Classification and Visualization Tool for Gait Analysis of Parkinson’s Disease

U Kit Pun, Huanying Gu, Ziqian Dong and N. Sertac Artan
New York Institute of Technology, 1855 Broadway, New York, NY, 10023 USA
Email: upun@nyit.edu, hgu03@nyit.edu, ziqian.dong@nyit.edu, nartan@nyit.edu

Abstract—There is no standard diagnostic test for Parkinson’s Disease. In this paper, we propose a data mining based statistical diagnostic method towards a standard and accurate diagnostic test. The result shows the proposed regression formula, which only requires patient’s stride data, can provide a 90% accuracy to diagnose PD patients.

Keywords: classification, visualization, Parkinson’s Disease

I. INTRODUCTION

Parkinson’s disease (PD) is a neurodegenerative disease which often affects patients’ movement. The exact cause of PD is still not fully understood. There is no standard diagnostic test for it and the diagnostic requires the expertise of a specialist. Currently, PD is diagnosed via various neurological examinations by specialists [1].

The most common symptoms of PD are tremor, gait disturbance, stiffness, and slowness. These symptoms together lead to the alteration of the patient’s gait [2]. Gait parameters have been used to differentiate PD and healthy subjects. A stride refers to the distance a subject moves in between two consecutive times the same heel leaves the floor. Every stride consists of a stance (heel in floor), and a swing (heel in motion). The stride actions of left and right heel occur simultaneously with an overlap [3]. Statistical analysis has been applied to study gait variability in patients with PD that uses stride interval parameters to form feature vector in classification approach [9].

In this paper, we analyze the gait data using classification techniques and provide a visualization tool to assist in the diagnosis of PD. The paper is organized as follows. Section II presents our classification and visualization methods. Section III presents results. Section IV presents discussions and future work. Section V concludes the paper.

II. METHODS

Figure 1 presents the flow chart of the classification approach to classify the severity of PD based on gait features. Our approach consists of the following four consecutive steps.

I. Data selection: The available databases are evaluated and the appropriate one is selected for our study.

II. Features selection: Data is cleaned and useful features are selected.

III. Visualization: Relevant information is identified based on the selected features.

IV. Formula integration: Results are integrated to form a formula, and this formula is evaluated.

Figure 1. The flow chart of the classification approach to classify the severity of PD based on gait features.

III. RESULT

A. Data Selection

This study does not include experimental procedures involving human subjects. We use the Hausdorff’s gait dynamics database to evaluate our proposed method [4]. This database contains stride records from 15 PD patients (44–80 yrs old) and 16 healthy control subjects (20–74 yrs old). The patients are instructed to walk for 5 minutes in a 77 m long hallway, which is equipped with force sensors. The data includes elapsed time (sec), left / right stride interval (sec), left / right swing interval (sec, % of stride), left / right stance interval (sec, % of stride), and double support interval (sec, % of stride). Outliers with more than 3 standard deviations (SDs) difference are removed to eliminate the data from the turns at the hallway’s ends. Additional information for each subject such as age and average gait speed [5, 6] is provided along with the Hoehn and Yahr scale for the PD group. In this scale, PD patients are given a score from 1 to 5 to indicate their stage of disease. A higher score corresponds to more advanced stages of the disease [2]. We use this score as the ground truth in this study.

In addition to the features in the database, we derive three features that we identify and calculate for this study as summarized in Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average stride interval (sec)</td>
<td>This is the average of left and right stride (left stride interval + right stride interval) / 2</td>
</tr>
<tr>
<td>Swing ratio difference (%)</td>
<td>This indicates the unbalance of left and right (left swing interval - right swing interval)</td>
</tr>
<tr>
<td>Stride length (m)</td>
<td>This is the length a subject can move in one-stride time, but not the length of one step. (Average stride interval * gait speed)</td>
</tr>
</tbody>
</table>

TABLE 1. ADDITIONAL FEATURES.
B. Features Selection

In order to extract useful data, we calculate the standard deviation (SD), the first, the second, and the third quartiles (Q₁, Q₂, Q₃) for each feature listed above except elapsed time (sec), since it is an ongoing value. We have a total of 60 feature values for each subject.

To identify the feature values which are most closely correlated with PD, the information gain (IG) for each feature is computed since IG can infer which feature is most useful for discriminating classes [7]; namely control subjects and PD patients. The formula for information gain (IG) calculation are given in (1) and (2).

\[ IG(T, a) = H(T) - \sum_{\forall v \in vals(a)} \frac{|x \in T|\{x_a = v\}}{|T|} \times H(\{x \in T|\{x_a = v\}) \]  

\[ H(T) = -\sum_{t=1}^{31} p(t \_a) \log_2 p(t \_a) \]  

Here, \( H \) indicates Shannon Entropy, \( P \) indicates probability, \( T \) indicates training set and \( x_a \in vals(a) \) indicates the value of the \( d \)th class label in the training set. As an example, the calculation of IG of stride length \( Q_2 \) is given below starting with the \( H(T) \) calculation. Since there are a total of 31 subjects (16 control subjects and 15 PD patients), \( p(t_1) = \frac{16}{31} \) and \( p(t_2) = \frac{15}{31} \), and then entropy can be given as,

\[ H(T) = - \left( \frac{16}{31} \log_2 \frac{16}{31} + \frac{15}{31} \log_2 \frac{15}{31} \right) = 0.9992. \]

Identified features are used to classify the subjects into control and PD groups with the sizes 16 and 15, respectively. We use stride length \( Q_2 \) to differentiate the subjects, which are sorted in descending order using the feature value. The first 16 subjects with greater values are classified into the first group, and the next 15 subjects belong to the second group. The control-PD ratio are (14 control: 2 PD) and (2 control: 13 PD), respectively. Therefore, we use 

\[ \sum_{\forall v \in vals(a)} \frac{|x \in T|\{x_a = v\}}{|T|} \times H(\{x \in T|\{x_a = v\}) \]

\[ = \frac{| \text{control} |}{|T|} \times H(\text{control}) + \frac{| \text{PD} |}{|T|} \times H(\text{PD}) \]

\[ = \frac{16}{31} \times \left[ - \left( \frac{14}{16} \log_2 \frac{14}{16} + \frac{2}{16} \log_2 \frac{2}{16} \right) \right] \]

\[ + \frac{15}{31} \times \left[ - \left( \frac{2}{15} \log_2 \frac{2}{15} + \frac{13}{15} \log_2 \frac{13}{15} \right) \right] \]

Then, the IG of stride length \( Q_2 \) is,

\[ 0.9992 - 0.5547 = 0.4445. \]

Table 2 lists the information gain for some features. Since our goal is to select 3 features with the highest IGs, we eliminate repeated selections from the same group. For example, double support interval \( Q_3 \) (%) and double support interval SD (%) both have high IGs. In this case, we only select the one with higher IG since they are generated from the same data set. The three features with largest IG are listed below and they are used for visualization.

1) Double support interval SD (%) — 0.6542
2) Stride length \( Q_2 \) (m) — 0.4445
3) Swing ratio difference SD (%) — 0.2906

C. Visualization

1 Visualization

The \( Q_1 \), \( Q_2 \), and \( Q_3 \) of the above features are used to generate their quartiles graphs (Figures 2). From this graph, control subjects overall have relatively greater values than PD patients in stride length (m), which means the control subjects can move farther in one-stride time. In double support interval (%) and swing ratio difference (%), Control subjects’ values are more concentrated, and the patients’ values are more dispersed, which may indicate that patients’ strides were more unstable and more unbalanced due to PD.

![Figure 2. Quartiles graphs for stride length, double support interval, and swing ratio difference for the control group and the PD patients. The numbers indicate the subject ID in each group.](image)

### Table 2. Result of IG for Some Features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Information Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double support interval (%) SD</td>
<td>0.6542</td>
</tr>
<tr>
<td>Double support interval (%) Q₁</td>
<td>0.0924</td>
</tr>
<tr>
<td>Double support interval (%) Q₂</td>
<td>0.1757</td>
</tr>
<tr>
<td>Double support interval (%) Q₃</td>
<td>0.2906</td>
</tr>
<tr>
<td>Left swing interval (%) Q₁</td>
<td>0.1757</td>
</tr>
<tr>
<td>Right swing interval (%) Q₁</td>
<td>0.1757</td>
</tr>
<tr>
<td>Stride length (m) Q₂</td>
<td>0.4445</td>
</tr>
<tr>
<td>Swing ratio difference (%) SD</td>
<td>0.2906</td>
</tr>
<tr>
<td>Stride time (sec) Q₂</td>
<td>0.0368</td>
</tr>
</tbody>
</table>

II Visualization

In order to show how the selected features interact with each other, we use the three selected features to generate a three-dimensional (3D) scatter plot (Figure 3). In the graph, each rectangle is drawn by one subject’s double support interval SD (%) and swing ratio difference SD (%). As all the points with value greater than 1 SD are removed to eliminate outliers, subjects’ values in these two aspects would all fit into their corresponding rectangles. Therefore, these rectangles indicate the value dispersion of each subject.
Figure 3. 3D scatter plots (grey data points: control subjects, black data points: PD patients).

Figure 3 demonstrates that:

1) The stride length Q2 (m) has non-overlapping distribution, clearly separates control subjects group and PD patients group. Whereas, double support interval SD (%) and swing ratio difference SD (%) do not have such a clear diversity.

2) There is one control subject’s data mixed into PD patients group. While comparing with given information, that data belongs to the second-oldest control subjects, which may indicate age as a possible factor affecting subjects’ stride data.

3) The squares (dispersion) of control subjects are overall smaller than the PDs’, which indicates that control subjects’ strides are more stable and balanced than PD patients as the previous section suggested.

4) All values tend to go in one direction, from control subjects to PD patients, starting from the grey cube as the arrow points out.

III Visualization

In this step, our goal is to find a clear regular pattern of subjects’ data distribution. In this part, we use the 3 features value selected for generating the 3D graph. In Figure 4, each point represents a subject. The distribution tends to spread out from controls to PDs. In order to find the origin, we use linear regression, which gives the plane with minimum distance to all the points [8]. It is suggested that a starting point of the trending direction can be found in the plane. In our case, our 3D linear regression is (3), as shown in Figure 4.

\[ y = -1.6488x_1 + 1.3730x_2 + 1.8934 \]  

Here, y is double support interval SD (%) since it has the largest IG. x1 is stride length Q2 (m), and x2 is swing ratio difference SD (%). Figure 4 shows that control subjects overall have greater stride length, Q2 (m) values and smaller swing ratio difference, SD (%) values than PD patients. Therefore, we use the largest stride length, Q2 (m) value and the smallest swing ratio difference, SD (%) value along all subjects to insert as x1 and x2 in (3), to find the corresponding point in the regression plane, which is our trend starting point (grey cube in Figure 4). As discussed earlier, all subjects’ points tend to spread out from this point (along the black line in Figure 4),

from control subjects to patients. When comparing the clinical information with our result, we found:

1) Younger control subjects (13 subjects, 20–57 yrs old) are all located closer to trend starting point.

2) Advanced PD patients (9 patients, in stage 3 and above) are located in the farther distance from the trend starting point.

3) Control subjects with age 61 and elder are located very close to PD patients with stage 1.5. It is hard to differentiate, as the result of II Visualization suggested, age could be a factor affecting stride data.

Therefore, we divide them into 3 groups: young control group (13 control subjects, 20–57 yr), vague group (3 control subjects, 61–74 yr & 6 PD patients, stage 1.5–2.5) and advance PD group (9 PD patients, stage 3–4), as shown in Figure 5.

D. Formula Integration

Our goal is to use the previous result to generate a regression formula. The first version is simply developed by the three features we selected (section 3.B).

\[ P(x) = \frac{S(x) \times D(x)}{L(x)} \]  

(4)

Figure 5. 3D plots - groups (white data points: young control group, grey data points: vague group, black data points: advance PD group).
In (4), \( P \) indicates subject’s Parkinson’s disease severity value, \( S, D, L \) indicates subject’s swing ratio difference SD (%), double support interval SD (%), stride length Q \(_2\) (m), respectively. From section 3.C, we show that stride length Q \(_2\) (m) has better distribution than others. The proportion (\( \text{Prop} \)) of these 3 features are indicated as in (5). Also, the subjects’ dispersing squares, which are generated by double support interval SD (%) and swing ratio difference SD (%), has good distribution, but it is a squared value. To simplify, we adopt the proportion relationship as (6).

\[
\text{Prop}(L) > \text{Prop}(S) \approx \text{Prop}(D) \quad (5)
\]

\[
\text{Prop}(L) \approx \sqrt{\text{Prop}(S) \times \text{Prop}(D)} \quad (6)
\]

Given (5) and (6), we further modify our regression formula as (7).

\[
P(x) = \frac{\sqrt{S(x) \times D(x)}}{2L(x)} \quad (7)
\]

Table 3 shows the severity for all subjects. 8 out of 9 PD patients with 3 or more H.Y. score locate in the first 8 order, and 14 out of 16 control subjects locate in the last 14 order. Only two control subjects have higher \( P(x) \) values than PD patients. Further investigation indicates that these two subjects are the eldest two in the control subjects group. As suggested in section 3.C, subjects’ ages would affect our result.

Although \( P(x) \) value gives a good overall differentiation of PD patients and control subjects, it is hard to differentiate PD patients in stage 3—4 and stage 1.5—2.5 by using \( P(x) \) value. Therefore, we define the \( P(x) \) value range to have a better expression as shown in Table 4. We group different H.Y. scale into a new 3 level scale to present PD’s severity. When we use this \( P(x) \) severity to evaluate the result in Table 3 again, one PD patient (park12) was misplaced in a lower level, and the 2 eldest control subjects were misplaced in level 1 PD group. 28 out of 31 subjects had been placed correctly. Based on this training set, prediction accuracy is up to 90%.

**Table 3. Severity Results; C1—C16, P1—P15 indicated control1—control16, P1—the 9 patients; H.Y. indicates Hoehn and Yahr scale.**

<table>
<thead>
<tr>
<th>P1</th>
<th>P14</th>
<th>P7</th>
<th>P13</th>
<th>P8</th>
<th>P1</th>
<th>P4</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(x)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H.Y.</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Age</td>
<td>74</td>
<td>57</td>
<td>64</td>
<td>79</td>
<td>64</td>
<td>77</td>
<td>74</td>
</tr>
<tr>
<td>P6</td>
<td>P15</td>
<td>P3</td>
<td>P5</td>
<td>P12</td>
<td>C12</td>
<td>P9</td>
<td>C11</td>
</tr>
<tr>
<td>P(x)</td>
<td>1.57</td>
<td>1.49</td>
<td>1.39</td>
<td>1.35</td>
<td>1.24</td>
<td>1.03</td>
<td>0.98</td>
</tr>
<tr>
<td>H.Y.</td>
<td>2</td>
<td>2.5</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1.5</td>
</tr>
<tr>
<td>Age</td>
<td>53</td>
<td>76</td>
<td>80</td>
<td>75</td>
<td>57</td>
<td>74</td>
<td>68</td>
</tr>
<tr>
<td>P2</td>
<td>C1</td>
<td>C5</td>
<td>C16</td>
<td>C13</td>
<td>C2</td>
<td>C10</td>
<td>C14</td>
</tr>
<tr>
<td>P(x)</td>
<td>0.91</td>
<td>0.90</td>
<td>0.89</td>
<td>0.87</td>
<td>0.84</td>
<td>0.78</td>
<td>0.70</td>
</tr>
<tr>
<td>H.Y.</td>
<td>1.5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>44</td>
<td>57</td>
<td>47</td>
<td>40</td>
<td>61</td>
<td>22</td>
<td>38</td>
</tr>
<tr>
<td>C6</td>
<td>C9</td>
<td>C3</td>
<td>C8</td>
<td>C15</td>
<td>C4</td>
<td>C7</td>
<td></td>
</tr>
<tr>
<td>P(x)</td>
<td>0.65</td>
<td>0.55</td>
<td>0.50</td>
<td>0.48</td>
<td>0.42</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td>H.Y.</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>30</td>
<td>32</td>
<td>23</td>
<td>22</td>
<td>20</td>
<td>52</td>
<td>22</td>
</tr>
</tbody>
</table>

**Table 4. P(x) Scale with Corresponding H.Y. Score.**

<table>
<thead>
<tr>
<th>Severity</th>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.Y.</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

IV. DISCUSSIONS AND FUTURE WORKS

The advantage of the PD regression formula is that subjects are only required to perform Hausdorff’s data collection procedure, rather than more sophisticated neurological tests to perform this quick diagnostic test. The method of generating PD regression formula was based on Hausdorff’s gait dynamics database. Although we believed that subjects’ age was a factor in our study, we still could not find the exact influence due to the limited size of the training data set. Also, other types of neuro-degenerative or musculoskeletal disease could also affect our analysis [5, 6]. We expect the result may vary and would be more accurate if we have more data. Our future plan is to develop our own data collection procedure to generate a larger training data set. It would be interesting to identify other factors that can influence subjects’ gait performance. Apart from this, our method of using data mining techniques to discover regression formula are not limited to PD only. Same procedure could be applied to other kinds of data to discover the regular pattern.

V. CONCLUSION

In our study, we performed a deep analysis on Hausdorff’s PD patients’ data. We took several steps on visualization in different aspects, and we formed a Parkinson’s disease severity formula with 90 % of correct diagnosis rate based on Hausdorff’s training set. This diagnosis only requires a simple quick test from the subject. After the required test environment is set up, the presence of the specialist through the whole test is not strictly necessary. We believed our study will conduct to a quick preliminary PD diagnosis.

REFERENCES


