reducing dimensionality of the initial feature space. The decomposition of the digit image of size 16x16 at one level of resolution results in four sub-images including one approximation (low-frequency) and three details (high-frequency) sub-images. Each sub-image is 8x8 of size. Only the coefficients of approximation sub-image obtained at this level of decomposition are considered as features. The idea behind the choice of this sub-image is that the approximation coefficients usually contain the most important information, and hence, they will constitute part of the extracted features.

2.3 Wavelet feature selection

Feature selection is important in many pattern recognition problems for finding an optimal or suboptimal subset of features out of original features. Selection of potential features allows reducing system complexity and processing time and therefore, achieves high recognition accuracy (performance).

The hybrid method presented here is based on principal component analysis followed by sequential forward selection (SFS) algorithm [20]. It chooses the features based on the characteristics of the data with taking into account the advantage of the two models. In the SFS approach, the best subset of features F is initialized as the empty set and in each step we add to F the feature that gives the highest correct classification rate along with the features already included in F. The process continues until the correct classification rate given by F and each of the features not yet selected does not increase.

In this paper, for handwritten digit recognition, the hybrid method is used to optimize the wavelet feature vector obtained in sub-section (2.2) since it contains a large number of components (8x8 values). Hence, the selected features are sent to a support vector machines classifier.

3 Support vector machines classifier

Support vector machines (SVMs) introduced by Vapnik [19] are a relatively new learning process influenced highly by advances in statistical learning theory and a sufficient increase in computer processing power in recent years. In the last ten years, SVMs have led to a growing number of applications such as texture classification and recently, they have been extended to multiclass applications like handwritten character recognition [1]. Before the discovery of SVMs, machines were not very successful in learning and generalisation tasks, with many problems being impossible to solve.

The purpose of SVM is to find an optimal linear classifier (optimal hyperplane) which generates the maximum margin between the two data sets in the feature space and theoretically this is based on the structural risk minimisation theory of statistical pattern recognition [19]. A further important concept in SVM is the transformation of data into a higher dimensional space for the construction of optimal separating hyperplane. SVM perform this nonlinear mapping into a higher dimensional feature space by means of a kernel function and then construct a linear optimal separating hyperplane between the two classes in the feature space. Those data vectors nearest to the constructed line in the transformed space are called the support vectors (SV) and contain valuable information required for classification.

The SVM takes an input vector $x \in \Re^d$ which is mapped into a higher dimension feature space $\mathcal{F}$ by $z = \phi(x)$ via a nonlinear mapping $\phi : \Re^d \rightarrow$ $\mathcal{F}$. This feature vector is classified to one of the two classes by linear classifier as:

$$y = f(x) = \operatorname{sgn}(\langle w, \phi(x) \rangle + b) , \qquad y \in \{-1,1\}$$

Where, the optimal separating hyperplane in the feature space is defined by the weight vector $w \in \mathcal{F}$ and a scalar $b \in \Re$.

Based on supervised learning, the parameters w and b are determined by using a training set composed of N data, $(x_1, y_1),...,(x_i, y_i)$ and $i = 1...N$.

To find the optimal hyperplane for separable data, we solve the quadratic optimization problem given by:

Minimise $\frac{1}{2}\|w\|^2 - C\sum_{i=1}^{N}\xi_i$

Subject to $y_i(\langle w, \phi(x_i) \rangle + b) \geq 1$, $i = 1...N$ (6)

Using Lagrange multipliers, it can be shown that the linear support vector training problem is reduced to:

Maximize

$$\max_{\alpha} \sum_{i}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (7)$$

Subject to $\sum_{i=1}^{N} \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$, $i = 1..N$

With α_i denoting a Langrange multipliers and C is a constant which controls the trade-off between the complexity of the decision function and the number of training examples misclassified.

The kernel function $K(,)$ describes the inner product $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$.

To build an SVM classifier, the user needs to tune C and choose a kernel function and its parameters. Some commonly used kernels include Gaussian, Radial Basis Functions, and Polynomials.

As mentioned before, the originally SVM was designed for binary classification. To extend it to multiclass problem is still an ongoing research issue. Currently the main approach for multiclass SVM is by constructing and combining several binary classifiers while the other is by directly considering all data in one optimization formulation. There are two methods included in the approach: One-against-all (OAA) and One-against-one (OAO).

4 Results and discussion

In this section, we present some experiments using the hybrid feature selection method in order to reduce and optimize the wavelet feature vector presented in section 2.2.

Some different wavelet families are considered such as: Haar wavelet, symlet wavelets (sym4), (sym8), coiflet wavelet (coif1) and Daubechies wavelet (db4) in order to compare their performance.

The training and testing data sets used for the experiment consists of handwritten digits extracted from the USPS database [13]. It contains grey scale digit images of size 16x16 pixels. Some samples are shown in figure 3. Hence, in this experiment, we have randomly extracted 60 digits per class to build the training set and 40 digits per class for testing. For classification, we used the SVM classifier [15], which is based on LIBSVM. The hyper parameters of the SVM classifier are fixed such as: $C = 10$ and gamma=0.256 for the RBF kernel function.

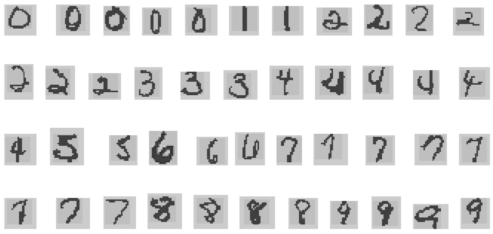

Fig. 4. Some samples of USPS database

After one level of decomposition of the digit image of size 16x16, the wavelet coefficients of approximation image of size 8x8 are obtained. Finally, the normalized coefficients are used as features. To reduce the dimension of this feature vector and hence, select the potential features, a hybrid feature selection strategy PCA-SFS is used in this paper. Table 1 summarizes the results found by PCA , SFS and PCA-SFS approaches using different types of wavelet and Figure 4 shows the trade-off the between recognition rates and the number of features selected in the case of the symlet wavelet transform (sym4).

Table 1. Results of recognition rates

Type of wavelet	Wavelet features	R. Rates	Wavelet features+PCA	R. Rates	Wavelt features+SFS	R. Rates	Wavelet features+PCA+SFS	R. Rates
Haar	64	95.00%	29	95.50%	36	96.25%	23	96.75%
Db4	64	95.50%	32	96%	43	97%	31	97.25%
Coifl	64	94.75%	39	96.75%	42	97%	25	97.25%
Sym4	64	96.50%	39	96.50%	36	96.75%	30	97.25%
Sym8	64	96.25%	35	96.75%	52	97%	29	97.25%

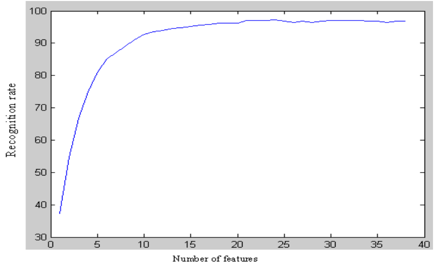

Fig.5. Performance of the PCA-SFS method

The results obtained in this experiment (Table 1) show that the proposed hybrid model based on PCA followed by SFS method outperform both the PCA and SFS methods. The recognition rate of 97.25% is reached for db4, coif1, sym4 and sym8 wavelets, but a smaller number of features are used only in the case of coif1 wavelet transform.

It can be observed that the results obtained in this experiment using the wrapper approach based on SFS technique combined with the SVM classifier are significantly higher than those obtained by considering all the initial features. Moreover, such results are obtained using a much smaller number of features. For example, by using the symlet wavelet (sym4), a subset of only 36 suitably selected features, among the whole set of 64, is sufficient for the SVM classifier to get 96.75% recognition rate, against 96.50% .

From these results, we can see also that the Db4 wavelet, the coif1 wavelet and the sym8 wavelet give the higher recognition rates which reach 97%, but the number of features is widely increased to reach 43, 42 and 52 respectively.

It has been also observed that the results obtained using the PCA approach are inferior to those of the SFS technique.

From this contestation, in practice, it is important to choose a compromise between feature number and achievable recognition rate for reducing the cost and complexity of the system.

4 Conclusion

In the present paper, we have proposed the hybrid model based on PCA-SFS method for wavelet feature subset selection. For the robustness of our approach, we have used the support vector machines (SVM) classifier, known for its performance, as the evaluation criterion. It searches for potential features aiming to improve the accuracy of the system.

The interesting results obtained here shown that this method was succeed in reducing both the number of features and error rate of the classifier.

Among the wavelets used here, it can be observed that the coif1 wavelet transform is more suitable for this application.

For comparison, in future work, we plan to study different approaches for feature selection such as genetic algorithm, sequential forward floating selection and sequential backward floating selection as well as to apply different schemes of representation for our problem.

References:

[1] C. Bahlmann, B. Haasdonk, H. Burkhardt, On-line Handwriting Recognition with Support Vector Machines - A Kernel Approach, *Publ. in Proc. of the 8th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2002, pp. 49–54.

[2] U. Bhattacharya, B.B. Chaudhuri, A majority voting schema for multiresolution recognition of handprinted numerals, *Proceeding of the seventh international conference on document analysis and recognition*, 2003.

[3] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence*, Vol.97, 1997, pp.245–271.

[4] M. Bressan, J. Vitria, On the Selection and Classification of Independent Features. *IEEE Transactions on PAMI*, Vol.25, N°.10, 2003, pp.1312-1317.

[5] S.E.N. Correia, J. Carvalho, Recognition of unconstrained handwritten numerals using biorthogonal spline wavelets, *Proceeding of the XII brazilian symposium on compter graphics and image processing*, 2000.

[6] S. Das, Filters, wrappers and a boosting-based hybrid for feature selection. *In Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 74–81.

[7] M. Dash, K. Choi, P. Scheuermann, H. Liu, Feature selection for clustering – a filter solution, *In Proceedings of the Second International Conference on Data Mining*, 2002, pp. 115–122.

[8] K. Fukunaga, W.L.G. Koontz, Application of the Karhunen-Loeve Expansion to Feature Selection and Ordering, *IEEE Transaction on Computer*, Vol.19, N°.4 ,1970, pp.311-318.

[9] A. Jain, D. Zongker, Feature selection: Evaluation, application, and small sample performance, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.19, N°.2, 1997 pp. 153–158.

[10] G. Kim, S. Kim, Feature selection using genetic algorithms for handwritten character recognition, *In 7th IWFHR*, Amsterdam-Netherlands, 2000, pp.103–112.

[11] R. Kohavi, G. John, Wrappers for feature selection. Artificial Intelligence, 1997, pp 273-324.

[12] P. Langley, Selection of relevant features in machine learning, *In Proceedings of the AAAI Fall Symposium on Relevance*, 1994, pp. 140–144.

[13] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner Gradient based learning applied to document recognition, *Proc. of the IEEE*, Vol.86, N°.11, 1998, pp. 2278–2324.

[14] S.W. Lee, C.H. Kim, H. Ma et al, Multiresolution recognition of unconstrained handwritten numerals with wavelet transform and multi-layer cluster neural network, *Pattern Recognition*, Vol.29, N°.12, 1996, pp.1953-1961.

[15] J. Ma, Y. Zhao, S. Ahalt, OSU SVM classifier matlab toolbox, 2002.

[16] S. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.11, 1989, pp.674-693.

[17] P. Pudil, J. Novovicova, J. Kittler, Floating Search Methods in Feature Selection, *Pattern Recognition Letters*, Vol.15, N°.11, 1994:119-1125.

[18] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, L. Lam, Computer Recognition of Unconstrained Handwritten Numerals, *Proceedings of the IEEE*, Vol. 80,N°.70, 1992:1162-1180.

[19] V. Vapnik, The Nature of Statistical Learning Theory, *Springer Verlag*, NewYork, 1995.

[20] A.W. Whitney, A direct method of nonparametric measurement selection. *IEEE Trans. Comput.* Vol.20, 1971, pp.1100-1103.

[21] E. Xiang, M. Jordan, R. Karp, Feature selection for high dimensional genomic microarray data, *In Proc 8th Int Conf Machine Learning*, Williams College, Massachusetts, 2001 .