Neurocomputing 75 (2012) 64–71

Contents lists available at ScienceDirect

Neurocomputing

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

A comparative study of feature extraction methods for the diagnosis of Alzheimer's disease using the ADNI database

F. Segovia, J.M. Górriz *, J. Ramírez, D. Salas-Gonzalez, I. Álvarez, M. López, R. Chaves,
The Alzheimer's Disease Neuroimaging Initiative[1]

Department of Signal Theory, Networking and Communications, University of Granada, Spain

ARTICLE INFO

Available online 30 July 2011

Keywords:
Computer-aided diagnosis system
Alzheimer's disease
PLS
GMM
Support vector machine
ADNI database

ABSTRACT

Several approaches appear in literature in order to develop Computed-Aided-Diagnosis (CAD) systems for Alzheimer's disease (AD) detection. Although univariate models became very popular and nowadays they are widely used, recent investigations are focused on multivariate models which deal with a whole image as an observation. In this work, we compare two multivariate approaches that use different methodologies to relieve the small sample size problem. One of them is based on Gaussian Mixture Model (GMM) and models the Regions of Interests (ROIs) defined as differences between controls and AD subject. After GMM estimation using the EM algorithm, feature vectors are extracted for each image depending on the positions of the resulting Gaussians. The other method under study computes score vectors through a Partial Least Squares (PLS) algorithm based estimation and those vectors are used as features. Before extracting the score vectors, a binary mask based dimensional reduction of the input space is performed in order to remove low-intensity voxels. The validity of both methods is tested on the ADNI database by implementing several CAD systems with linear and nonlinear classifiers and comparing them with previous approaches such as VAF and PCA.

1. Introduction

Alzheimer's Disease (AD) is one of the most severe and frequent neurodegenerative disorder in the elderly population and has dramatic health consequences as well as socio-economic implications. Furthermore, the incidence and prevalence of this disease is increasing due to the growth of the older population in developed nation.

Positron Emission Tomography (PET) is a non-invasive medical imaging modality that provides 3D maps modeling the glucose consumption rate of the brain. Since glucose consumption is related to the brain activity, PET images can be used for diagnosing several diseases, including AD. However, performing image classification via visual examination of these images can be subjective and prone to errors. For this reason, in recent years many research efforts have focused on developing a Computer-Aided Diagnosis (CAD) system for AD based on medical imaging [1–4].

The analysis of functional images using computers can be performed at several scales. On the one hand, the most familiar scale to the neuroimaging community concerns mass univariate statistical testing, which models data at the scale of individual voxel. In the early 2000s it was demonstrated [5] that a Statistical Parametric Mapping (SPM) [6] model could be a suitable model to describe the pattern of cerebral functional neurodegeneration. Nevertheless SPM suffers the inconveniences of local and univariate approaches and it was not developed specifically to study a single image, but for comparing groups of images [7]. Subsequently, the sum of abnormal t-values obtained by SPM in regions that were typically hypometabolic in AD has been proposed and used as an AD indicator, with a high accuracy rate [8]. Furthermore a recent study has shown that a voxel-based analysis of 18 FDG-PET increases the diagnostic accuracy and confidence for both AD and FrontoTemporal Dementia (FTD), particularly when findings in a clinical evaluation are not definitive, and physicians are not already highly confident with their clinical diagnosis [9].

On the other hand, multivariate approaches consider all voxels of the brain as a single observation. Recent advances in statistical classification and feature extraction techniques [10,11] have led to an intensive use of those methods. In addition, multivariate approaches are able to surmount the small sample size problem [12]. Most of these multivariate approaches use only a small set of voxels or regions to distinguish between pathological and control

* Corresponding author.
E-mail address: gorriz@ugr.es (J.M. Górriz).

[1] Data used in preparation of this article were obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf.

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.03.050

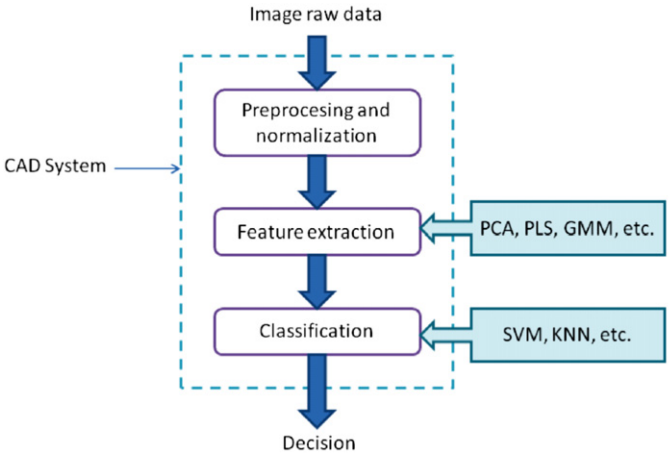


Fig. 1. General diagram of the parts that compose a CAD system (multivariate approach).

images. One of the simplest multivariate approaches for developing a CAD system for AD is the well-known Voxel-As Features (VAF) method [7]. This method separates AD patients and controls by means of a SVM classifier that is trained with all voxels (from SPECT images) with intensity value above a given threshold. Despite its simplicity, this method achieves results similar to other more sophisticated methods. In [3] a PCA-based feature extraction method is shown. The authors use the Fisher Discrimination Ratio to obtain the most important components obtained by applying a PCA algorithm to the functional images. Then, a Bayesian classifier is used to distinguish between controls and AD patients. Other multivariate approaches use some statistical measures as representation of the brain images. In [13], a reduced map of the brain is computed as the skewness of each m-by-m sliding block of the transaxial slices of the original image. After that, the voxels which present an extreme (very high or very low) Welch's t-statistic between both classes (controls and AD patients) are selected. The mean, standard deviation, skewness and kurtosis are calculated for selected voxels and these measures are chosen as features for three different classifiers: SVM, Decision Trees and Multivariate Normal Model.

In this paper, we analyze two novel methods for extracting relevant information from PET images in order to develop more accurate CAD systems for AD. On the one hand we use the Gaussian Mixture Model to parcel the Regions of Interests (ROIs) of the images. Once each ROI has been modeled with a Gaussian (with a given center, shape and weight) the feature vector for an image is computed as the activation of each ROI in that image. On the other hand, we use the score vectors computed by a Partial Least Squares (PLS) algorithm as features vectors. This approach is similar to PCA since both use the concept of latent variables, although PLS takes into account the image labels for the score vector extraction and, for this reason, this method obtains higher accuracy rates than PCA-based ones.

Fig. 1 shows a block diagram of a multivariate CAD system architecture. It consists of three parts: (i) the preprocessing and normalization methodology, (ii) the feature extraction technique and (iii) and the classification algorithm. The goal of this paper is to improve the feature extraction block in order to develop more accurate CAD systems.

2. Feature extraction based on the Gaussian mixture model

2.1. Gaussian mixtures

GMMs are one of the most statistically mature methods for classical clustering (see e.g. [14]), though they are also used intensively for density estimation [15–17]. The basic assumption of GMM for density estimation is that the given data $\mathbf{x}_i$, $i=1\dots N$ are drawn from a probability distribution $p(\mathbf{x})$, which is modeled by a sum of k Gaussians

$$p(\mathbf{x}) = \sum_{n=1}^{k} w_n f_n(\mathbf{x}|\theta_n) \tag{1}$$

where $f_n(\mathbf{x}|\theta_n)$ is the density of the Gaussian n with parameter vector θ_n and the w_n are weight factors or mixing proportions with $\sum_n w_n = 1$. The normal distributions $f_n(\mathbf{x}|\theta_n)$ in d dimensions are given by

$$f_n(\mathbf{x}|\theta_n \in \{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_n|}} \mathrm{e}^{(-1/2)(\mathbf{x}-\boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\mathbf{x}-\boldsymbol{\mu}_n)} \tag{2}$$

with expectation values $\boldsymbol{\mu}_n$ and covariance matrices $\boldsymbol{\Sigma}_n$. Geometrical features of the Gaussians can be varied by parametrization of the covariance matrices $\boldsymbol{\Sigma}_n$ using the eigenvalue decomposition [18]. For our purpose, we assume shape, volume and orientation of the Gaussians variable since the relevant activation areas (ROIs) could be located shapeless and with different sizes across the brain.

The parameters for GMM are estimated by means of the Maximum Likelihood Estimation (MLE). This procedure consists in adapting the parameters w_n, $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ in order to maximize the likelihood of a mixture model with k components:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\theta) \tag{3}$$

where $\boldsymbol{\theta} = \{\theta_n\}$, for $n=1,\dots,k$ and $\mathbf{x} = \{\mathbf{x}_i\}$, for $i=1,\dots,N$, which corresponds to the probability to observe the given samples $\mathbf{x}_i$, if independent and identically distributed random variables are assumed [15,19].

In order to simplify formulation, we can suppose that data is already grouped into a histogram with B bars at positions $\mathbf{x}_j$, $j=1,\dots,B$, and with heights I_j, the maximum likelihood estimation can be used in a modified way [20]. In addition, the gray-level of each coordinate is taken into account with the parameter I_j. In that case the total number of observations is given by $N = \sum_{j=1}^{B} I_j$ and the likelihood can be generalized to

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \prod_{j=1}^{B} [p(\mathbf{x}_j|\theta)]^{I_j} \tag{4}$$

as there are I_j observations of data points at $\mathbf{x}_j$. That methodology supposes a *parcellation* approach that uses Gaussians mixture models for density estimation of the intensity profile of a functional image.

In order to estimate the unknown parameters, the expectation-maximization (EM) algorithm is used [21,22]. Along the same lines as shown, for instance, in [14] we can write down the equations to update the unknown parameters w_n, $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$, where the relations are only modified by a weight factor $I(x_j)$:

$$w_n = \frac{1}{N} \sum_{j=1}^{B} I(\mathbf{x}_j) q_n(\mathbf{x}_j) \tag{5}$$

$$\boldsymbol{\mu}_n = \frac{1}{w_n N} \sum_{j=1}^{B} I(\mathbf{x}_j) q_n(\mathbf{x}_j) \mathbf{x}_j \tag{6}$$

$$\boldsymbol{\Sigma}_n = \frac{1}{w_n N} \sum_{j=1}^{B} I(\mathbf{x}_j) q_n(\mathbf{x}_j)(\mathbf{x}_j - \boldsymbol{\mu}_n)(\mathbf{x}_j - \boldsymbol{\mu}_n)^T \tag{7}$$

where the posterior probability, $q_n(\mathbf{x})$, is defined by $q_n(\mathbf{x}) = w_n f_n(\mathbf{x})/p(\mathbf{x})$. Starting with an initial guess for w_n, $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ the EM algorithm recursively applies Eqs. (5)–(7) until convergence is reached, i.e. the changes in the log-likelihood are smaller than a given threshold. Note that in this context, each $\mathbf{x}_j$ is a vector with the three coordinates of a voxel and $\mathbf{x}$ represents all voxels of a PET image.

The core idea of this method is to perform space quantization by populating it with Gaussian kernels whose linear combination approximates image intensity. The resulting kernel locations act as new "super-voxels" whose intensity is estimated by projecting (integrating) the image onto the kernel function. For further details about the arguments for applying GMM to functional images, see [19].

2.2. Model selection

A key question for GMM approach is the number of Gaussians used for modeling ROIs (parameter k in Eq. (1)). Note that all parameters of the model are estimated thought the MLE algorithm, but the number of Gaussian should be calculated manually. If k is large, the model will represent the image very well, thus it can be satisfactory reconstructed from the Gaussian. However, a large number of Gaussians will result in large feature vectors (as it is described in Section 2.3). Thus, we should find a balance between size of the feature vectors and ability of reconstruction (related with the model adjustment).

In order to determine this parameter, the reconstruction error, $\mathbf{E}_{rec}$, has been estimated for several configurations. Results are shown in Fig. 2:

$$\mathbf{E}_{rec} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(\mathbf{I}_i - \mathbf{I}_i^{rec})^2}{n}}}{I_{max}} \tag{8}$$

where $\mathbf{I}_i$ and $\mathbf{I}_i^{rec}$ is the i-th voxel of the original image and reconstructed image respectively, n is the number of voxels and I_{max} is the maximum intensity.

According to Fig. 2 the reconstruction error tends to stabilize when the number of Gaussians increases. In this work, we have used a model with $k=64$ Gaussians that leads to a good representation of functional images (since the reconstruction error is small) and still the dimensionality of the feature vectors is small enough. This configuration has an additional advantage: It allows initializing the model following a symmetric configuration ($4 \times 4 \times 4$ Gaussians).

2.3. Feature extraction procedure

The goal of using GMM is to parcel or delimit the ROIs on a functional image. Thus, the GMM procedure is applied only once over an image computed as the difference between controls and AD images. The first step consists of creating an image

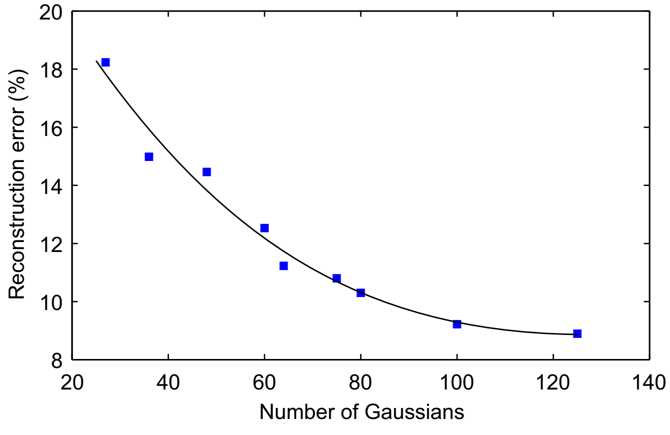


Fig. 2. Reconstruction error of a functional image vs the number of Gaussians used in the model.

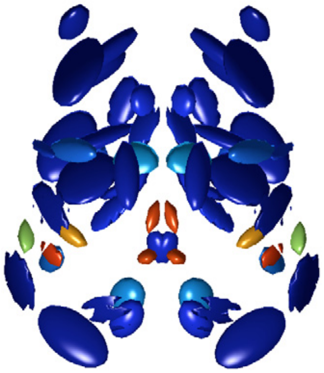


Fig. 3. Three-dimensional representation of the Gaussians obtained with the GMM-based method. Each ellipsoid represents a Gaussian and its color is related to the height of the Gaussian. Note that red ellipses match to regions that appear in literature as representative regions of AD [23]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\mathbf{M}_M = \mathbf{M}_{NC} - \mathbf{M}_{AD} \tag{9}$$

where $\mathbf{M}_{NC}$ and $\mathbf{M}_{AD}$ are the average image of all normal controls and AD images respectively. Then, the EM algorithm described above is applied to $\mathbf{M}_M$ to parcel the ROIs and model them as a set of Gaussians (see Fig. 3). Finally, the feature vector, $\mathbf{v} = (c_1, \ldots, c_k)$, for a given image is calculated from the Gaussian set obtained in the previous step where each c_n stands for the activation of the entire image for the Gaussian n and is computed as

$$c_n = h_n \sum_{i=1}^{V} I(\mathbf{x}_i) f_n(\mathbf{x}_i) \tag{10}$$

where $h_n = w_n / \sqrt{(2\pi)^3 |\boldsymbol{\Sigma}_n|}$ and f_n are the height and the density of the Gaussian n respectively. And $I(\mathbf{x}_i)$ is the intensity of the voxel whose coordinates are $\mathbf{x}_i$. Therefore, the number of features of an image is equal to the number of Gaussians (k parameter) of the GMM. However, a further reduction of the dimensionality may be performed by selecting only the Gaussians with higher height that correspond to regions where there are more differences between controls and AD images. The model selection criterion is chosen to be data-dependent since given a sufficient number k of Gaussians, all relevant activation areas would be included and/or modeled in them.

3. Feature extraction based on partial least squares

3.1. Partial least squares

PLS [24] is a statistical method for modeling relations between sets of observed variables by means of latent variables. It comprises regression and classification tasks as well as dimension reduction techniques and modeling tools. The underlying assumption of all PLS methods is that the observed data is generated by a system or process which is driven by a small number of latent (not directly observed or measured) variables. In its general form PLS creates orthogonal score vectors (also called latent vectors or components) by maximizing the covariance between different sets of variables. PLS can be naturally extended to regression problems. Both the predictor and predicted (response) variables are considered as a block of variables. PLS extracts the score vectors which serve as a new predictor representation and regresses the response variables on these new predictors. PLS can be also applied as a discrimination tool and dimension reduction method similar to Principal Component Analysis (PCA) [25,26]. After relevant latent vectors are extracted, an appropriate classifier can be applied.

Mathematically, PLS is a linear algorithm for modeling the relation between two data sets $X \subset \mathbb{R}^N$ and $Y \subset \mathbb{R}^M$. After observing n data samples from each block of variables, PLS decomposes the $n \times N$ matrix of zero-mean variables $\mathbf{X}$ and the $n \times M$ matrix of zero-mean variables Y into the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{11}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{12}$$

where the $\mathbf{T}$, $\mathbf{U}$ are $n \times p$ matrices of the p extracted score vectors (components, latent vectors), the $N \times p$ matrix $\mathbf{P}$ and the $M \times p$ matrix $\mathbf{Q}$ represent matrices of loadings and the $n \times N$ matrix $\mathbf{E}$ and the $n \times M$ matrix $\mathbf{F}$ are the matrices of residuals (or error matrices). The x-scores in $\mathbf{T}$ are linear combinations of the x-variables and can be considered as good "summaries" of the x-variables. Similarly, the y-scores in $\mathbf{U}$ are linear combinations of the y-variables and can be considered as good "summaries" of them [27]. Several algorithms have been proposed in the literature to implement the PLS model. In this paper, we use the SIMPLS algorithm [28].

The model structures of PLS and PCA are the same in the sense that the data are first transformed into a set of a few intermediate linear latent variables (components) and these new variables are taken into account. Essentially, the difference between PLS and PCA is that the former creates orthogonal weight vectors by maximizing the covariance between elements in $\mathbf{X}$ and $\mathbf{Y}$. Thus, PLS not only considers the variance of the samples but also considers the class labels. Fisher Discriminant Analysis (FDA) is, in this way, similar to PLS. However, FDA has the limitation that after dimensionality reduction, there are only c–1 meaningful latent variables, where c is the number of classes being considered.

3.2. Feature extraction procedure

First, a binary mask is applied to each image in order to remove the voxels that are not part of the brain. Only the voxels that have an intensity above 50% of maximum intensity in the image computed as the mean of all normal images, will be considered. Applying this mask to each image leads to a significant reduction of the input space. For example, for data used in this work, the initial number of voxels per image ($35 \times 48 \times 40 = 67\,200$) is reduced to 20 638.

Then, score PLS vectors are extracted (matrix $\mathbf{T}$ in Eq. (11)) and used as features. In this context, the $\mathbf{X}$ matrix contains the image-data: one row per each image and one column per each voxel, i.e. n–1 rows (where n is the size of the database) and 20 638 columns. And $\mathbf{Y}$ is a matrix of size n–1 × 1 that contains the labels. In order to avoid biased results, the PLS algorithm is run with all but one image (and their labels) of the database. The score vector for the remaining image is then computed through the weight matrix obtained. This process is repeated for all images of the database [7,3]. In other words, we applied a leave-one-out methodology to the feature extraction method with the purpose of avoiding that the label of a given image is taken into account to compute its score vector. Fig. 4 shows a diagram which describes this process. According to the PLS definition, the weight vector has as many components as images there are in the database minus two and this number will be the size of the feature vectors. However, as same as in PCA, a further reduction of the dimensionality is possible by truncating the feature vectors.

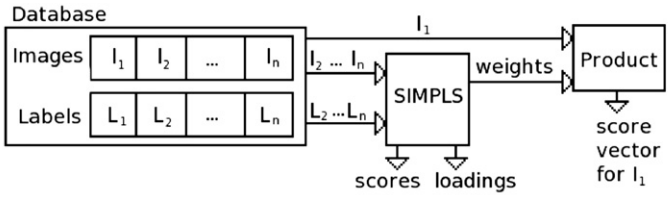


Fig. 4. Function diagram of the feature extraction method based on PLS. It represents the process followed to calculate the score vector of the first image of the database.

Table 1
Demographic details of the PET images used in this work. μ and σ stand of average and standard deviation respectively.

	Sex			Age		
	#	M	F	μ	σ	Range
NC	97	60	37	75.97	4.91	62–86
MCI	188	122	66	75.12	7.22	55–89
MCIc	23	18	5	73.97	7.35	57–85
AD	95	57	38	75.72	7.40	55–88

4. Database description

4.1. ADNI database

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a cooperative agreement grant whose primary goal is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and physicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

Data used in the preparation of this article consist of 403 Fludeoxyglucose (18^F-FDG) PET images collected from the ADNI Laboratory on NeuroImaging (LONI, University of California, Los Angeles). Participants enrollment was conditioned to some eligibility criteria. General inclusion/exclusion criteria was based on measures of disease severity, such as the Mini-Mental State Exam (MMSE) or Clinical Dementia Rating (CDR) were as follows:

- NORMAL control subjects: MMSE scores between 24 and 30 (inclusive), CDR of 0, non-depressed, non MCI, and non-demented. The age range of normal subjects will be roughly matched to that of MCI and AD subjects. Therefore, there should be minimal enrollment of normals under the age of 70.
- MCI subjects: MMSE scores between 24 and 30 (inclusive), a memory complaint, have objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia.
- Mild AD: MMSE scores between 20 and 26 (inclusive), CDR of 0.5 or 1.0, and meets NINCDS/ADRDA [29] criteria for probable AD.

Therefore, FDG PET data was separated into three different classes: Normal Control (NC), Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) images (see Table 1 for details). In addition, ADNI sites provide information about MCI who are stable after 2 years follow-up, and those who have converted into AD, i.e.

MCI converter (MCIc). However, it is worth noting that all images used in this paper are the first capture for each patient. Subsequent captures were only used for labeling purposes.

4.2. Image preprocessing

The PET data were first preprocessed according to the procedure described in http://adni.loni.ucla.edu/about-data-samples/image-data/ in order to remove differences due the scanner used for the acquisition. Subsequently, the images were normalized through a general affine model, with 12 parameters using the SPM software [6]. After the affine normalization, the resulting image was registered using a more complex non-rigid spatial transformation model. The nonlinear deformations were parametrized by a linear combination of the lowest-frequency components of the three-dimensional cosine transform bases [30]. A small-deformation approach was used, and regularization was by the bending energy of the displacement field, ensuring that the voxels in different FDG-PET images refer to the same anatomical positions in the brains. After spatial normalization, an intensity normalization was required in order to perform voxel intensity comparisons between different subjects. The intensity of the images was normalized to a value I_{max}, obtained averaging the 0.1% of the highest voxel intensities exceeding a threshold. The threshold was fixed to the 10th bin intensity value of a 50-bins intensity histogram, for discarding most low-intensity records from outside-brain regions, and preventing image saturation.

Once the images were normalized, they were downsampled with a factor of 2 yielding volumes of $35 \times 48 \times 40$ voxels. Thus, the computation time is reduced without loss of information since marks of AD are not at voxel level but at higher structures level (according to [23] those areas are the temporo-parietal region and the posterior cingulate that are significantly greater than the size of one voxel, especially with PET images where resolution is often higher than other image modalities).

5. Experiments and results

We have developed several CAD systems using the feature extraction methods described above and two SVM classifiers,[2] linear and nonlinear (RBF kernel with parameter $\sigma = 5$). Since our purpose is to distinguish between healthy subjects and AD patients, first we have trained the CAD systems with only controls and AD images (group 1). Second, we have also used MCI images by labeling MCI non-converters as controls and MCI converters as AD [31] (group 2). Thus, we can measure the ability of the systems to detect cases of dementia that will progress to AD.

Tables 2 and 3 show the statistical measures obtained using GMM and PLS. Those tables also show the performance of two existing approaches in the literature, i.e. VAF and PCA. VAF has been implemented as shown in [7] whereas the PCA method consists of using score vectors as feature vectors in a similar way as in [3]. Performance of the different feature extraction methods has been calculated via a k-fold cross-validation methodology (with k=5). Sensitivity and specificity of each test are defined as:

$$Sensitivity = \frac{TP}{TP+FN}, \quad Specificity = \frac{TN}{TN+FP}$$

where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives respectively. These probabilities reveal the ability to detect NOR/AD patterns thus, the best CAD system is the one that achieves the best trade-off between specificity and sensitivity. Since the number of samples in the second experiment (group 2) is unbalanced, positive likelihood ($Sensitivity/(1-Specificity)$) and negative likelihood ($(1-Sensitivity)/Specificity$) have been estimated in order to avoid improper assumptions. These measures are prevalence independent, i.e. they do not depend on the ratio of the classes.

Table 2
Statistical measures of performance of the proposed methods and the baseline approaches (VAF and PCA) for group 1 (controls vs AD).

	VAF	PCA	GMM	PLS	
Accuracy	80.21%	86.98%	87.50%	87.50%	SVM lin
Specificity	81.44%	85.57%	86.60%	90.72%	
Sensitivity	78.95%	88.42%	88.42%	84.21%	
PL	3.8686	7.3899	7.4789	5.7457	
NL	0.2351	0.1632	0.1516	0.1102	
Accuracy	41.67%	85.42%	90.63%	86.46%	SVM RBF
Specificity	51.55%	87.63%	90.72%	90.72%	
Sensitivity	31.58%	83.16%	90.53%	82.11%	
PL	0.7534	5.2030	9.5762	5.0697	
NL	1.5344	0.1488	0.1025	0.1130	

Table 3
Statistical measures of performance of the proposed methods and the baseline approaches (VAF and PCA) for group 2 (controls and MCI vs MCI converters and AD).

	VAF	PCA	GMM	PLS	
Accuracy	68.24%	77.42%	78.91%	76.92%	SVM lin
Specificity	78.60%	87.72%	90.88%	88.42%	
Sensitivity	43.22%	52.54%	50.00%	49.15%	
PL	1.3841	1.8484	1.8175	1.7389	
NL	0.4952	0.2337	0.1825	0.2356	
Accuracy	70.72%	78.16%	78.41%	76.18%	SVM RBF
Specificity	100.0%	94.39%	90.88%	92.98%	
Sensitivity	–	38.98%	48.31%	35.59%	
PL	1	1.5469	1.7580	1.4437	
NL	–	0.1440	0.1889	0.1972	

The accuracy of the different approaches depend on the size of the feature vector. The maximum size of the feature vectors is the number of Gaussian for the GMM-based method and the number of images there are in database minus two, for PLS-based algorithm. However, this number may be reduced by selecting only the most important Gaussians/components (i.e. Gaussians with higher height or the first PLS components). Fig. 5 shows the accuracy rates achieved with both approaches in function of number of Gaussians/components selected and compares them with the performance obtained using PCA and VAF.

6. Discussion

As it is shown in Section 5, the two methods analyzed in this work are valid approaches to develop CAD systems for AD and achieve better accuracy, sensitivity and specificity than previous approaches. The success rate achieved in the automatic diagnosis exceeds 90% for group 1. When MCI images are considered, the success rate decreases significantly due to the high variability of the MCI pattern (see Fig. 6). Despite that fact the GMM and PLS-based methods improve the baseline. The relevance of the classification results obtained is also confirmed by the ROC curves shown in Fig. 7 that measures the trade-off between sensitivity and specificity for varying the number of Gaussians of the PLS components.

It is worth noting that CAD systems are reproducing current medical knowledge since they have been trained with samples

[2] We used the SVM implementation included in the Bioinformatics Toolbox of Matlab.

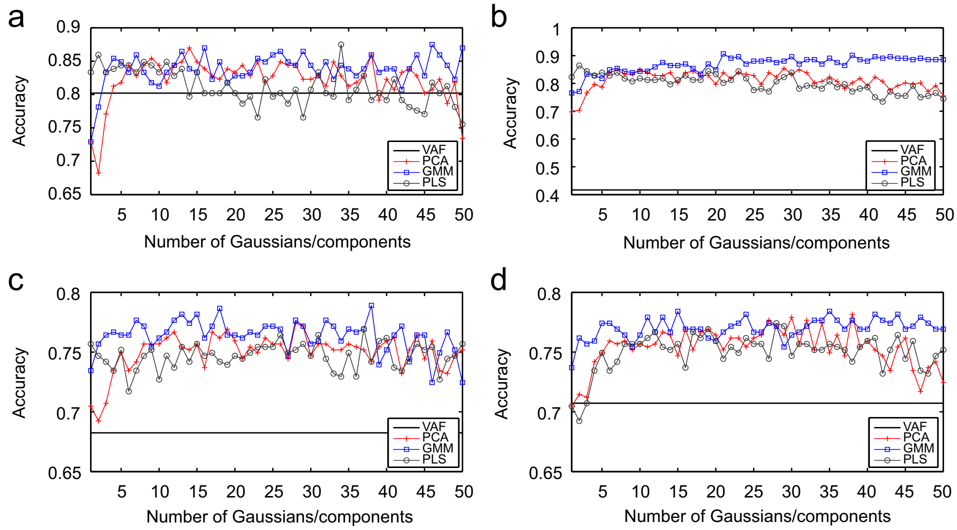


Fig. 5. Success rates obtained with the developed CAD systems in function of the number of components used (number of Gaussian for GMM-based method and number of components for PCA/PLS-based methods). Left column: CAD systems with a linear classifier. Right column: CAD systems with a nonlinear classifier. Top row: Experiments for data of group 1. Bottom row: Experiments for data of group 2.

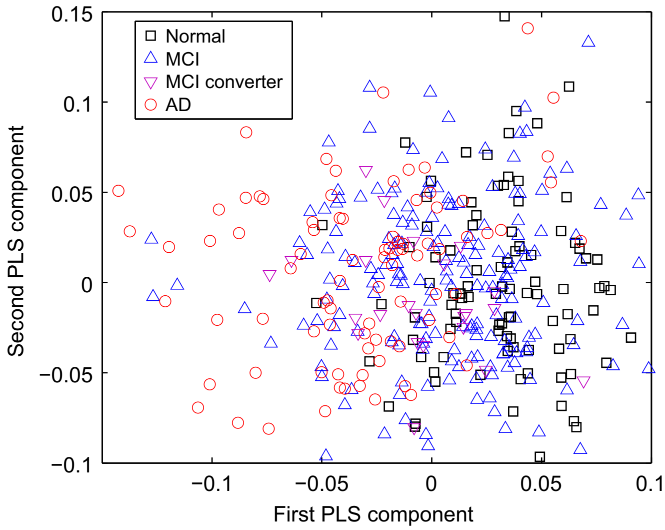


Fig. 6. Representation of all images of the database using only the first two PLS components. Note the high variability of the MCI pattern.

labeled by physicians. For this reason, statistical measures reported in this paper are an estimation about how a trained system is able to reproduce a medical diagnosis performed by experts. Therefore, some possible errors in the labeling process could render the classification task more difficult, especially considering that the labels were assigned based on the scores obtained by patients in cognitive tests (as MMSE and CDR) and they do not consider the information provided by the PET images for the database used.

In general, the GMM-based method yields better results than the PLS-based one. Parcellation of ROIs using GMM has proven to be an effective approach to extract features. In addition, this method has the advantage that the size of the features vectors does not depend on the number of samples of the database. On the other hand, PLS-based method performs better than PCA approach and execution time is lower than when GMM model.

In short, the feature extraction approaches proposed in this paper achieve good classification performance when a linear classifier is used. Furthermore, nonlinear classifiers also performs accurately, mainly when the GMM method is used. Generally, nonlinear classifiers require more samples or smaller feature vectors than the linear ones in order to yield good classification results. In our case, both experiments use the same number of samples, however, the GMM approach provides smaller feature vectors than PLS (specially, 64 vs 192, before pruning) and therefore, it performs better with nonlinear classifiers.

7. Conclusions

In this work, two features extraction methods to improve the classification of PET images to diagnosing Alzheimer's disease are presented. The proposed methodologies are two multivariate approaches which allow to reduce the dimension of the feature vector in order to surmount the small-size problem which arises in classification problems when the dimension of the feature vector is very high compared to the number of available samples. The first approach uses Gaussian Mixture Model in order to extract the most discriminant regions from PET images. The second method uses score vectors obtained through a Partial Least Squares algorithm as features. Score vectors are chosen following a criterion of maximum covariance between images and labels.

The GMM-based method first computes an image with differences between controls and AD images. Then, it models this image by means of a set Gaussian using the well-known EM algorithm. The Gaussians obtained represent the ROIs to distinguish normal and AD images and are used to calculate features. Finally, the feature vector for a given image contains a measure of the activation of the whole image for each Gaussian. Thus, we get as many features as Gaussians have been defined. On the other hand, the PLS-based method uses score vectors as features. This method performs an initial reduction of the input space by applying an intensity mask that discard low-intensity voxels. Both methods have been tested with two classifiers based on SVM, one linear and other nonlinear. The resulting CAD systems were trained using PET images from the ADNI database and the statistical performance of the methods were estimated using a *k*-fold cross-validation methodology. The presented methods yield peak accuracy rates of 90% when we distinguish between controls and AD images and, in general, outperform previous approaches such as the ones based on VAF or PCA [3].

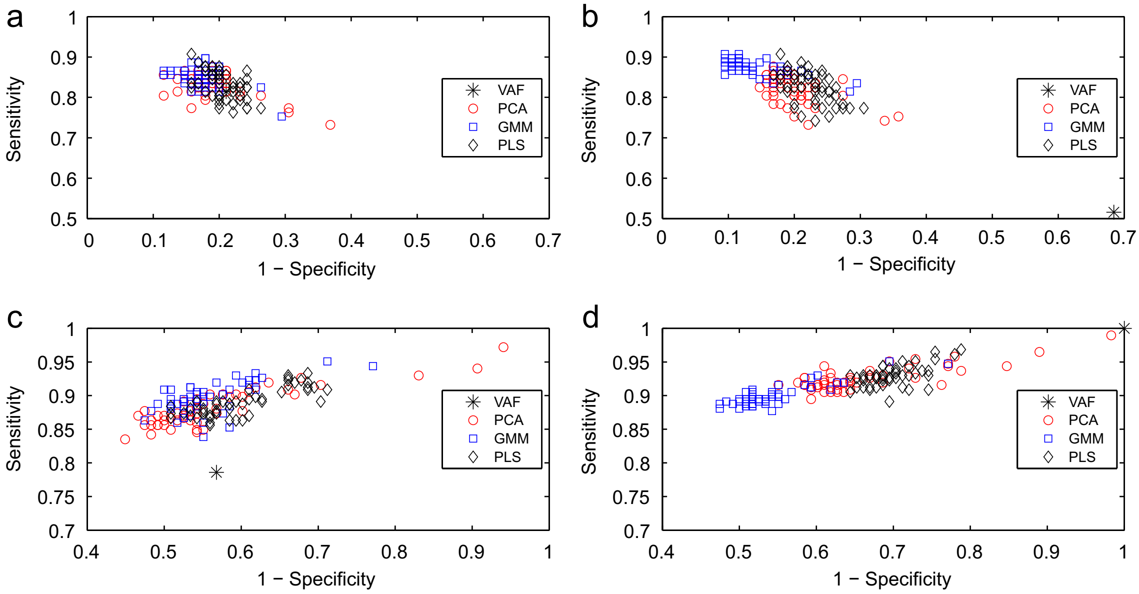


Fig. 7. ROC curves for statistical measures obtained with both methods analyzed and the baseline: (a) linear classifier, image group 1; (b) nonlinear classifier, image group 1; (c) linear classifier, image group 2; (d) nonlinear classifier, image group 2.

Acknowledgments

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 Projects (TIC-02566 and TIC-4530). Furthermore it was supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

[1] J.M. Górriz, A. Lassl, J. Ramírez, D. Salas-Gonzalez, C.G. Puntonet, E.W. Lang, Automatic selection of ROIs in functional imaging using Gaussian mixture models, Neuroscience Letters 460 (2) (2009) 108–111.

[2] F. Segovia, J.M. Górriz, J. Ramírez, D. Salas-Gonzalez, I. Álvarez, M. López, R. Chaves, P. Padilla, Classification of functional brain images using a GMM-based multi-variate approach, Neuroscience Letters 474 (1) (2010) 58–62. doi:10.1016/j.neulet.2010.03.010.

[3] M. López, J. Ramírez, J.M. Górriz, D. Salas-Gonzalez, I. Álvarez, F. Segovia, C. Puntonet, Automatic tool for the Alzheimer's disease diagnosis using PCA and Bayesian classification rules, IET Electronics Letters 45 (8) (2009) 389–391.

[4] P. Vemuri, J.L. Gunter, M.L. Senjem, J.L. Whitwell, K. Kantarci, D.S. Knopman, B.F. Boeve, R.C. Petersen, C.R. Jack Jr., Alzheimer's disease diagnosis in individual subjects using structural mr images: validation studies, Neuroimage 39 (3) (2008) 1186–1197.

[5] M. Signorini, E. Paulesu, K. Friston, D. Perani, G. Lucignani, A.D. Sole, M.D. Martin, G. Striano, F. Grassi, R. Frackowiak, F. Fazio, Assessment of 18f-fdg pet brain scans in individual patients with statistical parametric mapping. A clinical validation, Neuroimage 9 (1999) 63–80.

[6] 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.

[7] 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, in: Lecture Notes in Computer Science, vol. 3217, 2004, pp. 654–662.

[8] K. Herholz, E. Salmon, D. Perani, J.-C. Baron, V. Holthoff, L. Frölich, P. Schönknecht, K. Ito, R. Mielke, E. Kalbe, G. Zündorf, X. Delbeuck, O. Pelati, D. Anchisi, F. Fazio, N. Kerrouche, B. Desgranges, F. Eustache, B. Beuthien-Baumann, C. Menzel, J. Schröder, T. Kato, Y. Arahata, M. Henze, W.-D. Heiss, Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG PET, Neuroimage 17 (1) (2002) 302–316. doi:10.1006/nimg.2002.1208.

[9] N.L. Foster, J.L. Heidebrink, C.M. Clark, W.J. Jagust, S.E. Arnold, N.R. Barbas, C.S. DeCarli, R. Scott Turner, R.A. Koeppe, R. Higdon, S. Minoshima, FDG-PET improves accuracy in distinguishing frontotemporal dementia and Alzheimer's disease, Brain 130 (10) (2007) 2616–2635. doi:10.1093/brain/awm177.

[10] A. Abraham, E. Corchado, J.M. Corchado, Hybrid learning machines, Neurocomputing 72 (13–15) (2009) 2729–2730.

[11] E. Corchado, A. Abraham, A. de Carvalho, Hybrid intelligent algorithms and applications, Information Sciences 180 (14) (2010) 2633–2634.

[12] R.P.W. Duin, Classifiers in almost empty spaces, Proceedings of the 15th International Conference on Pattern Recognition, vol. 2, IEEE, 2000, pp. 1–7.

[13] D. Salas-Gonzalez, J.M. Górriz, J. Ramírez, I. Álvarez, M. López, F. Segovia, M. Gómez-Río, Skewness as feature for the diagnosis of Alzheimer's disease using SPECT images, in: IEEE International Conference on Image Processing, El Cairo, Egypt, 2009.

[14] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.

[15] G. McLachlan, D. Peel, Finite Mixture Models, John Wiley and Sons, New York, 2000.

[16] M. Aladjem, Projection pursuit mixture density estimation, IEEE Transactions on Signal Processing 53 (11) (2005) 4376–4383.

[17] J. Goldberger, S. Gordon, H. Greenspan, Unsupervised image-set clustering using an information theoretic framework, IEEE Transactions on Image Processing 15 (2) (2006) 449–458.

[18] J. Banfield, A. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics 49 (1993) 803–821.

[19] J.M. Górriz, F. Segovia, J. Ramírez, A. Lassl, D. Salas-Gonzalez, GMM based SPECT image classification for the diagnosis of Alzheimer's disease, Applied Soft Computing 11 (2) (2011) 2313–2325.

[20] G.J. McLachlan, P.N. Jones, Fitting mixture models to grouped and truncated data via the EM algorithm, Biometrics 44 (2) (1988) 571–578.

[21] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of Royal Statistical Society Series B 39 (1) (1977) 1–38. doi:10.2307/2984875.
[22] T.K. Moon, The expectation-maximization algorithm, IEEE Signal Processing Magazine 13 (6) (1996) 47–60.
[23] S. Minoshima, B. Goirdani, S. Berent, K. Frey, N. Foster, D. Khul, Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease, Annals of Neurology 42 (1) (1997) 85–94. doi:10.1002/ana.410420114.
[24] S. Wold, H. Ruhe, H. Wold, W.D. III, The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverse, Journal of Scientific and Statistical Computations 5 (1984) 735–743.
[25] D.V. Nguyen, D.M. Rocke, Tumor classification by partial least squares using microarray gene expression data, Bioinformatics 18 (1) (2002) 39–50.
[26] R. Rosipal, L.J. Trejo, Kernel PLS-SVC for linear and nonlinear classification, in: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), 2003, pp. 640–647.
[27] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, 2009.
[28] S. de Jong, Simpls: an alternative approach to partial least squares regression, Chemometrics and Intelligent Laboratory Systems 18 (3) (1993) 251–263. doi:10.1016/0169-7439(93)85002-X.
[29] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, E.M. Stadlan, Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group^{a} under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease, Neurology 34 (7) (1984) 939–944.
[30] J. Ashburner, K.J. Friston, Nonlinear spatial normalization using basis functions, Human Brain Mapping 7 (4) (1999) 254–266.
[31] R.F. Allegri, F.B. Glaser, F.E. Taragano, H. Buschke, Mild cognitive impairment: Believe it or not? International Review of Psychiatry 20 (4) (2008) 357–363. doi:10.1080/09540260802095099.

Fermin Segovia received the B.Sc. Degree in Computer Science in 2006, and the Master degree in Computer Engineering and Networks in 2009, both from the University of Granada, Spain. He is currently a Ph.D. student of the Department of Signal Theory, Networking and Communications at the University of Granada. His main research interests include Image Processing and Biomedical Applications.

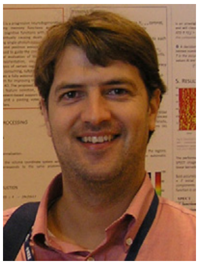
J.M. Górriz received the B.Sc. degrees in Physics and Electronic Engineering from the University of Granada, Granada, Spain, and the Ph.D. degrees from the Universities of Cádiz and Granada, Spain, in 2000, 2001, 2003, and 2006 respectively. He is currently an Associate Professor with the Department of Signal Theory, Networking, and Communications at the University of Granada. He has coauthored more than 170 technical journals and conference papers in these areas and has served as Editor in Chief for the Open Acoustics Journal, Bentham, since 2007.

J. Ramírez received the M.Sc. degree in Electronic Engineering in 1998, and the Ph.D degree in Electronic Engineering in 2001, all from the University of Granada. Since 2001, he is an Associate professor at the Department of Signal Theory Networking and Communications of the University of Granada (Spain). His research interest includes signal processing and biomedical applications including brain image processing, robust speech recognition, speech enhancement, voice activity detection, seismic signal processing and implementation of high performance digital signal processing systems. He has coauthored more than 150 technical journal and conference papers in these areas. He has served as reviewer for several international journals and conferences.

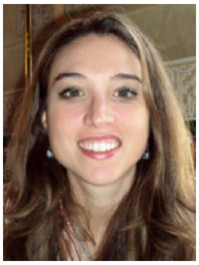
His present interests lie in the field of statistical signal processing and its application to speech and image processing.

M. López received the B.Sc. degree in Telecommunications Engineering from the University of Seville (Spain) in 2007 and the Master degree in Multimedia Technologies from the University of Granada (Spain) in 2008. She is currently a Ph.D. Student in the Department of Signal Theory, Networking and Communications at the University of Granada. Her research interests lie in the field of signal processing, in particular image processing and classification for biomedical applications.

Ignacio Álvarez received the B.Sc. Degree in Physics from the Complutense University of Madrid, Spain and the Ph.D. from the University of Granada, Spain in 2004 and 2009. Actually, he is working under a local fellowship in the Department of Signal Theory, Networking and Communications at the University of Granada. His present research interests include supervised learning, signal processing, independent component analysis and biomedical applications.



Diego Salas-Gonzalez was born in Málaga, Spain, in 1980. He obtained his M.Sc. and B.Phil. degrees in physics from the University of Granada in 2003 and 2005 respectively. He was a granted national researcher from the Ministry of Education and Science of Spain from 2004 to 2008. His research interests are in statistical signal processing, Bayesian inference and applications in bioinformatics and image processing.



R. Chaves received the B.Sc. degree in Telecommunications Engineering and the Master degree in Multimedia Technologies from the University of Granada, Spain in 2008 and 2009 respectively. She is a Ph.D. student in Classification Techniques of neurologic alterations at the University of Granada (SIPBA research group, Department of Signal Theory, Networking and Communications). Her research interests at present lie in the field of Signal Processing and Biomedical Applications.