Efficient Wavelet Adaptation for Hybrid Wavelet–Large Margin Classifiers

Julia Neumann Christoph Schnörr *
Gabriele Steidl

Dept. of Mathematics and Computer Science
University of Mannheim
D-68131 Mannheim, Germany
{jneumann,schnoerr,steidl}@uni-mannheim.de

February 2, 2005

Abstract

Hybrid wavelet – large margin classifiers have recently proven to solve difficult signal classification problems in cases where solely using a large margin classifier like, e.g., the Support Vector Machine may fail. In this paper, we evaluate several criteria rating feature sets obtained from various orthogonal filter banks for the classification by a Support Vector Machine. Appropriate criteria may then be used for adapting the wavelet filter with respect to the subsequent support vector classification. Our results show that criteria which are computationally more efficient than the radius – margin Support Vector Machine error bound are sufficient for our filter adaptation and, hence, feature selection. Further, we propose an adaptive search algorithm that, once the criterion is fixed, efficiently finds the optimal wavelet filter. As an interesting byproduct we prove a theorem which allows the computation of the radius of a set of vectors by a standard Support Vector Machine.

Keywords. filter design, feature selection, signal and image classification, Support Vector Machine, wavelets

*corresponding author: Fax +49 621 181 2744

1 Introduction

A persistent problem in signal and image classification concerns filter design for feature extraction and selection [1, 2, 3]. In most cases, this problem is addressed *irrespective of* the subsequent classification stage which may result in an unacceptably large classification error. In contrast, we are interested in an approach which takes the target classifier and the data into consideration for filter design and the selection of appropriate features. In [4], a hybrid architecture was introduced consisting of a wavelet transform with an energy map and a classifier applied to the resulting feature vectors. As target classifier the Support Vector Machine (SVM) [5] was suggested which is highly flexible and belongs to the most competitive approaches. For various applications, it was shown that the classification error depends on the filters used in the wavelet transform and that *jointly* designing both the filter stage and the classifier may considerably outperform standard approaches based on a *separate* design of both stages. In contrast to best basis methods [6], the wavelet itself was adapted while the structure of the basis remained fixed [7]. However, although there exist more sophisticated measures for estimating the classification ability of training sets, only the simple class centre distance was used to adapt the feature extraction step, i.e., the wavelet filter, to the subsequent SVM classifier. Moreover, the computation of the optimal filters was very expensive even with the proposed genetic algorithm [7].

This motivates our investigation of suitable adaptation criteria and the design of algorithms for their efficient optimisation which is addressed in the present paper. Obviously, the most appropriate measures to evaluate the classification ability of a training set are generalisation error bounds, the most common of which is the radius – margin bound for SVMs [5]. The direct application of this criterion to feature selection has been studied in [8], but since we take filter optimisation into account, we have to deal with more complex objective functions here. Hence, we focus on reasonable adaptation criteria that require less computational effort. In this paper, we evaluate five common criteria. Having selected one of those criteria, we propose a robust grid search heuristic for efficiently finding the global optimum over the resulting parameter space that succeeds in solving our problems in acceptable time.

Our results are relevant for image-, and arbitrary dimensional signal classification by utilising the standard tensor product design of wavelets.

This paper is organised as follows: first we introduce the hybrid wavelet – SVM archi-

tecture in Sec. 2. Next, in Sec. 3, we discuss a range of criteria that approximate the generalisation error. Moreover, we provide a theorem which simplifies the computation of the radius – margin error bound and which may also be interesting in other contexts. Sec. 4 contains a thorough numerical evaluation of the proposed criteria. In Sec. 5, we suggest an algorithm for effectively finding the optimal wavelet filters with respect to an arbitrary fixed criterion. Finally, we conclude and indicate further work in Sec. 6. The appendix gives an interesting relation between single class SVMs and support vector problems for novelty detection and clustering which also proves our theorem on the radius computation.

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.

2.1 Feature Extraction

Our feature extraction relies on filtering by a two–channel filter bank as illustrated in Fig. 1. Thereby, $H_0(z) := \sum_{k\in\mathbb{Z}} h_0[k]z^{-k}$ denotes the z–transform of the low–pass analysis filter coefficients $(h_0[k])_{k\in\mathbb{Z}}$ and $H_1(z) := \sum_{k\in\mathbb{Z}} h_1[k]z^{-k}$ the z–transform of the high–pass analysis filter coefficients $(h_1[k])_{k\in\mathbb{Z}}$, analogously for the synthesis filters. Moreover, $\boxed{2\uparrow}$ and $\boxed{2\downarrow}$ symbolise up– and downsampling by 2, respectively.

In this paper we are interested in *orthogonal* or *paraunitary* filter banks which can be characterised by the property

$$\begin{pmatrix} H_0(z) & H_0(-z) \\ H_1(z) & H_1(-z) \end{pmatrix} \begin{pmatrix} H_0(z^{-1}) & H_1(z^{-1}) \\ H_0(-z^{-1}) & H_1(-z^{-1}) \end{pmatrix} = 2\mathbf{I} \ .$$

Then the synthesis filters are given by $G_0(z) = H_0(z^{-1})$, $G_1(z) = H_1(z^{-1})$ and the orthogonality property ensures that $S(z) = \tilde{S}(z)$.

Fundamental for our application is another important property of orthogonal filter banks, namely the so–called *lattice factorisation* of the corresponding polyphase matrix. For details we refer to [9, Sec. 4.5]. Based on this factorisation every orthogonal filter pair (H_0, H_1) of length $2L + 2$ with at least one vanishing moment, i.e. $H_1(1) = 0$, is, up to filter translation and the sign of the high–pass filter, uniquely determined by a vector $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{L-1})$ consisting of L angles $\theta_l \in [0, \pi)$. In other words, there exists

a one–to–one correspondence between the π–periodic parameter space

$$\mathcal{P}_L := \{\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{L-1}) : \theta_l \in [0, \pi)\,,\ l = 0, \ldots, L-1\}$$

and the set of all orthogonal filters of length $2L + 2$ which is given in a constructive way by the lattice factorisation of the polyphase matrix of the filter bank. There exist alternative factorisations of orthogonal filter banks' polyphase matrices, see, e.g., [10].

In the following, we are interested in input signals $\mathbf{s} \in \mathbb{R}^l$ of length $l = k \cdot 2^d$ $(2 \nmid k)$ which are normalised with respect to the Euclidean norm, i.e. $\|\mathbf{s}\|_2 = \text{constant}$. In our application we only need the successively applied analysis filter bank. For the filtering by the d–level octave–band filter bank generated by $\boldsymbol{\theta}$, we define the filter operator

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

where $\mathbf{c^d}$ and $\mathbf{d^j} = (d_1^j, \ldots, d_{l/2^j}^j)$ $(j = 1, \ldots, d)$ are the subband coefficients as illustrated in Fig. 2. By [11, Sec. 3.3] there is a close relation between paraunitary filter banks and orthogonal wavelets and we refer to the filter operator $F_{\boldsymbol{\theta}}$ also as orthogonal wavelet transform which produces the wavelet coefficients $\mathbf{d^j}$. By the orthogonality of our filter bank, the mapping $F_{\boldsymbol{\theta}}$ is norm preserving with respect to the Euclidean norm , i.e., $\|F_{\boldsymbol{\theta}}\mathbf{s}\|_2 = \|\mathbf{s}\|_2$.

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

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

Note that in our experiments we always deal with input signals $\mathbf{s}$ having average value zero so that $\mathbf{e}^T\mathbf{c^d} = 0$. 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 [3] to represent the channel variance. Other Hölder norms may be used as well.

In summary the feature extraction process produces the feature vectors $\mathbf{x} := E_{\|\,\|}F_{\boldsymbol{\theta}}\mathbf{s}$ which depend on $\boldsymbol{\theta}$ and the chosen norm in $E_{\|\,\|}$ as illustrated in Fig. 3.

For later considerations it is important that by the norm preserving property of the orthogonal wavelet transform

$$\|\mathbf{x}\|_2 \le \|\mathbf{s}\|_2\ \ , \tag{2}$$

where we have equality if we use the Euclidean norm in $E_{\|\,\|}$ and $l = 2^d$. This implies that the feature vectors $\mathbf{x}$ lie within or on a sphere in $\mathbb{R}^d$ centred at the origin.

2.2 SVM 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 an SVM as classifier. For a detailed introduction to SVMs see [12].

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$ containing the feature vectors. We introduce a so–called *kernel* function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is square integrable, positive definite and symmetric. The kernel function K induces a *reproducing kernel Hilbert space* $\mathcal{H}_K := \overline{\text{span}\,\{K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}\}}$ of real valued functions on $\mathcal{X}$ with inner product satisfying

$$\langle f(\cdot), K(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_K} = f(\mathbf{x}) \quad \forall\, f \in \mathcal{H}_K \ .$$

Then, for $f = \sum_{j=1}^{N} c_j K(\mathbf{x}_j, \cdot) \in \mathcal{H}_K$, the norm $\|f\|_{\mathcal{H}_K}$ is given by

$$\|f\|^2_{\mathcal{H}_K} = \sum_{j,k=1}^{N} c_j c_k K(\mathbf{x}_j, \mathbf{x}_k) = \mathbf{c}^T \mathbf{K} \mathbf{c}$$

where $\mathbf{c} := (c_1, \ldots, c_N)^T$ and $\mathbf{K} := (K(\mathbf{x}_j, \mathbf{x}_k))_{j,k=1}^{N}$ is symmetric positive definite.

For a known training set

$$\mathcal{Z} := \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\} : i = 1, \ldots, n\} \tag{3}$$

of n associations, we are interested in the construction of a function $f \in \mathcal{H}_K$ such that $\text{sgn}(f)$ well predicts the class labels y. More precisely, the SVM intends to find $f \in \mathcal{H}_K$ as the solution of

$$\begin{aligned} \min_{f \in \mathcal{H}_K, \boldsymbol{\xi} \in \mathbb{R}^n} \ & C\left(\sum_{i=1}^{n} \xi_i\right) + \frac{1}{2}||f||^2_{\mathcal{H}_K} \\ \text{subject to} \quad & y_i f(\mathbf{x}_i) \geq \ 1 - \xi_i \ , \quad i = 1, \ldots, n \ , \\ & \xi_i \geq \qquad 0 \ , \quad i = 1, \ldots, n \end{aligned} \tag{4}$$

for some constant $C \in \mathbb{R}_+$ controlling the trade–off between the approximation error and the regularisation term. For the choice $C = \infty$, the resulting classifier is called *hard margin* classifier, otherwise *soft margin* classifier. By the *Representer Theorem* [13, 14], the minimiser of (4) has the form

$$f(\mathbf{x}) = \sum_{j=1}^{n} c_j K(\mathbf{x}, \mathbf{x}_j) \ . \tag{5}$$

In particular, the sum incorporates only our training vectors $\mathbf{x}_j$. Using this representation we set up the dual problem and obtain that f can be found by solving the following quadratic problem (QP) and setting $\mathbf{c} := \mathbf{Y}\boldsymbol{\alpha}$:

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \; & -\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Y}\mathbf{K}\mathbf{Y}\boldsymbol{\alpha} + \mathbf{e}^T\boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{0} \le \boldsymbol{\alpha} \le C\mathbf{e} \end{aligned} \tag{6}$$

where $\mathbf{Y} := \operatorname{diag}(y_1, \ldots, y_n)$ and $\mathbf{K} := (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$.

The *support vectors* (SVs) are those training patterns $\mathbf{x}_i$ for which the coefficients α_i in the solution of (6) do not vanish. Then the function f in (5) has a representation which only depends on the SVs. The so–called *margin* ρ is defined by

$$\rho := \|f\|_{\mathcal{H}_K}^{-1} = (\boldsymbol{\alpha}^T\mathbf{Y}\mathbf{K}\mathbf{Y}\boldsymbol{\alpha})^{-\frac{1}{2}} \quad . \tag{7}$$

Often it is more intuitive to work instead on $\mathcal{X}$ on a subspace of the Hilbert space ℓ_2 of square summable real valued sequences with inner product $\langle \mathbf{a}, \mathbf{b}\rangle_2 = \sum_{j=1}^\infty a_j b_j$ and norm $\|\mathbf{a}\|_2^2 = \sum_{j=1}^\infty a_j^2$. By the properties of our kernel K there exists a unique function, the so–called *feature map* $\boldsymbol{\phi} : \mathcal{X} \to \ell_2$, which is related to K by the property

$$K(\mathbf{x}_1, \mathbf{x}_2) = \langle \boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2)\rangle_2 \quad \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \quad . \tag{8}$$

Then our classifier f from (5) can be rewritten as $f(\mathbf{x}) = \langle \mathbf{f}, \boldsymbol{\phi}(\mathbf{x})\rangle_2$ with $\mathbf{f} = \sum_{j=1}^n c_j \boldsymbol{\phi}(\mathbf{x}_j)$. In other words, f becomes a linear function in the *feature space* $\boldsymbol{\phi}(\mathcal{X})$.

3 Criteria for Feature Adaptation

To steer our feature extraction process via the parameters $\boldsymbol{\theta}$ such that the subsequent SVM performance becomes optimal we need a criterion that

- measures the *generalisation error* of the SVM, i.e., the probability that $\operatorname{sgn}(f(\mathbf{x})) \neq y$ for a randomly chosen example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, 1\}$, and
- can be efficiently evaluated for different sets of $\boldsymbol{\theta}$–dependent feature vectors.

Although there exist many proven bounds for the error risk or its expectation in the literature (see, for example, [15, 16]), in essence, most of them rely either on the number of SVs (cf. [17, Theorem 5.2], [18] and [16, Sec. 5.2.1]) or on the size of the margin ρ separating the classes normalised by a measure of the feature vector variation such as

their radius (cf. [17, Theorem 5.2], [19]). In the following we start with this group of criteria. However, although they match the first criteria requirement they do not fulfil the second one. This motivates us to propose also simplified criteria and to compare their performance.

In our experiments we investigate five criteria: the radius–margin bound, the margin, the alignment, the class centre distance and the generalised Fisher criterion:

Radius – Margin Let the margin ρ be given by (7). Further let R be the radius of the smallest sphere in ℓ_2 enclosing all $\boldsymbol{\phi}(\mathbf{x}_j)$, i.e., the solution of

$$\begin{aligned} &\min_{\mathbf{a}\in\ell_2, R\in\mathbb{R}} \quad R^2 \\ &\text{subject to} \quad \|\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}\|_2^2 \leq R^2 \ , \quad j = 1, \ldots, n \ . \end{aligned} \tag{9}$$

Then the expectation of the quotient

$$\mathcal{C}_1(\boldsymbol{\theta}) := \frac{1}{n}\frac{R^2}{\rho^2} \tag{10}$$

forms an upper bound on hard margin SVMs' generalisation error [5, Theorem 10.6]. Therefore we consider a minimal value $\mathcal{C}_1$ as the ultimate criterion for a hard margin SVM classifier. At first glance the computation of ρ and R in $\mathcal{C}_1$ requires the solution of two structurally different optimisation problems (4) and (9). Fortunately, by the following theorem both ρ and R can be obtained by the same kind of QP (6). This is indeed very profitable since for standard SVMs (6), sophisticated algorithms are available in many implementations as, e.g., SVM*light* [20].

Theorem 1. *Let K be a kernel with $K(\mathbf{x}, \mathbf{x}) = \kappa$ for all $\mathbf{x} \in \mathcal{X}$. Then the optimal radius R in* (9) *can be obtained by solving* (6) *with $\mathbf{Y} = \mathbf{I}$. If $\boldsymbol{\alpha}$ is the solution of* (6) *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 $\boldsymbol{\beta} := \frac{\boldsymbol{\alpha}}{\mathbf{e}^T\boldsymbol{\alpha}}$.*

The proof of the theorem which also reveals an interesting relation to the SV problems used for clustering and novelty detection is given in the appendix.

Note that in the soft margin case, there also exists a radius margin bound. According to [21], the expectation of the generalisation error of the SVM is bounded from above by the expectation of the term

$$\frac{1}{n}\left(4R^2\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\xi_i\right)$$

where $\boldsymbol{\alpha}$ is the solution of (6) and $\boldsymbol{\xi}$ is defined by the primal problem (4).

Anyway, the computation of the bound still requires the solution of two QPs of the form (6) for each considered parameter vector $\boldsymbol{\theta}$ so that we look for simpler criteria.

Margin Due to (2), the radius R is bounded. This motivates to consider only the denominator of (10), i.e., to use a maximal

$$\mathcal{C}_2(\boldsymbol{\theta}) := \rho$$

as an 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.

Note that in the soft margin case, one can analogously use the SVM's optimisation criterion $C \sum_{i=1}^{n} \xi_i + \frac{1}{2}||f||^2_{\mathcal{H}_K}$ as an adaptation criterion.

However, the computation of ρ still requires the solution of one QP for each $\boldsymbol{\theta}$.

Alignment In [22, 23] the sample alignment

$$\hat{A}(\mathbf{K}_1, \mathbf{K}_2) := \frac{\langle \mathbf{K}_1, \mathbf{K}_2 \rangle_F}{\|\mathbf{K}_1\|_F \|\mathbf{K}_2\|_F}$$

with Frobenius inner product $\langle \cdot, \cdot \rangle_F$ and corresponding norm $\| \cdot \|_F$ was proposed as a measure of conformance between kernels. Especially, the kernel matrix $\mathbf{y}\mathbf{y}^T$, where $\mathbf{y}$ denotes the vector of class labels, is viewed as the optimal kernel matrix for two-class classification. This leads to maximising the criterion

$$\mathcal{C}_3(\boldsymbol{\theta}) := \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^T \rangle_F}{\|\mathbf{K}\|_F \|\mathbf{y}\mathbf{y}^T\|_F} = \frac{\mathbf{y}^T \mathbf{K} \mathbf{y}}{n \|\mathbf{K}\|_F} \tag{11}$$

which, by the inequality of Cauchy–Schwarz, only takes values in $[0, 1]$. A border case of an SVM is the Parzen window estimator. Note that by [22, Theorem 4], the generalisation accuracy of this classifier is bounded by a function of the alignment.

Class Centre Distance In all our experiments, the denominator in (11) doesn't influence the alignment much. Furthermore, supposing normed training vectors $\|\mathbf{x}_i\|_2 = c$ (as guaranteed by (2) when using the Euclidean norm for energy computation) and a Gaussian kernel

$$K(\mathbf{x}, \mathbf{y}) := e^{-\frac{||\mathbf{x}-\mathbf{y}||_2^2}{2\sigma^2}} \tag{12}$$

with large deviation $\sigma > 0$, the numerator in (11) 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 ($i = \pm 1$), this can be rewritten for $n_1 = n_{-1}$ as $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}\|_2^2$. Therefore we propose the criterion

$$\mathcal{C}_4(\boldsymbol{\theta}) := \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}\|_2$$

which can be simply evaluated and is also easily differentiable. It was successfully applied in [4]. While $\mathcal{C}_4$ only takes into account the mean values of the classes we are next looking for classes that are distant from each other and at the same time concentrated around their means.

Generalised Fisher Criterion A generalisation of $\mathcal{C}_4$ are measures using scatter matrices. Let

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

where $\boldsymbol{\mu} := \sum_{i \in \{-1,1\}} \frac{n_i}{n} \boldsymbol{\mu}_i$ denote the *within–class scatter matrix* and the *between–class scatter matrix*, respectively. We consider the generalised Fisher criterion

$$\mathcal{C}_5(\boldsymbol{\theta}) := \frac{\mathrm{tr}(\mathbf{S}_b)}{\mathrm{tr}(\mathbf{S}_w)} = \frac{\frac{n_1}{n}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}\|_2^2 + \frac{n_{-1}}{n}\|\boldsymbol{\mu}_{-1} - \boldsymbol{\mu}\|_2^2}{\frac{n_1}{n}\sum_{k=1}^d \sigma_{1k}^2 + \frac{n_{-1}}{n}\sum_{k=1}^d \sigma_{-1k}^2}$$

where σ_{ik}^2 is the marginal variance of class i along dimension k. For equiprobable classes, the criterion simplifies to

$$\mathcal{C}_5(\boldsymbol{\theta}) \propto \frac{\mathcal{C}_4^2(\boldsymbol{\theta})}{\sum_{k=1}^d (\sigma_{1k}^2 + \sigma_{-1k}^2)} \ .$$

4 Numerical Criteria Evaluation

Now we want to see how the proposed criteria and their theoretical relations behave when analysing real data.

We use three structurally different real data bases: The first are electro–physiological data sets aiming at the detection of ventricular tachycardia as in [4]. For each patient and class, eight heartbeats of length $l = 512$ from a single episode are used for classifier training. Some exemplary beats for a sample patient are shown in Fig. 4. The second

data base contains children's stride time records for the examination of gait maturation as in [24]. The task is to analyse whether the dynamics of walking still change for healthy children between the ages 3–4 (young, $n_1 = 11$) and 6–7 (middle, $n_{-1} = 20$). From the data available by [25], we use the first $l = 384$ strides. Sample time series are depicted in Fig. 5. The third group of data are texture images from the MeasTex collection [26]. We use single rows of length $l = 512$ of the corrugated iron images 'Misc.0002' and 'Misc.0003' to have two classes of one–dimensional data. Both images with normalised contrast as well as two exemplary rows are shown in Fig. 6. Here, the first 32 rows of each texture are used for classifier training, i.e., $n_1 = n_{-1} = 32$. We normalised all samples by $\|\mathbf{s}_i\|_2 = 1000$ and set their average value to zero. For the gait maturation data base, it is also possible to classify without prior normalisation as the overall variability may be a useful feature here. In this case, for the appropriate parameter value $\sigma = 1$, the criteria evaluations also yield qualitatively similar results.

For the feature extraction, we apply orthogonal filter banks with filters of length ≤ 6 which can be parameterised by the two–dimensional space

$$\mathcal{P}_2 = \{\boldsymbol{\theta} = (\theta_0, \theta_1) : \theta_l \in [0, \pi)\,,\ l = 0, 1\}\ .$$

For the classification, a hard margin SVM, i.e. $C = \infty$ in (6) with Gaussian kernel of width $\sigma = 100$ is used. Note that, for the heartbeat problem illustrated in Fig. 4, e.g., the highest alignment $\mathcal{C}_3$ for the optimally aligned wavelet (see Fig. 8 (c)) is achieved for a kernel width of about $\sigma = 80$ for the weighted Euclidean norm and $\sigma = 200$ for the Euclidean norm.

We start by an example that confirms the tests in [4] and shows that the wavelet choice may heavily influence the classification performance. For this, we visualise the training data $\mathbf{x}_i \in \mathbb{R}^9$ of the sample heart patient from Fig. 4 by extracting its principal two components. The Principal Components Analysis (PCA) projects the data from $\mathbb{R}^9$ to $\mathbb{R}^2$ so that most of the total variance of the data is retained. The results for the Haar wavelet ($\boldsymbol{\theta} = (0, 0)$), the Daubechies wavelet with three vanishing moments ([27], ($\boldsymbol{\theta} \approx (1.47, 0.50)$) and the optimal wavelet with respect to $\mathcal{C}_3$ ($\boldsymbol{\theta} \approx (2.04, 0.56)$) for the Euclidean norm in $E_{\|\,\|}$ are shown in Fig. 7. The variance still contained in the plots is approximately 90%, 75% and 92% of the total variance, respectively.

Neither the Haar wavelet, nor the Daubechies wavelet appear to make the training data linearly separable. Our optimal wavelet, on the other hand, well separates the data. Moreover, the classes are nicely clustered now. Indeed, for example for this patient with

two further test episodes, the classification error for the weighted norm varies from 0 to 56% for different wavelets. Also, the optimal $\boldsymbol{\theta}$ does not always lie in the same region for different patients.

Next we evaluate and compare the criteria discussed in Sec. 3. In Figures 8 to 11 the criteria values, ordered from the computationally most efficient to the most expensive one, are plotted using a linear grey scale except for the radius – margin bound $\mathcal{C}_1$ which is plotted on a logarithmic scale due to its large variation. Additionally, its larger values are clipped to the trivial error bound 1 (except for the gait problem in Fig. 9 (e)) to enhance the contrast. To assess the effect of the clipping, the distribution of the logarithm of the bound is indicated by a histogram. Light spots represent favourable criterion values and, hence, beneficial filter operators $F_{\boldsymbol{\theta}}$.

For all four problems, the overall impression is that all shown criteria are alike. Moreover, all criteria show a detailed π–periodic 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. We will address this problem in Sec. 5.

The class centre distance $\mathcal{C}_4$ and particularly the alignment $\mathcal{C}_3$ resemble the margin $\mathcal{C}_2$. 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. Apart from the different distribution of the values, it rates the features mostly like the margin. Moreover, the range of values of the radius – margin bound from 10 resp. 3% to over 100% again indicates the significance of the wavelet choice.

Finally, for specific signals there may be an important difference between using the Euclidean norm and its weighted version as exhibited by Figures 10 and 11.

Classification experiments. To demonstrate the impact of wavelet adaptation on classification, we compare adapted wavelets to both the Haar and Daubechies wavelet with three vanishing moments 'D3'. In addition to the results for the SVM listed in Table 1, we include in Table 2 also results for the Gaussian Bayes classifier with piecewise quadratic decision boundary (see, e.g., [28]).

These results show that wavelet adaptation may significantly improve classification performance. We note that the results for the heart data should be taken with care, due to the small sample size. Sixteen additional heartbeats were available as test data for each class, yet 480 for the texture data. Surprisingly, the Gaussian Bayes classifier shows

comparable performance, at least for the texture data. Moreover, wavelet adaptation proves to be favourable here as well.

For further experimental results, we refer to [4].

5 An Adaptive Grid Search Algorithm

After selecting an appropriate criterion $f \in \{\mathcal{C}_1, \ldots, \mathcal{C}_5\}$ for the wavelet adaptation, we still have to search in the parameter space $\mathcal{P}_L$ for the angle vector $\boldsymbol{\theta}$ optimising this criterion. Note that our parameter space is π–periodic, but many local optima exist as illustrated in Fig. 12. In this paper we restrict our numerical experiments to the maximisation of the simplest criterion, the class centre distance $f = \mathcal{C}_4$ and to the parameter space $\mathcal{P}_2$. Of course our method will work for the other criteria and higher dimensional parameter spaces as well. In [29] we have considered various constrained and unconstrained optimisation strategies to find $\boldsymbol{\theta}$, e.g. Sequential Quadratic Programming, a simplex search method and a restricted step Newton method (for an overview see, e.g., [30]). We developed the following adaptive grid search algorithm which appears to be the most efficient method:

We start with an equispaced coarse grid

$$\mathcal{G}_0 := \left\{ \boldsymbol{\theta}_{j,k} := \left(\frac{\pi j}{N}, \frac{\pi k}{N} \right) : j,k = 0, \ldots, N-1 \right\} \quad .$$

On $\mathcal{G}_0$ we compute the function values $f_{j,k} := f(\boldsymbol{\theta}_{j,k})$ and $f_{\max} := \max\limits_{j,k} f_{j,k}$. Now we consider the neighbourhood $I_{j,k}$ of the points $\boldsymbol{\theta}_{2j+1,2k+1}$ as depicted in Fig. 13 (a).

The bilinear interpolation polynomial on $I_{j,k}$ at the even indexed points

$$(\boldsymbol{\theta}_{2j,2k}, f_{2j,2k}), (\boldsymbol{\theta}_{2j,2k+2}, f_{2j,2k+2}), (\boldsymbol{\theta}_{2j+2,2k}, f_{2j+2,2k}), (\boldsymbol{\theta}_{2j+2,2k+2}, f_{2j+2,2k+2})$$

is given by

$$\begin{aligned} \hat{f}(\boldsymbol{\theta}_{2j,2k} + 2h \cdot (x,y)) \;&=\; f_{2j,2k} + (f_{2j+2,2k} - f_{2j,2k})\, x + (f_{2j,2k+2} - f_{2j,2k})\, y \\ &\quad + (f_{2j,2k} - f_{2j+2,2k} - f_{2j,2k+2} + f_{2j+2,2k+2})\, x\, y \quad (0 \le x, y \le 1) \end{aligned}$$

where $h := \boldsymbol{\theta}_{2j+1,\cdot} - \boldsymbol{\theta}_{2j,\cdot}$ is the grid width. Then $\hat{f}$ is a continuous function on the whole parameter space. If f is a concave function on $I_{j,k}$, then it is easy to check for $\boldsymbol{\theta} \in I_{j,k}$ that

$$f(\boldsymbol{\theta}) \ge \hat{f}(\boldsymbol{\theta})$$

so that $f - \hat{f}$ may be considered as measure for the concavity of our function. As local concavity is a necessary condition for a local maximum of a twice differentiable function we use the following refinement strategy: If

$$\frac{f(\boldsymbol{\theta}) - \hat{f}(\boldsymbol{\theta})}{f_{\max} - f(\boldsymbol{\theta})} > tolF \tag{13}$$

for at least one $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_{2j,2k+1}, \boldsymbol{\theta}_{2j+1,2k}, \boldsymbol{\theta}_{2j+1,2k+1}, \boldsymbol{\theta}_{2j+1,2k+2}, \boldsymbol{\theta}_{2j+2,2k+1}\}$ then we further refine the segment $I_{j,k}$ as shown in Fig. 13 (b), otherwise we leave the segment as it is. The quotient (13) balances the improvement towards the bilinear interpolation with the rating compared with the optimum. We apply our refinement strategy to all segments of $\mathcal{G}_0$ and end up with a new adaptively refined grid $\mathcal{G}_1$. Then we apply the procedure again on the refined grid $\mathcal{G}_1$ and so on until the finest segments have grid width $h \leq tolX$. Note that function evaluations are only necessary on the new grid points.

In the beginning of the algorithm's runtime, heavily concave sections with arbitrary function values will satisfy condition (13), in the end only concave sections which at the same time have high function values, i.e. possible maxima, will be refined.

In summary, we propose the following algorithm:

Algorithm 5.1: GRIDSEARCH$(f, tolF, tolX, N)$

local $grid, index, indexnew$
calculate f on $\left\{0, \frac{\pi}{N}..., \pi - \frac{\pi}{N}\right\}^2$
$grid \leftarrow \frac{\pi}{N}$
$index \leftarrow \left\{0, ..., \frac{N}{2} - 1\right\}^2$
while $(index \neq \varnothing) \wedge (grid > tolX)$
do
- $indexnew \leftarrow \varnothing$
- **for each** $(i, j) \in index$
  - **do**
    - **if** improvement towards bilinear interpolation of f on intermediate grid points in $([2i, 2i+2] \times [2j, 2j+2]) * grid$ /(current maximum − function value) $> tolF$
    - **then**
      - refine f on $([2i, 2i+2] \times [2j, 2j+2]) * grid$
      - $indexnew \leftarrow indexnew \cup \{2i, 2i+1\} \times \{2j, 2j+1\}$
- $grid \leftarrow grid/2$
- $index \leftarrow indexnew$

In our numerical examples we consider again heartbeats ('heart1' and 'heart2': SR versus VT), i.e. signals of length 512 of a form as depicted in Fig. 4, as well as texture

samples ('m2m3' from Fig. 6 and 'a0m0' as images 'Asphalt.0000' and 'Misc.0000' from [26] again). We have chosen the weighted Euclidean norm for $E_{\|\,\|}$. For the parameters of Algorithm 5.1, we have used the values $tolF = 4$, $tolX = \pi/512$ and $N = 16$. Note that the number of function evaluations and the optimal value found by the algorithm are sensible to the parameter $tolF$, but its value can be used for all problems as we apply an absolute criterion.

Fig. 14 demonstrates the final grid generated by the algorithm. One can already see that the region where f is evaluated rapidly gets smaller with each finer grid.

Table 3 presents the results for all four sample problems. Thereby, the optimum values given for comparison were computed by picking the maximum of $f = \mathcal{C}_4$ on the equispaced grid with grid width of $\pi/128$ leading to $128^2 = 16384$ function evaluations. So the heuristic finds the optimum or a very close value with only $2-3\%$ of the criterion evaluations.

6 Conclusions

We have addressed the problem how to efficiently adapt the feature extraction process by orthogonal filter banks and norm computation of the subband coefficients to the subsequent classification by an SVM. We have proposed several criteria for judging the discrimination ability of a set of feature vectors and have highlighted some connections between these criteria. A theorem was provided that simplifies the computation of the radius–margin error bound. We have numerically shown that simple adaptation criteria like the class centre distance and the alignment suffice to promisingly design filters for our hybrid wavelet–large margin classifiers with Gaussian kernels. Furthermore, we have presented an adaptive grid search algorithm that efficiently finds the optimal orthogonal filter bank for our applications. This grid search can easily be implemented due to its simplicity and provides a robust algorithm that does not depend on experienced parameter tuning.

For multi-class classification problems, which will be addressed in a forthcoming paper, SVMs may also be used. This is done most effectively by using a sequence of binary SVM classifiers. For the reduction of the multi-class problem to such a sequence there exist several more or less costly and reliable ways [31, 32]. Then the wavelet adaptation can be applied with a different wavelet for each binary classifier. As an alternative approach, e.g. the generalised Fisher criterion $\mathcal{C}_5$ naturally generalises to

multiple classes.

The classification of images and higher–dimensional signals works analogously according to the construction of multivariate wavelets by tensor products. Hence, in principle our results are relevant for higher dimensions as well. In practice, however, features extracted by tensor wavelets – no matter whether adapted or not – suffer from a lack of rotational invariance. In our future work, we will address this issue without resorting to overly redundant sets of basis functions and the corresponding loss of computational speed.

A An SVM Formulation for Radius Computation

This section derives a QP equivalence that may be used to efficiently compute the radius of the smallest sphere enclosing a set of points as already stated in Theorem 1.

The QP (9) to determine the radius R for the points $\boldsymbol{\phi}(\mathbf{x}_j)$ $(j = 1, \ldots, n)$ in feature space is in fact a special case of the problem

$$\begin{aligned} \min_{\mathbf{a}\in\ell_2, R\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^n} \; & R^2 + C\sum_{j=1}^{n}\xi_j \\ \text{subject to} \quad & \|\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}\|_2^2 \le R^2 + \xi_j \;, \quad j = 1, \ldots, n \;, \\ & \xi_j \ge 0 \;, \quad j = 1, \ldots, n \end{aligned} \tag{14}$$

considered in [33] for clustering. Therefore we will refer to (14) as *SV clustering problem*. We will show that (14) can be solved by a single–class SVM, i.e., an SVM classification problem with all points belonging to the same class. Then the matrix $\mathbf{Y}$ in (6) is the identity matrix so that (6) simplifies to

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \; & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{0} \le \boldsymbol{\alpha} \le \hat{C}\mathbf{e} \;. \end{aligned} \tag{15}$$

We will prove the following theorem which generalises Theorem 1 also including the soft margin case $C < \infty$:

Theorem 2. *Let K be a kernel with corresponding feature map $\boldsymbol{\phi}$ and with the property that $K(\mathbf{x}, \mathbf{x}) = \kappa$ for all $\mathbf{x} \in \mathcal{X}$. Then there exists $\hat{C} > 0$ such that the optimal radius R in* (14) *can be obtained by solving the dual problem* (15) *of a single–class SVM. More precisely, if $\boldsymbol{\alpha}$ is the solution of* (15)*, then*

$$R^2 = \kappa + \boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} - 2(\mathbf{K}\boldsymbol{\beta})_j \;, \tag{16}$$

where $\boldsymbol{\beta} := \frac{\boldsymbol{\alpha}}{\mathbf{e}^T\boldsymbol{\alpha}}$ and $j \in \{1, \ldots, n\}$ denotes some index with $0 < \beta_j < C$.

Note that $C = \hat{C} = \infty$ for our original problem (9).

Our proof proceeds in two steps: first we show that the SV clustering problem (14) is equivalent to a single–class SVM with additional bias term also included in the objective function. This SVM was used for novelty detection in [34] and is therefore called *SV novelty detection problem* in the following. Then we prove that the SV novelty detection problem is equivalent to an ordinary single–class SVM (15) without bias term.

A.1 Equivalence of the SV Clustering Problem and the SV Novelty Detection Problem

The equivalence is best shown considering the dual problems. For solving (14), we introduce the Lagrangian

$$\mathcal{L}(\mathbf{a}, R, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\mu}) := R^2 + C\mathbf{e}^T\boldsymbol{\xi} - \sum_{j=1}^{n} \beta_j (R^2 + \xi_j - \|\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}\|_2^2) - \boldsymbol{\mu}^T\boldsymbol{\xi}$$

with Lagrange multipliers $\boldsymbol{\beta}, \boldsymbol{\mu} \geq \mathbf{0}$. Setting the derivative of $\mathcal{L}$ with respect to R, $\mathbf{a}$ and $\boldsymbol{\xi}$ to zero, it follows

$$\begin{aligned} &\mathbf{e}^T\boldsymbol{\beta} = 1 \ , \\ &\mathbf{a} = \sum_{j=1}^{n} \beta_j \boldsymbol{\phi}(\mathbf{x}_j) \ , \qquad (17) \\ &\boldsymbol{\mu} = C\mathbf{e} - \boldsymbol{\beta} \ . \qquad (18) \end{aligned}$$

Using these equations, the Lagrangian yields the dual problem

$$\begin{aligned} \max_{\boldsymbol{\beta} \in \mathbb{R}^n} \ & \left(W(\boldsymbol{\beta}) := \sum_{j=1}^{n} \beta_j \|\boldsymbol{\phi}(\mathbf{x}_j)\|_2^2 - \sum_{j,k=1}^{n} \beta_j \beta_k \langle \boldsymbol{\phi}(\mathbf{x}_j), \boldsymbol{\phi}(\mathbf{x}_k) \rangle_2 \right) \\ \text{subject to} \quad & \mathbf{e}^T\boldsymbol{\beta} = 1 \ , \\ & \mathbf{0} \leq \boldsymbol{\beta} \leq C\mathbf{e} \ . \end{aligned}$$

By the relation (8) between the kernel function and the feature map, the function $W(\boldsymbol{\beta})$ can be rewritten as

$$W(\boldsymbol{\beta}) = \sum_{j=1}^{n} \beta_j K(\mathbf{x}_j, \mathbf{x}_j) - \sum_{j,k=1}^{n} \beta_j \beta_k K(\mathbf{x}_j, \mathbf{x}_k) \ .$$

In our applications we are mainly interested in isotropic kernels $K(\mathbf{x}, \mathbf{y}) = k(\|\mathbf{x} - \mathbf{y}\|_2)$, e.g. in the Gaussian kernel (12). These kernels have $K(\mathbf{x}, \mathbf{x}) = \kappa$ for some $\kappa > 0$ and all $\mathbf{x} \in \mathcal{X}$. Then $W(\boldsymbol{\beta})$ can be further simplified to

$$W(\boldsymbol{\beta}) = \kappa - \boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta}$$

so that we finally have to solve the dual optimisation problem

$$\begin{aligned} \min_{\boldsymbol{\beta} \in \mathbb{R}^n} \ & \boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} \\ \text{subject to} \quad & \mathbf{e}^T\boldsymbol{\beta} = 1 \ , \\ & \mathbf{0} \leq \boldsymbol{\beta} \leq C\mathbf{e} \ . \end{aligned} \qquad (19)$$

Note that this problem coincides with our optimisation problem (15) except for the first constraint $\mathbf{e}^T\boldsymbol{\beta} = 1$. The Kuhn–Tucker complementarity conditions for problem (14) are

$$\begin{aligned} \beta_j(R^2 + \xi_j - \|\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}\|_2^2) &= 0\ , \quad j = 1, \ldots, n\ , & (20)\\ \mu_j\xi_j &= 0\ , \quad j = 1, \ldots, n\ . & (21) \end{aligned}$$

For $0 < \beta_j < C$, equations (18) and (21) imply that $\mu_j > 0$ and thereby $\xi_j = 0$. Now it follows by (20) that R^2 can be computed as

$$\begin{aligned} R^2 &= \|\boldsymbol{\phi}(\mathbf{x}_j) - \mathbf{a}\|_2^2 \\ &\overset{(17),(8)}{=} K(\mathbf{x}_j, \mathbf{x}_j) + \sum_{i,k=1}^{n} \beta_i\beta_k K(\mathbf{x}_i, \mathbf{x}_k) - 2\sum_{k=1}^{n} \beta_k K(\mathbf{x}_j, \mathbf{x}_k) \\ &= \kappa + \boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} - 2(\mathbf{K}\boldsymbol{\beta})_j\ . \end{aligned}$$

Let us turn to the SV novelty detection problem investigated by Schölkopf et al. in [34]. We are looking for a decision function

$$f(\mathbf{x}) = a(\mathbf{x}) + b := \sum_{j=1}^{n} \beta_j K(\mathbf{x}, \mathbf{x}_j) + b \tag{22}$$

with bias term b which solves the modified single–class SVM problem

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^n, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^n} \ & C\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\|a\|_{\mathcal{H}_K}^2 + b = C\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} + b \\ \text{subject to} \quad & a(\mathbf{x}_j) + b \geq 1 - \xi_j\ , \quad j = 1, \ldots, n\ , \\ & \boldsymbol{\xi} \geq \mathbf{0}\ . \end{aligned} \tag{23}$$

Analogous to the SV clustering problem, we build the Lagrangian

$$\mathcal{L}(\boldsymbol{\beta}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) := C\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} + b - \sum_{j=1}^{n} \alpha_j(a(\mathbf{x}_j) + b - 1 + \xi_j) - \boldsymbol{\mu}^T\boldsymbol{\xi}$$

with Lagrange multipliers $\boldsymbol{\alpha}, \boldsymbol{\mu} \geq \mathbf{0}$. Setting the derivative of $\mathcal{L}$ with respect to b, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ to zero, it follows

$$\begin{aligned} \mathbf{e}^T\boldsymbol{\alpha} &= 1\ , & (24)\\ \boldsymbol{\alpha} = \boldsymbol{\beta} &\overset{(24)}{\Rightarrow} \mathbf{e}^T\boldsymbol{\beta} = 1\ , & (25)\\ \boldsymbol{\mu} &= C\mathbf{e} - \boldsymbol{\alpha}\ . & (26) \end{aligned}$$

Using these equations, the Lagrangian yields the dual problem

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^n} \ & \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} - 1 \\ \text{subject to} \quad & \mathbf{e}^T\boldsymbol{\beta} = 1 \ , \\ & \mathbf{0} \leq \boldsymbol{\beta} \leq C\mathbf{e} \ . \end{aligned} \tag{27}$$

This problem is obviously equivalent to the dual SV clustering problem (19). We summarise:

Lemma 3. *Let K be a kernel with corresponding feature map $\boldsymbol{\phi}$ and with the property that $K(\mathbf{x},\mathbf{x}) = \kappa$ for all $\mathbf{x} \in \mathcal{X}$. Then the optimisation problems* (14) *and* (23) *are equivalent in that they lead to the same dual problem* (19).

From the dual solution $\boldsymbol{\beta}$ of (19). *the primal solution $\mathbf{a}, R, \boldsymbol{\xi}$ of* (14) *may be obtained by* (17) *and* (16) *and the Kuhn–Tucker conditions* (20) *and* (21). *The optimal values $b, \boldsymbol{\xi}$ for problem* (23) *may be obtained by the Kuhn–Tucker complementarity conditions as well.*

This lemma was also proved in [34]. Further, Vapnik [5, Sec. 10.7] already showed that R^2 can be computed as described by (16) with problem (19) for hard margin ($C = \infty$).

At first sight, it is astonishing that although the quadratic optimisation problems for SV clustering and SV novelty detection are deviated from quite different initial problems (14) and (23), they are equivalent. The paper [35] provides a nice geometrical interpretation for that. For the hard margin case ($C = \infty$) and the Gaussian kernel, this further implies that

$$R^2 = b \ .$$

A.2 Equivalence of the SV Novelty Detection Problem and the Single–Class SVM without Bias Term

The previous subsection shows the equivalence of the SV clustering problem which can be used for radius computation to a modified SVM (23) with bias term. We will now show that this special problem is equivalent to a single–class SVM without bias term.

With $a(\mathbf{x})$ defined as in (22), the common single–class SVM is described by the problem

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^n,\boldsymbol{\xi}\in\mathbb{R}^n} \; & \hat{C}\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\|a\|^2_{\mathcal{H}_K} = \hat{C}\mathbf{e}^T\boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} \\ \text{subject to} \; & a(\mathbf{x}_j) \geq 1 - \xi_j, \quad j = 1,\ldots,n \;, \\ & \boldsymbol{\xi} \geq \mathbf{0} \;. \end{aligned} \tag{28}$$

By setting up the Lagrangian as above, the traditional single–class SVM leads to the dual quadratic problem (15).

Lemma 4. *There exists $\hat{C} > 0$ such that the SV novelty detection problem* (23) *with parameter C is equivalent to the standard SVM problem* (28) *with parameter $\hat{C}$ in that the solutions are derivable from one another. The dual solutions $\boldsymbol{\alpha}$ of* (15) *and $\boldsymbol{\beta}$ of* (27) *are related by*

$$\boldsymbol{\beta} = \frac{\boldsymbol{\alpha}}{\mathbf{e}^T\boldsymbol{\alpha}} \tag{29}$$

or conversely by $\boldsymbol{\alpha} = \frac{1}{1-b}\boldsymbol{\beta}$ with the primal variable b.

Proof. The proof consists of two parts. Firstly, the dual solution of the biased SVM (27) will be derived from the dual solution of the SVM without bias (15). Secondly, the primal solution of the unbiased SVM (28) will be derived from the primal solution of the biased SVM (23).

1. Suppose problem (15) is solved by $\boldsymbol{\alpha}$. With $a := \mathbf{e}^T\boldsymbol{\alpha} > 0$, set $\boldsymbol{\beta} := \frac{\boldsymbol{\alpha}}{a}$. Then $\boldsymbol{\beta}$ is valid in problem (27) if $C = \frac{\hat{C}}{a}$. Suppose that $\boldsymbol{\beta}$ is not the optimal solution of problem (27), i.e., there exists some $\tilde{\boldsymbol{\beta}}$ satisfying $\mathbf{e}^T\tilde{\boldsymbol{\beta}} = 1$, $\mathbf{0} \leq \tilde{\boldsymbol{\beta}} \leq C\mathbf{e}$ so that

$$\begin{aligned} \frac{1}{2}\tilde{\boldsymbol{\beta}}^T\mathbf{K}\tilde{\boldsymbol{\beta}} &< \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} \\ \Rightarrow \frac{1}{2}(a\tilde{\boldsymbol{\beta}})^T\mathbf{K}(a\tilde{\boldsymbol{\beta}}) &< \frac{1}{2}(a\boldsymbol{\beta})^T\mathbf{K}(a\boldsymbol{\beta}) \\ \Rightarrow \frac{1}{2}\tilde{\boldsymbol{\alpha}}^T\mathbf{K}\tilde{\boldsymbol{\alpha}} - a &< \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} - a \;, \end{aligned}$$

where $\tilde{\boldsymbol{\alpha}} := a\tilde{\boldsymbol{\beta}}$. Since $\tilde{\boldsymbol{\alpha}}$ fulfils $\mathbf{0} \leq \tilde{\boldsymbol{\alpha}} \leq \hat{C}\mathbf{e}$ and $a = \mathbf{e}^T\tilde{\boldsymbol{\alpha}}$ holds, this is a contradiction to the assumption that $\boldsymbol{\alpha}$ is the optimal solution of problem (15).

2. On the other hand, let $(\boldsymbol{\beta}, b, \boldsymbol{\xi}^{\boldsymbol{\beta}})$ be the optimal solution of the primal problem (23). Then $\boldsymbol{\alpha} := \frac{1}{1-b}\boldsymbol{\beta}$, $\boldsymbol{\xi}^{\boldsymbol{\alpha}} := \frac{1}{1-b}\boldsymbol{\xi}^{\boldsymbol{\beta}}$ is a valid solution for problem (28). Note that $b < 1$ due to the dual constraint $\mathbf{e}^T\boldsymbol{\beta} = 1$ and the Kuhn–Tucker conditions. Assume that $\tilde{\boldsymbol{\alpha}}$ is valid for (28) as well, then the vector $(\tilde{\boldsymbol{\beta}}, b, \boldsymbol{\xi}^{\tilde{\boldsymbol{\beta}}}) := ((1-b)\tilde{\boldsymbol{\alpha}}, b, (1-b)\boldsymbol{\xi}^{\tilde{\boldsymbol{\alpha}}})$ is valid for

problem (23). Now we obtain for $\hat{C} = \frac{C}{1-b}$

$$\begin{aligned}
&\frac{1}{2}\tilde{\boldsymbol{\alpha}}^T\mathbf{K}\tilde{\boldsymbol{\alpha}} + \hat{C}\mathbf{e}^T\boldsymbol{\xi}^{\tilde{\boldsymbol{\alpha}}} < \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} + \hat{C}\mathbf{e}^T\boldsymbol{\xi}^{\boldsymbol{\alpha}} \\
\Leftrightarrow\ &\frac{1}{2}((1-b)\tilde{\boldsymbol{\alpha}})^T\mathbf{K}((1-b)\tilde{\boldsymbol{\alpha}}) + (1-b)\hat{C}\mathbf{e}^T\left((1-b)\boldsymbol{\xi}^{\tilde{\boldsymbol{\alpha}}}\right) < \\
&\quad\frac{1}{2}((1-b)\boldsymbol{\alpha})^T\mathbf{K}((1-b)\boldsymbol{\alpha}) + (1-b)\hat{C}\mathbf{e}^T\left((1-b)\boldsymbol{\xi}^{\boldsymbol{\alpha}}\right) \\
&\quad\Leftrightarrow \frac{1}{2}\tilde{\boldsymbol{\beta}}^T\mathbf{K}\tilde{\boldsymbol{\beta}} + C\mathbf{e}^T\boldsymbol{\xi}^{\tilde{\boldsymbol{\beta}}} + b < \frac{1}{2}\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta} + C\mathbf{e}^T\boldsymbol{\xi}^{\boldsymbol{\beta}} + b\ .
\end{aligned}$$

Consequently, since $(\boldsymbol{\beta}, b, \boldsymbol{\xi}^{\boldsymbol{\beta}})$ is the optimal solution for problem (23), $\boldsymbol{\alpha}$ is the optimal solution of (28). □

So far, we have shown that for special values of C depending on the solution of the problem, the biased and unbiased single–class SVMs are equivalent. Anyway, as C is a tuning parameter that cannot be determined analytically, this condition does not restrain the equivalence. Especially, for $C = \infty$, the hard margin case, no condition with respect to the weight factor C has to be taken into account.

References

[1] T. Randen, J. H. Husøy, Filtering for texture classification: A comparative study, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (4) (1999) 291–310.

[2] P. Scheunders, S. Livens, G. Van de Wouwer, P. Vautrot, D. Van Dyck, Wavelet-based texture analysis, International Journal on Computer Science and Information Management 1 (2) (1998) 22–34.

[3] M. Unser, Texture classification and segmentation using wavelet frames, IEEE Transactions on Image Processing 4 (11) (1995) 1549–1560.

[4] D. J. Strauss, G. Steidl, Hybrid wavelet-support vector classification of waveforms, Journal of Computational and Applied Mathematics 148 (2002) 375–400.

[5] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, NY, USA, 1998.

[6] N. Saito, Local feature extraction and its application using a library of bases, Ph.D. thesis, Dept. of Mathematics, Yale University (Dec. 1994).

[7] D. J. Strauss, G. Steidl, W. Delb, Feature extraction by shape-adapted local discriminant bases, Signal Processing 83 (2) (2003) 359–376.

[8] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for SVMs, in: T. K. Leen, T. G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems 13, MIT Press, Cambridge, MA, USA, 2001, pp. 668–674.

[9] G. Strang, T. Nguyen, Wavelets and Filter Banks, Wellesley–Cambridge Press, Wellesley, MA, USA, 1996.

[10] P. Moulin, M. K. Mihçak, Theory and design of signal–adapted FIR paraunitary filter banks, IEEE Transactions on Signal Processing 46 (4) (1998) 920–929.

[11] M. Vetterli, J. Kovačević, Wavelets and Subband Coding, Signal Processing, Prentice Hall, Englewood Cliffs, NJ, USA, 1995.

[12] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, Cambridge, MA, USA, 2000.

[13] G. S. Kimeldorf, G. Wahba, Some results on Tchebycheffian spline functions, Journal of Mathematical Analysis and Applications 33 (1) (1971) 82–95.

[14] G. Wahba, Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, in: B. Schölkopf, C. J. C. Burges, A. J. Smola (Eds.), Advances in Kernel Methods — Support Vector Learning, MIT Press, Cambridge, MA, USA, 1999, Ch. 6, pp. 69–88.

[15] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Machine Learning 46 (1–3) (2002) 131–159.

[16] R. Herbrich, Learning kernel classifiers: theory and algorithms, MIT Press, Cambridge, MA, USA, 2002.

[17] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, NY, USA, 1995.

[18] S. Floyd, M. K. Warmuth, Sample compression, learnability, and the Vapnik-Chervonenkis dimension, Machine Learning 21 (3) (1995) 269–304.

[19] R. Herbrich, T. Graepel, A PAC-Bayesian margin bound for linear classifiers: Why SVMs work, in: T. K. Leen, T. G. Dietterich, V. Tresp (Eds.), Advances in Neural Information System Processing 13, MIT Press, Cambridge, MA, USA, 2001, pp. 224–230.

[20] T. Joachims, Making large–scale SVM learning practical, in: B. Schölkopf, C. J. C. Burges, A. J. Smola (Eds.), Advances in Kernel Methods — Support Vector Learning, MIT Press, Cambridge, MA, USA, 1999, Ch. 11, pp. 169–184.

[21] K. Duan, S. S. Keerthi, A. N. Poo, Evaluation of simple performance measures for tuning SVM hyperparameters, Tech. Rep. CD-01-11, Dept. of Mechanical Engineering, National University of Singapore (2001).

[22] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J. Kandola, On kernel-target alignment, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural

Information Processing Systems 14, MIT Press, Cambridge, MA, USA, 2002, pp. 367–373.

[23] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, M. I. Jordan, Learning the kernel matrix with semi-definite programming, in: C. Sammut, A. G. Hoffmann (Eds.), Proc. of the 19th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, USA, 2002, pp. 323–330.

[24] J. M. Hausdorff, L. Zemany, C.-K. Peng, A. L. Goldberger, Maturation of gait dynamics: stride-to-stride variability and its temporal organization in children, Journal of Applied Physiology 86 (3) (1999) 1040–1047.

[25] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.

[26] G. Smith, MeasTex image texture database and test suite, Available at `http://www.cssip.uq.edu.au/meastex/meastex.html`, version 1.1 (May 1997).

[27] I. Daubechies, Orthonormal bases of compactly supported wavelets, Communications on Pure and Applied Mathematics 41 (7) (1988) 909–996.

[28] R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd Edition, John Wiley & Sons, New York, NY, USA, 2000.

[29] J. Neumann, C. Schnörr, G. Steidl, Effectively finding the optimal wavelet for hybrid wavelet – large margin signal classification, Tech. Rep. TR-03-005, Comp. science series, Dept. of Mathematics and Computer Science, University of Mannheim (Mar. 2003).

[30] R. Fletcher, Practical Methods of Optimization, 2nd Edition, John Wiley & Sons, New York, NY, USA, 1987.

[31] M. Heiler, Optimization criteria and learning algorithms for large margin classifiers, Master's thesis, Dept. of Mathematics and Computer Science, University of Mannheim (Oct. 2001).

[32] C.-W. Hsu, C.-J. Lin, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks 13 (2) (2002) 415–425.

[33] A. Ben-Hur, D. Horn, H. T. Siegelmann, V. Vapnik, A support vector method for clustering, in: T. K. Leen, T. G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems 13, MIT Press, Cambridge, MA, USA, 2001, pp. 367–373.

[34] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, in: S. A. Solla, T. K. Leen, K.-R. Müller (Eds.), Advances in Neural Information Processing Systems 12, MIT Press, Cambridge, MA, USA, 2000, pp. 582–588.

[35] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson, Estimating the support of a high-dimensional distribution, Tech. Rep. 99-87, Microsoft Research, short version appeared in *Neural Computation*, 2001 (1999).

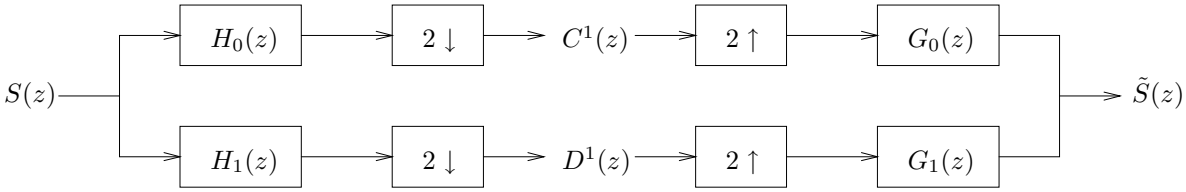


Figure 1: two–channel filter bank

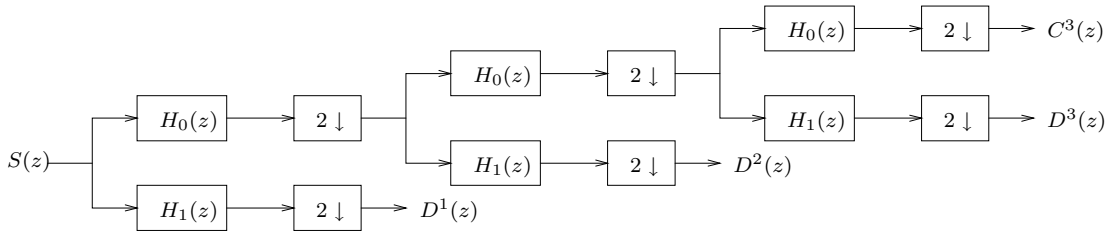


Figure 2: octave–band filter bank

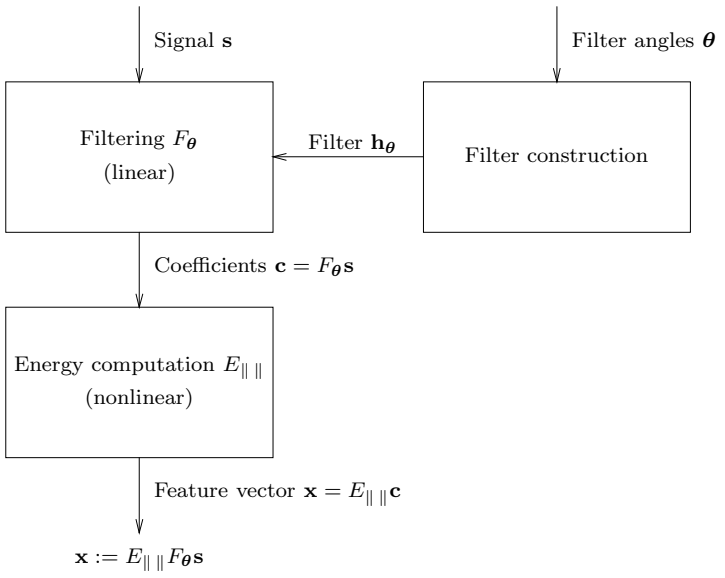


Figure 3: feature extraction process

Figure 4: Two–class problem (heartbeats: sinus rhythm (SR) and ventricular tachycardia(VT))

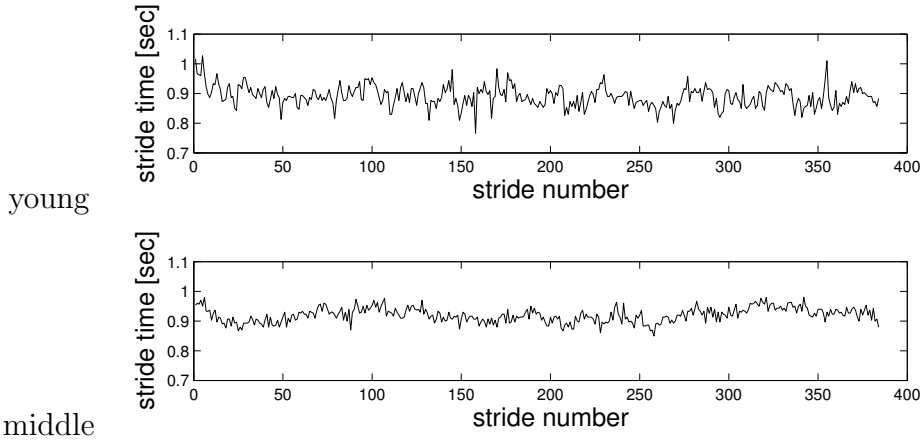


Figure 5: sample stride time records

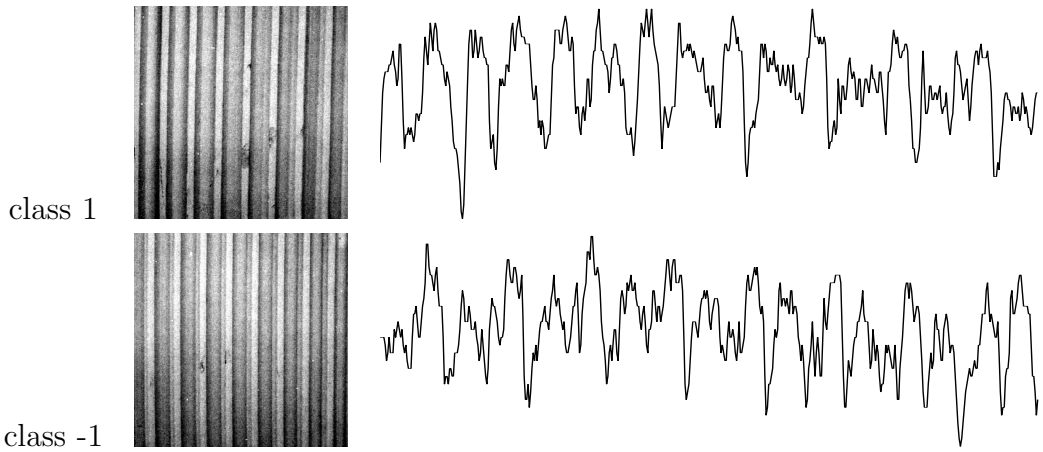
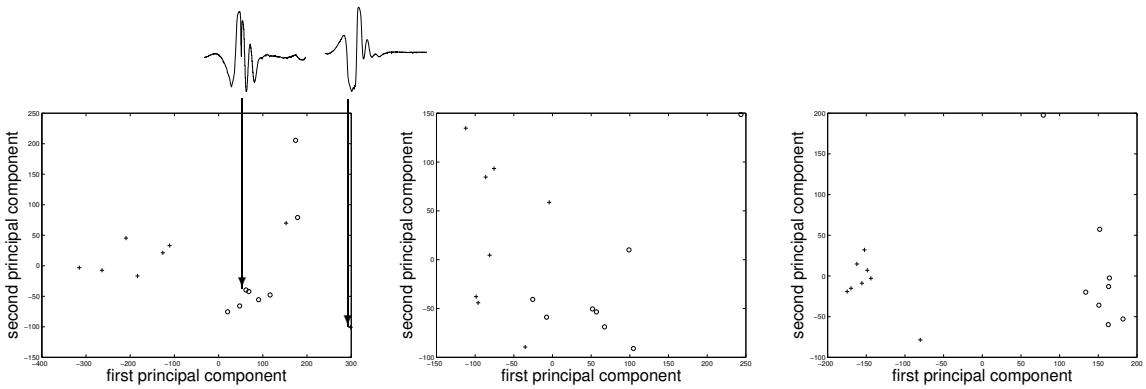
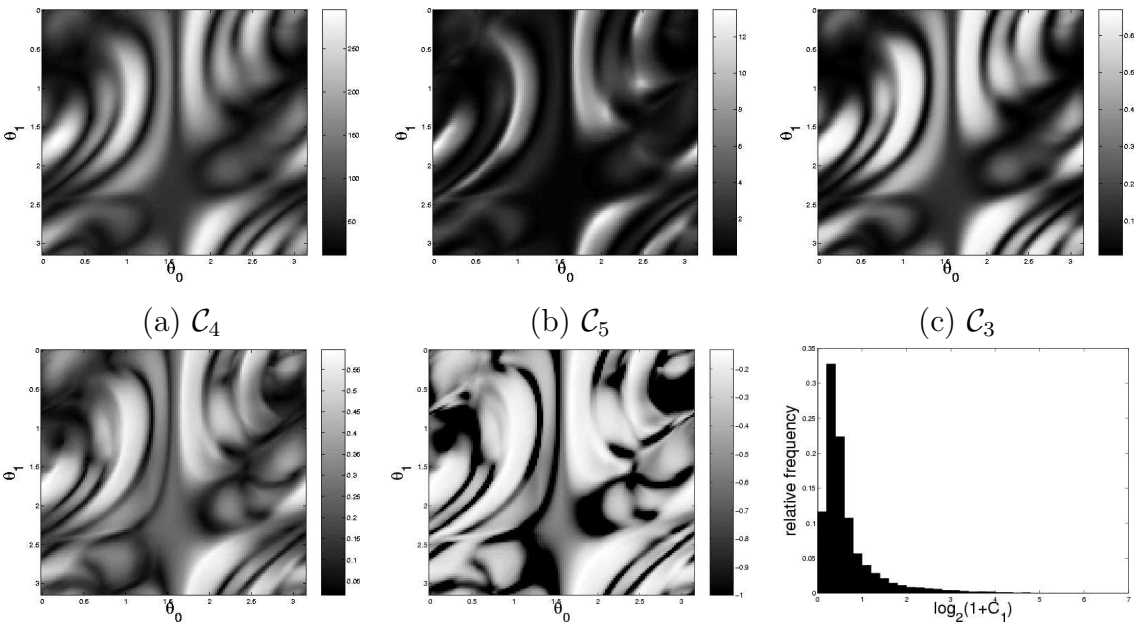


Figure 6: texture sample: linearly rescaled images and exemplary rows



(a) (b) (c)

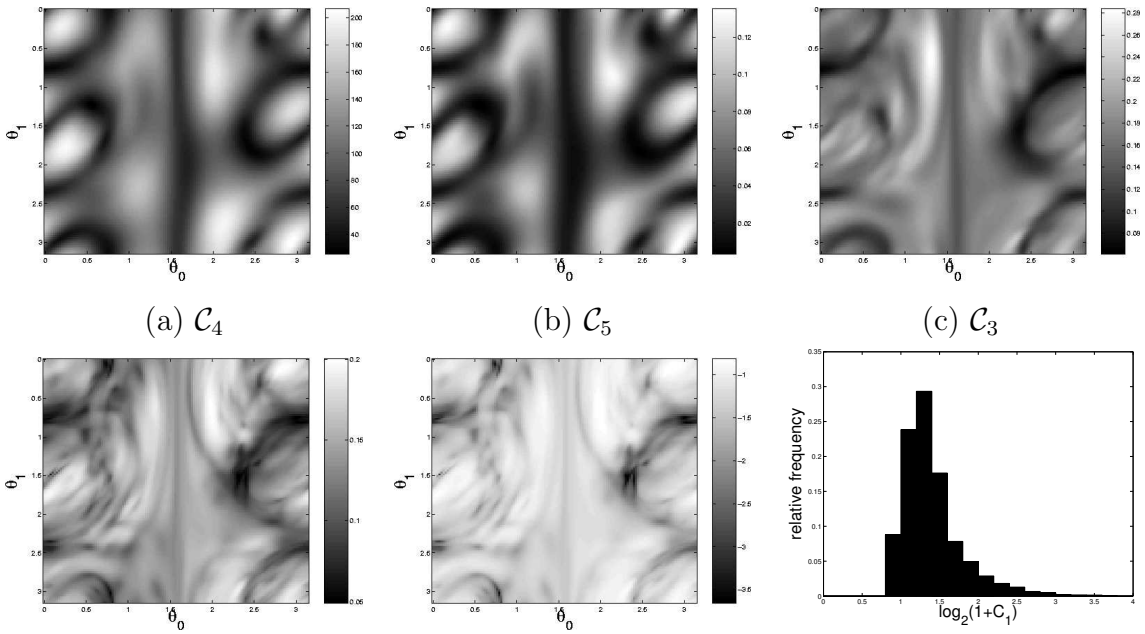
Figure 7: Principal components of training vectors for heartbeat classification with Euclidean norm in $E_{\|\,\|}$: (a) for the Haar wavelet, (b) for the Daubechies wavelet with three vanishing moments, (c) for the optimally aligned wavelet ($\mathcal{C}_3$)



(a) $\mathcal{C}_4$ (b) $\mathcal{C}_5$ (c) $\mathcal{C}_3$

(d) $\mathcal{C}_2$ (e) $-\log_2(1+\mathcal{C}_1)$ (f) histogram of (e)

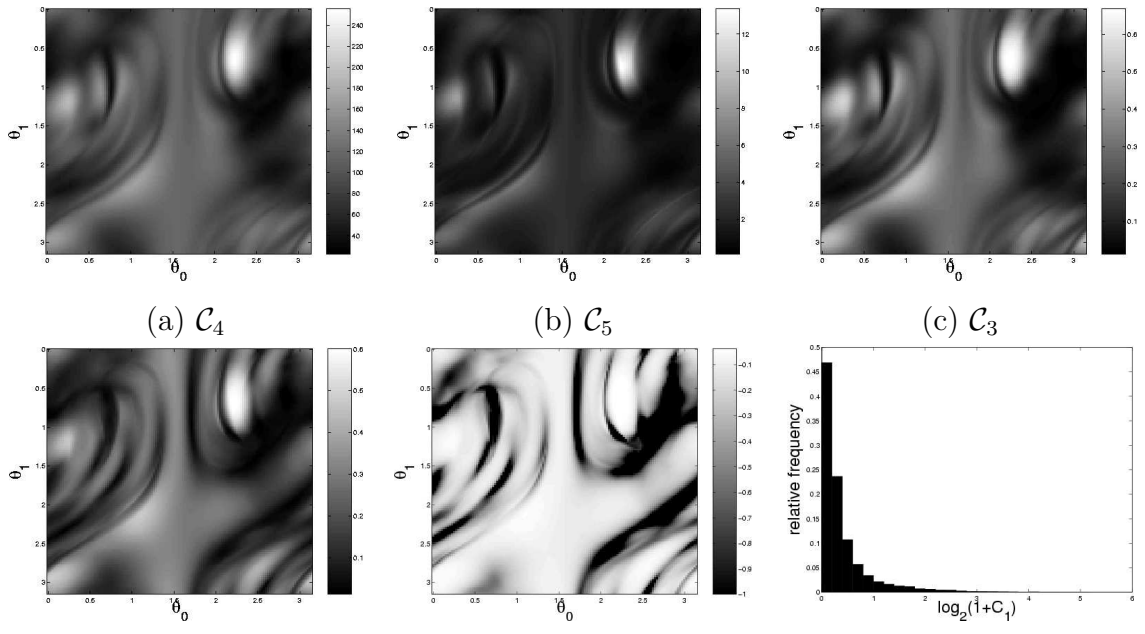
Figure 8: Criteria values for heartbeat classification with weighted Euclidean norm in $E_{\|\,\|}$; light spots represent favourable criterion values



(a) $\mathcal{C}_4$ (b) $\mathcal{C}_5$ (c) $\mathcal{C}_3$

(d) $\mathcal{C}_2$ (e) $-\log_2(1+\mathcal{C}_1)$ (f) histogram of (e)

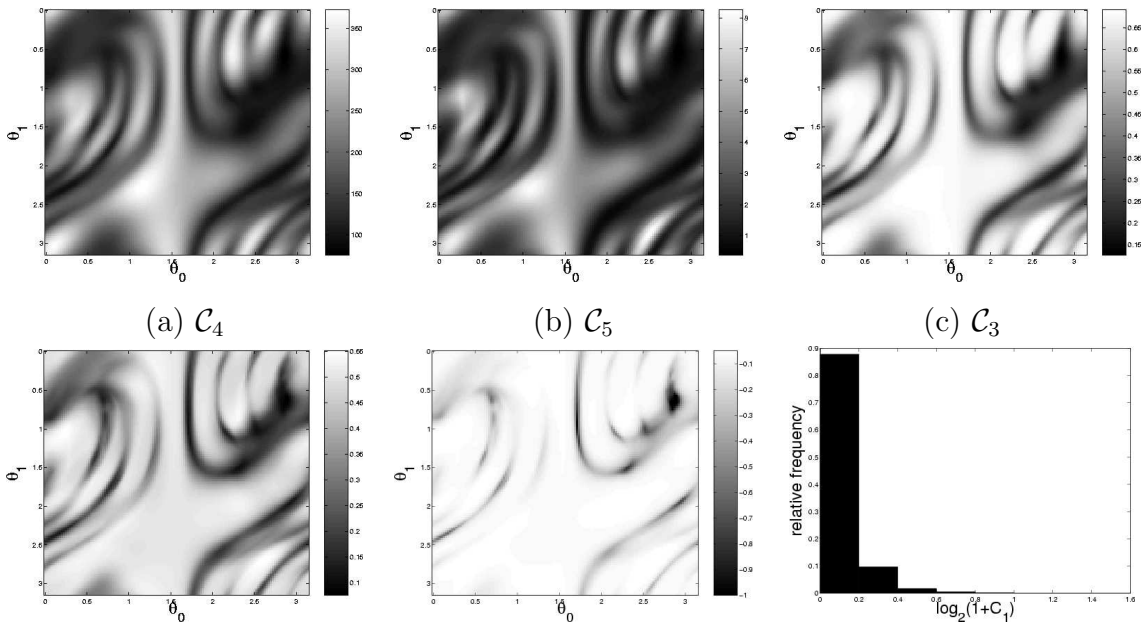
Figure 9: Criteria values for gait dynamics classification with Euclidean norm in $E_{\|\,\|}$; light spots represent favourable criterion values



(a) $\mathcal{C}_4$ (b) $\mathcal{C}_5$ (c) $\mathcal{C}_3$

(d) $\mathcal{C}_2$ (e) $-\log_2(1+\mathcal{C}_1)$ (f) histogram of (e)

Figure 10: Criteria values for texture row classification with weighted Euclidean norm in $E_{\|\,\|}$; light spots represent favourable criterion values



(a) $\mathcal{C}_4$ (b) $\mathcal{C}_5$ (c) $\mathcal{C}_3$

(d) $\mathcal{C}_2$ (e) $-\log_2(1+\mathcal{C}_1)$ (f) histogram of (e)

Figure 11: Criteria values for texture row classification with Euclidean norm in $E_{\|\,\|}$; light spots represent favourable criterion values

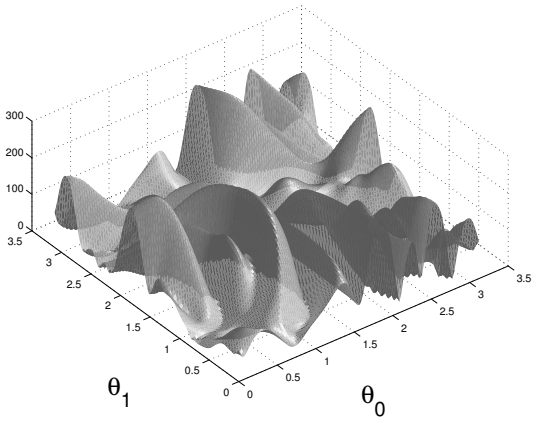


Figure 12: objective criterion for wavelet adaptation: example from Fig. 8 (a)

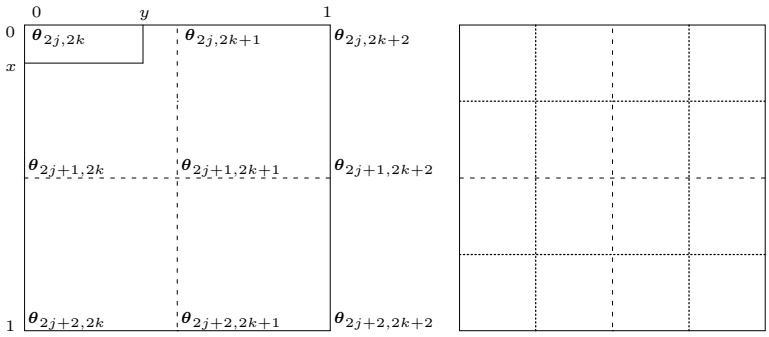


Figure 13: (a) segment $I_{j,k}$ of a coarse grid, (b) refined segment

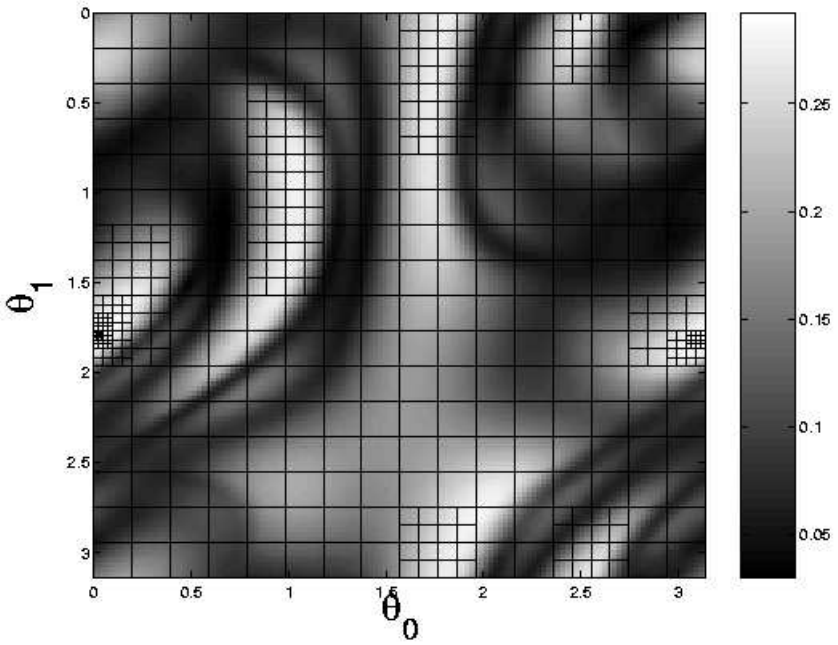


Figure 14: class centre distance for heartbeat classification with final grid used by Algorithm 5.1

		Support Vector Machine						
data set	energy	Haar	D3	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_3$	$\mathcal{C}_2$	$\mathcal{C}_1$
heart	weighted Euclidean	19	9	6	16	16	16	16
texture	weighted Euclidean	8	4	0	2	1	0	0
texture	Euclidean	1	2	0	0	0	0	0

Table 1: Classification error [%] for the Support Vector Machine; different wavelets (Haar, D3) and adaptation criteria $\{\mathcal{C}_i\}$

		Gaussian Bayes Classifier						
data set	energy	Haar	D3	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_3$	$\mathcal{C}_2$	$\mathcal{C}_1$
heart	weighted Euclidean	19	28	25	25	13	9	16
texture	weighted Euclidean	3	3	0	1	0	0	0
texture	Euclidean	3	3	0	0	0	0	0

Table 2: Classification error [%] for the Gaussian Bayes Classifier; different wavelets (Haar, D3) and adaptation criteria $\{\mathcal{C}_i\}$

	optimum	grid search	
problem	value	value	evals
heart1	0.2923	0.2923	538
heart2	0.2601	0.2601	392
a0m0	0.1670	0.1669	350
m2m3	0.2564	0.2564	364

Table 3: grid search results: returned maximal value and number of function evaluations