A multimodal Parkinson quantification by fusing eye and gait motion patterns, using covariance descriptors, from non-invasive computer vision

John Archila, Antoine Manzanera, Fabio Martinez

A multimodal Parkinson quantification by fusing eye and gait motion patterns, using covariance descriptors, from non-invasive computer vision

J. Archila[a], A. Manzanera[b], F. Martinez[a]

[a] *Biomedical Imaging, Vision and Learning Laboratory (BIVL^{2}ab), Universidad Industrial de Santander, Bucaramanga, Colombia,* famarcar@saber.uis.edu.co
[b] *Robotics & Autonomous Systems (U2IS), ENSTA Paris, Institut Polytechnique de Paris, France,* antoine.manzanera@ensta-paris.fr

Abstract

Background and Objective: Parkinson's disease (PD) is a motor neurodegenerative disease principally manifested by motor disabilities, such as postural instability, bradykinesia, tremor, and stiffness. In clinical practice, there exist several diagnostic rating scales that coarsely allow the measurement, characterization and classification of disease progression. These scales, however, are only based on strong changes in kinematic patterns, and the classification remains subjective, depending on the expertise of physicians. In addition, even for experts, disease analysis based on independent classical motor patterns lacks sufficient sensitivity to establish disease progression. Consequently, the disease diagnosis, stage, and progression could be affected by misinterpretations that lead to incorrect or inefficient treatment plans. This work introduces a multimodal non-invasive strategy based on video descriptors that integrate patterns from gait and eye fixation modalities to assist PD quantification and to support the diagnosis and follow-up of the patient. The multimodal representation is achieved from a compact covariance descriptor that characterizes postural and time changes of both information sources to improve disease classification.

Methods: A multimodal approach is introduced as a computational method to capture movement abnormalities associated with PD. Two modalities (gait and eye fixation) are recorded in markerless video sequences. Then, each modality sequence is represented, at each frame, by primitive features composed of

(1) kinematic measures extracted from a dense optical flow, and (2) deep features extracted from a convolutional network. The spatial distributions of these characteristics are compactly coded in covariance matrices, making it possible to map each particular dynamic in a Riemannian manifold. The temporal mean covariance is then computed and submitted to a supervised Random Forest algorithm to obtain a disease prediction for a particular patient. The fusion of the covariance descriptors and eye movements integrating deep and kinematic features is evaluated to assess their contribution to disease quantification and prediction. In particular, in this study, the gait quantification is associated with typical patterns observed by the specialist, while ocular fixation, associated with early disease characterization, complements the analysis.

Results: In a study conducted with 13 control subjects and 13 PD patients, the fusion of gait and ocular fixation, integrating deep and kinematic features, achieved an average accuracy of 100% for early and late fusion. The classification probabilities show high confidence in the prediction diagnosis, the control subjects probabilities being lower than 0.27 with early fusion and 0.3 with late fusion, and those of the PD patients, being higher than 0.62 with early fusion and 0.51 with late fusion. Furthermore, it is observed that higher probability outputs are correlated with more advanced stages of the disease, according to the H&Y scale.

Conclusions: A novel approach for fusing motion modalities captured in markerless video sequences was introduced. This multimodal integration had a remarkable discrimination performance in a study conducted with PD and control patients. The representation of compact covariance descriptors from kinematic and deep features suggests that the proposed strategy is a potential tool to support diagnosis and subsequent monitoring of the disease. During fusion it was observed that devoting major attention to eye fixational patterns may contribute to a better quantification of the disease, especially at stage 2.
Keywords: Parkinson, multimodal approach, temporal mean covariance, deep features, kinematic features

---

1. Introduction

Parkinson's Disease (PD) is a motor neurodegenerative disorder that affects more than 6.1 million people around the world [1]. Even worse, the disease is in geographic expansion, and there is no cure or effective treatment [1]. PD is related with dopamine deficiency, and it is mainly associated to motor disabilities, among others: postural instability, slowness, reduced steps, bradykinesia, tremor, and stiffness. Then, locomotion patterns constitute the basis of clinical protocols to establish, quantify, and monitor the PD level. In clinical practice, such patterns are evaluated independently during gait, evaluating postural configurations, and even doing exercises related to control and coordination. Parkinson detection, evaluation, and stage characterization are then carried out from personal therapy protocols and supported according to different diagnosis rating scales [2, 3]. For instance, the classical Hoehn and Yahr (H&Y) rating scale stratifies disease progression in five stages, considering physical capabilities such as gait, postural stability, and balance between others [4]. In H&Y, the disease is coarsely classified into five stages: from zero (no sign) to five, corresponding to the largest severity of the disease. However, in this rating scale, it is very difficult to discriminate among intermediate stages, for instance, some patients fail in pull test (corresponding to level three) but there are no evident tremor patterns (as in the second level)[3]. This fact results critical to define personalized treatments according to the progression of PD, causing the H&Y scale to fall into disuse. Nowadays, Unified Parkinson's Disease Rating Scale Motor Exami-

nation (UPDRS-ME), has become more important in clinical practice, evaluating among other patterns: gait, facial mobility, action tremor, bradykinesia, and hypokinesia [3]. Despite major sensitivity of UPDRS-ME scale, the evaluation and patient stratification highly depends on the expertise of the professional which could be prone to errors [5, 6].

A study with 50 patients and six different evaluators has evidenced a low level of agreement, which makes it difficult to determine the level of disease's motor function [5]. Automatic tools to support diagnosis then result fundamental to properly follow and personalize diagnosis and treatments. Such tools should integrate quantification and modelling of disease heterogeneity, allowing an earlier symptomatic diagnosis [7].

In such a sense the quantitative analysis of complementary disease patterns could improve the robustness of motor scale assessment, to detect and measure the progression of Parkinson with major sensitivity. Hence, the constant search for new biomarkers and the posterior integration with known patterns could be fundamental to better diagnose, quantify, characterize and monitor the disease. In the literature, alternatives have been proposed to capture motion patterns using acceleration and angular velocity components. These kinematics are in general captured from the lower limb during locomotion, using inertial measurement unit sensors [8]. These approaches nonetheless require multiple electrodes placed on each patient, which limits natural gestures during exercises. A more sophisticated alternative has proposed capture setups using Doppler-

based linear quantifier to capture motion tremor patterns [9]. This alternative overcomes invasive measurements and achieves a remarkable correlation of captured patterns with the disease, but requires sophisticated devices and capture setups that result difficult to deploy in clinical scenarios. Also, to support more sensitivity among stages, biomarkers such as sleep disorders, and eye movement have emerged to detect early disease stages [10]. The video analysis has also emerged as a potential alternative to observe and quantify abnormal locomotion patterns related with Parkinson's disease from different key processes, such as the gait, or focusing on standing out body parts like the arms, the face, and the eyes. Particularly, the eye movements have demonstrated to stand out abnormal patterns with the capability to describe the severity and disease progression, even in otherwise asymptomatic stages [11], [12]. Nevertheless, such experimental observations require sophisticated capture devices and protocols, limiting their effective integration on scale rating diagnoses [13].

This paper introduces a novel approach that captures movement abnormalities associated with PD from different sources, and assists in the disease quantification. The gait and eye fixation movement patterns are herein recorded and analyzed from video descriptors to characterize the disease. Each video sequence, at each modality, is represented by frame-covariance matrices that summarize responses of deep and kinematic features. These frame-covariances form a video manifold that codes motion pattern modalities in a compact representation. Then, a geometrical Riemmanian mean is computed

as a video descriptor. The same method is applied to both modalities to facilitate their interpretation, merging for any kind of (deep or kinematic) features. Such multimodal video descriptors are projected to a tangent plane to allow them to undergo linear operations. Then, the descriptor is mapped to a supervised machine learning strategy to obtain a PD classification. The proposed approach was validated from early (at the level of covariance descriptors) and late (at the level of output probabilities) integration of both modalities to better understand the capability of discrimination of the proposed video descriptors.

2. Related work

Parkinson's disease causes different motor and non-motor manifestations, at different stages, that could be recorded from different sensors, and focus on different human functionalities [10]. These manifestations allow to represent and quantitatively evaluate the disease progression or the effect of a particular treatment. The most common consequences are the motor disabilities developed progressively during the disease, which have been quantified from global locomotor observations such as bradykinesia, rigidity, hypokinesia, tremor, and others [10, 14]. Despite their importance, these patterns lack of sufficient sensitivity, for instance to the early disease detection, and also to precisely characterize and score the disease progression. Recently, eye movement has been proposed as a new complementary biomarker, where abnormal tiny motions are highly correlated with PD [15]. For instance, micro-tremors during

eye fixation is fully correlated with the disease. Then, in this work, global locomotion patterns and eye fixational movements were analyzed together to provide a rich dynamic description and better support disease characterization. The computation of both motor modalities is carried out from a non-invasive perspective, allowing to capture natural patient movements. In the next subsections, we briefly describe each modality and current state-of-the-art strategies used to capture and process the related motion patterns.

2.1. Gait analysis

Gait is a complex locomotion process that requires the coordination of neuromotor commands, muscle activation, and structural pose configurations to produce an optimal displacement. PD alterations produce unbalanced postures, requiring additional control and effort to mitigate tremor patterns, therefore producing non-optimal locomotion displacements. Gait movement patterns such as stride length, postural flexion, lack of swing arms and bradykinesia, currently constitute the main source of information to describe and quantify PD. In fact, most clinical diagnosis scales are based on such locomotion patterns.

Kinematic gait analysis, during a clinical routine, is achieved using sophisticated systems, e.g. optical marker-based frameworks to capture joint relationships during displacements [16], force plates to measure reaction forces [17] and acceleration [18]. Also, typical alternatives in motion analysis are based on inertial sensors attached to specific body parts to capture long kinematic patterns, during particular routines [19],[8]. Related to Parkinson's disease, these

alternatives have allowed collecting motion patterns that achieve a discrimination to control population. Despite these advantages, these methodologies result invasive and limit the natural gestures. To overcome such issue, Lin *et. al* proposed a microwave motion detector to recover tremor signs without invasive protocols [9]. These approaches nevertheless need a complex calibration process and their effectiveness depends on the proper localization of sensors. Regarding kinematic analysis, the structural body model is strongly simplified to a set of joints, which could be restrictive to understanding regional Parkinsonian patterns such as the tremor at the initial stages of the disease. Hence, new alternatives of video analysis could significantly improve the quantification of abnormal patterns associated with PD. In such line, Guayacan *et al.* introduced a 3D convolutional neural networks to classify and recover salient learned regions related to Parkinson's disease [20]. This approach recovers explainable maps but a further interpretation is needed to correlate such findings with clinical know patterns. Despite their relevance to express natural movements during locomotion, many existing markerless methods aim at reproducing classical marker models, which turns restrictive to compute global trajectories, and strongly simplifies such complex phenomenon.

2.2. Fixational oculomotor patterns

In recent studies, oculomotor patterns have been established as potential Parkinson's disease biomarkers, reporting a strong correlation with dopamine deficiency, which makes them candidates to support early detection and diagnosis of the disease [12]. To cha-

racterize such patterns, different experiments were proposed in the literature, allowing to measure the eye capability response and the control of eye movement [21]. For instance, an experiment has been carried out to evaluate the ocular fixation, *i.e.,* the ability to stabilize the gaze at a given point. Actually, it was found that in control patients, eyes register small involuntary movements called microsaccades at intervals of 1 to 2 Hz, while for Parkinson patients, the fundamental frequency of movements is around 5.7 Hz [22]. Hence, these kinds of eye patterns could be determinant to characterize PD patterns even in very early stages (including asymptomatic patients) [13].

The observation of abnormal patterns in eye movements are typically conducted using electro-oculography protocols, with the main objective to recover speed up movements as the saccades [23, 24]. This technique is limited to tracking movements only in the horizontal plane, reporting low sensitivity to capture abnormal fixational patterns, and any flicker induces noise in the signal, reducing the accuracy of the estimation. Alternatively, the video-oculography (VOG) allows to record bidimensional movements and estimate measures such as velocity, latency and other kinematic relationships related with dopamine deficiency [25], [26]. These works have revealed some significant differences between Parkinsonian and control patterns in the specific experiment of fast vergence eye movements, for which the latency of divergence and convergence was increased in PD subjects, using infrared video-oculography [27]. Nevertheless, the VOG is limited to monitoring global patterns and the quan-

tification of the kinematic variables of the eye. Furthermore, the required equipment is expensive, it requires a precise calibration protocol and it is invasive, by covering the entire ocular region.

Recently, some spatiotemporal relationships have been captured from the analysis of raw video sequences, computing dense pixel-wise motion-based representations that allow to characterize and distinguish some abnormal motion patterns on a particular population of study [28–30]. New markerless video schemes have also been introduced to magnify ocular video patterns, emphasizing fixation patterns [31]. These approaches have demonstrated capabilities to discriminate among Parkinson and Control population, but their analysis is dependent on video segmentation, manually performed from the video sequences. In fact, the use of video slices implies a geometric dependency on the eye shape. These approaches, however, can hardly represent small or short patterns, that could nonetheless be highly correlated with some stages of PD. Additionally, these patterns should be integrated with other Parkinsonian patterns to better quantify and enhance abnormal behaviors of patients.

2.3. Multimodal approaches on PD

In recent years the multimodal approaches have taken importance to complement disease patterns and perform a better analysis of the disease. For instance, typical Parkinsonian writing tests are complemented with speech patterns to provide decision tables that support disease characterization, but the final classification remains dependent on specialist expertise [32]. Gait, writing, and speech have been combined in a low dimensional feature space in [33], where

each selected modality, for each patient, was modeled using Gaussian mixtures. The Bhattacharyya distance was then used to predict MDS-UPDRS-III, based on a simple linear regression hypothesis. In [34] the writing, speech, and gait spectrograms were modeled with convolutional nets, and thereafter the embedding vectors were concatenated and used to predict the disease. This approach deals with asynchronous modalities integration, and the contribution of each modality was not investigated.

3. Materials and Methods

This work presents a novel multimodal methodology to capture and integrate movement abnormalities associated with Parkinson's disease by using as video representation a special Riemmanian manifold of temporal frame-covariance matrices.

The pipeline of the proposed approach is illustrated in figure 1. Firstly, videos that record particular modalities of interest are represented as a set of frame-covariance matrices. This set forms a special manifold that can be summarized by a Riemannian mean. The fusion of the two modalities can be performed at the video descriptor level (early fusion) or at the prediction level, after mapping in a supervised learning strategy as Random Forest.

3.1. Frame-level Representation

A markerless strategy is herein introduced by computing a spatially dense representation for each frame along the video sequence. The local representation refers to the set of features extracted for each frame I_t at time t, denoted F_t. The features in F_t aim at

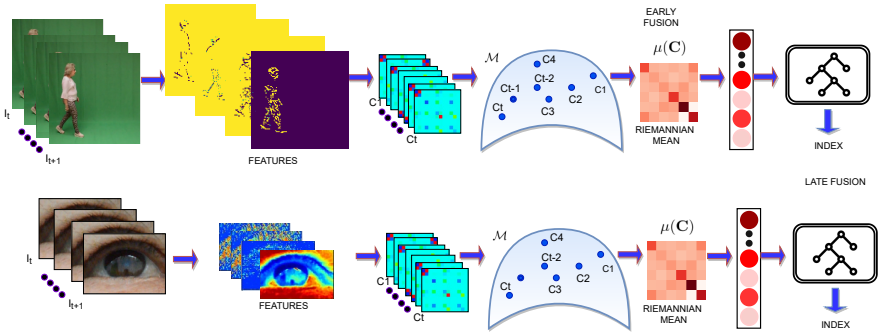

Figure 1: The pipeline of the proposed approach. Top: Gait. Bottom: Ocular fixation. In both modalities, we calculate the features for each frame along with the video, and then compute frame-level spatial covariance of the features. Finally, we summarize the information into a unique covariance matrix for the complete video (Riemann's mean). We propose two fusion approaches: early (concatenate the descriptors of each modality) and late (weight the probabilities of the two modalities to obtain a final probability).

enhancing relevant motion and characteristics that could be discriminative to PD, allowing thereafter a proper coding of abnormal patterns. Each frame I_t is then represented by a set of N features. $F_t = \{f_{(1,t)}, f_{(2,t)}, \ldots, f_{(n,t)}\}$. In this work, two different schemes were evaluated to characterize each frame: kinematic features and deep features calculated using pre-trained networks. On the one hand, kinematic features are computed from a dense optical flow field. On the other hand, taking advantage of the expressivity of deep representations, each frame is processed by a filter bank composed of the convolution kernels extracted from the first layers of a pre-trained convolutional network (see figure 2). The two next subsections detail these two sets of features.

3.2. Kinematic features from a Dense flow field

The optical flow is a 2d vector field that corresponds to the estimation of the apparent velocities of all pixels of the video between two consecutive frames. Such quantity is naturally relevant to characterize patients movement by computing kinematic local primitives on gait or eye fixation. To compute the optical flow, the video sequence is first pre-processed by computing a local entropy map to lower redundancy and enhance edges. The movement is then calculated using Farnebäck's method, [35] that uses a quadratic polynomial approximation of each pixel's neighborhood to estimate local velocity: $I_t(\mathbf{z}) \simeq \mathbf{z}^T \mathbf{A}_t \mathbf{z} + \mathbf{b}_t^T \mathbf{z} + c_t$, where $\mathbf{z} = (x, y)^T$ is the pixel position, and matrix $\mathbf{A}_t$, vector $\mathbf{b}_t$ and scalar c_t are estimated from the image I_t. The displacement vector $\mathbf{d}_t$ between I_t and I_{t+1} is then obtained as: $\mathbf{d}_t = -\frac{1}{2}\mathbf{A}_t^{-1}(\mathbf{b}_{t+1} - \mathbf{b}_t)$.

The obtained dense field $\mathbf{d}_t$ provides a rich and dense kinematic description of recorded video sequences. Hence, to characterize gait and eye motion patterns, a set of kinematic measures are extracted from the flow field. Specifically, we decompose the normalized velocity vector in unit tangential $\mathbf{T}_t$ and unit normal $\mathbf{N}_t$ components to obtain greater specificity of the velocity [36]. These two components allow determining a possible decrease in gait velocity in patients [37]. They also allow to determine the variation in eye movement focused on a fixed point, which could have a linear trend [38]. Similarly, scalar components of the acceleration are calculated as the magnitudes of the tangential and normal components of the acceleration, respectively denoted a_t^T and a_t^N. These components de-

termine the possible abrupt velocity changes due to the imbalances in the patient's gait [18]. Also, it is expected that the correlation between velocity and acceleration is lower in gait for patients with PD than in control subjects [37]. In the ocular fixation modality, the square wave jerks (SWJs) are inappropriate movements that occur through kinetic changes when the eye is dispersed from the fixed point. SWJs saccades have more frequency and magnitude in PD patients than control subjects [39]. In summary, kinematic features are encoded as the first order kinematics, corresponding to the horizontal and vertical components of the unit tangential vector $f_{(1,t)}$ and unit normal vector $f_{(2,t)}$ velocity, and as the second order kinematics, corresponding to the magnitude of tangential $f_{(3,t)}$ and normal acceleration $f_{(4,t)}$ [36]. The features based on the optical flow are illustrated on the left side of figure 2. Tangential and normal velocity is reflected in different parts of the body according to their components. The accelerations in $f_{(3,t)}$ and $f_{(4,t)}$ are less perceptible. However, they are visible on the feet and wrists.

3.3. Deep Features

The deep convolutional networks have recently demonstrated great capability to represent very complex visual patterns in classification and detection tasks, showing remarkable robustness to camera lens distortions, illumination changes or occlusions, among many others [40]. Nevertheless, proper end-to-end learning with these architectures requires a huge amount of data to carry out the training. Therefore, to achieve major flexibility modelling, we chose to represent each frame using learned features from pre-trained nets. These

features are then computed from convolutional deep pre-trained nets as an alternative and complementary description of each frame. So the first layers learn a set of kernel filters that provide a rich representation of images, including non linear relations achieved by activation functions. Specifically, these filters process an input image I_t from the video, that can be either an RGB frame (3 channels), or an optical flow map (2 channels), through a set of S learned convolution filters $\mathbf{\Phi} = \{\Phi_k\}_{1 \leq k \leq S}$ to form a set of S features $\mathbf{F_t^k} = a_k\left(I_t * \Phi_k\right)$, where a_k is a non linear activation function. For instance, for optical flow frames, the learned filters can compute acceleration related maps that may contribute to describe Parkinson disease patterns from video sequences.

Classically, the deep Convolutional Neural Networks (CNN) calculate features from layer to layer in such a way that, if the layer l has n_l neurons (i.e. calculates n_l features), and each neuron has $d_l \times d_l$ weights (i.e. calculates a $d_l \times d_l$ convolution), then the layer l actually computes n_l convolutions of size $d_l \times d_l \times n_{l-1}$, where n_{l-1} is the number of features (or channels) of the previous layer.

More recently, other CNN approaches have been proposed that perform separable convolution in depth, *i.e.* for each layer, n_{l-1} (depth-wise) convolutions of size $d_l \times d_l$ are first applied, and then, n_l (point-wise) convolutions of size n_{l-1} are applied. This produces the same number of features and the same data size, while reducing the computational cost by an order of magnitude [41, 42]. This decomposition allows less redundancy at the output compared to standard convolution [43]. Besides, the experiments on different

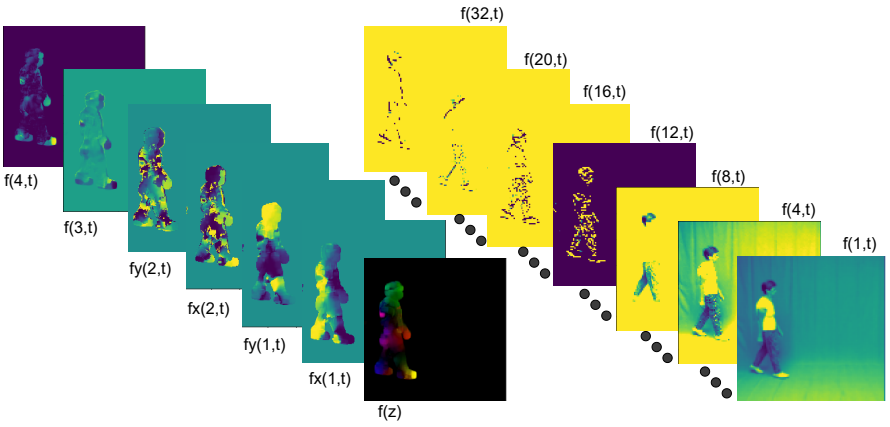

Figure 2: Left: Kinematic Features from the optical flow. The first order kinematics features are the horizontal and vertical components of the unit tangential $f_{(1,t)}$ and unit normal $f_{(2,t)}$ velocity. The second-order kinematics features are the magnitude of normal $f_{(3,t)}$ and tangential $f_{(4,t)}$ acceleration. Right: some of the 32 deep features coming from the fourth layer of MobileNet V2.

data sets show that a network with separable layers requires fewer data to achieve similar or better performance compared to the dense-layer architecture [43].

In this work we evaluated both the conventional and the separable architectures (such as MobileNet). Figure 2 (right) shows feature maps extracted from the gait modality. This figure highlights the expression of the different features within different parts of the body, and the co-variation of the different features, which justifies the interest of a covariance based representation.

3.4. Riemaniann space of covariance descriptors

The integration of gait and eye motion patterns is expected to provide a more sensitive description of Parkinson's patterns and a better quantification of the disease. Because computed video des-

criptors are represented as covariance matrices, a natural fusion can be done for the different modalities. A compact integration can be achieved by aggregating gait and eye descriptors. In this work two levels, early and late fusion were evaluated. Considering the fact that Parkinson's disease has a typical unilateral involvement, we considered one covariance descriptor for each eye, and one covariance for the gait, to recover the whole motion spectrum to characterize each patient. The description starts then by computing, for each frame t, a spatial covariance matrix C_t relative to the set of feature maps $F_t = \{f_{(1,t)}, \ldots, f_{(n,t)}\}$, where the n features can be the kinematic features, the deep features, or the union of all features. The covariance matrix is computed as:

$$C_t(i,j) = \mathbb{E}\left((f_{(i,t)} - \mathbb{E}(f_{(i,t)}))(f_{(j,t)} - \mathbb{E}(f_{(j,t)}))\right)$$

where the expectation $\mathbb{E}$ is calculated over the $W \times H$ points of each feature map $f_{(i,t)} \in \mathbb{R}^{W \times H}$, where W and H represent the width and height of the feature maps, respectively.

Then, a very compact representation is obtained for each frame, allowing to model complex patterns from a low temporal dimensional manifold: indeed, the covariance matrices lie on a half cone space, the actual dimension of the covariance matrix being given by $\dim(C) = \frac{n(n+1)}{2}$ where n is the number of features.

Such measures allow to summarize motion characteristics that may be typical of Parkinsonian patterns from a global (gait) and from a local (eye fixation) evaluation. Then, for a given video sequence, the frame feature maps are summarized as a sequence of spatial covariance matrices, represented as $\mathbf{C} = (C_1, C_2, C_3, \ldots, C_N)$.

These symmetric positive matrices C_i are part of a non Euclidean space which is a Riemmanian manifold $\mathcal{M}$[44], and then Euclidean metrics is not suitable to compute temporal statistics on $\mathbf{C}$.

To make such measures, each covariance point should be projected to a tangent plane to the manifold (logarithmic operation). Accordingly, a projected covariance could be mapped to the original Riemannian manifold, which corresponds to the exponential operation. Particularly, the mean in $\mathcal{M}$ of a set of covariance matrices $\mathbf{C}$ can be iteratively found by optimization, where the mean μ is the point (covariance matrix) with minimum distance ρ among the sample covariance matrices [44]. Thus, the geometrical mean can be expressed as:

$$\mu_{t+1} = \exp_{\mu_t}\left(\frac{1}{k}\sum_{i=1}^{k}\log_{\mu_t}(C_i)\right)$$

where μ_0 is the initial guess and μ_{t+1} it the $(t+1)$ approximation of the geometric mean. This expression requires in each iteration the computation of the matrix function $\log_{\mu_t}$ and $\exp_{\mu_t}$, expressed as:

$$\begin{aligned}\log_{\mu_t}(C_i) &= {\mu_t}^{\frac{1}{2}}\log\left({\mu_t}^{-\frac{1}{2}}C_i{\mu_t}^{-\frac{1}{2}}\right){\mu_t}^{\frac{1}{2}}\\ \exp_{\mu_t}(C_i) &= {\mu_t}^{\frac{1}{2}}\exp\left({\mu_t}^{-\frac{1}{2}}C_i{\mu_t}^{-\frac{1}{2}}\right){\mu_t}^{\frac{1}{2}}\end{aligned} \tag{1}$$

where ${\mu_t}^{\frac{1}{2}} = \exp(\frac{1}{2}\log(\mu_t))$ and $\log(\mu_t) = \sum_t \log(\lambda_t)\Lambda_t^T$ (in the same way: $\exp(\mu_t) = \sum_t \exp(\lambda_t)\Lambda_t^T$), where Λ and λ are the eigenvectors and eigenvalues of the matrix μ, respectively. Here exp and log are the corresponding functions of matrices that extend the real

exponential and logarithmic functions. As frame-covariance samples, the geometrical mean covariance has the dimension of $\mu_t \in \mathbb{R}^{d\times d}$, being symmetric ($\mu_t = {\mu_t}^\top$) and positive ($\det(\mu_t) > 0$). Therefore the final descriptor has the dimension of $\dim(\mu_t) = \frac{n(n+1)}{2}$ where n is the number of features. Finally, each video descriptor is defined as the Riemannian mean of frame-level covariance matrices. In this work, the implementation of the covariance mean has been computed as described in Algorithm 1.

Algorithm 1 Global video descriptor from intrinsic Riemannian mean

Require: $\mathbf{C} = (C_1, C_2, C_3, \ldots, C_N)$

1: start with: $\mu_0 = C_1$

2: **repeat**

3: $\quad X_k = \frac{1}{N}\sum_{i=1}^{N} \log_{\mu_k}(C_i)$

4: $\quad \mu_{k+1} = \exp_{\mu_k}(X_k)$

5: **until** $||X_k|| < \varepsilon$

Ensure: $[\mu_{k+1}]$

This global statistic provides a compact representation that summarizes the main tendencies observed along the different phases of the action and naturally reduces noise or error artifacts that could suddenly appear in one frame. Furthermore, because the global descriptor ignores the temporal relations between the different frames, it is also invariant to the phase of the action.

3.5. Parkinson prediction using Covariance descriptors

The covariance mean, that represents a global measure for any video, can be used as a patient signature to quantify the level of Parkinson's disorder or to automatically classify between Parkinson

and Control motion patterns. For this classification, a supervised strategy can be implemented to learn patterns from classes and to build a disease space, where new samples can be projected to be automatically labeled with a particular class. Nonetheless, these supervised algorithms generally operate under a Euclidean metric. To project each covariance into a Euclidean space, a logarithmic projection was carried out as $\log(C_i) = \Sigma \log(\lambda_i)\Sigma^T$, that defines a space with reference to the identity [44].

In this work was implemented the Random Forest as a supervised strategy due to the demonstrated effectiveness to represent very complex problems over discrete space, to address overfitting problems, and to be less sensitive to atypical data [45, 46]. Specifically in this work, a set of Riemannian mean descriptors of study subjects $C_1, C_2, ...C_i$ with the disease stage notations $y \in \{0, 1\}$ are used to learn boundaries between Parkinson and control. Then, the Random forest defines a set of decision trees, into a Bootstrap aggregating strategy through an ensemble learning, that allows to obtain multiple classifications and maximize the expected mean prediction. For so doing, the strategy randomly selects a set of covariance features to build different tree versions. Each tree is constructed using the (CART) technique based on a recursive procedure that seeks to obtain the division of the data that minimizes the label variance from each node [47]. The final prediction is made by averaging the predictions of the individual trees, as follows: $\hat{y} = \sum_{i=1}^{B} \frac{\theta_i}{B}$, where B is the number of trees.

3.6. Fusion modalities

The different Parkinsonian observations can be fused once the videos are described by the set of frame covariances, forming the special sequence manifold. In this work, two different levels of fusion are proposed, evaluated over the two motions of interest, *i.e.,* gait and eye fixation, described as follows:

3.6.1. Early Fusion

In the early fusion approach, we propose a joint representation (J^e) of Riemann's descriptors: in ocular fixation (C^e) (each eye separately) and gait (C^g) per *i-patient*. $J_i^e = [C_i^{e_{left}}, C_i^{e_{right}}, C_i^g]$

This descriptor represents the covariation among selected features to represent videos (kinematic and/or deep features) in the different actions of the same person. The formed descriptor J_i is then used to build a classification space with the Random forest strategy. Then, during training, different tree versions can be formed by grouping different features from different modalities. Finally, those hybrid random trees are used for classifying an unknown subject from his eyes and gait sequences.

3.6.2. Late Fusion

A second fusion alternative proposed in this work was to learn independent classification spaces using each modality separately. In such case, each mean covariance, for gait $C_i^g \rightarrow RF^g$ and for eyes $(C_i^{e_{left}}, C_i^{e_{right}}) \rightarrow RF^e$ is used to learn independent modality trees, resulting in two different random forest models: (RF^g, RF^e).

In such case, each of the specialized random forest provides its own probability of disease. Then, we model the resulting probability

P_f as a linear combination of the probabilities of each classifier by: $P_f = wP_g + (1 - w)P_e$, where P_e and P_g are the ocular fixation and gait probabilities respectively, and w is a modality importance weight in the final disease prediction.

4. Experimental setup

4.1. Data

A total of 26 participants was included in this study: 13 control subjects (average age of 72.2 ± 6.1) and 13 PD patients (average age of 72.3 ± 7.4). The PD patients were diagnosed in the second or third stage of the disease by a physician using standard protocols of the Hoehn-Yahr scale. This study was approved by the Ethics Committee of Universidad Industrial de Santander and written informed consent was obtained. Regarding the motion modalities, the following protocols were applied:

- For eye fixational recording, the patients observed a fixed spotlight projected on a screen with a dark background, for an average duration of 6 seconds. The eye region was manually cropped (210×140 pixels) to obtain the sequences of interest.
- For gait, markerless sagittal-plane videos were recorded with a spatial resolution of 520×520 pixels and a temporal resolution of $60\,fps$. The locomotion was recorded along a 6 meter displacement, for an average duration of 6 seconds. For each participant, one video for gait and one video for each eye were recorded, resulting in a total dataset of 78 videos.

4.2. Parameters tuning

The proposed approach was adjusted at different stages to optimize the representation w.r.t. description and quantification of Parkinsonian disease patterns. According to each stage, the following parameters were set:

- **Kinematic features.** In both modalities the video sequences were processed with the Farnebäck optical flow with 5 scales and 3×3 window size, to obtain velocity fields at each frame. From each computed frame was then computed a total of 6 kinematics, namely the horizontal and vertical components of the tangential and normal unit velocity vectors, along with the tangential and normal acceleration magnitudes.

- **Deep features.**

  The deep features were taken from the first layers output using two different pre-trained nets: the VGG16 (standard convolutions) and the MobileNet V2 (depth wise separate convolutions). For VGG net was selected the first to fourth deep layers, that count from 64 (kernel size of 3×3) to 128 (kernel size of 3×3) filters. Then, each activation output from this net has spatial size from (112×112) to (224×224). For MobileNetV2, the second to fifth layer were selected. Each layer counts a total of 32 filters and activation output with spatial size of (112×112).

- **Riemann descriptor.** A covariance mean is calculated for each video. Three descriptors are obtained for each patient,

for each eye and for the gait independently. The resulting covariance descriptor, for each video, is formed by only 6 scalar motion features and/or 32 deep features. The descriptor fusing all modalities corresponds to a vector whose dimension can vary from 108 to 4 332.

- **Random Forest.** The classifier was trained using the bootstrap aggregating strategy with optimization metrics based on entropy criterion. A comprehensive evaluation of different sets of trees and numbers of samples per leaf in the Random Forest was carried out to get the best classification results in each modality and using different types of fusion.

4.3. Experimental configuration

To evaluate the performance of the proposed approach a cross-validation leave-one-patient-out was implemented with the multimodal dataset. In such a scheme at each iteration, one patient is left out to test and the remaining ones (25 subjects in our particular experiment) are used for training. For these experiments, the Parkinsonian patients correctly classified were counted as true positive (TP) and the correct control patients were identified as true negative (TN). Then a set of metrics was used to fully understand the performance of the approach in its different configurations. The metrics herein implemented are the sensitiviy ($sen = \frac{TP}{TP+FN}$), specificity ($spec = \frac{TN}{FP+TN}$), accuracy ($acc = \frac{TP+TN}{TP+FP+FN+TN}$), precision ($prec = \frac{TP}{TP+FP}$), the F1-score ($F_1 = \frac{2\times prec\times sen}{prec+sen}$) and Matthews correlation coefficient (MCC), defined as:

$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$. Also, the confusion matrices (TP, TN, FP, FN) were calculated to assess the effectiveness of the classifier by giving the same weighting to each of the four groups [48–50].

5. Results

The proposed approach was firstly evaluated with respect to the features capabilities to describe each motion mode, to find the best configuration to proceed with the further multimodal analysis. Also, two different versions of multimodal information fusion were evaluated. The next subsections summarize the reported results in each of the considered evaluation phases.

5.1. Feature evaluation

In this work a frame-level representation from kinematic (velocity and/or acceleration, computed from a dense optical flow) and/or deep (using the first layer output from two different nets: VGG16 and/or MobileNetV2) features was considered. These features were evaluated independently over each motion mode to select the best configuration based on their capability in this task. Table 1 summarizes the individual performances of the different features.

The best PD prediction capability of deep features (DFs) has been reported for gait sequences, which may be associated with recovering particular postures during locomotion. The features extracted from the fourth layer of MobileNet resulted in the best features for the two motion modes, achieving an average accuracy of 0.96 ± 0.19 and 0.84 ± 0.36, for gait and eye fixation, respectively.

Regarding kinematic patterns, very compact covariance descriptors of 36 scalar values were obtained by integrating the velocity and acceleration patterns. Table 1 also illustrates the performance of such kinematic patterns in two different versions, using only velocities, and accelerations. For kinematics, the complete descriptor results in the best representation option in both modes, while remaining extremely compact in size.

Hence, from this study, the DFs computed from the fourth layer of MobileNet (DF-Mobilenet/4th) and the complete kinematic descriptor (KF-vel-acel) were selected. A more exhaustive evaluation was then carried out for these two configurations. The results are reported in Table 2. DFs achieved a remarkable performance for the different metrics considered. Regarding fixational eye patterns, the kinematic features KFs had higher MCC and accuracy scores, but DFs had a major specificity.

Table 3 displays the confusion matrices for each modality. In ocular fixation, the covariation of DFs presents a classification error of 23.1% in Parkinsonian patients and 7.7% in control subjects. However, the covariation of the kinematic characteristics of eye movements reduces the classification error of patients to zero but doubles the classification error of control subjects. This fact may be associated to the capability of kinematic patterns to recover tiny microtremor in Parkinsonian patients but introducing artifacts in control subjects. The complementarity of the two types of feature is also observed in the gait modality, so that the DFs classified with zero error in the control subjects, but with 7.7% error in the PD patients.

Similarly, the KFs reduce the error in the patient classification to zero but increase the classification error in 15.4% of control subjects. In fact, the kinematic locomotion of control subject results highly variable and therefore such representation may be unable to cover the whole spectrum of possible movements. In contrast, the Parkinsonian patients have locomotion signatures that can be properly recovered from kinematic descriptors but with increasing variability of postural configurations, representing a limitation for a deep feature representation.

To better illustrate the discriminatory behavior of the motion descriptors constructed, a low-dimensional space was built from a projection of resultant covariance matrices using KFs ans DFs features. The first three components were plotted using principal component Analysis (PCA). Figure 3 illustrates the resultant projections in a 3D geometrical space for descriptors that correspond to eye, gait, and the fusion of both modalities. As expected, the low-dimensional representation of such geometrical means is useful for analyzing the grouping of points labeled with the same class. It appears that a discrimination rule can be easily implemented for the three descriptors.

5.2. Early fusion classification

A first multimodal motion integration was achieved by concatenating (early fusion) the gait covariance descriptor and the eye fixation covariance descriptors (one for each eye). This fusion from mean covariance matrices allows a straightforward integration of the main features that characterize Parkinsonian patterns during the

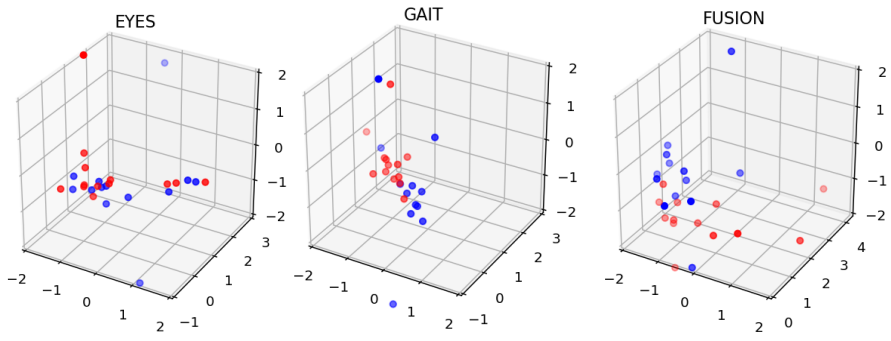

Figure 3: Projection over the three principal components of sample descriptors, for eyes, gait and fusion modalities. Red and blue points represent Parkinson and Control patients, respectively.

sequence. In this study, gait and eye integration was validated using KFs and DFs. Also, it was considered a descriptor that integrates both features (KF-DF). The KF covariances have a dimension of 6×6 (KF). The KF-DF covariances have a dimension of 38×38 (KF-DF), for each video descriptor of gait, eyes, and fusion.

Table 4 summarizes the scores achieved from this early integration using different sets of features. For all the studied subjects, the proposed approach achieved a perfect score by using a mixed representation of kinematic and deep features. The postural representation obtained from DFs together with the kinematic description was sufficient to distinguish PD and control patients.

Furthermore, covariance coding is a highly interpretable descriptor that can be used to recover salient information for each motion modality. Finally, the independent use of deep or kinematic features already achieves a good performance.

A more detailed analysis was carried out by computing the con-

fusion matrices for the different kinds of fusion (see in table 6). Regarding independent sets of features, the deep features achieve a better integration with only one false negative. For KFs, two PD subjects were incorrectly classified as controls. This behavior could be associated with patients reporting small changes in gait because of the early stage of the disease.

A feature importance analysis was conducted to measure the contribution of each feature in the descriptor, ordered with respect to the impurity reduction at each split during the Random Forest training. In this experiment a full descriptor that included KFs and DFs was considered. The total descriptor dimension corresponds to the three covariance matrices (one for gait and two for eyes), i.e. $3 \times 38 \times 38 = 4\,332$. In Figure 4, the classification performance of the proposed descriptor was illustrated by selecting an incremental set of the most important features, according to the ranking performed by the Random Forest classifier. In such cases, using only 20% of the most important features, the proposed descriptor achieved an average score of 65%. From the observed results, we can hypothesize that each feature contributes approximately equally to the final prediction.

5.3. Late fusion classification

The second multimodal integration proposed corresponds to building a discrete space classification for each modality independently, and then to fusing the probabilities obtained in each space. Each discrete space is obtained from a Random Forest, and each predicted sequence is projected onto the respective modality to obtain a cor-

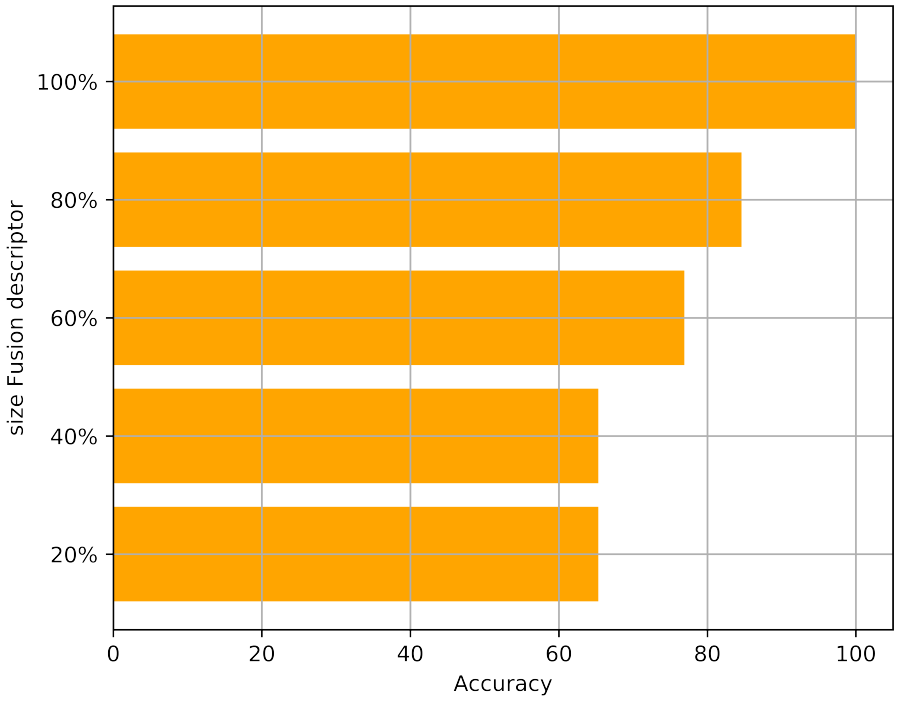

Figure 4: Feature importance analysis using Random forest with the general descriptor that includes kinematic and deep features. In the experiment was defined sets with the percentage of the most important features.

responding probability of PD. The final probability P_f is the linear weighting of the ocular fixation probability P_e and gait probability P_g, as: $P_f = wP_g + (1-w)P_e$.

In the parameter study, the mode weight parameter w was varied from 0.1 to 0.9, and the best score was obtained by setting w to 0.4, which means 40% for the gait and 60% for the eye fixation modalities. This experiment is summarized in Figure 5 which plots the distributions of probability prediction for PD patients, for different values of w. For $w = 0.4$, the prediction for patients diagnosed with

the disease has remarkable confidence and the outliers from the distribution reveal a probability prediction higher than 0.5. In contrast, $w < 0.4$ induces a significant variability in the probability prediction, with at least one example reported as false-negative (outlier point lower than 0.5). On the other hand, $w > 0.4$ increases the median of the predicted probability, which highlights the eye fixation contribution, but with false-labeled samples. Two outliers (that still obtain a positive prediction) with $w = 0.4$ correspond to patients in an early stage of the disease. This analysis highlighted the potential of eye fixational patterns as early PD biomarkers, while postural gait motion acts as a complementary cue, and can strengthen the disease analysis and quantification.

Table 5 summarizes the results of the proposed late fusion using different configurations of frame-level features to compute the Riemannian mean. This multimodal integration also results successful to discriminate between PD and control patients. In addition, the configuration of the classifier from a rich representation of both features is more effective than a single feature type to achieve perfect scores. A more detailed analysis can be obtained from the confusion matrices (see Table 6). The use of only DFs shows comparable results with respect to early fusion. Nevertheless, late fusion obtains two false positives, *i.e.,* control patients that were classified as PD patients. Finally, a comprehensive experiment was carried out to analyze the probability outputs from the random forest classifier in early and late fusion for each patient. Figure 6 displays the

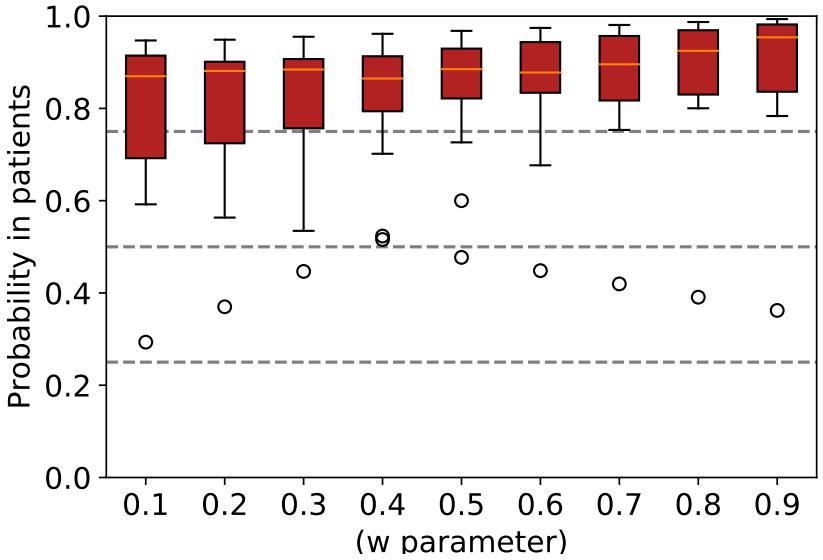

Figure 5: Distribution of probability predictions for Parkinson patients, for different w values. For $w = 0.4$, the outliers get a probability prediction higher than 0.5, achieving a proper classification. In contrast, all other values of w induce at least one false-negative.

PD probability, as outputted by the Random Forest classifier for each patient in the two fusion schemes. Control patients remain on probabilities lower than 0.3, which is a fairly confident index for binary classification. In addition, for PD patients, the probabilities are generally close to one. Interestingly, the three patients (p4, p5, and p6), in the early stage of the disease (second stage, according to the annotation of an expert following the H&Y scale), had gait locomotion patterns very similar to control subjects of the same age. Typically, patient p5 (second stage) had a lower probability than the patient p1 that had been categorized as third stage, according to the H&Y scale. For such patients, the oculomotor description appears

to be the most discriminant with respect to the PD. For the other patients, the gait modality offers a greater contribution to the final probability because the gait patterns show more pronounced impairments than in the early stages. However, these patients had a higher PD probability in the early fusion scheme, which could indicate that early covariance integration leads to a better representation of the disease and could be more effective.

6. Discussion

A novel approach was introduced for fusing motion modalities captured in markerless video sequences. In this study, a population of 26 patients, distributed as Parkinson (13 patients) and Control (13 patients), was considered. For each patient multiple sequences were recorded during gait locomotion from the sagittal view, without any invasive device. In addition, eye fixation patterns were recorded with a standard camera, and with weakly controlled conditions. Each frame of the sequence was first represented with DFs and/or KFs, computed from pre-trained convolutional networks and a dense optical flow, respectively. In this work, the DFs were computed from conventional (VGG) and separable architectures (MobileNetV2). Regarding KF, the horizontal and vertical components of the unit tangential vector and unit normal vector velocity were considered, as well as the magnitude of tangential and normal acceleration. Then, frame-level covariance coding was carried out to represent instantaneous posture and kinematics, which were thereafter summarized in a Riemannian temporal mean covariance. In

multiple experiments, the proposed approach showed remarkable results, in configurations for early fusion (average accuracy of 100%) and late fusion (average accuracy of 100%). In the best configurations, the proposed approach combines KFs and DFs to achieve a more robust representation at the frame level. In addition, eye fixation had a high discrimination power and was therefore weighted with more importance in multimodal fusion.

These mean covariance matrices are very compact, ranging from 6 to 38 features. They are used as motion descriptors that can be fused (early or late) through a random forest classifier. A main limitation on covariance representation occurs when the prediction is based on only KFs or DFs, for both types of fusion. In such case, the kinematic information properly models Parkinsonian patterns in both modes but results insufficient to cover the whole variability for control subjects. In contrast, the proposed approach obtained excellent results by integrating both types of KFs and DFs in a total population of 26 subjects recorded three times, including 13 patients diagnosed with PD. It turns out that complementary features can deal with proper modelling of disease patterns, but also covering the control population. According to the random forest probabilistic results, early fusion was the best option, achieving an accuracy of 100%, using 38 features. Thus, the combination of both types of features significantly improves the differentiation between the motor patterns of control subjects and patients with PD.

Nowadays, the quantification of patients strongly depends on medical expertise based only on coarse motor scales and reported

indices. These scales only consider strong motion changes, which limits the sensitivity to monitor the progression of the disease or to make an early diagnosis, and often produces high variance in final scores associated with a particular patient [5]. To overcome this issue, an approach is proposed to better monitor the disease, taking advantage of the markerless quantification of known patterns such as gait, but also integrating new biomarkers of the disease, such as tremor from the eyes. The integration of these motions results naturally from covariance frames, and markerless capture may offer potential applications in other types of non-controlled scenario. As demonstrated by the results, the combination of DFs and KFs make it possible to distinguish between control and PD patterns, with extremely small size descriptors (between 108 and 4 332), making real-time recognition possible, which is promising for clinical scenarios.

Multimodal approaches have been previously reported to analyze PD better: for instance, integration of walking patterns, speech signal analysis, and controlled writing experiments [32–34]. Nonetheless, these methodologies depend on sophisticated capture devices and tedious experiments, making their use in routine clinical practice difficult. More recently, markerless strategies based on deep learning representations from video-sequences have emerged [20]. In the latter work, an average accuracy of 90% was reported for the task of classifying PD with respect to the control population. These strategies have characterized gait patterns from end-to-end learning representations, but remain dependent on a huge and balanced training set to compute spatio-temporal patterns that discriminate

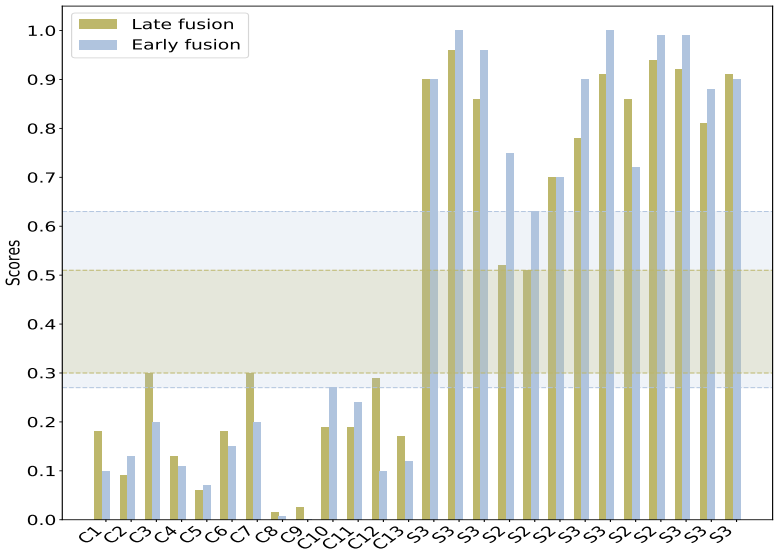

Figure 6: Probabilities of Parkinson with 13 control subjects (C) and 13 patients (P) considering kinematics and deep features in the two fusion's types.The wider horizontal blue stripe shows a higher confidence range for early fusion than for late fusion (horizontal green stripe).

patients with PD from the control population. Similarly, Lin *et. al* proposed a microwave motion detector to characterize tremor patterns using a non-invasive device, but it required complex calibration processes and protocols to capture motion signs [9]. Other studies have reported the sensitivity of the disease associated with eye motion patterns. In these studies for instance, in a population of 112 patients and only two controls (two from 60 control subjects), an ocular tremor with an average fundamental frequency of 5.7 Hz and an average magnitude of 0.27 in the horizontal plane

and 0.33 in the vertical plane has been found. This shows the potential characteristics of PD biomarkers related to eye patterns [22]. These fundamentals have been explored to propose video strategies that recover and learn eye fixation patterns, making possible the representation of disease in weakly controlled scenarios [31]. In this case an approach proposed in previous work [31] achieved an average accuracy of 95%, in a population with 13 control subjects and 13 patients. Despite the remarkable results, this approach is limited by focusing on eye analysis and, losing other signs that may complement disease characterization.

Our work is based on a simple video system with a standard camera that provides objective information to the specialist in supporting the diagnosis and treatment of the disease. The proposed approach proved effective in classification tasks by achieving perfect classification scores in multimodal configurations. In addition, it is robust in assigning categorical probabilities to positive classified patients. Furthermore, many of the classic patterns are only detectable in advanced stages of the disease, restricting the analysis to the advanced Parkinson population. In contrast, the proposed approach considers gait patterns but also uses new eye fixational patterns, which have recently shown a major sensitivity to very early stages of the disease.

In summary, the proposed method integrates a local and global representation from KFs and DFs, which are effectively combined into covariance matrices to form a special Riemmanian manifold for each video sequence. The Riemannian mean from the manifold is

easily integrated among different sequences and motion modes. The validation reported in this work should be extended in future studies to evaluate the ability of the proposed approach to distinguish different levels of the disease. In addition, patient diagnosis should be carried out on more sophisticated scales, such as the UPDRS-ME, to better correlate symptoms, to address the inter-experts variability, and to define a relevant inclusion of this tool within a clinical routine.

Despite the remarkable results of deep learning networks, these representations remain dependent on a large set of data to address the variability of samples, with a stratified condition among classes of the problem. In the biomedical context, meeting such requirements is difficult and leads to unnatural implementations to support routine treatments. For instance, the deep representation proposed by Guayacan et. al [20] achieved PD discrimination through an end-to-end learning scheme from 3D video analysis. However, this approach has reported average accuracy around 90%. These results show some limitations in operating with spatiotemporal maps, which may be associated with insufficient data for training. In addition, this type of approach has restrictions to include additional modalities. In contrast, recent advances in representation of deep learning strategies are exploited using DFs that can generalize the representation of input images without requiring any additional training. The proposed scheme has the advantage of being robust in PD discrimination, but also very compact (video descriptors with a size between 108 and 4 332 scalar values). In addition, the proposed approach can integrate several motion modalities without any change

in the pipeline of the method, which may benefit to a broader analysis from the clinical domain.

The proposed approach was validated through a limited study of 26 subjects, due to difficulties to acquire data from more patients. The main issue around data is the quantification of populations with comparable demography characteristics that allow to address methodologies related with the capability to discriminate disease patterns. Hence, as perspectives is proposed a further validation with larger datasets and stratified patients according to the progression of the disease. This validation may be useful to discover new multimodal patterns that enhance the sensitivity of clinical scales like the UPDRS and may better measure disease progression or the effectiveness of a particular treatment. Also, the inclusion of new modalities may enrich disease representation and impact as a tool to support early diagnosis.

Acknowledgments

The authors acknowledge the Vicerrectoriá de Investigación y Extensión (VIE) of the Universidad Industrial de Santander for supporting this research work by the project: *Cuantificación de patrones locomotores para el diagnóstico y seguimiento remoto en zonas de dificil acceso*, with SIVIE code 2697.

References

[1] V. L. Feigin, E. Nichols, e. a. Alam, Tahiya, Global, regional, and national burden of neurological disorders, 1990–2016:

a systematic analysis for the Global Burden of Disease Study 2016, The Lancet Neurology 18 (5) (2019) 459–480. doi:10.1016/S1474-4422(18)30499-X.
URL http://www.sciencedirect.com/science/article/pii/S147444221830499X

[2] Q. Bai, T. Shen, B. Xu, Q. Yu, H. Zhang, C. Mao, C. Liu, S. Wang, Quantification of the motor symptoms of parkinson's disease, in: 2017 8th International IEEE/EMBS Conference on Neural Engineering (NER), 2017, pp. 82–85. doi:10.1109/NER.2017.8008297.

[3] J. S. Perlmutter, Assessment of parkinson disease manifestations, Current protocols in neuroscience Chapter 10 (2009) Unit10.1. doi:10.1002/0471142301.ns1001s49.
URL https://europepmc.org/articles/PMC2897716

[4] M. Hoehn, M. Yahr, Parkinsonism: onset, progression, and mortality. 1967, Neurology 57 (10 Suppl 3) (2001) S11—26.
URL http://intl.neurology.org/cgi/content/full/57/10_suppl_3/S11

[5] B. Post, M. P. Merkus, R. M. A. de Bie, R. J. de Haan, J. D. Speelman, Unified parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?, Movement disorders : official journal of the Movement Disorder Society 20 (12) (2005) 1577—1584. doi:10.1002/mds.20640.
URL https://doi.org/10.1002/mds.20640

[6] Metric properties of nurses' ratings of parkinsonian signs with a modified Unified Parkinson's Disease Rating Scale, Neurology 49 (6) (1997) 1580–1587. arXiv:https://n.neurology.org/content/49/6/1580.full.pdf, doi:10.1212/WNL.49.6.1580.
URL https://n.neurology.org/content/49/6/1580

[7] A. W. Michell, S. J. G. Lewis, T. Foltynie, R. A. Barker, Biomarkers and Parkinson's disease, Brain 127 (8) (2004) 1693–1705. arXiv:https://academic.oup.com/brain/article-pdf/127/8/1693/842701/awh198.pdf, doi:10.1093/brain/awh198.
URL https://doi.org/10.1093/brain/awh198

[8] S. Aghanavesi, J. Westin, F. Bergquist, D. Nyholm, H. Askmark, S. M. Aquilonius, R. Constantinescu, A. Medvedev, J. Spira, F. Ohlsson, et al., A multiple motion sensors index for motor state quantification in parkinson's disease, Computer methods and programs in biomedicine 189 (2020) 105309.

[9] C.-H. Lin, J.-X. Wu, J.-C. Hsu, P.-Y. Chen, N.-S. Pai, H.-Y. Lai, Tremor class scaling for parkinson disease patients using an array x-band microwave doppler based upper limb movement quantizer, IEEE Sensors Journal.

[10] Jones Rachel, Biomarkers: casting the net wide, Nature 466 (7310) (2010) S11–S12. doi:https://doi.org/10.1038/466S11a.

[11] Anderson Tim J., MacAskill Michael R., Eye movements in patients with neurodegenerative disorders, Nature Reviews Neurology 9 (2) (2013) 74–85. doi:https://doi.org/10.1038/nrneurol.2012.273.

[12] M. S. Ekker, S. Janssen, K. Seppi, W. Poewe, N. M. de Vries, T. Theelen, J. Nonnekes, B. R. Bloem, Ocular and visual disorders in Parkinson's disease: Common but frequently overlooked, Parkinsonism & Related Disorders 40 (2017) 1–10. doi:10.1016/j.parkreldis.2017.02.014.
URL http://www.sciencedirect.com/science/article/pii/S1353802017300640

[13] A. Larrazabal, C. García Cena, C. Martínez, Video-oculography eye tracking towards clinical applications: A review, Computers in Biology and Medicine 108 (2019) 57–66. doi:10.1016/j.compbiomed.2019.03.025.
URL http://www.sciencedirect.com/science/article/pii/S0010482519301040

[14] A. Mirelman, P. Bonato, R. Camicioli, T. D. Ellis, N. Giladi, J. L. Hamilton, C. J. Hass, J. M. Hausdorff, E. Pelosin, Q. J. Almeida, Gait impairments in Parkinson's disease, The Lancet Neurology 18 (7) (2019) 697–708. doi:10.1016/S1474-4422(19)30044-4.
URL http://www.sciencedirect.com/science/article/pii/S1474442219300444

[15] G. T. Gitchel, P. A. Wetzel, A. Qutubuddin, M. S. Baron,

Experimental support that ocular tremor in Parkinson's disease does not originate from head movement, Parkinsonism & Related Disorders 20 (7) (2014) 743–747. doi:10.1016/j.parkreldis.2014.03.028.
URL http://www.sciencedirect.com/science/article/pii/S1353802014001400

[16] E. Mirek, J. L. Kubica, J. Szymura, S. Pasiut, M. Rudzińska, W. Chwała, Assessment of Gait Therapy Effectiveness in Patients with Parkinson's Disease on the Basis of Three-Dimensional Movement Analysis, Frontiers in Neurology 7 (2016) 102. doi:10.3389/fneur.2016.00102.
URL https://www.frontiersin.org/article/10.3389/fneur.2016.00102

[17] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, V. Venkatraman, Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease, Future Generation Computer Systems 83 (2018) 366–373. doi:10.1016/j.future.2018.02.009.
URL http://www.sciencedirect.com/science/article/pii/S0167739X1731631X

[18] E. Rastegari, S. Azizian, H. Ali, Machine learning and similarity network approaches to support automatic classification of parkinson's diseases using accelerometer-based gait analysis, in: Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

[19] A. R. Anwary, H. Yu, M. Vassallo, An automatic gait feature extraction method for identifying gait asymmetry using wearable sensors, Sensors 18 (2) (2018) 676.

[20] L. C. Guayacán, E. Rangel, F. Martínez, Towards understanding spatio-temporal parkinsonian patterns from salient regions of a 3d convolutional network, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 3688–3691.

[21] A. W. Przybyszewski, M. Kon, S. Szlufik, A. Szymanski, P. Habela, D. M. Koziorowski, Multimodal learning and intelligent prediction of symptom development in individual parkinson's patients, Sensors 16 (9) (2016) 1498.

[22] G. T. Gitchel, P. A. Wetzel, M. S. Baron, Pervasive ocular tremor in patients with parkinson disease, Archives of neurology 69 (8) (2012) 1011–1017.

[23] O. Rascol, M. Clanet, J.-L. Montastruc, M. Simonetta, M. Soulier-Esteve, B. Doyon, A. Rascol, Abnormal ocular movements in parkinson's disease: evidence for involvement of dopaminergic systems, Brain 112 (5) (1989) 1193–1214.

[24] M. Vidailhet, S. Rivaud, N. Gouider-Khouja, B. Pillon, A.-M. Bonnet, B. Gaymard, Y. Agid, C. Pierrot-Deseilligny, Eye movements in parkinsonian syndromes, Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 35 (4) (1994) 420–426.

[25] A. H. Clarke, Laboratory testing of the vestibular system, Current opinion in otolaryngology & head and neck surgery 18 (5) (2010) 425–430.

[26] A. Khosla, D. Kim, Optical Imaging Devices: New Technologies and Applications, CRC Press, 2017.

[27] J. Hanuška, C. Bonnet, J. Rusz, T. Sieger, R. Jech, S. Rivaud-Péchoux, M. Vidailhet, B. Gaymard, E. Růžička, Fast vergence eye movements are disrupted in parkinson's disease: a video-oculography study, Parkinsonism & Related Disorders 21 (7) (2015) 797–799.

[28] J. Naruniec, M. Wieczorek, S. Szlufik, D. Koziorowski, M. Tomaszewski, M. Kowalski, A. Przybyszewski, Webcam-based system for video-oculography, IET Computer Vision 11 (2) (2016) 173–180.

[29] S. Adhikari, D. E. Stark, Video-based eye tracking for neuropsychiatric assessment, Annals of the New York Academy of Sciences 1387 (1) (2017) 145–152.

[30] T. B. Carson, S. Z. Sutton, Application for smart phone or related devices for use in assessment of vestibulo-ocular reflex function, uS Patent App. 15/569,472 (Oct. 18 2018).

[31] I. Salazar, S. Pertuz, W. Contreras, F. Martínez, A convolutional oculomotor representation to model parkinsonian fixational patterns from magnified videos, Pattern Analysis and Applications 24 (2) (2021) 445–457.

[32] H. N. Pham, T. T. T. Do, K. Y. Jie Chan, G. Sen, A. Y. K. Han, P. Lim, T. S. Loon Cheng, Q. H. Nguyen, B. P. Nguyen, M. C. H. Chua, Multimodal detection of parkinson disease based on vocal and improved spiral test, in: 2019 International Conference on System Science and Engineering (ICSSE), 2019, pp. 279–284. doi:10.1109/ICSSE.2019.8823309.

[33] J. C. Vasquez-Correa, T. Bocklet, J. R. Orozco-Arroyave, E. Nöth, Comparison of user models based on gmm-ubm and i-vectors for speech, handwriting, and gait assessment of parkinson's disease patients, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6544–6548. doi:10.1109/ICASSP40776.2020.9054348.

[34] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, E. Nöth, Multimodal assessment of parkinson's disease: A deep learning approach, IEEE Journal of Biomedical and Health Informatics 23 (4) (2019) 1618–1630. doi:10.1109/JBHI.2018.2866873.

[35] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Scandinavian conference on Image analysis, Springer, 2003, pp. 363–370.

[36] A. Saleh, M. A. Garcia, F. Akram, M. Abdel-Nasser, D. Puig, Exploiting the kinematic of the trajectories of the local descriptors to improve human action recognition., in: VISIGRAPP (3: VISAPP), 2016, pp. 182–187.

[37] S. Okuda, S. Takano, M. Ueno, Y. Hara, Y. Chida, T. Ikkaku, F. Kanda, T. Toda, Gait analysis of patients with parkinson's disease using a portable triaxial accelerometer, Neurology and Clinical Neuroscience 4 (3) (2016) 93–97.

[38] M. Rucci, M. Poletti, Control and functions of fixational eye movements, Annual Review of Vision Science 1 (2015) 499–518.

[39] J. Otero-Millan, R. Schneider, R. J. Leigh, S. L. Macknik, S. Martinez-Conde, Saccades during attempted fixation in parkinsonian disorders and recessive ataxia: from microsaccades to square-wave jerks, PLoS One 8 (3) (2013) e58535.

[40] S. Hijazi, R. Kumar, C. Rowen, Using convolutional neural networks for image recognition, Cadence Design Systems Inc.: San Jose, CA, USA (2015) 1–12.

[41] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861.

[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[43] L. Sifre, S. Mallat, Rigid-motion scattering for image classification, Ph. D. thesis.

[44] P. T. Fletcher, S. Joshi, Riemannian geometry for the statistical analysis of diffusion tensor data, Signal Processing 87 (2) (2007) 250–262.

[45] J. Ali, R. Khan, N. Ahmad, I. Maqsood, Random forests and decision trees, International Journal of Computer Science Issues (IJCSI) 9 (5) (2012) 272.

[46] M. Z. Alam, M. S. Rahman, M. S. Rahman, A random forest based predictor for medical data classification using feature ranking, Informatics in Medicine Unlocked 15 (2019) 100180.

[47] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, Classification and regression trees, CRC press, 1984.

[48] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC genomics 21 (1) (2020) 6.

[49] J. Liu, B. Kantarci, C. Adams, Machine learning-driven intrusion detection for contiki-ng-based iot networks exposed to nsl-kdd dataset, in: Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning, 2020, pp. 25–30.

[50] V. Abromavičius, D. Plonis, D. Tarasevičius, A. Serackis, Two-stage monitoring of patients in intensive care unit for sepsis prediction using non-overfitted machine learning models, Electronics 9 (7) (2020) 1133.

Table 1: Comparison of the accuracy obtained by each feature alone, in each modality

Features	**Acc**		**Descriptor**
	Gait	**Eye**	**Size**
DF-VGG/1st	0.923	0.807	4096
DF-VGG/2nd	**0.961**	0.807	4096
DF-VGG/3rd	0.923	**0.846**	4096
DF-VGG/4th	0.923	**0.846**	16384
DF-MobileNetV2/2nd	0.923	0.769	1024
DF-MobileNetV2/3rd	0.923	0.807	1024
DF-MobileNetV2/4th	**0.961**	**0.846**	1024
DF-MobileNetV2/5th	0.884	0.769	1024
KF-vel	0.576	0.730	16
KF-vel-acel	**0.923**	**0.923**	36

Table 2: Scores in ocular fixation (Eye) and gait modality using only deep features (DF) or only kinematic features (KF)

	Eye-DF	**Eye-KF**	**Gait-DF**	**Gait-KF**
sen	0.769	**1**	0.923	**1**
spec	**0.926**	0.846	**1**	0.846
prec	**0.909**	0.866	**1**	0.866
acc	0.846	**0.923**	**0.961**	0.923
F1-s	0.824	**0.928**	**0.959**	0.928
MCC	0.700	**0.856**	**0.925**	0.856

Table 3: Confusion matrices per modality using only kinematic (KF) or deep features (DF), for Parkinson (PK) and Control (C) subjects.

	Eye-DF		**Eye-KF**		**Gait-DF**		**Gait-KF**	
	PK	**C**	**PK**	**C**	**PK**	**C**	**PK**	**C**
PK	10(76.9%)	3(23.1%)	13(100%)	0	12(92.3%)	1(7.7%)	13(100%)	0
C	1(7.7%)	12(92.3%)	2(15.4%)	11(84.6%)	0	13(100%)	2(15.4%)	11(84.6%)

Table 4: Early fusion scores for the two modalities, using kinematic (KF), deep (DF) or joint features (KF-DF)

	Early Fusion KF-DF	**Early Fusion DF**	**Early Fusion KF**
sen	1	0.923	0.846
spec	1	1	1
prec	1	1	1
acc	1	0.961	0.923
F1-s	1	0.959	0.916
MCC	1	0.925	0.856

Table 5: Late fusion scores for the two modalities, using kinematic (KF), deep (DF) or joint features (KF-DF)

	Late Fusion KF-DF	**Late Fusion DF**	**Late Fusion KF**
sen	1	0.923	0.846
spec	1	1	0.923
prec	1	1	0.916
acc	1	0.961	0.884
F1-s	1	0.959	0.879
MCC	1	0.925	0.771

Table 6: Confusion matrices for the different fusion modes, using kinematic (KF), deep (DF), or joint (KF-DF) features, for Parkinson (PK) and Control (C) subjects.

	Early Fusion (KF-DF)		**Early Fusion (DF)**		**Early Fusion (KF)**		**Late Fusion (KF-DF)**		**Late Fusion (DF)**		**Late Fusion (KF)**	
	PK	**C**	**PK**	**C**	**PK**	**C**	**PK**	**C**	**PK**	**C**	**PK**	**C**
PK	13(100%)	0	12(92.3%)	1(7.7%)	11(84.6%)	2(15.4%)	13(100%)	0	12(92.3%)	1(7.7%)	11(84.6%)	2(15.4%)
C	0	13(100%)	0	13(100%)	0	13(100%)	0	13(100%)	0	13(100%)	1(7.7%)	12(92.3%)