Feasible Adaptation Criteria for Hybrid Wavelet – Large Margin Classifiers

Julia Neumann, Christoph Schnörr, and Gabriele Steidl

Dept. of Mathematics and Computer Science
University of Mannheim, D-68131 Mannheim, Germany
http://www.cvgpr.uni-mannheim.de, http://kiwi.math.uni-mannheim.de
{jneumann,schnoerr,steidl}@uni-mannheim.de

Abstract. Hybrid wavelet – large margin classifiers have recently proven to solve difficult signal classification problems in cases where merely using a large margin classifier like, e.g., the Support Vector Machine may fail. The features for our hybrid classifier are selected from the outputs of *all* orthonormal filter banks of fixed length with respect to criteria measuring class separability and generalisation error.
In this paper, we evaluate a range of such adaptation criteria to perform feature selection for hybrid wavelet – large margin classifiers. The two main points we focus on are (i) approximation of the radius – margin error bound as the ultimate criterion for the target classifier, and (ii) computational costs of the approximating criterion for feature selection relative to those for the classifier design.
We show that by virtue of the adaptivity of the filter bank, criteria which are more efficient than computing the radius – margin are sufficient for wavelet adaptation and, hence, feature selection. Our results are relevant for image– and arbitrary–dimensional signal classification by utilising the standard tensor product design of wavelets.

1 Introduction

Motivation. A persistent problem in signal and image classification concerns filter design for feature extraction and selection [6, 7, 11]. In most cases, this problem is addressed *irrespective of* the subsequent classification stage. However, using 'off–the–shelf' filters like Daubechies' wavelets [2] may result in an unacceptably large classification error. Fig. 1 shows a typical example for a difficult signal classification problem.

In this context, our approach is to take the target classifier and data into consideration for filter design and the selection of appropriate features. Given a sample set of labelled patterns, the main idea is to *adapt* the filter bank based on a criterion measuring class separability and generalisation error to obtain the *optimal features* for the particular problem under consideration.

It has recently been shown for a number of difficult applications that *jointly* designing both the filter stage and the classifier in this way may considerably outperform standard approaches based on a *separate* design of both stages [10].

Fig. 1. Two–class problem (heart beats: sinus rhythm (SR) and ventricular tachycardia(VT)): Choosing standard wavelets for feature extraction may result in a classification error up to 31%!

This motivates the investigation of suitable adaptation criteria which is summarised in the present paper.

Problem statement. The target classifier in our hybrid approach is the Support Vector Machine (SVM) which is well known to belong to the most competitive approaches and has favourable properties from the perspective of optimisation during the learning stage [12]. Accordingly, a suitable criterion for feature selection is the radius – margin bound which captures the generalisation error [12]. The direct application of this criterion to feature selection has been studied in [13].

In the hybrid approach studied here, however, the objective function with respect to the filters is quite complex and can be minimised by exhaustive search only. In contrast to related work [3], this is nevertheless computationally feasible and efficient in our case, due to the lattice factorisation of orthonormal filter banks (see Sec. 2 and [5, Sec. 5.3]).

On the other hand, determining the optimally adapted filter bank requires many evaluations of the objective function. This is no longer computationally feasible if the objective function is based on a criterion the evaluation of which is as time consuming as the design of the classifier itself! Since this holds for the radius – margin bound as criterion of our target classifier – each evaluation requires to solve two quadratic programs! –, approximations of this criterion have to be investigated which are suitable for the overall design of the hybrid approach.

Organisation of the paper. We summarise the hybrid architecture in Sec. 2. Next, in Sec. 3, we discuss a range of criteria in view of the problems stated above, along with a confirmation and illustrations by numerical evaluations in Sec. 4.

2 Hybrid Wavelet – SVM Architecture

In this section we briefly introduce our hybrid architecture for feature extraction and subsequent classification of the resulting feature vectors.

Feature extraction. Our feature extraction process consists of two steps, namely filtering by an orthogonal two-channel octave band filter bank, and energy computation of the resulting coefficients in the different frequency bands.

We deal with input signals $\mathbf{s} \in \mathbb{R}^N$, where N is a power of 2. Fundamental for our filter adaptation process is that any orthogonal two-channel filter bank with filters of length $2L+2$ is determined by L angles $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{L-1}) \in [0, \pi)^L$ by the so–called lattice decomposition of the corresponding polyphase matrix [9, Theorems 4.6 and 4.9]. Filtering by the d-level octave band filter bank given by $\boldsymbol{\theta}$ can then be considered as orthogonal discrete wavelet transform

$$F_{\boldsymbol{\theta}} : \mathbb{R}^N \to \mathbb{R}^N \ , \ \mathbf{s} \mapsto \left(\mathbf{c^d}, \mathbf{d^d}, \ldots, \mathbf{d^1}\right) \ ,$$

which maps the input signal $\mathbf{s}$ to its wavelet coefficients $\mathbf{d^j} = (d^j_1, \ldots, d^j_{N/2^j})$ in the jth frequency band, $j = 1, \ldots, d$. The mapping $F_{\boldsymbol{\theta}}$ is norm preserving with respect to the Euclidean norm $\|\cdot\|_2$, i.e., $\|F_{\boldsymbol{\theta}}\mathbf{s}\|_2 = \|\mathbf{s}\|_2$.

To generate a handy number of features that still make the signals well distinguishable, we introduce the energy operator

$$E_{\|\,\|} : \mathbb{R}^N \to \mathbb{R}^d \ , \ \left(\mathbf{c^d}, \mathbf{d^d}, \ldots, \mathbf{d^1}\right) \mapsto \left(\|\mathbf{d^d}\|, \ldots, \|\mathbf{d^1}\|\right) \ .$$

As possible norms for $E_{\|\,\|}$ we consider besides the Euclidean norm the weighted Euclidean norm $\sqrt{\frac{1}{n}\sum_{i=1}^n c_i^2}$, which was proposed by Unser [11] to represent the channel variance. Other Hölder norms may be used as well.

In summary our feature extraction process produces the feature vectors $\mathbf{x} := E_{\|\,\|}F_{\boldsymbol{\theta}}\mathbf{s}$. For later considerations it is important that the norm preserving property of the orthogonal wavelet transform implies

$$\|E_{\|\,\|}F_{\boldsymbol{\theta}}\mathbf{s}\|_2 \le \|\mathbf{s}\|_2 \ . \tag{1}$$

In our experiments we deal w.l.o.g. with input signals $\mathbf{s}$ with fixed Euclidean norm and average value zero and apply the full wavelet decomposition, i.e., $N/2^d = 1$. Then it is easy to check that $\mathbf{c}^d = 0$. Now (1) implies that the feature vectors lie within a sphere in $\mathbb{R}^d$ centred at the origin. Moreover, if we use the Euclidean norm in $E_{\|\,\|}$, then we have equality in (1).

Classification. To rate a set of feature vectors according to their classification ability, it is essential to take into account the classifier in use. We intend to apply a SVM as classifier. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$ containing the feature vectors. Given a training set $\mathcal{Z} := \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\} : i = 1, \ldots, n\}$ of n associations, we are interested in the construction of a real valued function f defined on $\mathcal{X}$ such that $\operatorname{sgn}(f)$ well predicts the class labels y. Let $\mathbf{y} := (y_1, \ldots, y_n)$ denote the vector of class labels and let $\mathbf{Y} := \operatorname{diag}\mathbf{y}$. We introduce a so–called *kernel* function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is square integrable, positive definite and symmetric. In our applications we will use Gaussian kernels $K(\mathbf{x}, \mathbf{y}) := e^{-\frac{||\mathbf{x}-\mathbf{y}||_2^2}{2\sigma^2}}$ where $\sigma > 0$. With the kernel K we associate the kernel matrix $\mathbf{K} := (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$. Then the standard SVM finds f as linear combination

$$f(\mathbf{x}) = \sum_{i=1}^n y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) \ , \tag{2}$$

where the coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ are given by the solution of the quadratic optimisation problem (QP)

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{YKY}\boldsymbol{\alpha} + \mathbf{e}^T\boldsymbol{\alpha} \quad \text{subject to} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{e} \ . \tag{3}$$

For $C = \infty$ the resulting classifier is called *hard margin SVM*, otherwise *soft margin SVM*. The *support vectors* (SVs) are those training patterns $\mathbf{x}_i$ for which the coefficients α_i in the solution of (3) do not vanish. Then the sum (2) involves only SVs. The *margin* separating the classes is defined by $\rho := (\boldsymbol{\alpha}^T \mathbf{YKY}\boldsymbol{\alpha})^{-\frac{1}{2}}$. Note that (3) originates from the unconstrained optimisation problem

$$\min_{f \in \mathcal{H}_K} C\sum_{i=1}^{n}(1 - y_i f(\mathbf{x}_i))_+ + \frac{1}{2}||f||^2_{\mathcal{H}_K}, \qquad (\tau)_+ := \begin{cases} \tau & \text{if } \tau \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{H}_K$ denotes the reproducing kernel Hilbert space associated with K. For details see [12].

3 Criteria for Feature Adaptation

To steer the parameters $\boldsymbol{\theta}$ in our feature extraction process according to the subsequent SVM classifier we want to find a measure that allows for fast comparison of different sets of feature vectors based on maximising the SVM performance. In this paper, we restrict our attention to hard margin SVMs for simplicity. All results can be formulated for soft margin SVMs as well [4]. Possible criteria for adaptation are obtained by bounds for the *generalisation error*, i.e., the probability that $\text{sgn} f(\mathbf{x}) \neq y$ for a randomly chosen example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, 1\}$. In our experiments we investigate five criteria:

Radius – Margin. Since the expectation of the quotient

$$\mathcal{C}_1 := \frac{1}{n}\frac{R^2}{\rho^2} \tag{4}$$

forms an upper bound on the SVM generalisation error [12, Theorem 10.6] we consider a minimal value $\mathcal{C}_1$ as the ultimate criterion for the SVM classifier. Here R is the radius of the smallest sphere in $\mathcal{H}_K$ enclosing all $K(\cdot, \mathbf{x}_j)$, i.e., the solution of

$$\min_{a \in \mathcal{H}_K, R \in \mathbb{R}} R^2 \quad \text{subject to} \quad \|K(\cdot, \mathbf{x}_j) - a\|^2_{\mathcal{H}_K} \leq R^2 \ , \quad j = 1, \ldots, n \ . \tag{5}$$

In [4] we proved that (5) can be also solved by the QP (3):

Proposition 1. *Let K be a kernel with $K(\mathbf{x}, \mathbf{x}) = \kappa \ \forall \mathbf{x} \in \mathcal{X}$. Then the optimal radius R in* (5) *can be obtained by solving* (3) *with $\mathbf{Y} = \mathbf{I}$. If $\boldsymbol{\alpha}$ is the solution of* (3) *and j an index of a SV, then $R^2 = \kappa + \boldsymbol{\beta}^T \mathbf{K}\boldsymbol{\beta} - 2(\mathbf{K}\boldsymbol{\beta})_j$, where $\boldsymbol{\beta} := \frac{\boldsymbol{\alpha}}{\mathbf{e}^T\boldsymbol{\alpha}}$.*

However, the computation of ρ and R in (4) still requires the solution of two QPs for each parameter vector $\boldsymbol{\theta}$.

Margin. Due to (1), the radius R is bounded. This motivates to consider only the denominator of (4), i.e., to use a maximal $\mathcal{C}_2 = \rho$ as objective criterion. Indeed, our experiments indicate that if training and test data have the same underlying distribution, the margin behaves much like the classification error.

Alignment. As a measure of classification ability for kernel problems, the alignment

$$\mathcal{C}_3 := \frac{\mathbf{y}^T \mathbf{K} \mathbf{y}}{n \|\mathbf{K}\|_F} \tag{6}$$

with Frobenius norm $\|\cdot\|_F$ was proposed in [1]. By [1, Theorem 4], the generalisation accuracy of the expected Parzen window estimator which is related to an SVM is bounded by a function of the alignment.

Class Centre Distance. In all our experiments, the denominator in (6) doesn't influence the alignment much. Furthermore, supposing normed training vectors $\|\mathbf{x}_i\|_2 = c$ and a Gaussian kernel with large deviation σ, the numerator in (6) is approximately proportional to $\mathbf{y}^T (\langle \mathbf{x}_i, \mathbf{x}_j \rangle)_{i,j=1}^n \mathbf{y}$. Introducing the class means $\boldsymbol{\mu}_i := \frac{1}{n_i} \sum_{y_j = i} \mathbf{x}_j$ with class cardinalities n_i ($i = \pm 1$), for $n_1 = n_{-1}$ this can be rewritten as $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}\|_2^2$. The criterion $\mathcal{C}_4 := \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}\|_2$ can be simply evaluated and is also easily differentiable. It was successfully applied in [10].

Scatter Measures. While $\mathcal{C}_4$ only takes into account the mean values of the classes we are now looking for classes that are distant from each other and at the same time concentrated around their means. A generalisation of $\mathcal{C}_4$ are measures using scatter matrices. We consider the generalised Fisher criterion

$$\mathcal{C}_5 := \frac{\operatorname{tr}(\mathbf{S}_b)}{\operatorname{tr}(\mathbf{S}_w)} = \frac{\frac{n_1}{n} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}\|^2 + \frac{n_{-1}}{n} \|\boldsymbol{\mu}_{-1} - \boldsymbol{\mu}\|^2}{\frac{n_1}{n} \sum_{k=1}^d \sigma_{1k}^2 + \frac{n_{-1}}{n} \sum_{k=1}^d \sigma_{-1k}^2}, \quad \boldsymbol{\mu} := \sum_{i \in \{-1,1\}} \frac{n_i}{n} \boldsymbol{\mu}_i$$

where σ_{ik}^2 is the marginal variance of class i along dimension k and

$$\mathbf{S}_w := \frac{1}{n} \sum_{i \in \{-1,1\}} \sum_{y_j = i} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T, \quad \mathbf{S}_b := \sum_{i \in \{-1,1\}} \frac{n_i}{n} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

denote the *within–class scatter matrix* and the *between–class scatter matrix*, respectively. For equiprobable classes, $\mathcal{C}_5$ is proportional to $\mathcal{C}_4^2 / \sum_{k=1}^d (\sigma_{1k}^2 + \sigma_{-1k}^2)$.

4 Numerical Evaluation

So far we have proposed several criteria for judging the discrimination ability of a set of feature vectors and have shown some connections between the criteria. We now want to see how these links show up when analysing real data.

We use two structurally different real data sets: The first electro–physiological data set originates from the detection of ventricular tachycardia as in [10]. For

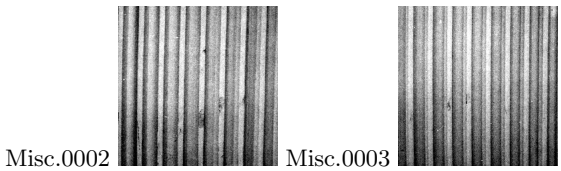


Fig. 2. texture sample: linearly rescaled images

each patient, eight beats from a single episode are used for classifier training. Some exemplary beats for a sample patient are shown in Fig. 1. The second group of data are the texture images from the MeasTex collection [8]. We use single rows of the corrugated iron images 'Misc.0002' and 'Misc.0003' shown in Fig. 2 to have one–dimensional data as in the first data set. Here, the first 32 rows of each texture are used for classifier training. We normalised all samples by $\|\mathbf{s}_i\|_2 = 1000$ and set their average value to zero.

We apply orthogonal filter banks with filters of length ≤ 6 which can be parameterised by two angles $\boldsymbol{\theta} = (\theta_0, \theta_1) \in [0, \pi)^2$. The parameter space was discretised with 128 angles per dimension. For the classification, a hard–margin SVM with Gaussian kernel of width $\sigma = 100$ is used. The highest alignment $\mathcal{C}_3$ is achieved with $\sigma \approx 150$ and $\sigma \approx 80$ for the Euclidean and the weighted Euclidean norm, respectively.

To control the filter design, we generate plots that show the values of the five criteria subject to the two–dimensional parameter space. The values are plotted using a linear grey scale except for the radius – margin bound which is plotted on a logarithmic scale due to its large variation. Additionally, the larger values are clipped to the trivial error bound 1 to enhance the contrast. To assess the effect of the clipping, the distribution of the logarithm of the bound is indicated by a histogram. The resulting images are shown in Fig. 3, where the plots (a) – (e) are ordered from the simplest and computationally most efficient criterion to the most expensive one.

For all three problems, the overall impression is that all shown criteria are alike. Moreover, all criteria show a detailed structure for the parameter space. This indicates that effectively finding the optimal wavelet according to the chosen criterion is not easy even for the simple criteria. The class centre distance and particularly the alignment resemble the margin. That is, the wavelets that generate a high class centre distance or alignment also guarantee a large margin. Although the scatter criterion $\mathcal{C}_5$ also takes into account the variances, it doesn't seem to be superior to the simplest criterion $\mathcal{C}_4$.

The radius – margin bound $\mathcal{C}_1$ covers a large range of values from 10 resp. 3% to 100%. This indicates the significance of the wavelet choice which is also emphasised in Fig. 4. Apart from the different distribution of the values, the radius – margin bound rates the features mostly like the margin.

For specific signals there may be an important difference between using the Euclidean and the weighted Euclidean norm as exhibited by Fig. 3–2 and 3–3.

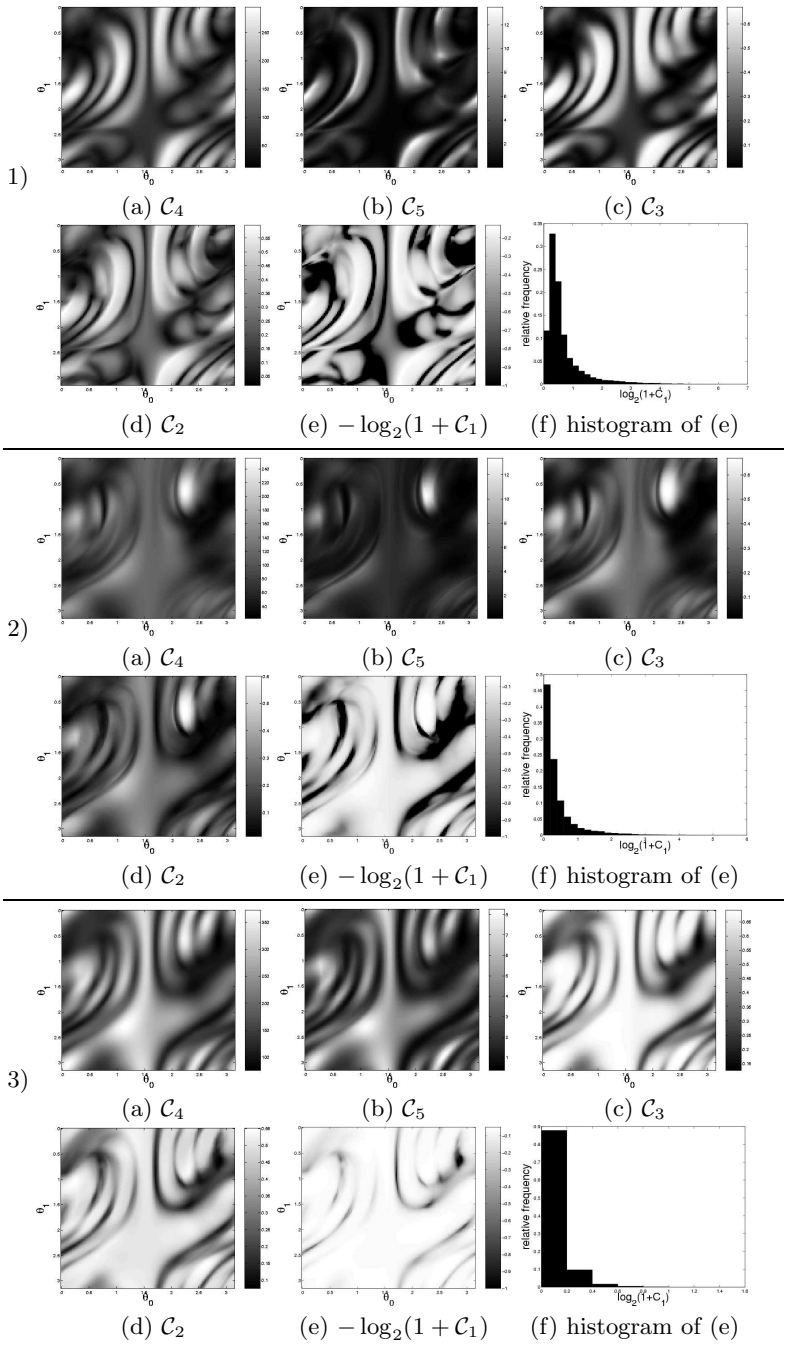
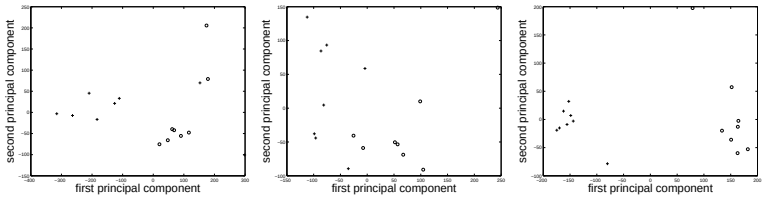


Fig. 3. Criteria values for 1) heartbeat classification with weighted Euclidean norm in $E_{\|\,\|}$, 2) texture row classification with weighted Euclidean norm in $E_{\|\,\|}$, 3) texture row classification with Euclidean norm in $E_{\|\,\|}$; light spots represent favourable criterion values and, hence, beneficial filter banks



(a) (b) (c)

Fig. 4. Principal components of training vectors for heartbeat classification with Euclidean norm in $E_{\|\,\|}$: (a) for the Haar wavelet, (b) for the Daubechies wavelet with three vanishing moments, (c) for the optimally aligned wavelet ($\mathcal{C}_3$); these results show that wavelet adaptation may considerably improve class separability

The plots show that simple adaptation criteria suffice to promisingly design filters for hybrid wavelet – large margin classifiers with Gaussian kernels.

Acknowledgements. This work is funded by the DFG, Grant Sch 457/5-1.

References

1. N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In *NIPS*, volume 14, pages 367–373. The MIT Press, 2002.
2. I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41:909–996, 1988.
3. E. Jones, P. Runkle, N. Dasgupta, L. Couchman, and L. Carin. Genetic algorithm wavelet design for signal classification. *IEEE TPAMI*, 23(8):890–895, 2001.
4. J. Neumann, C. Schnörr, and G. Steidl. Feasible adaptation criteria for hybrid wavelet – large margin classifiers. Technical Report TR-02-015, Dept. of Mathematics and Computer Science, University of Mannheim, 2002.
5. J. Neumann, C. Schnörr, and G. Steidl. Effectively finding the optimal wavelet for hybrid wavelet – large margin signal classification. Technical Report TR-03-005, Dept. of Mathematics and Computer Science, University of Mannheim, 2003.
6. T. Randen and J. H. Husøy. Filtering for texture classification: A comparative study. *IEEE TPAMI*, 21(4):291–310, Apr. 1999.
7. P. Scheunders, S. Livens, G. Van de Wouwer, P. Vautrot, and D. Van Dyck. Wavelet-based texture analysis. *International Journal on Computer Science and Information Management*, 1(2):22–34, 1998.
8. G. Smith. MeasTex image texture database and test suite. Available at `http://www.cssip.uq.edu.au/meastex/meastex.html`, May 1997. Version 1.1.
9. G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley, 1996.
10. D. Strauß and G. Steidl. Hybrid wavelet-support vector classification of waveforms. *Journal of Computational and Applied Mathematics*, 148:375–400, 2002.
11. M. Unser. Texture classification and segmentation using wavelet frames. *IEEE Transactions on Image Processing*, 4(11):1549–1560, 1995.
12. V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
13. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In *NIPS*, pages 668–674. The MIT Press, 2000.