Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of...

Article *in* Information Sciences · July 2013

DOI: 10.1016/j.ins.2009.05.012

CITATIONS
84

READS
416

7 authors, including:

Javier Ramírez
University of Granada
325 PUBLICATIONS **3,754** CITATIONS
SEE PROFILE

Juan M Gorriz
University of Granada
345 PUBLICATIONS **2,972** CITATIONS
SEE PROFILE

Miriam López
University of Granada
78 PUBLICATIONS **1,225** CITATIONS
SEE PROFILE

Manuel Gomez-Rio
Hospital Universitario Virgen de las Nieves
83 PUBLICATIONS **1,249** CITATIONS
SEE PROFILE

Information Sciences 237 (2013) 59–72

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features ☆

J. Ramírez [a,*], J.M. Górriz [a], D. Salas-Gonzalez [a], A. Romero [a], M. López [a], I. Álvarez [a], M. Gómez-Río [b]

[a] Dpto. Teoría de la Señal, Telemática y Comunicaciones, Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
[b] Department of Nuclear Medicine, Hospital Universitario Virgen de las Nieves, Avda. de las Fuerzas Armadas, 2, Granada, Spain

ARTICLE INFO

Article history:
Received 9 July 2008
Received in revised form 22 April 2009
Accepted 21 May 2009
Available online 30 May 2009

Keywords:
Support vector machines
Computer-aided diagnosis
Alzheimer's type dementia
Feature extraction

ABSTRACT

Alzheimer's disease (AD) is the most common cause of dementia in the elderly and affects approximately 30 million individuals worldwide. With the growth of the older population in developed nations, the prevalence of AD is expected to triple over the next 50 years while its early diagnosis remains being a difficult task. Functional imaging modalities including Single Photon Emission Computed Tomography (SPECT) and positron emission tomography (PET) are often used with the aim of achieving early diagnosis. However, conventional evaluation of SPECT images often relies on manual reorientation, visual reading of tomographic slices and semiquantitative analysis of certain regions of interest (ROIs). These steps are time consuming, subjective and prone to error. This paper shows a fully automatic computer-aided diagnosis (CAD) system for improving the early detection of the AD. The proposed approach is based on image parameter selection and support vector machine (SVM) classification. A study is carried out in order to finding the ROIs and the most discriminant image parameters with the aim of reducing the dimensionality of the input space and improving the accuracy of the system. Among all the features evaluated, coronal standard deviation and sagittal correlation parameters are found to be the most effective ones for reducing the dimensionality of the input space and improving the diagnosis accuracy when a radial basis function (RBF) SVM is used. The proposed system yields a 90.38% accuracy in the early diagnosis of the AD and outperforms existing techniques including the voxel-as-features (VAF) approach.

1. Introduction

Emission Computed Tomography (ECT) has been widely used in biomedical research and clinical practice during the last three decades. ECT differs from many other medical imaging modalities such as magnetic resonance imaging (MRI) in producing a mapping of physiological functions instead of imaging anatomical structures. In this way, tomographic radiopharmaceutical imaging provides in vivo three-dimensional maps of a pharmaceutical labeled with a gamma ray emitting

☆ This work was partly supported by the MICINN under the PETRI DENCLASES (PET2006-0253), TEC2008-02113, NAPOLEON (TEC2007-68030-C02-01) and HD2008-0029 projects and the Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía, Spain) under the Excellence Project (TIC-02566).
* Corresponding author. Tel.: +34 958240842; fax: +34 958240831.
E-mail address: javierrp@ugr.es (J. Ramírez).

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2009.05.012

radionuclide. The distribution of radionuclide concentrations are estimated from a set of projectional images acquired at many different angles around the patient [5].

Single Photon Emission Computed Tomography (SPECT) is an ECT imaging technique developed in the 1960s, but not widely used in clinical practice until the 1980s. It is a noninvasive, three-dimensional functional imaging modality that provides clinical information regarding biochemical and physiologic processes in patients. SPECT images are produced by the disintegration of the nucleus of a radioisotope that leads to the emission of a gamma photon with a random direction and uniformly distributed in the sphere surrounding the nucleus. If the photon does not suffer a collision with electrons or other particles with-in the body, its trajectory will be a straight line or "ray". In order to discriminate the direction of incidence using a photon detector external to the patient, a physical collimation is required. Typically, a collimator is placed prior to the detector in such a manner that photons incident from all but a single direction are blocked by the plates. This guarantees that only photons incident from the desired direction will strike the photon detector. SPECT is essential for imaging the brain with either regional cerebral blood flow (rCBF) agents or brain receptors, and for imaging myocardial perfusion. For the past two decades, brain SPECT has become an important diagnostic and research tool in nuclear medicine [2,32]. The ultimate value of this technology depends on good technique for image acquisition and proper data reconstruction [38,50].

Alzheimer's disease (AD) is a progressive neurodegenerative disorder first affecting memory functions and then gradually affecting all cognitive functions with behavioral impairments and eventually causing death. Its diagnosis is based on the information provided by a careful clinical examination, a thorough interview of the patient and relatives, and a neuropsychological assessment. A SPECT rCBF study is frequently used as a complimentary diagnostic tool in addition to the clinical findings [24,43,23]. However, in late-onset AD there are minimal perfusion in the mild stages of the disease, and age-related changes, which are frequently seen in healthy aged people, have to be discriminated from the minimal disease-specific changes. These minimal changes in the images make visual diagnosis a difficult task that requires experienced explorers. Even with this problem still unsolved, the potential of computer-aided diagnosis (CAD) has not been explored in this area.

Several approaches for designing CAD systems of the AD can be found in the literature [33,26,45]. The first family is based on the analysis of regions of interest (ROI) by means of some discriminant functions. The second approach is the statistical parametric mapping (SPM) [15] software tool and its numerous variants. SPM is widely used in neuroscience. It was not developed specifically to study a single image, but for comparing groups of images. SPM has been designed as a univariate approach since the classical multivariate techniques such as MANCOVA [47] require the number of observations (i.e. scans) to be greater than the number of components (i.e. voxels) of the multivariate observation. The importance of multivariate approaches is that the effects due to activations, confounding effects and error effects are assessed statistically, both in terms of effects at each voxel, and interactions among voxels [15]. On the other hand, statistical learning classification methods have not been explored in depth for AD CAD, quite possibly due to the fact that images represent large amounts of data and most imaging studies have relatively few subjects (generally < 100) [26,46,44].

Since their introduction in the late seventies [51], Support vector machines (SVMs) marked the beginning of a new era in the learning from examples paradigm [6,27]. SVMs have focussed recent attention from the pattern recognition community due to a number of theoretical and computational merits derived from the Statistical Learning Theory (SLT) [52,53] developed by Vladimir Vapnik at AT&T. Moreover, recent developments in defining and training statistical classifiers make it possible to build reliable classifiers in very small sample size problems [12] since pattern recognition systems based on SVM circumvent the *curse of dimensionality*, and even may find nonlinear decision boundaries for small training sets. These techniques have been successfully used in a number of applications [7,49] including voice activity detection (VAD) [13,14,37,40,39,56,21,20], content-based image retrieval [48], texture classification [29] and medical imaging diagnosis [16,28,57,35].

This paper shows a complete CAD system for the early detection of the Alzheimer' type dementia (ATD) by SPECT image classification. The proposed method combining SVM and advanced feature extraction schemes is developed with the aim of reducing the subjectivity in visual interpretation of SPECT scans by clinicians, thus improving the detection of the AD in its early stage. The paper is organized as follows. Section 2 provides a background on SVM classification. Section 3 summarizes the key ROIs where the disease becomes observable in its early stage. Section 4 shows the acquisition setup, reconstruction algorithm as well as the template-based spatial normalization techniques used for obtaining an accurate and anatomically standardized model of the functional brain activity provided by SPECT images. Section 5 defines first- and second-order statistics that are evaluated for building the classifier based on their discrimination ability. Finally, Section 6 shows the experiments that were conducted in order to evaluate the proposed SVM classifier as a diagnostic tool for the early detection of the AD.

2. Support vector machines

Support vector machines [6,52,53] are widely used for pattern recognition in a number of applications by its ability to learn from experimental data. The reason is that SVM often performs better than other conventional parametric classifiers [27,19]. SVM separate a given set of binary labeled training data by means of a hyperplane that is maximally distant from the two classes (known as the maximal margin hyperplane). The objective is to build a function $f : R^N \rightarrow \{\pm 1\}$ using training data that is, N-dimensional patterns $\mathbf{x}_i$ and class labels y_i:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l) \in R^N \times \{\pm 1\}, \tag{1}$$

so that f will correctly classify new unseen examples $(\mathbf{x}, y)$.

Linear discriminant functions define decision hypersurfaces or hyperplanes in a multidimensional feature space, that is:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0, \tag{2}$$

where $\mathbf{w}$ is known as the weight vector and w_0 as the threshold. The weight vector $\mathbf{w}$ is orthogonal to the decision hyperplane and the optimization task consists of finding the unknown parameters w_i, $i = 1, \ldots, N$, defining the decision hyperplane.

Let x_i, $i = 1, 2, \ldots, l$, be the feature vectors of the training set, X. These belong to either of the two classes, ω_1 or ω_2. If the classes were linearly separable the objective would be to design a hyperplane that classifies correctly all the training vectors. The hyperplane is not unique and the selection process focusses on maximizing the generalization performance of the classifier, that is, the ability of the classifier, designed using the training set, to operate satisfactorily with new data. Among the different design criteria, the maximal margin hyperplane is usually selected since it leaves the maximum margin of separation between the two classes. Since the distance from a point $\mathbf{x}$ to the hyperplane is given by $z = |g(x)|/||w||$, scaling $\mathbf{w}$ and w_0 so that the value of $g(x)$ is +1 for the nearest point in ω_1 and -1 for the nearest points in ω_2, reduces the optimization problem to maximizing the margin: $2/||w||$ with the constraints:

$$w^Tx + w_0 \geqslant 1, \quad \forall x \in \omega_1, w^Tx + w_0 \leqslant 1, \quad \forall x \in \omega_2, \tag{3}$$

or equivalently, minimizing the cost function $J(w) = (1/2)||w||^2$ subject to:

$$y_i(w^Tx_i + w_0) \geqslant 1, \quad i = 1, 2, \ldots, l. \tag{4}$$

Thus, designing the classifier leads to a nonlinear (quadratic) optimization task subject to a set of linear inequality constraints. By using the optimization methodology adopted by Karush-Kuhn-Tucker [6], the solution $\mathbf{w}$ is found to be a linear combination of $N_s \leqslant l$ feature vectors named support vectors and the optimum hyperplane is called the support vector machine. The support vectors are the training vectors that are closest to the linear classifier since lie on either of the two hyperplanes, i.e. $w^Tx + w_0 = \pm 1$. On the other hand, the optimization process with inequality constraints guarantee any local minimum is also global and unique so that the optimal maximal margin hyperplane defining the support vector machine is unique.

For non-separable classes, the optimization process needs to be modified in an efficient and elegant manner. In mathematical terms, the maximal margin hyperplane for non-separable data is selected by minimizing the cost function:

$$J(w, w_0, \xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i, \tag{5}$$

subject to the constraints:

$$y_i[w^Tx_i + w_0] \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0 \quad i = 1, 2, \ldots, l., \tag{6}$$

where the variables ξ_i are known as slack variables. Note that, the goal of the optimization task is to make the margin as large as possible and reduce the number of points with $\xi > 0$. The parameter C is a positive constant that controls the relative influence of the two competing terms.

When no linear separation of the training data is possible, SVM can work in combination with kernel techniques so that the hyperplane defining the SVM corresponds to a nonlinear decision boundary in the input space. If the data is mapped to some other (possibly infinite dimensional) Euclidean space using a mapping $\Phi(x)$, the training algorithm only depends on the data through dot products in such an Euclidean space, i.e. on functions of the form $\Phi(x_i) \cdot \Phi(x_j)$. If a "kernel function" K is defined such that $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, it is not necessary to know the Φ function during the training process. In the test phase, an SVM is used by computing dot products of a given test point $\mathbf{x}$ with $\mathbf{w}$, or more specifically by computing the sign of

$$f(x) = \sum_{i=1}^{N_S}\alpha_i y_i \Phi(s_i) \cdot \Phi(x) + w_0 = \sum_{i=1}^{N_S}\alpha_i y_i K(s_i, x) + w_0, \tag{7}$$

where s_i are the support vectors.

Thus, the use of kernels in SVM enables to map the data into some other dot product space (called feature space) F via a nonlinear transformation $\Phi : R^N \rightarrow F$ and perform the above linear algorithm in F. Fig. 1 illustrates this process where the 2-D input space is mapped to a 3-D feature space. Note that, all the points belonging to a given class remain at a given side of the separating hyperplane and the data become linearly separable. In the input space, the hyperplane corresponds to a nonlinear decision function whose form is determined by the kernel. There are three common kernels that are used by SVM practitioners for the nonlinear feature mapping:

- Polynomial

$$K(x, y) = [\gamma(x \cdot y) + c]^d. \tag{8}$$

- Radial basis function (RBF)

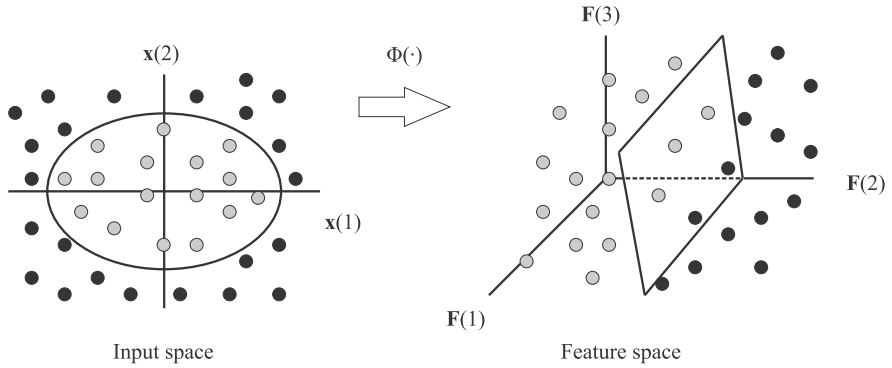


Fig. 1. Effect of mapping the input space to the feature space where the separation boundary becomes linear.

$$K(x,y) = \exp(-\gamma||x-y||^2). \tag{9}$$

- Sigmoid

$$K(x,y) = \tan h(\gamma(x\cdot y)+c). \tag{10}$$

Thus, the decision function is nonlinear in the input space

$$f(x) = \mathrm{sgn}\left\{\sum_{i=1}^{N_s}\alpha_i y_i K(s_i,x) + w_0\right\} \tag{11}$$

and the parameters α_i are the solution of a quadratic optimization problem that are usually determined by Quadratic Programming (QP) or the well known Sequential Minimal Optimization (SMO) algorithm [36]. Many classification problems are separable in the feature space and are able to obtain better results by using RBF kernels instead of linear and polynomial kernel functions [9,17].

3. Diagnosis of Alzheimer's type dementia by means of SPECT

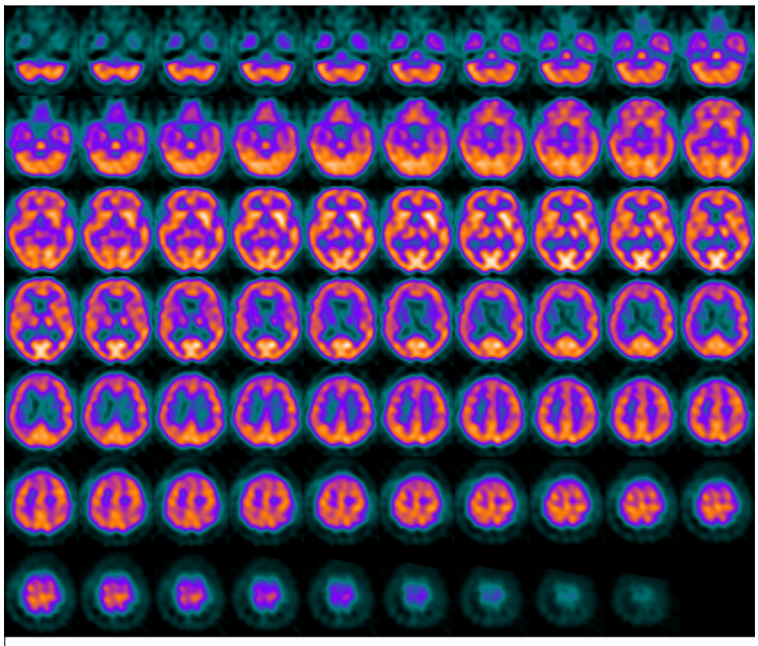
Functional SPECT imaging providing information about the rCBF have been found to be a valuable aid for the early diagnosis of the AD [18]. Fig. 2 shows typical brain perfusion patterns of a normal subject and a patient affected by AD. Although many studies exist no final agreement has been achieved for the best regions of the brain to be quantified when diagnosing AD:

- Many studies have shown the temporo-parietal region to be practical for the early detection of the disease in patients that are no longer characterized by specific cognitive impairment but by general cognitive decline [10]. Although bilateral temporo-parietal abnormalities, with or without other regional defects, are known as the predominant pattern for AD, they appear to be neither sensitive nor specific of the early AD.
- Perfusion deficits in posterior cingulate gyri and precunei regions are probably more specific and more frequent in early AD than temporo-parietal deficits [30].
- Hypo-perfusion in the medial temporal lobe and hippocampus is not found in mild AD due to the difficulties of imaging these deep brain structures [4].

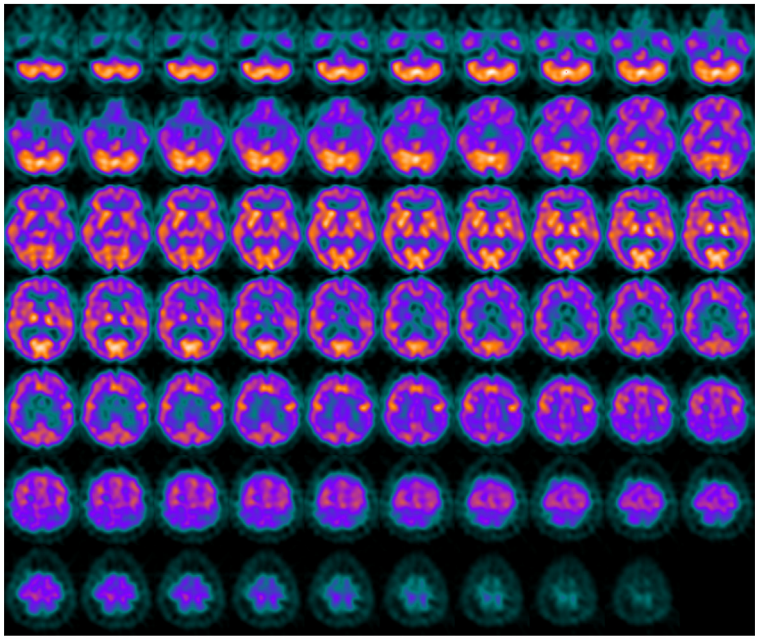
As a conclusion, not all the information found in a complete SPECT scan will be of interest for the diagnosis of the disease during its early stage. An study of the regions in the brain that are more effective for diagnosing the ATD is conducted in Section 5.

4. SPECT image acquisition and preprocessing

The ultimate value of a CAD system strongly depends on effective techniques for image acquisition, proper data reconstruction and image registration [1,42]. After introducing all the necessary knowledge and tools for building the diagnosis system, this section shows the image acquisition setup and preprocessing steps of the SPECT scans that are needed prior to defining the classifier.



(a)



(b)

Fig. 2. Typical perfusion patterns of: (a) a normal subject, and (b) a patient affected by Alzheimer's type dementia.

4.1. Image acquisition

Each patient is comfortably positioned on the imaging couch with the head "immobilized" in a radiolucent head holder. The patient's head is fixed so that a line perpendicular to the detector head runs 5°–7° cephalad to the canthomeatal line. In addition, the detector is positioned as close to the brain's patient as possible, preferably with a radius of rotation of 14 cm or less from the surface of the collimator to the center of the patient's brain. The patients are injected with a gamma emitting technetium-99 m labeled ethyl cysteinate dimer (^{99m}Tc-ECD) radiopharmaceutical and the SPECT scan is acquired by a three-

head Picker Prism 3000 gamma camera. A total of 180 projections are taken for each patient with a 2-degree angular resolution. Finally, images of the brain cross sections are reconstructed from the projection data using the filtered backprojection (FBP) algorithm described below in combination with a Butterworth noise removal filter.

4.2. Image reconstruction

Cross-sectional images of the brain can be reconstructed from projection data [31,54,25,5,8]. Assumuing ideal conditions, projections are a set of measurements of the integrated values of some parameter of the object along a line path. If the object is represented by a two-dimensional function $f(x,y)$, the line integral computed over the line $x\cos\theta + y\sin\theta = t$ is defined as:

$$P_\theta(t) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - t)dx\,dy, \tag{12}$$

where $P_\theta(t)$ is known as the Radon transform of the function $f(x,y)$. Note that, it is related with the sum of radioactive counts recorded in any time interval at point t when the detector is at angle θ.

The key to tomographic imaging is the *Fourier Slice Theorem* which relates the measured projection data to the 2-D Fourier transform of the object cross section. Thus, the Fourier transform $S_\theta(w)$ of a parallel projection $P_\theta(t)$ of an image $f(x,y)$ taken at angle θ and defined to be:

$$S_\theta(w) = \int_{-\infty}^{+\infty} P_\theta(t)\exp(-j2\pi wt)dt, \tag{13}$$

gives a slice of the two-dimensional Fourier transform:

$$F(u,v) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\exp(-j2\pi(ux+vy))dx\,dy, \tag{14}$$

subtending an angle θ with the u-axis", that is,

$$S_\theta(w) = F(u = w\cos\theta,\ \ v = w\sin\theta). \tag{15}$$

The above result is the key of straight ray tomography and indicates that by having projections of an object function at a number of angles $\theta_1,\theta_2,\ldots,\theta_k$ and taking the Fourier transform of them, the values of $F(u,v)$ can be determined on radial lines. In practice, only a finite number of projections are taken so that the function $F(u,v)$ is only known along a finite number of radial lines.

Projection data used in this study are reconstructed using the filtered backprojection (FBP) algorithm that is easily derived from the Fourier Slice Theorem. The image of the cross section $f(x,y)$ of an object is obtained by:

$$f(x,y) = \int_0^{\pi} Q_\theta(x\cos\theta + y\sin\theta)d\theta, \tag{16}$$

where

$$Q_\theta(t) = \int_{-\infty}^{+\infty} S_\theta(w)|w|\exp(j2\pi wt)dw. \tag{17}$$

The FBP algorithm then consists of two steps: the filtering part, which can be visualized as a simple weighting of each projection in the frequency domain, and the backprojection part.

A major drawback of FBP is the undesired amplification of the high frequency noise and its impact on the quality of the reconstructed image. These effects are caused by the filtering operation or multiplication of $S_\theta(w)$ by $|w|$ in Eq. (17). In order to attenuate the high frequency noise amplified during FBP reconstruction, several window functions have been proposed. The reconstruction method is then described by Eqs. (16) and (17) and is normally redefined by applying a frequency window which returns to zero as the frequency tends to π. Among the most common window functions used for FBP reconstruction are: (i) sinc (Shepp-Logan filter), (ii) cosine, (iii) Hamming and, (iv) Hanning window functions. However, even when the reconstruction noise is kept low using a noise controlled FBP approach, the noise captured by the acquisition system needs to be filtered out to improve the quality of the reconstructed images. In this way, the preprocessing stage of most automatic SPECT image processing systems often incorporates prefiltering, reconstruction and postfiltering to minimize the noise acquired by the gamma camera as well as the noise amplified during FBP reconstruction.

4.3. Image registration

The complexity of brain structures and the differences between brains of different subjects make necessary the normalization of the images with respect to a common fixed-size template. This step allows us to compare the voxel intensities of the brain images of different subjects.

The SPECT images used in this work are first spatially normalized using the SPM software [15] in order to ensure that a given voxel in different images refer to the same anatomical position in the brain. The normalization method assumes a general affine model with 12 parameters [55] and a cost function which presents an extreme value when the template and the image are matched together. For each voxel $x = (x_1, x_2, x_3)$ in an image, the affine transformation into the coordinates $y = (y_1, y_2, y_3)$ is expressed by a matrix multiplication $y = Mx$.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}. \tag{18}$$

The objective function CF to be optimized is the mean squared difference between the source and the template:

$$cf = \sum_i (f(Mx_i) - g(x_i))^2, \tag{19}$$

where f and g denote the source image and the template, respectively. Once the image is normalized by means of an affine transformation, it is registered using a more complex non-rigid spatial transformation model. The deformations are parameterized by a linear combination of the lowest-frequency components of the three-dimensional cosine transform bases [3]. A small-deformation approach is used and regularization is achieved by the bending energy of the displacement field. After the spatial normalization, a $95 \times 69 \times 79$ voxel representation of each subject is obtained. Each voxel represents a brain volume of $2.18 \times 2.18 \times 3.56\ \text{mm}^3$.

Fig. 3 shows an example of the operation of the normalization process on SPECT images. Left column shows arbitrary source images in the dataset, central column shows the template used for image registration, and finally the corresponding normalized images are shown in the right column. It is clearly shown that the transformed image matches the shape of the template.

Finally, intensity level of the SPECT images is normalized to the maximum intensity, which is individually computed for each volume by averaging over the 3% of the highest voxel intensities following a procedure similar to [41].

5. Discriminant statistics of Alzheimer's type dementia

A major problem associated with pattern recognition systems is the so-called *curse of dimensionality*, that is, the number of available features for designing the classifier can be very large compared with the number of available training examples. Under these conditions, the difficulty of an estimation problem increases drastically with the dimension of the space, since

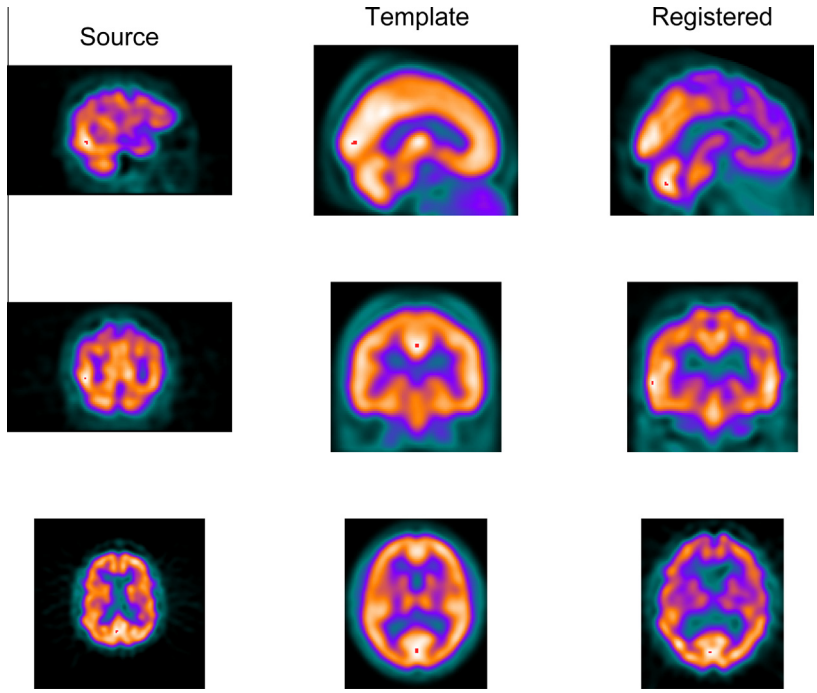


Fig. 3. Three SPECT images. *Left column*: source image. *Central column*: template. *Right column*: transformed image.

one needs an exponential increasing number of patterns to sample the space properly. This well known statement induces some doubts about where it is a good idea to go to a high dimensional feature space for learning when using SVM. However, Statistical Learning Theory tell us that the contrary can be true: learning in high dimension feature space can be simpler if a low-complexity, i.e. simple *class* of decision rule (e.g. linear classifiers) is used. As a conclusion, not the dimensionality but the complexity of the function *class* matters [34]. Moreover, recent developments in defining and training statistical classifiers make it possible to build reliable classifiers in very small sample size problems [12] since pattern recognition systems based on SVM circumvent the curse of dimensionality, and even may find nonlinear decision boundaries for small training sets.

Independently of the classifier to be used, there are clear motivations for reducing the dimensionality of the input space to a reasonable minimum:

(1) reduction of the computational cost of the training and testing algorithms,
(2) elimination of correlation between input data, and
(3) selection of the most discriminant set of input data.

Next sections review the definition of the image statistics that have been evaluated in this work for designing the CAD system.

5.1. First-order statistics

Let I be the random variable representing the N_g intensity levels of the image and $P(I)$ its first-order histogram. Its moments m_i and central moments $\hat{m}_i$ are defined by:

$$m_i = E[I^i] = \sum_{I=0}^{N_g-1} I^i P(I), \hat{m}_i = E[(I - E[I])^i] = \sum_{I=0}^{N_g-1} (I - m_1)^i P(I). \tag{20}$$

The most frequent used central moments are the mean $\mu = m_1 = E[I]$ and variance $\sigma^2 = \hat{m}_2$. Other parameter that results from the first-order histogram is the entropy:

$$H = -E[\log_2 P(I)] = -\sum_{I=0}^{N_g-1} P(I)\log_2 P(I), \tag{21}$$

which is a measure of histogram uniformity. The closer to the uniform distribution, the higher the entropy.

5.2. Second-order statistics

First-order statistics provide information related to the intensity level distribution of the image, but they do not give any information about relative positions of the various intensity levels with-in the image. This information can be extracted from the second-order statistics, where the pixels are considered in pairs.

Second-order statistics derived from the co-occurrence matrix [22] are defined as follows. Let d be the relative distance measured in pixel numbers. The orientation ϕ is quantized in four directions: horizontal, diagonal, vertical and antidiagonal (0°, 45°, 90° and 135°). For each combination of d and ϕ a two-dimensional histogram is defined. In this work, only the dependence of horizontally adjacent pixels ($d = 1$, $\phi = 0$) is considered being its spatial dependence matrix or co-occurrence matrix defined as:

$$P = (I(m,n) = I_1, I(m \pm d, n) = I_2) = \frac{\#\text{ of pairs at distance } d \text{ with values } (I_1, I_2)}{\text{total } \#\text{ of possible pairs}}. \tag{22}$$

From the definition of the co-occurrence matrix, a number of second-order statistics parameters can be defined:

- Angular second moment

$$ASM = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (P(i,j))^2. \tag{23}$$

- Contrast

$$CON = \sum_{n=0}^{N_g-1} n^2 \sum_{i=0,j=0,|i-j|=n}^{N_g-1} P(i,j). \tag{24}$$

- Inverse difference moment

$$IDF = -\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1}\frac{P(i,j)}{1+(i-j)^2}. \tag{25}$$

- Entropy

$$H_{xy} = -\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1}P(i,j)\log_2 P(i,j). \tag{26}$$

- Correlation

$$COR = \frac{1}{\sigma_x\sigma_y}\left[\sum_{i,j=0}^{N_g-1}ijP(i,j) - \mu_x\mu_y\right]. \tag{27}$$

The proposed feature selection method aims at automatically finding the most discriminant statistics of the SPECT images and the ROIs without having knowledge of the disease but analyzing the set of available images in a database.

5.3. Feature selection

In order to evaluate different sets of image parameters for the early detection of the AD, a study was carried out to assess the values the statistics take for the different classes, say ω_1, ω_2, and if they differ significantly. The main objective of the study is to combine different parameters and end up with the "best" input vector for classification. The analysis considered first- and second-order statistics of sagittal, coronal and transversal slices of the brain while the Fisher linear discriminant ratio (FDR) defined by:

$$FDR = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} \tag{28}$$

was used as class separability measure, where μ_1 and μ_2 denote the with-in class mean value of the input feature, respectively, and σ_1^2 and σ_2^2 their variances.

Fig. 4 shows an example of the procedure for selecting the most discriminant slices based on the computation of the FDR when the standard deviation and correlation are considered as inputs for classification. The first row of each figure is a 2-D image showing the value of the corresponding statistic for each patient and each of the x (sagittal), y (coronal) and z (transversal) slices. Note that, normal and ATD subjects are grouped and separated by an horizontal black line to easily show the differences in the value of the statistic among the two classes. The value of the FDR is also plotted below each 2-D feature image. The most discriminant slices are those which yields the maximum value of the FDR and identify the ROIs (slices) for detecting the AD as discussed in Section 3. This study was extended to all the first and second-order statistics that were defined previously in Section 5. It can be concluded that not all the slices in the volume element provide the same discriminant value. Among all the statistics evaluated, the standard deviation of the y (coronal) slices and correlation of the x (sagittal) slices yielded the maximum value of the FDR and were found to be the most discriminant input vectors of the AD.

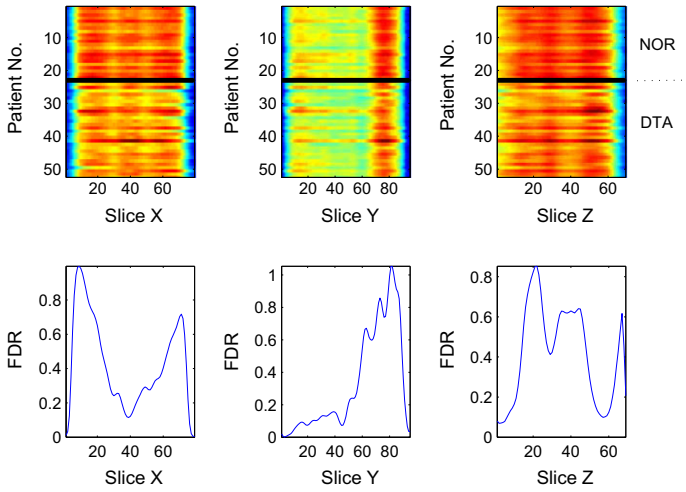
6. Evaluation results

This section shows the experimental results carried out in order to evaluate the performance of the classification system and its utility as a CAD system for the early AD. First, a baseline system based on the voxel-as-features (VAF) [45,46,44] paradigm is implemented for reference. Second, the experimental results that were conducted to evaluate the proposed system are shown.

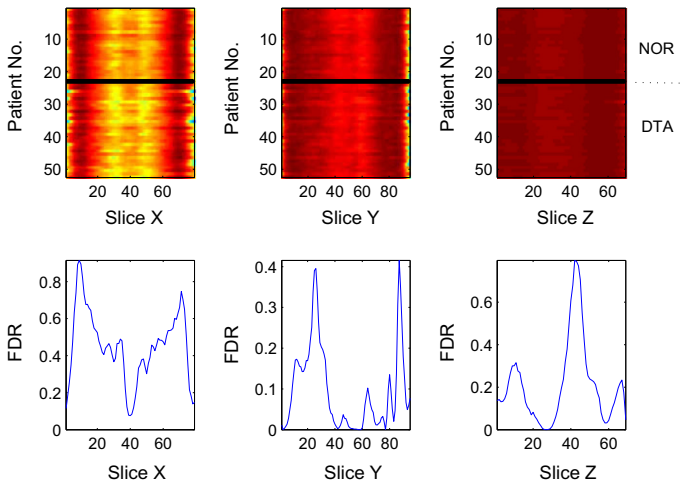
The SPECT images used in this work were initially labeled by experienced clinicians of the "Virgen de las Nieves" hospital (Granada, Spain) using 4 different labels: normal (NOR) for patients without any symptoms of ATD and possible ATD (ATD-1), probable ATD (ATD-2) and certain ATD (ATD-3) to distinguish between different levels of the presence of typical characteristics for ATD. In total, the database consists of 52 patients: 23 NOR, 13 ATD-1, 12 ATD-2 and 4 ATD-3. The latter three labels were combined and only two classes (NOR and ATD) were finally used.

6.1. Baseline voxel-as-features classification

A SVM classifier based on the VAF paradigm [45,46,44] was developed for reference. The dimension of the $95 \times 69 \times 79$-voxel volume representing the rCBF of each subject was reduced by averaging $n \times n \times n$ voxels. A SVM-based classifier using the voxel intensities as features was trained and tested by means of a leave-M-out ($M = 5$) cross validation strategy. The classifier is trained with all but M images of the database. The remaining images, which are not used to define the classifier, are then categorized. In this way, all SPECT images are classified and the success rate is computed from a number of correctly



(a) Standard deviation



(b) Correlation

Fig. 4. Standard deviation, correlation and FDR values of sagittal, coronal and transversal sections for normal subjects and ATD patients.

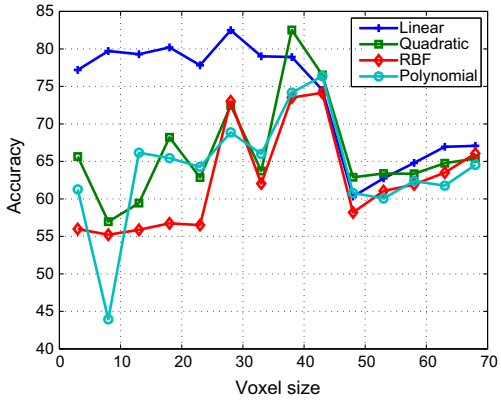


Fig. 5. Performance of the VAF-based SVM classifier evaluated using a leave-5-out cross validation strategy.

classified subjects. Fig. 5 shows the accuracy of the VAF-based SVM classifier as a function of the voxel size n for different kernels including linear, quadratic, RBF and polynomial kernels. Among all the kernels, the linear kernel is the one that yields the best accuracy for a VAF system in discriminating normal subjects from patients affected by ATD given the high dimension of the input space [11]. The performance of the linear kernel is reduced as the voxel size increases motivated by two reasons: (i) dimensionality of the input space is progressively reduced and, (ii) information is lost after averaging large volume elements. As a conclusion, the accuracy of the SVM-based VAF system proposed in [16] (below 85% for a linear kernel) needs to be improved for developing a more accurate CAD system for the early detection of the ATD.

The results above are in agreement with the SVM statistical learning theory and the ability of linear classifiers to effectively group l points of a high dimension space in two classes [11]. It can be clearly shown that linear hyperplanes can separate effectively two classes in high dimension feature spaces (in fact, this is the key of SVMs). Let us consider l points in the N-dimensional feature space. It is assumed that the points are *well distributed* so that there is no subset of $N+1$ points that lie on an $(N-1)$-dimensional hyperplane. The number $O(l,N)$ of groupings that can be formed by $(N-1)$-dimensional hyperplanes to separate the l points in two classes is given by [11]:

$$O(l,N) = 2\sum_{i=0}^{N}\binom{l-1}{i} \tag{29}$$

where

$$\binom{l-1}{i} = \frac{(l-1)!}{(l-1-i)!i!} \tag{30}$$

Thus, the probability of grouping l points in the N-dimensional feature space in two linearly separable classes is given by:

$$P_l^N = \frac{O(l,N)}{2^l} = \begin{cases} \frac{1}{2^{l-1}}\sum_{i=0}^{N}\binom{l-1}{i}, & l > N+1, \\ 1, & l \leqslant N+1. \end{cases} \tag{31}$$

Fig. 6 shows the probability P_l^N as a function of N/l. Note that, for low dimension spaces $(N/l < 0.3)$, P_l^N it is almost zero, that is, linear SVM performs poorly in discriminating the two classes. However, when the dimension of the space increases, the probability of any two groups of the l points being linearly separable approaches unity. As a conclusion, high dimensional spaces defined for reduced voxel sizes are more effectively classified by linear kernels as shown in Fig. 5. Moreover, as the voxel size increases, the dimension of the feature space decreases and the performance of linear kernels decreases when compared to RBFs, quadratic and polynomial kernels as a result of the decrease of the probability P_l^N. On the other hand, if we are given l points to be classified, then mapping into a higher dimension space increases the probability of locating then in linearly separable SVM two class groupings.

6.2. SVM training and testing results

Aiming at reducing the dimensionality of the input space and further improving the performance of the CAD system by means of more effective kernels, a SVM-based classifier was developed using the most discriminant set of input statistics obtained in Section 5.3, i.e. standard deviation of the coronal slices and the correlation of the sagittal slices. Dimensionality of the input space is reduced by considering only the image statistics of the slices with normalized FDR exceeding a given threshold. Fig. 7 shows the accuracy of the CAD system and the dimension of the input space as a function of the threshold value when a RBF kernel is used. Note that, the accuracy of the system increases up to 90% as the threshold increases. The

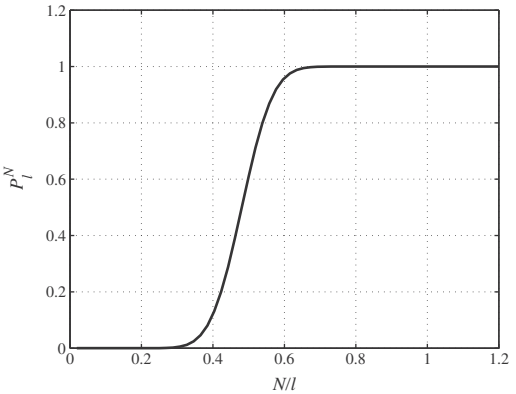


Fig. 6. Probability of grouping l points in the N-dimensional feature space in two linearly separable classes.

best results are obtained for a two-dimensional feature vector consisting of the standard deviation and correlation of the coronal and sagittal slices with the highest value of the FDR as shown in Fig. 4. These results are in agreement with the previous discussion about the dimensionality of the space. In high dimensional feature spaces, RBF kernels perform poorly. Meanwhile, reducing the dimensionality of the input space by selecting the most discriminant image statistics and ROIs (slices) in the volume improves the accuracy of the system. The benefits are obtained as a result of mapping the low dimensional input space into a high dimension feature space where the data becomes linearly separable.

Fig. 8 shows the training patterns, their associated class labels as well as the support vectors defining the SVM classification rule when linear, quadratic, RBF and polynomial kernels are used for mapping into the feature space. It is clearly shown that reducing the dimensionality of the input space to a two-coefficient input space yields high discrimination accuracy. Among all the experiments carried out, RBF kernel functions yielded the best results with a 90.38% classification

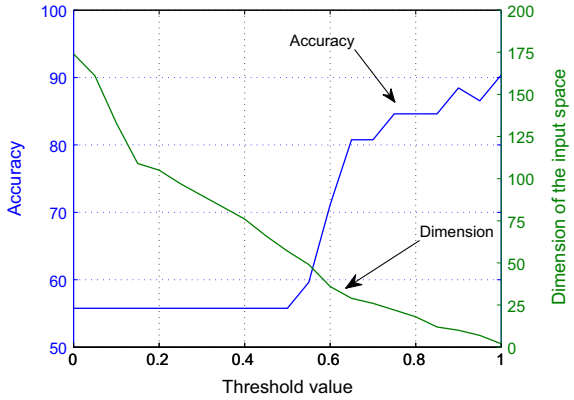


Fig. 7. Accuracy and dimension of the input space for a RBF SVM system trained using the standard deviation and correlation of the slices with the normalized FDR exceeding a given threshold.

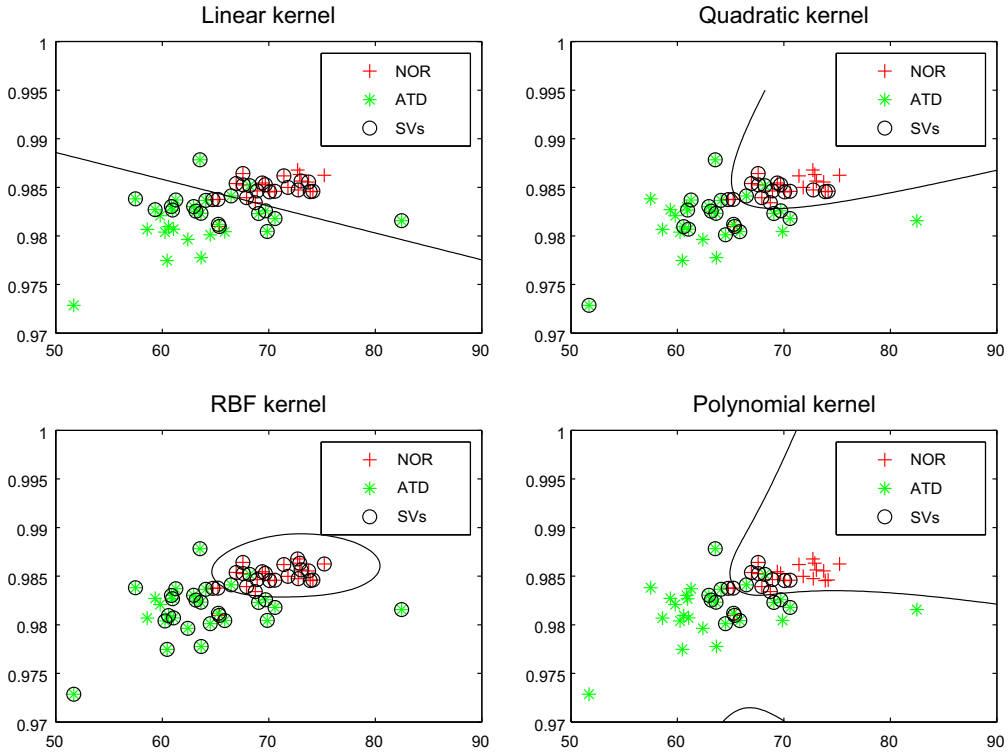


Fig. 8. Decision functions in the input space for a two-dimensional input vector consisting of the most discriminant coronal standard deviation and sagittal correlation slices.

accuracy (sensitivity of 93.10% and specificity of 86.96%). Meanwhile, linear kernels, that yield the best results in a high dimension input space (VAF approach) [16] as shown is Section 6.1, yielded just a 84.62% classification accuracy. Thus, the proposed system yielded a significant improvement over the VAF approach where the high dimension of the input space makes unnecessary a nonlinear mapping into the feature space.

7. Conclusions

This paper showed a fully automatic computer-aided diagnosis system for improving the early detection of the AD. The proposed approach is based on image parameter selection and support vector machine (SVM) classification. The system was developed by exploring the most discriminant set of input data including first and second-order statistics of sagittal, coronal and transversal sections of the human brain. It was found that the most discriminant image parameters of the AD are the coronal standard deviation and the sagittal correlation. Moreover, reducing the dimensionality of the input space by means of a FDR-based feature selection process leaded to a two-coefficient input vector that yielded high discrimination accuracy specially when a RBF kernel is used. The proposed CAD system taking advantage of RBF kernels yielded a 90.38% accuracy (sensitivity of 93.10% and specificity of 86.96%) in the early diagnosis of the ATD, thus outperforming recently developed VAF-based approaches where the high dimension of the input space makes linear SVM the most effective when compared to quadratic, RBF and polynomial kernels.

References

[1] U.R. Acharya, M. Sankaranarayanan, J. Nayak, C. Xiang, T. Tamura, Automatic identification of cardiac health using modeling techniques: a comparative study, Information Sciences 178 (23) (2008) 4571–4582.
[2] I. Álvarez, J.M. Górriz, J. Ramírez, D. Salas-González, M. López, C.G. Puntonet, F. Segovia, Alzheimer's diagnosis using eigenbrains and support vector machines, Electronics Letters 45 (7) (2009) 342–343.
[3] J. Ashburner, K.J. Friston, Nonlinear spatial normalization using basis functions, Human Brain Mapping 7 (4) (1999) 254–266.
[4] H. Braak, E. Braak, Neuropathological stageing of Alzheimer-related changes, Acta Neuropathologica 82 (4) (1991) 239–259.
[5] P.P. Bruyant, Analytic and iterative reconstruction algorithms in SPECT, The Journal of Nuclear Medicine 43 (10) (2002) 1343–1358.
[6] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (2) (1998) 121–167.
[7] X. Cao, Y. Xu, D. Chen, H. Qiao, Associated evolution of a support vector machine-based classifier for pedestrian detection, Information Sciences 179 (8) (2009) 1070–1077.
[8] E.S. Chornoboy, C.J. Chen, M.I. Miller, T.R. Miller, D.L. Snyder, An evaluation of maximum likelihood reconstruction for SPECT, IEEE Transactions on Medical Imaging 9 (1) (1990) 99–110.
[9] P. Clarkson, P. Moreno, On the use of support vector machines for phonetic classification, in: Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing, vol. 2, 1999, pp. 585–588.
[10] J.J. Claus, F. van Harskamp, M.M.B. Breteler, E.P. Krenning, I. de Koning abd, J.M. van der Cammen, A. Hofman, D. Hasan, The diagnostic value of SPECT with tc 99 m HMPAO in Alzheimer's disease. a population-based study, Neurology 44 (3) (1994) 454–461.
[11] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Transactions on Electronic Computers 14 (3) (1965) 326–334.
[12] R.P.W. Duin, Classifiers in almost empty spaces, in: International Conference on Pattern Recognition (ICPR), vol. 2 (2), 2000, pp. 4392–4395.
[13] D. Enqing, L. Guizhong, Z. Yatong, Z. Xiaodi, Applying support vector machines to voice activity detection, in: 6th International Conference on Signal Processing, vol. 2, 2002, pp. 1124–1127.
[14] D. Enqing, Z. Heming, L. Yongli, Low bit and variable rate speech coding using local cosine transform, in: Proc. of the 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering (TENCON'02), vol. 1, 2002, pp. 423–426.
[15] K.J. Friston, J. Ashburner, S.J. Kiebel, T.E. Nichols, W.D. Penny, Statistical Parametric Mapping: The Analysis of Functional Brain Images, Academic Press, 2007.
[16] G. Fung, J. Stoeckel, SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information, Knowledge and Information Systems 11 (2) (2007) 243–258.
[17] A. Ganapathiraju, J.E. Hamaker, J. Picone, Applications of support vector machines to speech recognition, IEEE Transactions on Signal Processing 52 (8) (2004) 2348–2355.
[18] I. Goethals, C. van deWiele, D. Slosman, R. Dierckx, Brain SPETC perfusion in early Alzheimer disease: where to look?, European Journal of Nuclear Medicine 29 (8) (2002) 975–978
[19] J. Górriz, J. Ramírez, A. Lassl, D. Salas-Gonzalez, E. Lang, C. Puntonet, I. Álvarez, M. López, M. Gómez-Río, Automatic computer aided diagnosis tool using component-based SVM, in: Nuclear Science Symposium Conference Record, 2008, pp. 4392–4395.
[20] J.M. Górriz, J. Ramírez, J. Segura, C. Puntonet, An effective cluster-based model for robust speech detection and speech recognition in noisy environments, The Journal of the Acoustical Society of America 120 (2006) 470–481.
[21] J.M. Górriz, J. Ramírez, E. Lang, C. Puntonet, Hard C-means clustering for voice activity detection, Speech Communication Journal 48 (2006) 1638–1649.
[22] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man and Cybernetics 3 (6) (1973) 610–621.
[23] R. Higdon, N.L. Foster, R.A. Koeppe, C.S. DeCarli, W.J. Jagust, C.M. Clark, N.R. Barbas, S.E. Arnold, R.S. Turner, J.L. Heidebrink, S. Minoshima, A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging, Statistics in Medicine 23 (2004) 315–326.
[24] J.M. Hoffman, K.A. Welsh-Bohmer, M. Hanson, FDG PET imaging in patients with pathologically verified dementia, Journal of Nuclear Medicine 41 (11) (2000) 1920–1928.
[25] H.M. Hudson, R.S. Larkin, Accelerated image reconstruction using ordered subsets of projection data, IEEE Transactions on Medical Imaging 13 (4) (1994) 601–609.
[26] K. Ishii, A.K. Kono, H. Sasaki, N. Miyamoto, T. Fukuda, S. Sakamoto, E. Mori, Fully automatic diagnostic system for early- and late-onset mild Alzheimer's disease using FDG PET and 3D-SSP, European Journal of Nuclear Medicine and Molecular Imaging 33 (5) (2006) 575–583.
[27] T. Joachims, Text categorization with support vector machines: learning with many relevant features, Lecture Notes in Computer Science 1398 (1998) 137–142.
[28] I. Kalatzis, D. Pappas, N. Piliouras, D. Cavouras, Support vector machines based analysis of brain SPECT images for determining cerebral abnormalities in asymptomatic diabetic patients, Medical Informatics and the Internet in Medicine 28 (3) (2003) 221–230.
[29] K.I. Kim, K. Jung, S.H. Park, H.J. Kim, Support vector machines for texture classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (11) (2002) 1542–1550.

[30] D. Kogure, H. Matsuda, T. Ohnishi, T. Asada, M. Uno, T. Kunihiro, S. Nakano, M. Takasaki, Longitudinal evaluation of early Alzheimer disease using brain perfusion SPECT, The Journal of Nuclear Medicine 41 (7) (2000) 1155–1162.
[31] K. Lange, R. Carson, EM reconstruction for emission and transmission tomography, Journal of Computer Assisted Tomography 8 (1984) 306–312.
[32] M. López, J. Ramírez, J.M. Górriz, D. Salas-González, I. Álvarez, F. Segovia, C.G. Puntonet, Automatic tool for the Alzheimer's disease diagnosis using PCA and bayesian classification rules, Electronics Letters 45 (8) (2009) 389–391.
[33] S. Minoshima, K.A. Frey, R.A. Koeppe, N.L. Foster, D.E. Kuhl, A diagnostic approach in Alzheimer's disease using three dimensional stereotactic surface projections of fluorine-18-FDG PET, Journal of Nuclear Medicine 36 (7) (1995) 1238–1248.
[34] K.R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based learning algorithms, IEEE Transactions on Neural Networks 12 (2) (2001) 181–201.
[35] B. Pang, D. Zhang, K. Wang, Tongue image analysis for appendicitis diagnosis, Information Sciences 175 (3) (2005) 160–176.
[36] J.C. Platt, Advances in Kernel Methods – Support Vector Learning, chap. Fast Training of Support Vector Machines using Sequential Minimal Optimization, MIT Press, 1999.
[37] F. Qi, C. Bao, Y. Liu, A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech, in: International Symposium on Chinese Spoken Language Processing, 2004, pp. 77–80.
[38] J. Ramírez, J.M. Górriz, M. Gómez-Río, A. Romero, R. Chaves, A. Lassl, A. Rodríguez, C.G. Puntonet, F. Theis, E. Lang, Effective emission tomography image reconstruction algorithms for SPECT data, Lecture Notes in Computer Science 5101 (2008) 741–748.
[39] J. Ramírez, P. Yélamos, J.M. Górriz, C.G. Puntonet, J.C. Segura, SVM-enabled voice activity detection, Lecture Notes in Computer Science 3972 (2006) 676–681.
[40] J. Ramírez, P. Yélamos, J.M. Górriz, J.C. Segura, SVM-based speech endpoint detection using contextual speech features, Electronics Letters 42 (7) (2006) 877–879.
[41] P. Saxena, D.G. Pavel, J.C. Quintana, B. Horwitz, An automatic thresholdbased scaling method for enhancing the usefulness of Tc-HMPAO SPECT in the diagnosis of Alzheimers disease, Medical Image Computing and Computer-Assisted Intervention – MICCAI, Lecture Notes in Computer Science 1496 (1998) 623–630.
[42] D.D. Sha, J.P. Sutton, Towards automated enhancement, segmentation and classification of digital brain images using networks of networks, Information Sciences 138 (1–4) (2001) 45–77.
[43] D.H. Silverman, G.W. Small, C.Y. Chang, Positron emission tomography in evaluation of dementia: regional brain metabolism and long-term outcome, Journal of the American Medical Association 286 (17) (2001) 2120–2127.
[44] J. Stoeckel, N. Ayache, G. Malandain, P.M. Koulibaly, K.P. Ebmeier, J. Darcourt, Automatic classification of SPECT images of Alzheimer's disease patients and control subjects, Lecture Notes in Computer Science 3217 (2004) 654–662.
[45] J. Stoeckel, G. Fung, SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information, in: Proc. of the Fifth International Conference on Data Mining (ICDM05), 2005, pp. 410–417.
[46] J. Stoeckel, G. Malandain, O. Migneco, P.M. Koulibaly, P. Robert, N. Ayache, J. Darcourt, Classification of SPECT images of normal subjects versus images of Alzheimer's disease patients, Lecture Notes in Computer Science 2208 (2001) 666–674.
[47] B.G. Tabachnick, L.S. Fidell, Computer-Assisted Research and Design Analysis, Pearson Education, 2000.
[48] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (7) (2006) 1088–1099.
[49] H.-H. Tsai, D.-W. Sun, Color image watermark extraction based on support vector machines, Information Sciences 177 (2) (2007) 550–569.
[50] S. Vandenberghea, Y. D'Asselera, R.V. de Wallea, T. Kauppinenb, M. Koolea, L. Bouwensa, K.V. Laerec, I. Lemahieua, R. Dierckx, Iterative reconstruction algorithms in nuclear medicine, Computerized Medical Imaging and Graphics (25) (2001) 105–111.
[51] V.N. Vapnik, Estimation of Dependences Based on Empirical Data, Springer-Verlag, New York, 1982.
[52] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin, 1995.
[53] V.N. Vapnik, Statistical Learning Theory, John Wiley and Sons Inc., New York, 1998.
[54] Y. Vardi, L.A. Shepp, L. Kaufman, A statistical model for positron emission tomography, Journal of the American Statistical Association 80 (389) (1985) 8–20.
[55] R.P. Woods, Spatial transformation models, in: I.N. Bankman (Ed.), Handbook of Medical Imaging, Academic Press, San Diego, 2000, pp. 465–490 (Chapter 29).
[56] P. Yélamos, J. Ramírez, J.M. Górriz, C.G. Puntonet, J.C. Segura, Speech event detection using support vector machines, Lecture Notes in Computer Science 3991 (2006) 356–363.
[57] S.-M. Zhou, J.Q. Gan, F. Sepulveda, Classifying mental tasks based on features of higher-order statistics from EEG signals in brain-computer interface, Information Sciences 178 (6) (2008) 1629–1640.