Parametrization of textural patterns in 123I-ioflupane imaging for the automatic detection of Parkinsonism

F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, M. Moreno-Caballero, M. Gómez-Río, and The Parkinson’s Progression Markers Initiative

Citation: Medical Physics 41, 012502 (2014); doi: 10.1118/1.4845115
View online: http://dx.doi.org/10.1118/1.4845115
View Table of Contents: http://scitation.aip.org/content/aapm/journal/medphys/41/1?ver=pdfcov
Published by the American Association of Physicists in Medicine
Parametrization of textural patterns in $^{123}$I-ioflupane imaging for the automatic detection of Parkinsonism

F. J. Martinez-Murcia, a) J. M. Górriz, and J. Ramírez
Signal Processing and Biomedical Applications Research Group, Department of Signal Theory, Networking and Communications, University of Granada, Granada 18071, Spain

M. Moreno-Caballero and M. Gómez-Río
Department of Nuclear Medicine, Virgen de las Nieves Hospital, Granada 18071, Spain

The Parkinson’s Progression Markers Initiative b)

(Received 22 July 2013; revised 2 November 2013; accepted for publication 22 November 2013; published 17 December 2013)

Purpose: A novel approach to a computer aided diagnosis system for the Parkinson’s disease is proposed. This tool is intended as a supporting tool for physicians, based on fully automated methods that lead to the classification of $^{123}$I-ioflupane SPECT images.

Methods: $^{123}$I-ioflupane images from three different databases are used to train the system. The images are intensity and spatially normalized, then subimages are extracted and a 3D gray-level co-occurrence matrix is computed over these subimages, allowing the characterization of the texture using Haralick texture features. Finally, different discrimination estimation methods are used to select a feature vector that can be used to train and test the classifier.

Results: Using the leave-one-out cross-validation technique over these three databases, the system achieves results up to a 97.4% of accuracy, and 99.1% of sensitivity, with positive likelihood ratios over 27.

Conclusions: The system presents a robust feature extraction method that helps physicians in the diagnosis task by providing objective, operator-independent textural information about $^{123}$I-ioflupane images, commonly used in the diagnosis of the Parkinson’s disease. Textural features computation has been optimized by using a subimage selection algorithm, and the discrimination estimation methods used here makes the system feature-independent, allowing us to extend it to other databases and diseases.

© 2014 American Association of Physicists in Medicine. [http://dx.doi.org/10.1118/1.4845115]

Key words: Parkinson’s disease, $^{123}$I-ioflupane, computer aided diagnosis, Haralick texture features, support vector machines

1. INTRODUCTION

Parkinsonian syndrome (PS), also known as Parkinsonism, is a neurological syndrome characterized by tremor, hypokinesia, rigidity, and postural instability. It is considered as the second most common neurodegenerative disease, with a prevalence of 1%–3% in the population over 65 years of age. A wide range of etiologies may lead to the PS, while the most common cause is the neurodegenerative condition called Parkinson’s disease (PD). This disease originates due to the progressive loss of dopaminergic neurons of the nigrostriatal pathway, which connects the substantia nigra to the striatum. As a result, the dopamine content of the striatum decreases, and consequently, dopamine transporters (DAT) are lost. Other possible causes include some toxins, a few metabolic diseases, and a handful of non-PD neurological conditions (atypical Parkinsonian syndromes), such as multiple system atrophy (MSA), progressive supranuclear palsy (PSP), or corticobasal degeneration (CBD).

As the PD is related to a loss of dopamine transporters in the nigrostriatal pathway, the study of its status by means of brain imaging techniques has set a breakthrough in the diagnosis process, particularly in the case of Parkinsonian syndromes. $^{123}$I-ioflupane (best known by its trade name DaTSCAN) is a tracer derived from cocaine analogue, which binds to the dopamine transporters in the striatum. Then, the density of these transporters is measured using single photon emission computed tomography (SPECT). As a result, images show a reduced uptake of the tracer in the striatum in patients with PS. The analysis of DaTSCAN images is commonly performed by visual inspection, using a predefined rating or the analysis of region of interest (ROI). Thus, new automatic and quantitative methods of analysis are desirable.

Automatic classification of medical imaging is a wide research field, and many advances have been performed in detecting neurodegenerative diseases such as Alzheimer’s disease, with MRI, SPECT, or PET images, and PS using DaTSCAN images. A baseline approach to this problem is the voxels-as-features (VAFs) based method, commonly used as an estimation of the performance of visual analysis performed by experts. Further approaches include feature selection techniques using discriminant analysis, feature reduction methods such as independent component analysis (ICA), principal component analysis (PCA), factor analysis (FA), single value decomposition (SVD), and a wide range of classifiers, including multivariate analysis, naive Bayes, and support vector machines (SVM).
In this work, a new approach to the feature reduction task is proposed, extracting several features that could reveal some characteristics of the DaTSCAN imaging. Particularly the computation of 12 Haralick features\(^2\) that characterize some characteristics of the DaTSCAN imaging. Particularly, a task is proposed, extracting several features that could reveal influential information in the co-occurrence matrix (GLCM) that has been used successfully in 3D image segmentation.\(^2\) Then, five different methods to estimate each feature’s discrimination ability are proposed, selecting the most relevant features for the classification process. This procedure makes the system feature-independent and extensible to other databases. Finally, a linear kernel support vector machine (SVM), that has been successfully applied to many other tasks,\(^1\)\(^5\),\(^1\)\(^9\),\(^2\)\(^0\),\(^2\)\(^2\) is used to perform image classification.\(^3\)

This paper is organized as follows. First, in Sec. 2, the methodology followed to analyze the images (preprocessing, image subvolume selection, Haralick texture features computation, and discrimination-based feature selection and classification) is explained. Then, in Sec. 3 the databases and the validation procedures used are presented, and some experiments are proposed to evaluate different possibilities. In Sec. 4, the results for each experiment and database are exposed and discussed, and finally, in Sec. 5, the conclusions of using this texture approach are drawn.

2. METHODOLOGY

In this paper, a new method for the automatic detection of Parkinsonism is proposed. This method consists of four stages:

- **Image preprocessing.** First, images are spatially normalized according to a template,\(^2\)\(^6\) and then, they are intensity normalized using an intensity normalization to the maximum strategy.\(^2\)\(^0\)

- **Feature extraction.** A threshold is used to establish a subvolume of the normalized images, which will be useful for both enhancing the performance of our system and preventing the introduction of nonuseful, but influential information in the co-occurrence matrix computation. This threshold ranges from 0% (whole image) to 50% of the maximum intensity value. Twelve Haralick texture features are extracted from these subvolumes.\(^2\)\(^3\)

- **Feature selection.** Then, the most discriminant features are selected as an input to the classifier. The discrimination ability is computed using the Fisher’s discriminant ratio (FDR), Bhattacharyya distance (BD), Student’s t-test (t-test), Mann-Whitney-Wilcoxon’s U-test (MWW), and relative entropy (RE).\(^2\)\(^7\)\(^–\)\(^3\)\(^0\) This will make the system feature-independent, selecting exclusively the most promising features.

- **Classification.** The selected features are used as an input to a linear kernel SVM classifier.\(^3\)\(^1\)

2.A. Image preprocessing

To be able to compare the images with each other, some assumptions should be made. First, we must consider that the same voxel coordinate in each image corresponds to the same anatomical position. Then, we should assume that the same intensity value represents the same uptake value in areas of specific activity. To do so, a spatial and intensity normalization process is applied to each image in the databases.

2.A.1. Spatial normalization

In order to establish a statistically fair comparison among images, the voxels in the same coordinate should represent the same anatomical position in the brain. To this purpose, a spatial normalization is needed. The spatial normalization procedure applied is based on a template which represents an standard anatomical space. The normalization method assumes a general model that can be used to perform an elastic deformation of the raw images to fit the template. The SPM 8 software\(^2\)\(^6\) used in our databases assumes a general affine model with 12 parameters and a Bayesian framework that maximizes the product of the prior function (which is based on the probability of obtaining a particular set of zooms and shears) and the likelihood function, derived from the residual squared difference between the template and the processed image.

2.A.2. Intensity normalization

In order to establish comparisons between the uptake value in areas of specific activity (binding to dopaminergic transporters) and areas of nonspecific activity (vascular activity) between subjects, some kind of intensity normalization is required. The method used in this work is based on obtaining a scalar intrinsic parameter from the image, \(I_p\), and then dividing the whole original, spatially normalized \(I_o\) by this parameter:

\[
I = \frac{I_o}{I_p},
\]

where \(I\) is the resulting spatially and intensity normalized image. Although different normalization strategies can be used, such as integral normalization, in Ref. 20, the most stable choice appeared to be the normalization to the maximum strategy, which will be used in the rest of the paper.

In this approach, the \(I_p\) value is obtained using a 100-bin histogram of intensities in the image. Since the choice of the maximum intensity can produce outliers in the resulting image, and the choice of a lower value can produce saturation, the average value of the 95th bin of the histogram is established as the \(I_p\) value. This normalization will be applied to our three test databases, along with a nonintensity-normalized case, which will be analyzed to assess the real contribution of this intensity normalization step to the final results.
2.B. Feature extraction

2.B.1. Volume selection

The volume to which further calculations are applied could be defined as the whole image, but in order to prevent the inclusion of outliers (such as the intensity differences between the brain intensity and the background), we have devised a volume selection algorithm based on a scalar intensity threshold, \( I_{th} \).

This algorithm allows us to select the biggest box-shaped volume that contains only brain voxels. To do so, a mask is computed over the average image \( I_{mean} \) with the given threshold, and the coordinates of the smallest subvolume that wraps the whole mask are extracted as detailed in Table I, allowing the obtaining of our subimage. An example of the resulting subimage is depicted in Fig. 1.

2.B.2. Gray level co-occurrence matrix

A co-occurrence matrix is a square matrix that counts the number of repetitions of co-occurring values over an image, given an offset that measures the distance from the central voxel to the neighbour voxels. In this work, we use a modification of the widely used co-occurrence matrix defined in Ref. 25 that generalizes this matrix to a three-dimensional space. For two different gray levels \( i \) and \( j \), the co-occurrence matrix \( C \) is defined over a \( n \times m \times k \) three-dimensional image \( I \), given an offset \( \Delta \) as follows:

\[
C_{\Delta}(i, j) = \sum_{p=1}^{n,m,k} \begin{cases} 
1, & \text{if } I(p) = i \text{ and } I(p+\Delta) = j \\
0, & \text{otherwise},
\end{cases}
\]

where \( i \) and \( j \) are different gray levels, \( p = (x, y, z) \) is the spatial position where \( x = 1..n, y = 1..m, z = 1..k \), and \( \Delta = (d, d, d) \) is the offset vector, which has been set as dependent only on the distance \( d \). In this work, the 3D-GLC matrix is computed over the subimage defined in Sec. 2.B.1. To obtain better performance, we have performed a rescaling of the original data into 16 gray levels.

2.B.3. Haralick texture features

After computing the 3D-GLC matrix, 12 Haralick texture features\(^{24} \) are derived from the matrix of probabilities \( P_{\Delta} = C_{\Delta}/\sum_{i,j} C_{\Delta}(i, j) \) (a normalized version of the GLC matrix) using the following expressions:

- **Energy**
  \[
  \sum_{i,j}(P_{\Delta}(i, j))^2
  \]

- **Entropy**
  \[- \sum_{i,j} P_{\Delta}(i, j) \times \log(P_{\Delta}(i, j))\]

- **Correlation**
  \[\frac{\sum_{i,j} (i-\mu_i)(j-\mu_j)P_{\Delta}(i, j)}{\sigma_i\sigma_j}\]

- **Contrast**
  \[\sum_{i,j} (i-j)^2 P_{\Delta}(i, j)\]

- **Variance**
  \[\frac{1}{s} \sum_{i,j} (P_{\Delta}(i, j) + j P_{\Delta}(i, j))\]

- **Sum mean**
  \[\sum_{i,j} (i-j)^2 \times P_{\Delta}(i, j)\]

- **Inertia**
  \[\sum_{i,j} (i+j - \mu_i - \mu_j)^3 \times P_{\Delta}(i, j)\]

- **Cluster shade**
  \[\sum_{i,j} (i+j - \mu_i - \mu_j)^3 \times P_{\Delta}(i, j)\]

- **Cluster tendency**
  \[\sum_{i,j} (i+j - \mu_i - \mu_j)^3 \times P_{\Delta}(i, j)\]

- **Homogeneity**
  \[\sum_{i,j} P_{\Delta}(i, j)\]

- **Max probability**
  \[\max_{i,j} P_{\Delta}(i, j)\]

- **Inverse variance**
  \[\sum_{i,j} P_{\Delta}(i, j)\]

Thirteen spatial directions are considered to compute the GLC matrix (see Ref. 25). Furthermore, these GLC matrices can be calculated taking into account all the voxels that are...
at a distance \( d \) from the central voxel, so we have calculated one matrix for each distance \( d = 1 \ldots 10 \). Therefore, one GLC matrix in each of the 13 spatial directions and each value of \( d \) makes \( 13 \times 10 = 130 \) matrices for each image. From each one, 12 Haralick texture features are computed, so we obtain a feature vector \( \mathbf{x}_0 \) that contains 1560 Haralick features that characterizes each image. A lower number of features is desirable in the clustering task, so in Sec. 2.C, some measures are described to reduce the dimension of \( \mathbf{x}_0 \) by selecting the most discriminant features.

2.C. Discrimination measures

In this work, we have first tested the performance of using only one of the Haralick texture features, but there is a potential benefit in using all texture features in all directions. In this last case, the most useful features can be selected by related supervised learning methods, widely used as a non-iterative method. Therefore, we define the vector \( \Omega \), that contains all samples of the given textural feature for each patient belonging to class \( c \) (where \( c \) stands for either 1 or 2, normal control or PD). For each texture feature, the different aforementioned discrimination measures (S) is computed as follows:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR</td>
<td>( S_{FDR} = \frac{(\bar{G}_1 - \bar{G}_2)^2}{\sigma_1^2 + \sigma_2^2} )</td>
</tr>
<tr>
<td>t-test</td>
<td>( S_t = \frac{\bar{G}_1 - \bar{G}_2}{\sigma_1 \sqrt{n_1} + \sigma_2 \sqrt{n_2}} )</td>
</tr>
<tr>
<td>Bhattacharyya</td>
<td>( S_B = -\ln \sum_{i=1}^n \sqrt{(\Omega_i(1)\Omega_i(2))} )</td>
</tr>
<tr>
<td>Relative entropy</td>
<td>( S_{RE} = \sum_{i=1}^n (\frac{\Omega_i(1)}{\Omega_i(2)}) \ln (\frac{\Omega_i(1)}{\Omega_i(2)}) )</td>
</tr>
<tr>
<td>MWW</td>
<td>( S_{MWW} = \min \left( R_i - \frac{\omega_0 + i}{\omega_0 + 1} \right) )</td>
</tr>
</tbody>
</table>

where \( \bar{G}_c \) is the mean of \( \Omega_c \), \( \sigma_c^2 \) its variance, \( \sigma_{\Omega_1,\Omega_2}^2 \) is an estimator of common standard deviation, and \( n_i \) is the number of samples in each class. It is important to note that in the case of relative entropy, we must consider only the positive elements of \( \Omega \), to avoid dividing by zero.

Once we have obtained the vector \( s \) that contains values of \( S \) for each feature, we create a vector \( \mathbf{x}_\text{rank} \) by sorting \( \mathbf{x}_0 \) in descending order of discrimination ability, using the values in \( s \). Finally, first \( N \) features are selected from \( \mathbf{x}_\text{rank} \) as follows:

\[
\mathbf{x} = \mathbf{x}_\text{rank}(1:N).
\]

2.D. Support vector machines

Support vector machines (SVM) (Ref. 31) are a set of related supervised learning methods, widely used as a non-probabilistic binary classifier. In our system, we will use a basic linear kernel SVM classifier, minimizing the hyperplane margin using the sequential minimal optimization (SMO) algorithm.33 Despite linear kernel, SVM achieves poorer performance than other more specific kernels (like Gaussian or quadratic), it performs very well on datasets that have many attributes but a smaller number of samples (in this case, we handle up to 1560 features, and about 100–200 subjects per database) without losing generalization properties. Moreover, as discussing the ability of different classifiers is not the purpose of this paper, linear SVM are a widely used option that allow us to compare to other methods such as Ref. 22.

3. TEST AND VALIDATION

3.A. Databases

3.A.1. Virgen de la Victoria database

The first database used in this paper (from now on, VDLV database) is supplied by the Virgen de la Victoria hospital in Málaga (Spain). The images were obtained after a period of between 3 and 4 h after the intravenous injection of 185 MBq (5 mCi) of DaTSCAN, with prior thyroid blocking with Lugol’s solution. The tomographic study (SPECT) with Ioflupane/FP-CIT-I-123 was performed using a General Electric gamma camera, Millennium model, equipped with a dual head and general purpose collimator. A 360° circular orbit was made around the cranium, at 3° intervals, 60 images with a duration of 35 s per interval, 128 × 128 matrix. Image reconstruction was carried out using filtered back-projection algorithms without attenuation correction, application of a Hanning filter (frequency 0.7) and images were obtained with transaxial cuts.

The images were interpreted by three nuclear medicine specialists, with marking of the clinical orientation. Visual assessment was established by exclusively considering the normal/abnormal criterion and after arriving at a consensus report between the three specialists, i.e., whether the FP-CIT SPECT allowed differentiation of a group of conditions with presynaptic involvement from others in which their integrity is assumed, without trying to assign them to different clinical groups within the set of pathological studies. A study was considered to be normal when bilateral, symmetrical uptake appeared in caudate and putamen nuclei, and abnormal when there were areas of qualitative reduced uptake in any of the striatal structures.

A total of \( N = 208 \) subjects (100 affected subjects and 108 controls, with a prevalence PKS-normal of 48.08%), randomly selected from the total studies performed in this center until December 2008 and referred to it because of a movement disorder, were included in the study. A more detailed description of the database can be found in Ref. 34.

3.A.2. Virgen de las Nieves database

Another database that we have used to test our system is supplied by the Virgen de las Nieves hospital in Granada (Spain) (from now on, VDLN database). A total number of
$N = 118$ images (73 with dopaminergic deficit and 45 healthy controls, with a prevalence of 61.86%) are obtained after the injection of 185 MBq of DaTSCAN on previously thyroid-blocked patients, and acquired using a three head Picker Prism 3000 gamma camera.

Patients were derived to the nuclear medicine service between 2007 and 2012, after a Parkinsonian clinical diagnosis, to perform a DaTSCAN analysis of the nigrostriatal system. To establish an accurate gammagraphic diagnosis, the following criteria were stated:

- Visual interpretation: a subjective assessment by the observer physician, mainly by analyzing the total intensity in both striatum, possible asymmetries between them and the existent relationship between the striatal and the background brain activity.
- Exploration quantification, using a delimitation of the regions of interest (ROI), computing the radiopharmaceutical activity in counts per pixel for each marked area. Two variations were used: ROI Q1 that quantifies the total activity in each striatum and the occipital region, along with three other regions for every striatum (head, body, and tail); and ROI Q3 that measures exclusively the total activity in each striatum plus a occipital region determination.

The final gammagraphic diagnosis was defined first by the visual interpretation of the data, using the quantification as a supporting method.

### 3.A.3. Parkinson’s progression markers initiative database

The last database used in the preparation of this paper was obtained from the Parkinson’s Progression Markers Initiative (PPMI) (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org.

The images in this database were imaged 4 + 0.5 h after the injection of between 111 and 185 MBq of DaTSCAN. Subjects were also pretreated with saturated iodine solution (10 drops in water) or perchlorate (1000 mg) prior to the injection. All subjects had a supplied $^{57}$Co line marker affixed along the canthomeatal line, which will facilitate subsequent image processing and allow the core lab to accurately distinguish left and right in the face of multiple image file transfers. These markers are only evident in the $^{57}$Co window and hence do not contaminate the $^{123}$I-DaTSCAN brain data.

Raw projection data are acquired into a $128 \times 128$ matrix stepping each $3^\circ$ for a total of 120 projection into two 20% symmetric photopeak windows centered on 159 KeV and 122 KeV with a total scan duration of approximately 30–45 min. Other scan parameters (collimation, acquisition mode, etc.) are selected for each site. The images of both the subject’s data and the cobalt striatal phantom are reconstructed and attenuation corrected, implementing either filtered back-projection or an iterative reconstruction algorithm using standardized approaches. After the processing, images used are spatially and intensity normalized, and of a $91 \times 109 \times 91$ size.

A total number of $N = 269$ DaTSCAN images from this database were used in the preparation of the paper. 158 suffering from PD and 111 normal controls, with a prevalence of 58.7%.

### 3.B. Validation techniques

The proposed methodology has been tested on the three previously described databases, using a cross-validation method called leave-one-out to extract several performance parameters: accuracy, sensitivity, specificity, positive likelihood (PL), and negative likelihood (NL). This method achieves an almost unbiased error estimate, however it might be affected by the database topology, specially in databases where the number of samples is low. This is one of the reasons why we used different databases to test our method.

Along with the widely used accuracy (which is prevalence dependent) and the medicine standards sensitivity and specificity, we use the positive and negative likelihood ratios (PL and NL), leading to a set of parameters that allow us to compare between different tests performed on different databases. Likelihood ratios are also frequently used in clinical medicine, where values of PL greater than 5 or NL values less than 0.2 can be applied to the pretest probability of a patient having the disease tested for to estimate a posttest probability of the disease state existing. A positive result for a test with an PL of 8 adds approximately 40% to the pretest probability that a patient has a specific diagnosis.

To provide a better understanding of the accuracy distribution (mean, range, and skewness), we have made use of the box plot (Fig. 3). In these plots, the box represents the range between the first and third quartiles of each group of data, which is known as the inter-quartile range (IQR), and the median of each group is indicated by the line that divides the box. The extreme values (within 1.5 times the inter-quartile range from the upper or lower quartile) are the ends of the lines extending from the IQR. Points at a greater distance from the median than 1.5 times the IQR are plotted individually, representing potential outliers.

### 3.C. Experiments

- **Experiment 1.** The 3d co-occurrence matrices are computed for the central and the $d$ surrounding voxels, $d$ ranging from 1 to 10, considering all 13 spatial directions as in Ref. 25. Then, 12 Haralick features are obtained from these matrices, and used as a feature vector, using two different approaches. In the first case, we use the features computed solely at a distance $d$ from the central voxel, our “single distance approach.” In the second one, we use all the features computed from the distance 1 to $d$, in a sort of “cumulative distance approach.”

- **Experiment 2.** In the last case, we consider a total number of 1560 features, which are each of the 12 Haralick features at 10 distances, in all 13 directions together. Then, the $N$ most discriminant ($N$ ranging from 1% to 100% of the total number of features), by means of five different selection criteria, are set as the feature vector.
4. RESULTS AND DISCUSSION

4.A. Experiment 1

In Experiment 1, the influence and effect of using each of the 12 Haralick texture features on the results is tested. GLC matrices are computed at a distance $d$ ranging from 1 to 10 voxels from the central one as in Ref. 38, but additionally to that previous work, the subvolume selection algorithm was included to optimize the textural features computation, and the system is tested against the three aforementioned databases using preprocessed images. As discussed later in this section, the optimum subvolume will have a smaller size than $40 \times 40 \times 50$, and so, the maximum value of $d = 10$ corresponds to at least 20% of the brain subvolume selected at that point; lower frequency textural changes can be neglected for diagnosis. Furthermore, as the voxel size of all databases is approximately $2 \times 2 \times 2$ mm, the maximum textural changes are computed within a $20 \times 20 \times 20$ mm area. This is approximately half the size of the striatum, which is enough to correctly extract the textural features of the area.

The computed GLC matrices are used to extract the Haralick texture features in two different ways: a “single distance approach,” which only considers one type of feature using only the matrices at a distance $d$ from the central voxel—and using all the spatial directions—and a “cumulative distance approach” which considers one type of feature too, but this time using all matrices in distances ranging from 1 to $d$.

As commented before, we will use a normalization to the maximum strategy to normalize all the images in our three databases. After this, the 3D-GLC matrices will be computed over the selected subvolumes, given an intensity threshold $I_{th}$ (see Sec. 2.B.1). The optimum value of $I_{th}$ should be high enough to avoid introducing background voxels in the subvolume selected, yet adequately low to select the biggest subvolume containing only brain voxels. This should lead to the best performance, since the 3D GLC matrices (and the Haralick texture features) would have enough information, and would not include nonbrain textural patterns.

To better illustrate the influence of the subvolume selection process, Fig. 2 shows the resulting subvolumes of applying different intensity thresholds, particularly for $I_{th} = (0.15, 0.30, 0.45) \times I_{max}$. As we can check in that picture, values of $I_{th} > 0.30 \times I_{max}$ obtain the maximum subvolume that selects only brain voxels, which is supposed to lead to better estimations of the GLC matrices, and therefore, to better textural features.

The behavior of each of the Haralick texture features can also be analyzed using a box plot (see Sec. 3.B), to show both numerical accuracy values and the properties (robustness, parameter independence) of using each one. Figure 3 depicts all 130 accuracy results of the single approach for each feature extracted from a subimage that uses $I_{th} = 0.30 \times I_{max}$ at each distance $d$ (ranging from 1 to 10) in each of the 13 spatial directions. In this case, we can observe that best performance is obtained with the cluster tendency in all databases. Good values are also achieved using homogeneity, contrast, and correlation. This behavior is consistent along all three databases, which allow us to propose cluster tendency as the best feature to characterize PD patterns.

Finally, to characterize the ability of using one single feature at the same time, we show its performance at the defined operation point (Using an $I_{th} > 0.30 \times I_{max}$ and a value of $6 < d < 8$) in Table II.

When performing at the operation point, Table II shows that the most effective approach is the single distance based one. Furthermore, while this approach uses always 13 values of a given texture feature (one per spatial direction), the cumulative approach uses $d \times 13$ values to train and test the classifier. Even in that case, the cumulative approach would need to include far more features (by increasing the value of $d$) to provide a good classification performance. This makes the single approach computationally more efficient than the cumulative. Moreover, a higher amount of features does not make the cumulative a better approach, since a number of nonrelevant features is introduced into the classification process. However, both prove the ability of the Haralick texture features to characterize DaTSCAN images.

4.B. Experiment 2

For Experiment 2, all features computed in the aforementioned experiments (the 13 direction vectors and 10 distances used to compute the 3D co-occurrence matrix, and the 12 Haralick texture features extracted from these matrices)
are used as an input to the classifier. But, in order to reduce dimensionality, we use the measures of discrimination ability proposed in Sec. 2.C to rank these features in a descending order of ability in distinguish PD patterns from normal controls, selecting the first $N$.

First, the impact of our volume selection threshold $I_{th}$ (see Sec. 2.B.1) on the quality of the resulting Haralick features, and thus, the accuracy of the experiment, will be analyzed. As commented before, best results should be obtained when taking into account the biggest volume of the brain containing

![Box Plot of accuracy distribution for each Haralick Texture Feature (PPMI)](image1)

![Box Plot of accuracy distribution for each Haralick Texture Feature (VDLN)](image2)

![Box Plot of accuracy distribution for each Haralick Texture Feature (VDLV)](image3)

**FIG. 3.** Box plot of all 130 accuracy values computed for each feature, using the “single approach,” at 10 distances $d$ (ranging from 1 to 10) and 13 spatial directions, for (a) PPMI database, (b) VDLN database, and (c) VDLV database. The cross marks represent the outliers.

### Table II. Accuracy values obtained at the operation point, using cluster tendency as a feature. The distance $d$ used to compute the GLC matrix is also displayed.

<table>
<thead>
<tr>
<th>Database - approach</th>
<th>$I_{th}$</th>
<th>$d$</th>
<th>Feature</th>
<th>Acc.</th>
<th>Sens.</th>
<th>Spec.</th>
<th>PL</th>
<th>NL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPMI - cumulative</td>
<td>30</td>
<td>8</td>
<td>Cluster tendency</td>
<td>0.952</td>
<td>0.973</td>
<td>0.937</td>
<td>15.37</td>
<td>0.029</td>
</tr>
<tr>
<td>PPMI - single</td>
<td>30</td>
<td>6</td>
<td>Cluster tendency</td>
<td>0.952</td>
<td>0.964</td>
<td>0.943</td>
<td>16.92</td>
<td>0.038</td>
</tr>
<tr>
<td>VDLN - cumulative</td>
<td>30</td>
<td>6</td>
<td>Cluster tendency</td>
<td>0.906</td>
<td>0.911</td>
<td>0.904</td>
<td>9.50</td>
<td>0.098</td>
</tr>
<tr>
<td>VDLN - single</td>
<td>30</td>
<td>6</td>
<td>Cluster tendency</td>
<td>0.907</td>
<td>0.911</td>
<td>0.904</td>
<td>9.50</td>
<td>0.098</td>
</tr>
<tr>
<td>VDLV - cumulative</td>
<td>35</td>
<td>7</td>
<td>Cluster tendency</td>
<td>0.899</td>
<td>0.879</td>
<td>0.920</td>
<td>10.99</td>
<td>0.130</td>
</tr>
<tr>
<td>VDLV - single</td>
<td>35</td>
<td>6</td>
<td>Cluster tendency</td>
<td>0.923</td>
<td>0.907</td>
<td>0.940</td>
<td>15.12</td>
<td>0.098</td>
</tr>
</tbody>
</table>
only brain voxels, and thus, eliminating the background. This corresponded, in Sec. 4.A, to values of $I_{th} > 0.30 \times I_{max}$. Regarding all databases, we obtain Fig. 4, in which average values of accuracy for every value of $I_{th}$ are plotted. These average values are computed by averaging all 50 accuracy values that correspond to each value of $I_{th}$. The accuracy values are obtained using each of the 5 proposed selection methods, and using each value of percentage of features selected (ranging from 1 to 100%, by steps of 10%, of the total amount of 1560 features). This is very helpful to discover underlying tendencies that appear when varying $I_{th}$; tendencies that otherwise might not be so clear.

A difference in performance across different databases is very noticeable. It might be influenced by several factors. First, the number of subjects in the database, which usually influences the final performance due to the so-called small sample size problem.39 This problem shows that the performance of a system decreases when the dimension of the feature space (this time, the number of texture features) is very large compared to the number of training samples. It can be seen that the best performance is achieved by the PPMI database (269 subjects), followed by VDLV (208) and ending with the 118-image VDLN database, which could easily be related with this problem. SVM linear classifiers are used in our system in an effort to minimize this problem, since they are more effective in high dimensional spaces. 31 Also, there is an important difference between the acquisition protocols of PPMI and both VDLV and VDLN databases. There exists a small, yet significant, loss of information in VDLV and VDLN databases, due to a common practice in medicine, by which only the striatum areas are scanned. This practice introduces artifacts in the computation of the textural features that degrade the performance of our system. Anyway, it shows that our method, when properly trained, achieves encouraging generalizable results, which also could be improved using bigger datasets, for example, a mixture of all the databases used here. Other influential database-dependent features are addressed in Sec. 4.C.

Regarding the general behavior of the system, for two out of three databases there is a clear maximum in accuracy for an $I_{th} = 0.30 \times I_{max}$ (see Fig. 1), while the remaining one obtains similar results along a wide range of $I_{th}$. Furthermore, best average values are obtained using the normalized database, although the PPMI case is slightly different, probably due to the attenuation correction preprocessing. The chosen value of $I_{th}$ leaves all no-brain voxels out of the textural features computation. Therefore, the computed values correspond only to the internal brain textural changes, and thus, to the textural patterns of the disease, leading to a better performance.

As results suggest, the use of our volume selection strategy with an intensity threshold between 0.25 and 0.45 seems potentially helpful in all cases, which supports the visual analysis that we performed in Sec. 4.A. Moreover, the use of intensity normalized images using the normalization to the maximum algorithm has also a good impact on the performance of the system. In this context, Fig. 5 analyzes the behavior of our system using each of the discrimination-based ranking methods. On these three figures, the values of the computed average accuracy (using the values for intensity thresholds of 0.10 to 0.45) are plotted over the percentage of selected features (previously ranked from the most to the least discriminant, following the specified criteria) for the three databases.

Some conclusions about the amount of features that each of our discrimination-ranking methods need can be drawn from these graphs. Methods that obtain their maximum accuracy using less than 50% of the features can be considered of great help, as they perform a significant feature reduction. Therefore, methods like MWW can no longer be considered, as it needs more than 50% of features to obtain good results.
Fig. 5. Average accuracy computed for each selection criteria, using all accuracy values for intensity thresholds of 0.10 to 0.45. These values are plotted over N, the number of features selected using some of the ranking criteria defined in Sec. 2.C (where N ranges from 1% and 100% of the 1560 total Haralick features calculated). These values correspond to the images of the (a) PPMI database, (b) VDLN database, and (c) VDLV database (Experiment 2).

The opposite behavior is given by BD and RE that obtain their maximum average accuracy using the first 10% of features. FDR and t-test need a higher amount, but still less than 50%.

The aforementioned behavior corresponds to an average behavior in accuracy. To take a deeper look at the different evaluation parameters and different selection criteria, results obtained at the operation point ($I_{th} = 0.30 \times I_{max}$) with each selection criteria are shown on Table III.

This table confirms that the Mann-Whitney-Wilcoxon method can be no longer considered, as it needs more than a 50% of the features to obtain poorer results than all others. Regarding the remaining methods, we observe that those that needed a lower amount of features (BD and RE) obtain here lower values of accuracy that those that needed a higher amount (t-test and FDR). So, the choice of the best method is, in this case, a matter of trade-off between the computer performance (the number of features to estimate) and the accuracy needed. As in clinical practice, sensitivity (and PL) is the parameter that reveals the ability of the system in detecting positives (in the context of clinical diagnosis, affected subjects), we can conclude that FDR and t-test are the best discrimination-ranking methods to use in this task, although all other methods reveal the ability of our system in the PD detection with an relevant performance (over 90% of sensitivity in most cases).

For comparison purposes, we have established a baseline method proposed in Illan et al.,\textsuperscript{20} where a VAFs approach with linear SVM, using different normalization strategies were tested. Two additional methods have been compared with our proposed system in order to check the performance versus state-of-the-art algorithms. These methods have been an asymmetrical single value decomposition (SVD) (Ref. 40) that applied SVD on both sides of the brain (since PD often appears only in one hemisphere), and an empirical mode decomposition (EMD) (Ref. 23) using different independent mode functions (IMF), particularly the IMF-3. Table IV compares all the aforementioned methods.

In Table IV, performance values for the single approach at the operation point are shown for Experiment 1 (using the five selected features in Sec. 4.A) and Experiment 2. This operation point implies the range of values that we previously clarified for Experiment 1: an $I_{th} = 0.30 \times I_{max}$, using distances $d = (6, 7)$. Moreover, values at the operation point for the Experiment 2 (using relative entropy, FDR or Student’s t-test) are also shown, using the same $I_{th}$ as the previous one and the first 35% of all features (a good value not only for FDR and t-test, but also for relative entropy), to be able to compare. All performance values are computed using the PPMI database. This is done this way because it contains less incomplete brains (due to a usual practice in medicine, in which only the striatum areas are scanned), and contains a higher number of patients, making the results less prone to the small sample size problem and better generalizable when comparing to other systems.
These values are compared with the VAF, SVD, and EMD based, previously mentioned, systems. We can observe that the performance values obtained with Experiment 1 are very similar to other state-of-the-art methods, such as the proposed in Refs. 23 and 40, whereas the methodology used in Experiment 2 outperform all previously used methods. Particularly, as we have already mentioned, the use of either FDR or t-test to select the most discriminant features gives us results over a PL of 26 and sensitivity over 99%, which proves the ability of some Haralick textures, and the combination of them, in characterizing the different Parkinson’s disease patterns, and the robustness of the proposed methods against variations in the input parameters.

### 4.C. Discussion

In this section, we discuss the impact on the final result of using different Haralick texture features, the distance at which the GLC matrix has been calculated and how the way the different experiments combine these two aspects affect the performance of the system.

First, Experiment 1 analyzes the behavior of the 12 different Haralick texture features using the computed 3D-GLC matrix, and two different approaches: the single approach, which took the Haralick features in all spatial directions only at a distance \( d \) from the central voxel, and the cumulative approach, which used all distances ranging from 1 to \( d \). Regarding Table II, we obtained that for both approaches best values were obtained using a distance \( 6 < d < 8 \) in all databases. We have tried to eliminate some of the database differences, comprising image sizes, acquisition techniques, by using spatial and intensity normalization, and so, the range of good performance of \( d \) can give us an idea of the “smoothness” of our images, defined as the frequency at which the vast majority of the information is contained. As previously shown, the range \( 6 < d < 8 \) voxels corresponds to distances of \( 12 < d < 16 \) mm, which should be enough to characterize the textural changes of the striatum. This experiment gives us the idea of how well can the Haralick texture features characterize the textural changes in DaTSCAN images.

Second, the similar distribution of performance values that we saw in Fig. 3 among all three databases is a very good indication of how cluster tendency models all databases. As cluster tendency measures the grouping of pixels that have similar gray-level values, it can be expected that the striatum (two strongly cluster shaped volumes, see Fig. 1) will result in high cluster tendency levels in healthy patients, and lower levels in diseased patients, leading to the high performance of the use of this Haralick feature in all databases.

Third, the introduction of the image subvolume selection algorithm has improved a lot the computation of the different Haralick texture analysis, improving the performance results in almost every case (see Fig. 4). It is due to the elimination of outliers in the computation of the Haralick texture features, avoiding the inclusion of nonbrain voxels in the computation.
that could affect the texture of the image, making the estimation of the texture features more accurate.

And finally, the introduction of a selection strategy based on discrimination measures has improved the performance on our three databases, as it selects not only one textural feature, but the most discriminative features irrespective of their definition, distance \(d\), or spatial direction. It is easy to assume that, although cluster tendency revealed as the most discriminative feature, other features can perform as well for some \(d\). Therefore, this strategy would put together all the best features, practically improving the performance on all databases.

To better illustrate the latter, in Fig. 6, the total number of each of the 12 Haralick texture features contained in the first 100 most discriminant features selected, depending on the method used, is depicted. The figure reveals that images from different origins can contain some textural differences, making some textural features more discriminant than others. This might be due mainly to a common procedure in the Parkinson’s disease diagnosis, where only the regions of interest (striatum areas) are scanned, resulting in 3D images where most of the brain is missed. The missing areas can clearly affect the textural features computation for some databases, as results suggested in Tables II and III. Other smaller differences can be due to the reconstruction algorithms used, filtering, attenuation correction, etc.

Although the different distribution of the most discriminant features might be seen as a drawback of our system, it is, in fact, one of its most important strengths: the ability of selecting the most suitable features regardless of the database used, the image modalities, and probably, the disease analyzed. Therefore, our system proves to be adaptable to different conditions and particularities of each database, while maintaining robustness against changes in the input parameters, and can also be considered potentially extensible to other uses in clinical practice.

5. CONCLUSIONS

In this work, a new approach to the computer aided diagnosis of the Parkinson’s disease (PD) based on the Haralick texture features has been proposed. It is based on the intensity normalization of the images, the extraction of

\[
\text{Energy, Entropy, Correlation, Contrast, Variance, \(S_{\text{umMean}}, \text{Inertia, ClusterShade, ClusterTendency, Homogeneity, MaxProba}, \text{InVariance, bhattacharyya, entropy, FDR, ttest, wilcoxon.}
\]
subimages of interest and the computation of the 3D gray level co-occurrence matrices, then extracting Haralick texture features from those. These textural features are ranked by means of their estimated discrimination ability, using five different measuring methods, and the first \( N \) features are taken as a feature vector for a SVM classifier.

Therefore, along with the novel use of textural features in the diagnosis of Parkinson’s disease, two major contributions have been stated: the influence of our image subvolume selection method based on an intensity threshold for an optimum computation of the Haralick texture features, and the use of the proposed discrimination estimation methods to select the most promising features to use with the classifier.

On the one hand, the use of a subvolume instead of the whole image improves the calculation of the Haralick texture features by eliminating nonbrain textural changes, which has a clear impact on the final results. Furthermore, the smaller size of the subvolume helps to reduce the computation load, which is also desirable when using large amounts of data. Therefore, once this advantage has been proven, new fully automatic subvolume selection algorithms can be implemented to eliminate the dependence on the \( I_{th} \) parameter.

On the other hand, the proposed discrimination-ranking methods (especially the Fisher’s discriminant ratio and Student’s \( t \)-test) make the feature selection automatic, depending exclusively in the discrimination estimation algorithms for each feature (within the test loop), but not in values of \( d \) neither values of \( N \). This makes the system fully extensible to any database, since it selects only the most discriminative texture features that model different diseases in different ways without a previous knowledge.

Using the leave-one-out cross-validation technique over these three databases, the system achieves results up to a 97.4% of accuracy, and 99.1% of sensitivity, with positive likelihood ratios over 27 on the standard widely used in Parkinson’s disease PPMI database.

**ACKNOWLEDGMENTS**

PPMI—a public-private partnership—is funded by The Michael J. Fox Foundation for Parkinson’s Research and funding partners, including Abbott, Biogen Idec, F. Hoffman-La Roche Ltd., GE Healthcare, Genentech, and Pfizer Inc. This work was partly supported by the MICINN under the TEC2008-02113 and TEC2012-34306 projects and the Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía, Spain) under the Excellence Projects P07-TIC-02566, P09-TIC-4530, and P11-TIC-7103.

**Author to whom correspondence should be addressed. Electronic mail: fjesusmartinez@ugr.es**

**Data used in the preparation of this paper were obtained from the Parkinson’s Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). As such, the investigators within PPMI contributed to the design and implementation of PPMI and/or provided data but did not participate in the analysis or writing of this report. PPMI investigators include (complete listing at PPMI site).**


