

Translation invariant classification of non-stationary signals

Vincent Guigue, Alain Rakotomamonjy and Stéphane Canu

PSI CNRS FRE2645

INSA de Rouen

Avenue de l'universite, 76801 Saint Etienne du Rouvray, France

Abstract

Non-stationary signal classification is a complex problem. This problem becomes even more difficult if we add the following hypothesis: each signal includes a discriminant waveform, the time location of which is random and unknown. This is a problem that may arise in Brain Computer Interfaces (BCI) or in electroencephalogram recordings of patients prone to epilepsy. The aim of this article is to provide a new graph-based representation for classifying this kind of signals. This representation characterizes the waveform without reference to the absolute time location of the pattern in the signal. We will show that it is possible to create such a signal description using graphs on a time-scale or time-frequency signal representation. The definition of an inner product between graphs is then required to implement kernel methods algorithms like Support Vector Machines. Our experimental results shows that this approach is very promising and performs very well on real-word datasets.

Key words: Kernel methods, signal, classification, wavelet, time-frequency

1 Introduction

The classification of non-stationary signals is a common and difficult problem in signal processing. Classical statistical descriptors like mean, or Fourier coefficients are not efficient descriptors for such data where time-dependent informations are needed. When dealing with non-stationary signals, usual approaches consist in using as a signal descriptors the raw time-series, time-scale representations (TSRs) or time-frequency representations (TFRs). However, all these solutions lead to an inefficient high dimensional classification problem.

Furthermore, if the discriminative part of the signal is a transient signal, the position of which is unknown and variable, then a translation invariant classifier is required. Such a situation occurs, for instance in a classical Brain Computer Interface (BCI) problem such as the P300 speller paradigm (Farwell and Donchin, 1988) or in electroencephalogram recordings (EEGs) of epileptic patients (Durka, 2004).

TSRs or TFRs, combined with adapted distance measures, have been widely used to classify non-stationary signals. Various strategies have been implemented within this context. Davy (2000) proposes to optimize the TFR with regards to the classification results. The modulation frequency defined by Sukittanon et al. (2003) can be seen as a model of each TFR frequency. Hory (2002) propose to model the TFR using a mixture of χ^2 distributions, to focus on discriminant patterns and Michel et al. (2000) use a graphical structure to characterize the pattern skeleton of the TFR. Some other approaches consists in modeling the TSR : Mallat (1997) introduced a TSR based on wavelet maxima and Boashash et al. showed that the singularity influence cone of TSR is an interesting feature for signal classification (Boashash et al., 1987). Saito and Coifman (1994) have optimized the TS representation with respect to a discriminative criterion by selecting the best wavelet basis among several bases. The idea has also been extended to the optimization of the wavelet analytic function by Lucas et al. (2002). Crouse et al. (1998) obtain good results for signal classification based on the Hidden Markov Model (HMM) for each TSR scale. As far as the classification task is concerned, Davy et al. (2002) showed the interest of using Support Vector Machines (SVMs) to deal with non-stationary signals.

When the time-frequency or time-scale representations are considered as images, the problem of classifying signals boils down to a problem of image classification. Then, it is possible to take advantage of all the tools developed for such framework. For instance, Greenspan et al. (2001) fits a Gaussian Mixture Model (GMM) on images which provides a compact description and leads to good classification results of different classes of images.

In this paper, our objective is to present a novel approach that is able to classify non-stationary signals which are composed of a discriminant pattern that occurs at unknown time. Formally, the pattern classification setting is the following. We have at our disposal a set of signal examples of which the class labels are known. This set is denoted as $\{x_i(t), y_i\}_{i=1, \dots, \ell}$ where $x_i(t)$ is a finite energy signal and $y_i = \{1, -1\}$ its label. Suppose furthermore that each $x_i(t)$ can be written as :

$$x_i(t) = \Gamma(t - \tau)m_{y_i}(t - \tau) + b(t) \tag{1}$$

where $\Gamma(t)$ is a step function, $b(t)$ a gaussian white noise, τ a continuous

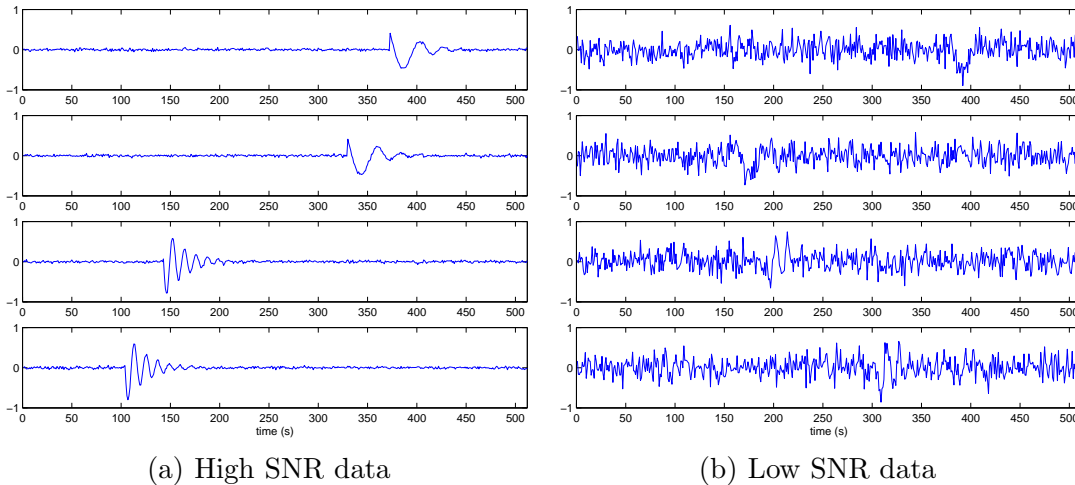


Fig. 1. Example of signals in a translation-invariant signal classification problem. The four signals on top belong to class 1, the four bottom ones belong to class -1.

random variable which depends to an unknown probability density function and $m_{y_i}(t)$ is an unknown but discriminant pattern. For instance, Figure (1) presents two classes of such signals where each signal is composed of a transient pattern that which is randomly located. Based only on this set of examples, we want to build a decision function that is able to predict the most correctly as possible the class label of a new signal $x(t)$.

For achieving this goal, we propose to use a time-frequency or a time-scale representation for coping with non-stationary signals. Then we reduce the representation dimension, by means of Gaussian modeling. Finally, the building of a graph between Gaussian models introduces comparative time information, which enables us to deal with the translation invariance problem. Indeed, the absolute time reference is suppressed in the graph representation. Then, we use a state-of-the-art algorithms like SVMs and k-nearest neighbors (k-nn) to classify the signals. Since we are using a graph representation of the signal, we consider a graph inner product to compare the representations. Our implementation relies on a modified version of the graph kernel defined by Kashima et al. (2003).

Section 2 presents our approach concerning the bi-dimensional representation (TFR or TSR). Section 3 deals with the building of the graph to provide a compact description of the bi-dimensional representation. We will focus on the definition of the graph kernel and its alternatives in section 4. Finally, we will compare the results of the different algorithms for classifying synthetic and real signals (section 5). We give our conclusions in section 6.

2 Translation-covariant signal description

Usually, signal discrimination problems are addressed by firstly extracting features from each input signal $x_i(t)$ and then by learning a classifier according to these extracted features. In our case, we have to cope with two difficulties related the signal's characteristics : the non-stationarity and the random-located discriminant patterns.

Representing non-stationary signals is commonly done through time-frequency representations (TFRs) or time-scale representations (TSRs) since both representations are able to depict the signal's frequency evolution over time . So in this paper, we also restrict ourselves to these two cases. However, in order to face with the random-located discriminant patterns, among all the possible time-frequency or time-scale representations r , we have to consider only those ones that are time-covariant. Namely, the representation r of a signal $x(t)$ has to satisfy the equation :

$$r[x(t - \tau)] = r_\tau[x(t)] \quad (2)$$

where r_τ is the translated representation of $r[x(t)]$. Time-covariant representation of the signal is one of the key feature of our classification scheme. In fact, such representations make the discrimination problem easier since in absence of noise, representations of same class signals are similar up to a translation τ . Hence, if afterwards we are able to extract features from these representations that do not depend on the time origin, then we can obtain a translation-invariant representation of the signal.

In this present section, we are going to present some of the most common time-frequency and time-scale representations that satisfy the time-covariance property.

2.1 Time-frequency and time-scale representation of non-stationary signals

This paragraph introduces two of the most frequent tools for analyzing non stationary signals.

2.1.1 Time-frequency representations

Time-frequency analysis and representations are known to be well-suited to non-stationary signals (Qian, 2001). In fact, with time-frequency represen-

tations, it is possible to follow the spectral properties variation of a signal with respects to time. Among all possible TFR, we restrict ourselves to the Cohen's class because of its intrinsic properties (Cohen, 1995). The Cohen's class time-frequency representation of a signal $x(t)$ can be written as :

$$C_x^\phi(t, f) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \phi(s - t, \nu - f) \mathcal{W}_x(s, \nu) ds d\nu$$

where ϕ is the so-called time-frequency kernel and $\mathcal{W}_x(s, \nu)$ is the Wigner-Ville transform of x :

$$\mathcal{W}_x(s, \nu) = \int_{-\infty}^{+\infty} x\left(s + \frac{\tau}{2}\right) x\left(s - \frac{\tau}{2}\right) e^{-j2\pi\nu\tau} d\tau$$

From these equations, we can state that any Cohen's class time-frequency representations can be obtained by filtering the Wigner-Ville representation through the kernel ϕ and several well-known representations fit into this framework like the Smoothed Pseudo Wigner-Ville distribution or the optimal Gaussian kernel representation (Baraniuk and Jones, 1993).

As we stated before, Cohen's class of TFRs exhibits several properties but the one that is the most interesting for us is that the resulting representation is time and frequency shift covariant.

2.1.2 Time-scale representations

Time-scale representation is a decomposition of signal over elementary functions that are well concentrated in time and frequency. The TSR is obtained by projecting the signal over a family of function, the elements of which are linked by a scale factor and a translation factor. Time-scale representations are also denoted as wavelet transform because of the particular property of the analytic function ψ . Given a wavelet ψ , the time-scale representation of a signal $x(t)$ is

$$P_x(b, a) = \{\langle x, \psi_{a,b} \rangle\}$$

where $a \in \mathbb{R}_+^*$ and $b \in \mathbb{R}$ and

$$\langle x, \psi_{a,b} \rangle = \int_{-\infty}^{+\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt \quad (3)$$

The squared coefficients $P_{a,b}^2$ correspond to the local energy of the signal at a

given scale a and time b .

In order to cope with the problem of obtaining a time-covariant representation of a signal, we need a wavelet transform that satisfies this property. Since orthogonal wavelet transforms do not verify such property, we have to rely on a continuous wavelet transform (Mallat, 1997).

2.2 An unified representation

For both time-frequency and time-scale representations, the time variable and scale or frequency variables are actually discretized leading to a representation of the form :

$$C_x^\phi(t_i, f_i) \quad \text{or} \quad P(b_i, a_i)$$

where t_i and b_i are both time variables and f_i and a_i are respectively the frequency and scale variables. Note that a relationship between the scale and frequency parameter can be deduced from the spectral representation of the wavelet ψ . From this point of view, one can consider a TFR or a TSR as a bi-dimensional discrete representation of a signal which maps a given time and scale or frequency pair (t_i, s_i) to a value z_i which represents either the time-frequency or time-scale signal value. z_i can also be related to a local energetic value of the signal. Hence, in the following given a signal $x(t)$, we will denote as the representation of signal the set of triplets:

$$r[x(t)] = \{t_i, s_i, z_i\}_{i=1}^k$$

where k is the number of time-scale or time-frequency pairs in the representation.

Let us note that the choice of time-frequency or time-scale representations depend essentially on the application at hand. Hence, for each given problem, model selection has to be performed before stating that one representation method performs better than the other. Figure 2 compares covariant representations obtained from time-frequency and time-scale transform. In both cases, two families of signals are depicted as images where the color at a given point (t_i, s_i) is related the value z_i . One can see that the TFR and TSR are radically different even when considering the same class of signal.

3 Graph modeling of TF and TS representations

The aim of this section is to introduce a new method based on graphs for modeling the TF or TS structure of a signal. Assuming that the discriminant

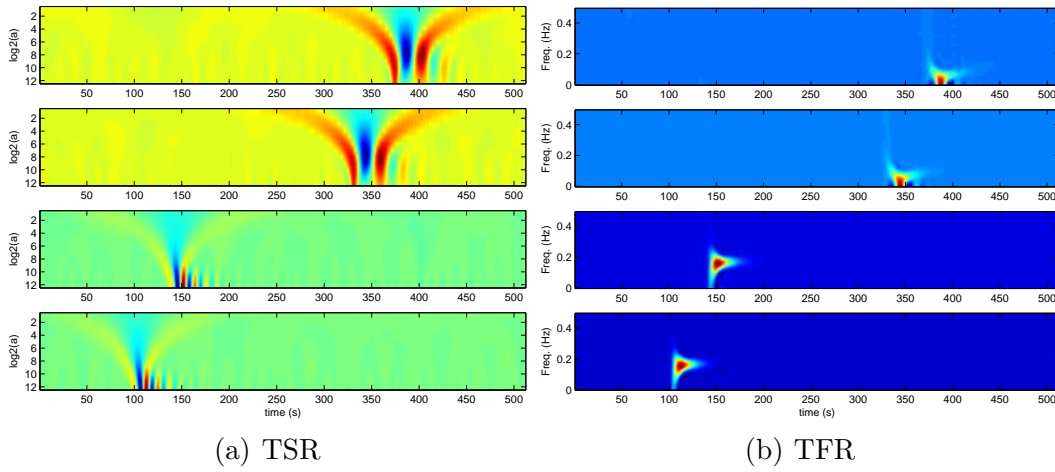


Fig. 2. TSR (left illustration) and TFR (right illustration) of two classes of signals. The two signals on top belong to class 1, the two signals below belong to class -1.

patterns of the signal have a signature in the TS or TF plane, this graph will focus on the structure of such signature.

3.1 Building the graph

Although time is important to describe the pattern, it is also a penalty factor due to the random location of the patterns. Our solution consists in filtering discriminant information by using comparative time between selected regions of the TS or TF plane. The idea is to move from a point representation (using absolute coordinates) to a vector representation (which focus on the comparative positions of points).

From the representation $r[x(t)] = \{t_i, s_i, z_i\}_{i=1}^k$ of a signal $x(t)$, we want to build a new representation $r_2[x(t)]$ which relates the time position of each triplet with regards to any other triplets. In this way, we suppress the absolute time information in the representation while still keeping some time information on the TF or TS structure of the signal. This novel representation r_2 can be represented as a graph for which the set of nodes H is defined as :

$$H = \{h_i\}_{i=1}^k$$

with $h_i = \{s_i, z_i\}$ and for which the set of edges is $A = \{a_{h_i h_j}\}_{i,j=1}^k$ with $a_{h_i h_j} = \Delta_{t_{ij}} = t_i - t_j$.

Note that each node represents the scale or frequency location s_i , and the weight information z_i . It is also important to note that this graph is a fully connected graph and thus, it is composed of k nodes and $k(k-1)$ edges. We will see in section 4 that the computational complexity of graph inner product is closely related to node cardinality. Hence, dealing with such large graphs

would be intractable as soon as the signal length increases or the number of scale analysis becomes large. Consequently, we need a graph representation with fewer nodes that capture most of the discriminant information of the TS or TF representation.

3.2 Dimensionality reduction

Several works dealing with signal classification from time-scale or time-frequency representation consider that discriminant informations are brought by high coefficient values of the representation. Although, in some particular classification problems this hypothesis is not appropriate, we also assume in this work that discriminant informations lie in the high energy regions of the representation. Modeling these regions will enable us to reduce the graph dimension while keeping most of the information (Hory, 2002). This idea has already been largely considered in the signal processing community. For instance, Donoho (1995) showed that suppression of small coefficients in orthogonal wavelet transform reduces the noise while preserving the original signal. In the same flavor, Mallat and Zhong (1992) showed that the local maxima in the continuous wavelet transform can accurately be used for reconstructing a denoised version of a signal. Hence, this high energy modeling will have 2 effects. The first one is to reduce the size of the graph representation and the second one is to reduce the noise in the signal representation.

Hence, we propose to reduce the dimensionality of the graph by considering a given high energy region of the TF or TS representation as a single node. This allows us to reduce the number of nodes by increasing the amount of information in each node.

Similarly to Greenspan et al (Greenspan et al., 2001), the idea is to consider n_g regions of high energy and to model each of these regions as a gaussian of mean $\mu = [t_G \ s_G]^T$ and covariance matrix Σ . In this way, the vector μ_g gives some information on the time and scale or frequency location of the gaussian (and thus the region) whereas Σ captures its shape information. Note that in such way, we have related the three following points : a high energy region of the time-frequency plane, a gaussian model, and a node of the graph. Hence, each single node of the graph is labelled with informations coming from its related region and gaussian model. The node labels that we have used are the following :

- the scale parameter s_G of the Gaussian model. This label is useful since it gives information on the scale of frequency position of the associated region.

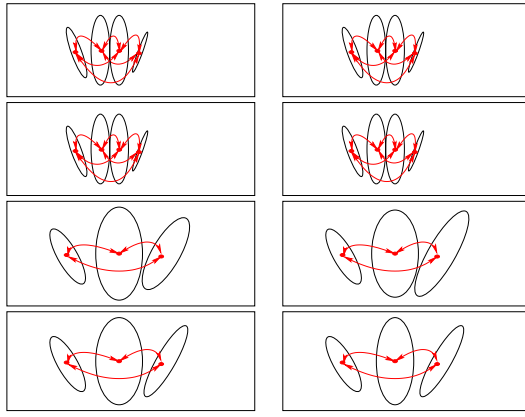


Fig. 3. The four graphical representations on top belong to class 1, the four bottom ones belong to class -1.

- the covariance matrix of the Gaussian :

$$\Sigma_G = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \quad (4)$$

This label gives information on the ellipsoidal shape of the gaussian.

- the mean value m_G of all coefficients in the region. Note that by construction all the coefficients of a given region have the same sign so that this mean value cannot vanish.

Other features can be extracted from each region and be used as a node label. For instance, it is possible to consider the energy or the relative energy of each region. However, the question that remains open is whether a given feature would be discriminative or not.

Again in this representation, the edge label between two gaussian clusters of mean vector $\mu_1 = [t_{G_1} \ s_{G_1}]^T$ and $\mu_2 = [t_{G_2} \ s_{G_2}]^T$ is given as $\Delta_t = t_{G_1} - t_{G_2}$. In other words, we still keep as edge labels the comparative time position between gaussian models of different high energy regions.

Figure (3) show the Gaussian modeling of the time-scale representations of eight different signals. In this case, the TS representations have been transformed in respectively a 4-node and 3-node graphs where each node represents a high energy region of the representation.

3.3 Heuristics for gaussian modeling of TF or TS representations

This paragraph introduces the algorithm that we used for building the Gaussian modeling of the TF or TS plane.

Given the set of triplets $\{t_i, s_i, z_i\}_{i=1}^k$ representing the time-frequency plane associated to a signal $x(t)$, our objective is to find n_g gaussian that models the shape of TSR high energy regions. We consider in the sequel that the number of gaussian n_g in the model is a user-defined parameter. The algorithm that we propose is an iterative algorithm that fits a single Gaussian model at a time. The Gaussian model of the j^{th} region is defined by :

$$G_j(\nu) = \exp\left(-\frac{1}{2}(\nu - \mu_j)^T \Sigma_j^{-1}(\nu - \mu_j)\right) \quad (5)$$

Basically, for a single gaussian, the modelling heuristic is the following. First, we look for the maximum of the TF or TS representation, then we grow a region from this maximum to a point where the z_i value reach a given percentage of the maximum. This region is then removed and we proceed in the same way for the next gaussian model. More formally, this translates in the following algorithm.

General algorithm For $j = 1, \dots, n_g$ times do :

- (1) find the absolute maxima among $|z_i|$,
- (2) spread from this point to define a region,
- (3) compute the Gaussian parameters (μ_j, Σ_j) of the region i ,
- (4) put all region coefficients to 0.

Spreading The aim of this step is to define an homogeneously decreasing or increasing high energy region for which coefficient values satisfy :

$$|z_i| \geq \eta |z_{max}| \quad \tau \in [0, 1]$$

where η is a user-defined parameter. The idea is to look iteratively for the regions in the neighborhood of $\{t_{max}, s_{max}, z_{max}\}$ that satisfy such conditions. In a nutshell, the idea is to grow a the region according to time and then to change scale and to proceed in the same way from the point of maximal value in that scale. This leads to the following algorithm :

Given an absolute maxima z_{max} of coordinates $\{t_{max}, s_{max}\}$ and a threshold $z_{lim} = \eta |z_{max}|$:

- (1) Initialization : set $t^* = t_{max}$, $s^* = s_{max}$, $R = \emptyset$.
- (2) find all triplets $\{t_i, s_i, z_i\}$ such as $s_i = s^*$, $z_i > z_{lim}$,
- (3) find the time continuous segment D which contains t^* among these triplets,
- (4) $R = R \cup D$
- (5) set $t^* = t_i$ such as $z_i = \max_{t \in D} z$,
- (6) define $s^* = s^* - 1$,
- (7) if $z^* > z_{lim}$ go to 2., else, get out of the loop.

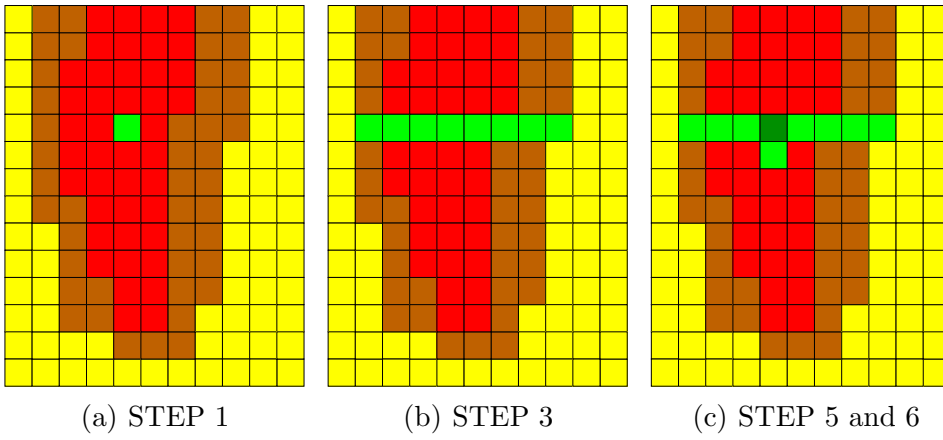


Fig. 4. Illustrations of the spreading procedure in the Gaussian modeling algorithm: (STEP 1) Initialization : find the coordinate (t^*, s^*) of the absolute maxima among the whole time scale plane. STEP 3 : spread from (t^*, s^*) to find the continuous segment D which gathers all the coefficients satisfying $z_i > z_{lim}$. STEP 5 and 6 : jump to the next scale, that is to say find the new coordinate (t^*, s^*) for the next scale.

It is necessary to carry out this loop twice to describe the entire high energy region : first towards the lower scale or frequency, then towards the higher scale or frequency (use $s^* = s^* + 1$ in step 5).

Gaussian parameters computation

Once the high energy region R is defined, the parameters of the modelling gaussian can be computed as follows :

$$\mu_j = \frac{\sum_{i \in R} \nu_i z_i}{\sum_{i \in R} z_i} \quad \text{where} \quad \nu_i = [t_i \ s_i]^T \quad (6)$$

$$\Sigma_j = \frac{\sum_{i \in R} z_i (\nu_i - \mu)(\nu_i - \mu)^T}{\sum_{i \in R} z_i} \quad (7)$$

The computational complexity of this algorithm is $O(k)$. Indeed, processing the full gaussian modelling needs at most a single sweep of all the triplets of the time-frequency or time-scale representation. And as we will see in the sequel, the computational complexity limitation of our approach is essentially due to the graph's inner product. Another advantage of this heuristic is that the gaussian model it produces is unique since the algorithm does not depend on any random initialization.

4 Inner product for signal representations

One of our primary objective is to classify signal by means of a state of the art classification algorithm based on kernel methods such as the Support Vector Machines (SVMs). In order to take advantage all the powerful machinery of SVMs, we need to define an inner product between signals through a kernel function. Such a kernel function is usually chosen or specifically designed so that it captures most of the discriminant information of the data to be classified. This section describes different possible kernels for signals with an emphasis on kernels that would satisfy the property of translation invariance. We will first define an inner product between graphs that is able to cope with our novel translation-invariant representation of signals. Then we will propose some possible alternative kernels that address the problem of translation-invariant classification.

4.1 Inner product for graph modeling of TS or TF representations

The novel graph-based representation of signal we introduced in the previous section imposes us to use a kernel on graphs. Many recent works have been devoted to the construction of such graph kernels (Gaertner, 2003). In this work, we have considered the kernel on labeled graphs proposed by Kashima et al. (2003).

Suppose we want to compute the inner product $K(G, G')$ between two graphs G, G' , the idea of Kashima et al. is to compare the label sequences generated by the two *synchronized* random walks of both graphs.

Suppose we have a graph G with n_g labeled nodes $\{h_i\}_{i=1}^{n_g}$ and a set of edges $\{a_i\}$ as described in section (3.2). A random walk on that graph produces a sequence of nodes denoted as a path, which depends on an initial probability $p_s(h)$ distribution on $\{h_i\}$, a transition probability from node h_{i-1} to node h_i $p(h_i|h_{i-1})$ and the probability $p_q(h_\ell)$ to stop the random walk at node ℓ . In our case, we have not used any prior knowledge for setting these probability distributions. Hence, $p_s(h)$ is an uniform distribution, $p(h_i|h_{i-1})$ is an uniform distribution over all adjacent nodes of the current one, and for all h_ℓ , $p_q(h_\ell)$ is chosen to be a constant equal to $1/n_g$.

From a given path obtained from the random walk, we can define the label sequence as a alternative sequence of node labels and edge labels :

$$h = (n_{x_1}, e_{x_1, x_2}, n_{x_2}, \dots, n_{x_\ell})$$

Then, the inner product between two label sequences h and h' of same length ℓ can be defined by as :

$$K(h, h') = K_n(h_1, h'_1) \prod_{i=2}^{\ell} K_e(h_{2i-2}, h'_{2i-2}) K_n(h_{2i-1}, h'_{2i-1})$$

whereas if the two sequence lengths are different then we define $K(h, h') = 0$. Note that the label sequence kernel is actually the product of each label kernel value and thus, it needs the definition of a kernel on nodes K_n and a kernel on the edge K_e .

Then the kernel between two graphs G and G' is defined as the expectation of the kernel between sequence $K(h, h')$ over all possible paths of all lengths :

$$K(G, G') = \sum_h \sum_{h'} K(h, h') p(h|G) p(h'|G')$$

An efficient and more detailed computation of $K(G, G')$ is given in the paper of Kashima et al. (2003). Although the Kashima's implementation is more efficient than a brute force implementation, the complexity of the kernel computation is high since it needs the solving of a linear system of size $((|G||G'|)^2)$, with $|G|$ the number of nodes in graph G . This is another reason for decreasing drastically the number of nodes in each graph.

This graph kernel requires the definition of kernels between edges and nodes labels. Since in our case, these labels respectively belong to \mathbb{R} and \mathbb{R}^d , we have chosen K_e and K_n to be Gaussian kernels.

4.2 Alternative translation-invariant representations and inner products.

In order to validate our graph-based approach and to evaluate its contributions with respects to classical approaches, we have benchmarked it against other approaches that are able to deal with the translation-invariant problem. This paragraph describes the alternative methods that we have used.

4.2.1 Classical representations

Raw signals

To justify a complex and computationally expensive approach, we have considered as a baseline method which uses the simplest available description of a signal, namely the raw time-series. This experiment enabled us to compare recognition rate as well as computational time between simple and complex

methods. In this case, we can consider as the representation of the signal $x(t)$

$$r[x(t)] = x(t)$$

Common statistical descriptors

Simple features can be extracted from a signal. Many of them can be translation-invariant. Examples of such statistical descriptors are given in the thesis of Healey (2000) on emotion recognition from physiological signals. The features she proposed are composed of mean and standard deviation over raw and normalized data, amplitude of the signal, variation between the first and the last point of the signal. A power spectral analysis based on Fourier transform is also considered. Haselsteiner (2000) have also shown the interest of carrying out two power spectral analyses for EEG signal classification problem : one on thin frequency bands and one on large frequency band. Although very simple, these features can lead to very good recognition rate in some classification problem especially if a careful variable selection is performed (Guigue et al., 2003).

4.2.2 Translation invariant kernels

In this paragraph, we propose some alternative kernels that are able to cope with the translation-invariance hypothesis by matching the characteristics of two signal representations.

Translation arrangement

Given two signal x_1 and x_2 , it is clear that a simple inner product $\langle r[x_1], r[x_2] \rangle$, where the representation r is whether the raw time-serie or a TF representation, would not be able to face the hypothesis of a random-located discriminant pattern .

A way for dealing with such random localisation is to consider a set of translated representation of either x_1 or x_2 . So let us define $\{r_\tau[x_2]\}_{\tau \in \Omega}$ as the set of the τ -translated representation of x_2 and Ω as the set of possible values of τ . Then, we define the inner product of x_1 and x_2 as the largest inner product between the representation of x_1 and the set $r_\tau[x_2]$. Formally, this translates into the following translation invariant inner product :

$$\langle x_1, x_2 \rangle = \max_{\tau \in \Omega} (k(r[x_1], r_\tau[x_2])) \quad (8)$$

This kernel can be related to other kernels which address the invariance problem. The idea of using translated version of the signal representation comes from the virtual SVM approach (Chapelle and Schölkopf, 2002) whereas look-

ing for the most “similar” representation of a signal x_1 with all the translated representation of x_2 gives equation (8) the flavor of a tangent distance kernel (Simard et al., 1998; Haasdonk and Keysers, 2002). From a signal processing point of view, equation (8) can also be interpreted in the following way. If r is the raw time-series representation of the signal and $k(x, y)$ the usual L_2 inner product of finite energy signals then, the inner product between two finite energy signals x_1 and x_2 is defined as the maximal intercorrelation value between x_1 and x_2 .

Bag of vectors

In order to suppress the time dependency of the graph representation, this method consider only the nodes of the graph representation. The set of edges Δ_t are no more taken into account. Hence, we consider that the signal representation is only the set of labelled nodes where labels are actually features vector. Consequently, the similarity of such representations can be measured by the similarity of the sets of node labels. For this purpose, we need a kernel for sets of vectors. Several works have dealt with such kernels. For instance, Wallraven et al. (2003) have designed a kernel that tries to match pairs of vectors from a set \mathcal{S}_1 and \mathcal{S}_2 .

$$K_{match}(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{2}(K(\mathcal{S}_1, \mathcal{S}_2) + K(\mathcal{S}_2, \mathcal{S}_1)) \quad (9)$$

with :

$$K(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{n_{g_1}} \sum_{i=1}^{n_1} \max_{j=1, \dots, n_{g_2}} (k(V_i^1, V_j^2)) \quad (10)$$

where n_{g_1} and n_{g_2} are respectively the cardinality of set \mathcal{S}_1 and \mathcal{S}_2 and V_i^k is the i^{th} vector in \mathcal{S}_k .

Because of the max function, kernels in equation (8) and (10) are not positive definite. However the following approximation can be used when $k(x, y)$ is a Gaussian kernel:

$$\max_{y \in \mathcal{Y}} (k(x, y)) \approx \frac{1}{|\mathcal{Y}|} \sum_{j=1}^{|\mathcal{Y}|} \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

In fact, if the bandwidth σ is narrow enough, the maximum of $k(x, y)$ will occur when $x \approx y$. In other cases, $k(x, y) \approx 0$.

Hence, equations (8) and (10) respectively become :

$$\langle x_1, x_2 \rangle = \sum_{\tau \in \Omega} \exp \left(-\frac{\|r_\tau[x_1(t)] - r[x_2(t)]\|^2}{2\sigma^2} \right) \quad (11)$$

$$\langle x_1, x_2 \rangle = \frac{1}{n_{g_1} \cdot n_{g_2}} \sum_{i=1}^{n_{g_1}} \sum_{j=1}^{n_{g_2}} \exp \left(-\frac{\|V_i^1 - V_j^2\|^2}{2\sigma^2} \right) \quad (12)$$

We can note that the resulting kernel between set of vector given in equation (12) is equivalent to a multiple instance kernel as defined by Gärtner et al. (2002).

5 Results

This section presents the results of some experimental analyses we carried out in order to validate our algorithm for translation-invariant classification of non-stationary signals. After having described the data we used for simulation, we present the performance results obtained from the different representations and kernels described above.

5.1 Building the simulated data

In many signal classification applications, data are composed of a pattern and noise. The pattern shape is characteristic, whereas its position is unknown and random. For instance, when we try to recognize a response to a stimulus in an electroencephalogram (EEG) (Farwell and Donchin, 1988), the time position of this response is unknown. Again, while detecting or recognizing epileptic patients, EEG spikes occur at any time of the signal.

Hence, we worked on artificial data that present such characteristics. This toy problem consider two classes of signal for which the discriminative patterns $m(t)$ (exponential decreasing chirps as plotted in Figure 1) are generated according to:

$$m(t) = e^{-\alpha t} \cos((u + vt)t + \phi) \quad (13)$$

Hence, the instantaneous frequency of $m(t)$ increases in time while the amplitude decreases. Then, according to the signal model given in equation (1), a signal $x(t)$ is of the form :

$$x(t) = m_{u,v}(t - \tau)\Gamma(t - \tau) + b(t) \quad (14)$$

where $\Gamma(t)$ is the step function, $b(t)$ a gaussian white noise and τ a random value. Each signal class is characterized by a particular pair $\{u, v\}$. In this work, we have chosen the following parameters : In class $y_i = 1$, we have $u = 1 \cdot 10^{-3}, v = 2 \cdot 10^{-3}$ and in class $y_i = -1$ we have $u = 5 \cdot 10^{-4}, v = 6 \cdot 10^{-4}$. For each signal, τ is drawn according to a uniform distribution on the length of the signal interval. For both classes, we have set $\alpha = 0.05$ and $\phi = 1$.

5.2 Classification experiments and results

According to the above description, we have built a learning dataset composed of 400 signals (with equal distribution between classes) of 1024 samples each. For this experiment, we have used a time-scale representation of the signal. Hence, we have analyzed each signal over 5 octaves with three divisions per octave through the continuous wavelet transform. Three Gaussians have then used for modeling each time-scale plane. The tolerance threshold is arbitrarily set to 50% of the spot maximum.

All SVMs and kernel parameters have been optimized by means of a cross-validation procedure on a set of 200 signals per class. As a result, we have $C = 1000$ and $\sigma = 30$ for raw signals representation, $C = 150$ and $\sigma = 1$ for TSR representations. For translation arrangements and for both representations, we have kept these parameter settings. Graph kernels K_e and K_n are Gaussians, we set $\sigma_a = \sigma_n = 0.5$ whereas $C = 300$ is the best parameter value for SVM classification. Same settings are used for the bag of vectors kernel, using only K_n ($\sigma_n = 0.5$).

The misclassification rates are given in table 1 for high SNR signals and table 2 for low SNR signals. Results are averaged over 30 runs and evaluated on a thousand of signals. Performances for two learning set sizes, 1 and 400, are reported in these tables. The aim of the single signal learning set experiment is twofold : it will measure the algorithm performance over a small test set and show the translation invariance skill of the method.

The classification problem on high SNR signals is trivial. However, combining them with the single signal learning set enable us to measure the translation invariance capability of each method. Graph and bag of vectors kernel achieved 100 % correct classification and demonstrate their abilities to describe the discriminant pattern whatever its position in the signal. Power Spectral Analysis (PSA) enables the statistical description to obtain good results. This is natural because PSA is a time independent analysis.

The low SNR experiment points out the noise sensitivity of the statistical representation which is due to the lack of time description. Graph kernel combined with SVM is the best method for this problem, since it yields the lowest

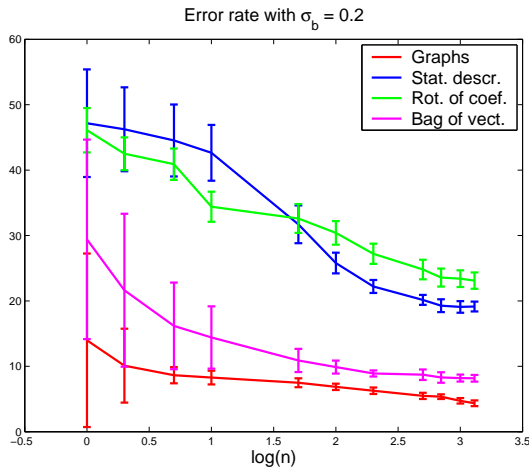


Fig. 5. Classification performance results using different kernels and representations with respects to the number of signals of a given class in the learning set.

misclassification rate and the lowest variance of the results. Graph kernel outperforms bag of vectors kernel by more than 3% whereas the only difference between the two methods is the time information brought by the graph edge labels. In the single signal learning set case, for the low SNR case, the performance difference between these two methods further increases and is of the order of 15%. We have also analyzed the importance of the training set size. This emphasizes the importance of the edge labels.

Figure (5) presents classification results for SVMs in a low SNR case. We can see that for all kernels as the training set size increases the error rate decreases. However, using our graph-based kernel performance results is good even for very low training set size. This suggests that this kernel is able to capture discriminant informations without the need of several learning examples with patterns located at different time. Again we can note that the graph kernel performs better than the bag of vectors kernel.

5.3 Classifying EEG signals

Our algorithm for translation-invariant signal classification has also been tested on real data. The problem we consider is the following : we want to classify some EEG recordings in order to diagnose the epilepsy of patients. Epileptic form activity is characterized in a EEG recordings by the presence of spikes which generally confirms the diagnostic of epilepsy. Hence our aim is to learn a decision function which recognizes epileptic spikes against other EEG signals due to other brain activities.

The dataset we used has been obtained from the authors of (Latka et al., 2005) and contains 30 EEG recordings with epileptic spikes and 27 other signals with

Learn. \ Test.	1/1000	400/1000		
Classifiers	1-nn = SVM	1-nn	3-nn	SVM
Coef.	48.01% \pm 3.81	20.19% \pm 1.15	22.02% \pm 1.33	19.12% \pm 1.02
Raw sig.	48.19% \pm 3.52	41.45% \pm 3.12	44.34% \pm 2.78	38.54% \pm 1.96
Stat. descr.	6.12% \pm 4.30	5.62% \pm 2.12	5.67% \pm 2.18	5.79% \pm 2.22
T.A. (coef.)	7.5% \pm 2.01	5.23% \pm 1.54	5.83% \pm 1.30	5.49% \pm 1.25
T.A. (sig.)	28.75% \pm 2.28	15.12% \pm 1.99	19.29% \pm 2.17	14.91% \pm 1.93
Bag of vect.	0% \pm 0	0% \pm 0	0% \pm 0	0% \pm 0
Graph	0% \pm 0	0% \pm 0	0% \pm 0	0% \pm 0

Table 1

Misclassification rate on the test set, average over 30 runs. High SNR ($\sigma_b = 0.02$). |Learn.|\|Test.| : number of data in learning and test set for each class, Coef. and Raw sig. : coefficients and raw signals used as descriptors, Stat. descr.: classical statistical descriptors, T.A. : translation arrangements on coefficients and raw signals, Bag of vect. : method based on bag of vectors, Graph : graph kernel.

Learn. \ Test.	1/1000	400/1000		
Classifiers	1-nn = SVM	1-nn	3-nn	SVM
Coef.	49.59% \pm 4.74	24.76% \pm 1.53	25.02% \pm 1.47	31.14% \pm 1.43
Raw sig.	49.99% \pm 0.84	45.08% \pm 2.01	45.34% \pm 1.87	47.08% \pm 0.97
Stat. descr.	47.16% \pm 8.22	40.54% \pm 2.12	39.64% \pm 2.29	19.27% \pm 0.98
T.A. (coef.)	46.12% \pm 3.39	17.17% \pm 1.54	18.61% \pm 1.24	23.57% \pm 1.38
T.A. (sig.)	49.73% \pm 1.15	43.17% \pm 1.99	43.53% \pm 2.07	42.27% \pm 1.63
Bag of vect.	29.40% \pm 15.24	11.9% \pm 1.13	10.96% \pm 0.97	8.29% \pm 0.80
Graph	13.98% \pm13.26	6.66% \pm 0.64	6.71% \pm 0.65	5.25% \pm0.35

Table 2

Misclassification rate on the test set, average over 30 runs. Low SNR ($\sigma_b = 0.2$). |Learn.|\|Test.| : number of data in learning and test set for each class, Coef. and Raw sig. : coefficients and raw signals used as descriptors, Stat. descr.: classical statistical descriptors, T.A. : translation arrangements on coefficients and raw signals, Bag of vect. : method based on bag of vectors, Graph : graph kernel.

artifacts (note that 3 outliers have been removed from the original data set). Signal lengths have been reduced to 1024 samples. Examples of these two classes of signal have been reported in Figure (6). Before experiments, all the signals have been normalized so that their maximal value is 1. A time-scale representation of the signal through a continuous wavelet transform with a

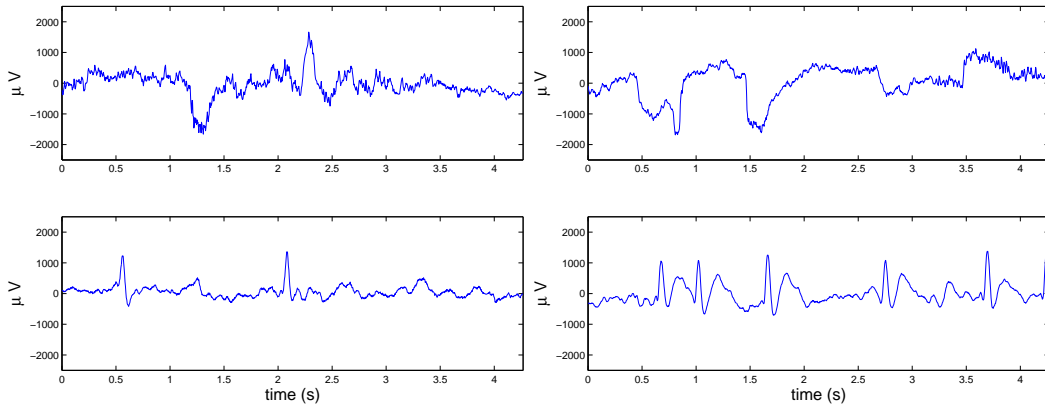


Fig. 6. Examples of EEG recordings without epileptic spikes (top) and with a single epileptic spike (bottom left) or with several spikes (bottom right).

Mexican hat wavelet has been used as a time-covariant signal representation. Since it is well known that EEG spikes are high frequency signals, the continuous wavelet transform has been set so that it focuses on high frequencies (Durka, 2004).

For the classification experiments, we have compared a translation arrangement kernel on the time-series and our graph kernel using SVMs. The decision functions have been obtained by using 5 randomly drawn signals of each class in the training set. The performances have been evaluated on the remaining signals. In order to get rid of the influence of the random draw, the results we present are averaged over 1000 different trials.

For both cases, the parameter settings have been chosen in order to achieve the best performance on the test set. For the translation arrangement kernel, we have built a set of translated version of each signal with a translation step of 10 samples, a Gaussian kernel with $\sigma = 3$ and a SVM parameter $C = 1$. For the graph kernel, we have decided to use as node labels the parameters defined in section (3). The gaussian kernel parameter for the node and edge kernel has been set to $\sigma = 0.2$ while the SVM parameter C has been fixed to 1.

The results we achieve are very interesting since the graph kernel error rate is as low as 2.8% whereas the translation arrangement kernel yields to 4% error rate. Furthermore, if a Gaussian white noise of standard deviation $\sigma = 0.5$ is added to each normalized signal, then performances of graph kernel and translation arrangement kernels are respectively 8% and 15%.

In summary, this experiment shows us that the translation-invariant kernel we propose is able to perform very well when model selection is appropriately performed. Furthermore, it is more robust to noise than a translation arrangement kernel. Note that in this case, the time complexity of a translation kernel is rather high since it needs the computation of about 100 translated signals.

6 Conclusions

Non vectorial descriptors open new perspectives in various fields. In the case of non-stationary patterns, randomly located in a signal, the graph-based representation we proposed enables us to keep a time description without absolute reference. The results point out the interest of such a description and its efficiency. The comparison with bag of vectors kernel is the best illustration of the contribution of graph representation and kernel for increasing performances. In fact, even if the shapes of the Gaussians are more discriminants than their locations (Fig. 3), graph kernel outperformed the bag of vectors algorithm by more than 3%.

Furthermore the approach we proposed in this paper can be extended to other domains that signal classification. Since the graph representation has been built from a two-dimensional representation, we believe that it would be possible to apply this approach to image classification or to object recognition in images.

In the context of signal classification, the perspectives of this work are three fold. First, we plan to optimize the time-frequency or time-scale representation with respects to a discriminant measure. Then, we should explore different ways to improve the modeling of the TS or TF Representation. A Gaussian mixture model can be obtained by using Expectation Maximization algorithm. Currently, Gaussians are selected according to their energy level. We need to define a criterion (like Fisher's one) to determine which part of the plane (or which Gaussian) are discriminants. This will enable us to treat the case where discriminant information resides in low energy part of the plane. Finally, the problem of model selection in graph kernel is an important point to address. We aim at finding an automatic procedure for selecting the parameters of our graph-based representation of signals.

Acknowledgements

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

Baraniuk, R. G., Jones, D. L., 1993. Signal-dependent time-frequency analysis using a radially gaussian kernel. *Signal Processing* 32, 263–284.

- Boashash, B., Lovell, B. C., White, L. B., 1987. Time-frequency analysis and pattern recognition using singular value decomposition of the wigner-ville distribution. In: SPIE Conf. On Advanced Algorithms and Architectures for Signal Processing. No. 826. San Diego, pp. 104–114.
- Chapelle, O., Schölkopf, B., 2002. Incorporating invariances in nonlinear support vector machines. In: Advances in Neural Information Processing Systems. Vol. 14. pp. 609–616.
- Cohen, L., 1995. Time Frequency Analysis. Prentice Hall.
- Crouse, M., Nowak, R., Baraniuk, R., 1998. Wavelet-based statistical signal processing using hidden markov models. IEEE Transactions on Signal Processing 46 (4), 886–902.
- Davy, M., 2000. Noyaux optimisés pour la classification dans le plan temps-fréquence - proposition d'un algorithme constructif et d'une référence bayésienne basée sur les méthodes mcmc - application au diagnostic d'enceintes acoustiques. Ph.D. thesis, Université de Nantes.
- Davy, M., Gretton, A., Doucet, A., Rayner, P., Dec. 2002. Optimised support vector machines for nonstationary signal classification. IEEE Signal Processing Letters 9 (12), 442–445.
- Donoho, D., 1995. De-noising by soft-thresholding. IEEE Transactions on Information Theory 41, 613–627.
- Durka, P., 2004. Adaptive time-frequency parametrization of epileptic eeg spikes. Physical Review E 69, 1–8.
- Farwell, L., Donchin, E., 1988. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. In: Electroencephalography and Clinical Neurophysiology. Vol. 70. pp. 510–523.
- Gaertner, T., 2003. A survey of kernels for structured data. In: SIGKDD Explorations. Vol. 5.
- Greenspan, H., Goldberger, J., Ridel, L., 2001. In: Computer Vision and Image Understanding. Vol. 84. pp. 384–406.
- Guigue, V., Rakotomamonjy, A., Canu, S., 2003. Reconnaissance d'émotions par méthodes à noyaux. In: 19e colloque GRETSI sur le Traitement du Signal et des Images. Vol. 1. Paris, France, pp. 69–76.
- Gärtner, T., Flach, P., Kowalczyk, A., Smola, A., 2002. Multi-instance kernels. In: ICML. pp. 179–186.
- Haasdonk, B., Keysers, D., 2002. Tangent distance kernels for support vector machines. In: icpr. Vol. 2. p. 20864.
- Haselsteiner, E., 2000. Neural based methods for time series classification. Ph.D. thesis, Technischen Universität Graz.
- Healey, J., 2000. Wearable and automotive systems for the recognition of affect from physiology. Ph.D. thesis, MIT.
- Hory, C., 2002. Mixtures of chi2 distributions for the interpretation of a time frequency representation. Ph.D. thesis, INPG.
- Kashima, H., Tsuda, K., Inokuchi, A., 2003. Marginalized kernels between labeled graphs. In: 20th International Conference on Machine Learning. AAAI Press, pp. 321–328.

- Latka, M., Was, Z., Kozik, A., West, B., 2005. Wavelet analysis of epileptic spikes. *Physical Review E Submitted*, 1–10.
- Lucas, M., Doncarli, C., Hitti, E., Dechamps, N., May 2002. Wavelet optimization for classification. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Orlando, Florida, USA.
- Mallat, S., 1997. *A Wavelet Tour Of Signal Processing*. Academic Press.
- Mallat, S., Zhong, S., 1992. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Anal. and Mach. Intell.* 14 (7), 710–732.
- Michel, O., Flandrin, P., Hero, A., 2000. Automatic extraction of time-frequency skeletons with minimal spanning trees. In: *ICASSP*. Vol. 1. pp. 89–92.
- Qian, S., 2001. *Introduction to Time Frequency and Wavelet Transforms*. Prentice Hall.
- Saito, N., Coifman, R., 1994. Local discriminant bases. In: *Wavelet Applications in Signal and Image Processing, Proc. SPIE 2303*. pp. 2–14.
- Simard, P., Cun, Y. L., Denker, J., Victorri, B., 1998. Transformation invariance in pattern recognition, tangent distance and tangent propagation. *Lecture Notes in Computer Science 1524*, 239–274.
- Sukittanon, S., Atlas, L., Pitton, J., McLaughlin, J., 2003. Non-stationary signal classification using joint frequency analysis. In: *ICASSP*. Vol. 6. pp. 453–456.
- Wallraven, C., Caputo, B., Graf, A., 2003. Recognition with local features : the kernel recipe. In: *ICCV 2003 Proceedings*. Vol. 2. IEEE Press, pp. 257–264.