Applying Machine Learning to Gait Analysis Data for Disease Identification

Ranveer JOYSEEREE[a,b,*] , Rami ABOU SABHA[b,c] and Henning MUELLER[b,d]

[a] *Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland*
[b] *University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland*
[c] *Saint Joseph University, Beirut, Lebanon*
[d] *Medical Informatics, University Hospitals & University of Geneva, Switzerland*

Abstract. A machine-learning framework to identify the specific disease afflicting certain patients using only gait analysis data is presented. Classifying such data into disease types consumes valuable clinical time that may be better spent. Effective classification also facilitates its future retrieval. To prove the feasibility of the approach, we applied it to the simpler case of identifying the disease class of patients with a view to extending the method to specific diseases in future work. The patients benefiting from this framework suffer from Neurological and Neuromuscular Diseases (NND), or Juvenile Idiopathic Arthritis (JIA). Standard clinical gait information of healthy individuals, and NND/JIA patients was sourced from hospitals participating in MD-PAEDIGREE. To classify the data into one of the three categories: healthy, NND, and JIA, certain parameters were carefully selected from them and used to train Random Forest (RF), boosting, Multilayer Perceptron (MLP), and Support Vector Machine (SVM) classifiers. Cross-validation was used to test the effectiveness of our approach and it yields a classification accuracy of 100% for RF, SVM, and MLP classifiers and 96.4% for boosting. Training and testing for all the classifiers took mere milliseconds, providing opportunities for real-time applications. To extend the method to the identification of specific illnesses, more discerning features from the gait data are currently being investigated. Moreover, a larger dataset is being gathered. Finally, we are attempting to reduce the number of features used for classification in order to further decrease computation time and algorithm complexity.

Keywords. *Medical informatics, gait classification, machine learning, support vector machines, neural networks.*

Introduction

Gait analysis has been carried out for decades[1] and it involves the measurement and detailed study of quantities associated with human locomotion.[2] Currently, these quantities are collected using a range of sensors and are saved in an appropriate format for future reference.[3]

Analysis of the locomotion of patients can greatly help clinicians in the diagnosis of the type of disease afflicting the former. Accurate diagnoses allow the patient to

* Further author information: send correspondence to Ranveer Joyseeree (ranveer.joyseeree@hevs.ch)

receive the appropriate care as soon as possible, which minimises their suffering and allows them to enjoy an improved quality of life. However, the accuracy comes at the price of valuable clinical time. Automating the process of classifying gait data can therefore allow clinicians to better invest their time in other care-giving activities.

Quantitative analysis and characterisation of gait data would also allow clinicians to determine similarity between patients and to retrieve archived patient data that closely match the data of new patients. Such comparisons between past and present cases may help them perform a better differential diagnosis of the patient at hand.

The main purpose of this study is to facilitate the automatic classification of gait data in terms of disease type. Successful classification will facilitate the retrieval of archived gait information to complement new studies. To achieve those objectives, a machine learning algorithm is proposed. It is tested on images of healthy individuals, patients affected by Neurological and Neuromuscular Diseases (NND), and those affected by Juvenile Idiopathic Arthritis (JIA).

If it is found to be effective, our proposed framework will clearly have a significant positive impact on the well-being of those patients. As a first step towards reaching the stated objectives and to prove that the concept is viable, we propose, in this paper, an initial approach that tackles the simpler problem of identifying the disease class (JIA, NND, or healthy) instead of specific diseases. This is presented next.

1. Literature review

Zheng et al.[4] used Random Forests (RF)[5] and KStar to discriminate between neuro-degenerative diseases. The maximum accuracy reported was 94.02%. Yang et al.[6] use SVM to classify a number of neurodegenerative diseases based on gait. The reported maximum accuracy is 93.96%. SVM was also used by Begg et al.[7] for automated recognition of young-old gait types with a reported average success rate of 83.3%. Finally, Chan et al.[8] employed Multilayer Perceptron (MLP),[9] KStar[10] and Support Vector Machines (SVM)[11] to distinguish between walking up or down stairs and between younger and older adults. They reported accuracy of 95.7% in determining the former and 80.6% in determining the latter.

Our method compares favourably with all of the above studies since it reports a higher accuracy. It is also applicable to more machine learning algorithms at once, namely RF, SVM, MLP, and Boosting.

2. Sample and setting of study

Through the MD-PAEDIGREE project,[12] gait analysis data files were collected from participating hospitals across Europe after having obtained ethics approval. The patients who agreed to be recruited for the project provided informed consent for their gait analysis data to be used for research purposes. Patient metadata such as date of birth, values such as electromyography (EMG) readings, three-dimensional (3D) data relating to the coordinates of markers place on the body of the patient, and so on were saved in a binary C3D file format.[13]

3. Methods

Using the bio-medical toolkit (b-tk),[14] the contents of these files were analysed. **Figure 1** is an example of how C3D files may be visualised using b-tk. The white dots on the black background represent the position of markers on a 3D grid and their visibility can be toggled on and off using the menu on the right. Various events such as heel strike are viewable on the timeline at the bottom.

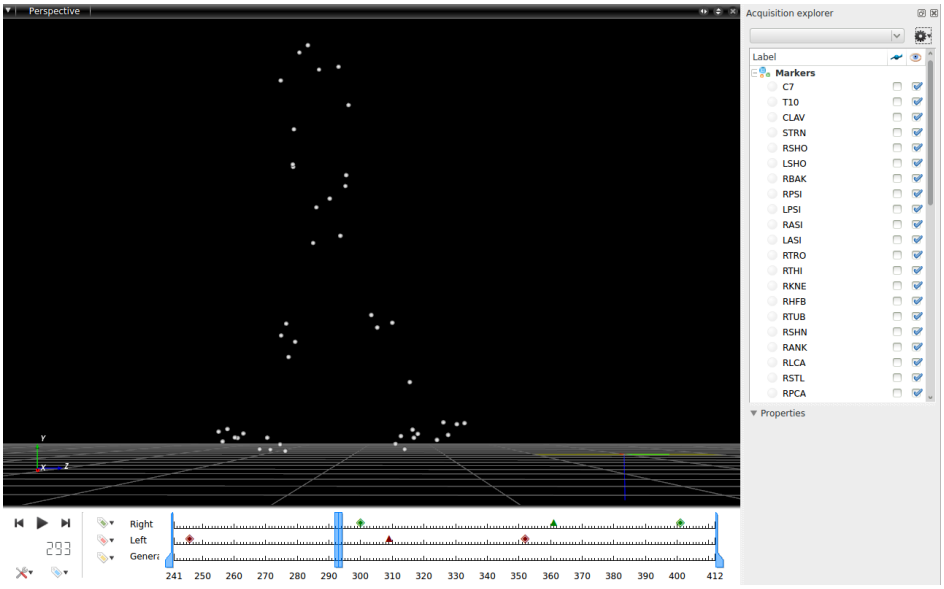

Figure 1. Visualisation of a c3d file using b-tk. The white dots on the black background represent the position of markers on the body of the patient in 3D grid space. Their visibility may be toggled on and off using the menu on the right. Events such as heel-strike may be viewed on the timeline at the bottom.

3.1 Parameter selection

Using the work of Givon et al.,[15] we identified 19 features/parameters that can potentially help us to effectively discriminate between the three labels: healthy, NND, and JIA. The parameters are listed below:

'Stride length', 'Right step length', 'Left step length', 'Stride time', 'Right step time', 'Left step time', 'Speed', 'Right foot speed', 'Left foot speed', 'Cadence', 'Double support', 'Left stance', 'Right stance', 'Left load response', 'Right load response', 'Left pre--swing', 'Right pre--swing', 'Left single support', and 'Right single support'.

3.2 Classification

For each gait analysis file, a row vector was constructed and populated with the values for the above 19 parameters. Following a cross--validation strategy, one of the constructed vectors was set aside for testing the machine learning algorithm used while the remaining ones were stacked together for training. Training and testing were carried out in C++ using OpenCV[16] with Random Forests (RF), boosting,[17] MLP, and SVM.

4. Results

A total of 28 gait analysis data files were collected through MD-PAEDIGREE. Eleven of them correspond to healthy individuals; ten suffer from an NND-related ailment while the remaining seven suffer from a JIA-related illness. Cross-validation was carried out on them, leading to 28 sets of results. They are presented in **Table 1**.

Table 1. Presented here are the results for RF, SVM, MLP and the boosting classifiers on the dataset of 28 healthy, NND and JIA patients. A 1 represents successful classification while 0 represents a misclassification.

File	Label	RF Result	SVM Result	MLP Result	Boosting Result
1	NND	1	1	1	1
2	NND	1	1	1	1
3	JIA	1	1	1	1
4	Healthy	1	1	1	1
5	NND	1	1	1	1
6	JIA	1	1	1	1
7	Healthy	1	1	1	1
8	NND	1	1	1	1
9	JIA	1	1	1	1
10	Healthy	1	1	1	1
11	NND	1	1	1	1
12	JIA	1	1	1	1
13	Healthy	1	1	1	1
14	NND	1	1	1	1
15	JIA	1	1	1	1
16	Healthy	1	1	1	1
17	NND	1	1	1	1
18	JIA	1	1	1	1
19	Healthy	1	1	1	1
20	NND	1	1	1	1
21	JIA	1	1	1	1
22	Healthy	1	1	1	1
23	Healthy	1	1	1	1
24	NND	1	1	1	0
25	Healthy	1	1	1	1
26	NND	1	1	1	1
27	Healthy	1	1	1	1
28	Healthy	1	1	1	1

In addition, for RF, 100 trees were utilised and the maximum tree depth was 10. The number of active variables used during tree-building was set to 4. For SVM, the kernel type was set to linear and C-Support Vector Classification was used, with the C value set to 1. For MLP, back-propagation was utilised with a maximum number of iterations of 300. Finally, for boosting, the boost type was set to real AdaBoost,[18] 100 weak classifiers were used, and the maximum tree depth was set to 5. Those parameters were chosen because they gave the best classification results.

The classification algorithms ran on a workstation running Ubuntu 14.04 LTS, with 16 gigabytes of RAM, and having an Intel™ Xeon® CPU with 8 cores running at 3.40 gigahertz. In both cases, the time taken to complete training and testing was of the order of milliseconds. From **Table 1**, we can observe a classification accuracy of 100% for RF, SVM, and MLP and 96.4% for the boosting algorithm. The approach is clearly effective in detecting the disease class of patients using only gait analysis data.

5. Discussion and Conclusions

A novel approach to predicting the disease class of patients using only their gait analysis information was presented. On a dataset of 28 patients, classification accuracies of 100% for RF, SVM, and MLP and 96.4% for the boosting algorithm were observed, respectively. The computation time for training and testing in all cases was in the order of milliseconds. The initial results obtained for the proposed method are, therefore, really promising and we argue that with future work focusing on identifying discerning features in the gait data will help extend the current method to classifying specific diseases instead of just disease classes.

We also recognise the need for a much larger dataset of gait analysis files for a better assessment of our method. This should not pose any major problems as the MD-PAEDIGREE project is ongoing and we expect to receive many more datasets through it. We also intend to optimize the number of features/parameters used for classification for faster computation times and lower algorithm complexity.

References

1. Vreeland RW, Sutherland DH, Dorsa JJ, Williams LA, Collins CC, Schottsteadt ER. A three-channel electromyograph with synchronized slow-motion photography. Ire Trans Biomed Electron.1961;8(1):4-6.
2. Davis RB III. Clinical gait analysis. IEEE Eng Med Biol Mag. 1988;7(3):35-40.
3. Garrido-Castro JL, Medina-Carnicer R, Martinez-Galisteo A. Design and evaluation of a new three dimensional motion capture system based on video. Gait Posture. 2006;24(1):126-129.
4. Zheng H, Yang M, Wang H, McClean S. Machine learning and statistical approaches to support the discrimination of neuro-degenerative diseases based on gait analysis. In: Intelligent Patient Management. Springer Berlin Heidelberg; 2009. p. 57-70.
5. Breiman L. Random forests. J Mach Learn. 2001;45(1):5-32.
6. Yang M, Zheng H, Wang H, McClean S. Feature selection and construction for the discrimination of neurodegenerative diseases based on gait analysis. Third IEEE Int Conf Pervasive Comput Technol Healthc. 2009;p. 1-7.
7. Begg RK, Palaniswami M, Owen B. Support vector machines for automated gait classification. IEEE Trans Biomed Eng, 2005;52(5):828-838.
8. Chan H, Yang M, Zheng H, Wang H, Sterritt R, McClean S, et al. Machine learning and statistical approaches to assessing gait patterns of younger and older healthy adults climbing stairs. Seventh IEEE Int Conf Nat Comput. 2011;1:588-592.
9. Ruck DW, Rogers SK, Kabrisky M, Oxley ME, Suter BW. The multilayer perceptron as an approximation to a bayes optimal discriminant function. IEEE Trans Neural Netw. 1990;1(4):296-298.
10. Cleary JG, Trigg LE. K*: An Instance- based Learner Using an Entropic Distance Measure. Proceedings of the 12th Int Conf Mach Learn, 1995; p. 108-114.
11. Cortes C, Vapnik V. Support-vector networks. J Mach Learn. 1995;20(3):273-297.
12. Model-Driven Paediatric European Digital Repository [Internet]. md-paedigree.eu – a clinically-driven and strongly VPH-rooted project. [updated 2015 February 6th; cited 2015 February 6th]. Available from: http://www.md-paedigree.eu/
13. The 3D Biomechanics Data Standard [Internet]. C3D.ORG. [updated 2014 September 1st; cited 2015 February 6th]. Available from: http://www.c3d.org/
14. Google [Internet] b-tk – Biomechanical Toolkit – Google Project Hosting. [updated 2015 February 6th; cited 2015 February 6th]. Available from: https://code.google.com/p/b-tk
15. Givon U, Zeilig G, Achiron A. Gait analysis in multiple sclerosis: Characterization of temporal-spatial parameters using GAITrite functional ambulation system. Gait Posture. 2009;29(1):138-142.
16. SourceForge [Internet]: SourceForge.net – Project Web Hosting – Open Source Software. [updated 2015 February 6th; cited 2015 February 6th]. Available from: http://opencvlibrary.sourceforge.net/
17. Schapire RE. The strength of weak learnability. J Mach Learn, 1990;5(2):197-227.
18. Freund Y and Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55:119–139.