DATA MINING TOOLBOX FOR GAIT ANALYSIS IN CHILDREN WITH CEREBRAL PALSY

[1]Yann-Aël Le Borgne, [1]Gianluca Bontempi, [2]Bruno Bonnechère, [2,3]Patrick Salvia,
[4]Marc Degaelen and [2,3]Serge Van Sint Jan
[1]Machine Learning Group, Faculty of Sciences, Université Libre de Bruxelles, Belgium
[2]Laboratory of Anatomy, Biomecanics and Organogenesis, Faculty of Medecine, Université Libre de Bruxelles, Belgium
[3]Functional Evaluation Center, Faculty of Medecine, Université Libre de Bruxelles, Belgium
[4]Hôpital Universitaire des Enfants Reine Fabiola, Brussels, Belgium; Email: yleborgn@ulb.ac.be

SUMMARY
Gait analysis in children with cerebral palsy (CP) is a complex task which requires analyzing and comparing large amounts of clinical and functional data. We propose to facilitate this task by relying on data mining tools, whose goal is to highlight the statistically relevant relationships existing in gait analysis data. The tools are developed by a multidisciplinary team involving clinical and data mining experts, and made available online by means of an interactive WEB interface.

INTRODUCTION
Gait analysis is the instrumented measurement of the movement patterns that make up walking. Gait data are recorded from motion analysis systems (e.g, Vicon, BTS), force platforms (e.g., AMTI, Kistler), foot pressure devices (e.g., RsSCAN), and electromyography systems (e.g., MotionLab). During a gait analysis exam, tens of temporal variables are recorded (kinematics, forces, moments and powers for different joints along the three different spatial dimensions, and EMG data for different muscles). Gait data are usually recorded for several gait cycles during a gait exam in order to account for the variability of a single subject's gait pattern. Gait analysis exams therefore generate vast amounts of data, which altogether allow to follow at a fine-grained level the spatiotemporal dynamics of a gait pattern.

Gait analysis of cerebral palsy (CP) children is particularly important for deciding whether a medical/surgical treatment may improve the condition of the child [1]. The analysis and interpretation of gait movements patterns of CP children is however particularly difficult. First, the etiology of CP gait patterns remains in many cases unclear. Second, gait variables exhibit complex nonlinear relationships, which cannot easily be visually identified [2].

We propose to assist the clinicians in their analysis of the CP gait patterns by means of a *data mining* toolbox. Data mining is a subfield of statistics and data analysis which aims at extracting statistically relevant relationships from large amounts of data [3]. The purpose of the toolbox is to provide clinical users with tools for managing groups of patients, compare populations by means of state-of-the-art clinical gait indices, and extract clinically relevant relationships from gait data. The toolbox is made available for testing by means of an online WEB interface, where 1828 gait analysis exams of a wide range of conditions (types of CP, GMFCS levels, presence of orthoses, …) can be analyzed and compared. To the best of our knowledge, the WEB interface and the range of tools we propose are the first of their kinds in the research efforts targeted at CP gait analysis.

METHODS
Three types of data mining tools were developed, allowing to manage, compare, and classify the kinematic and kinetic data of control and CP children.

The group management tool allows to create groups of subjects and to visualize their kinematic and kinetic data. Different criteria are proposed to the user for the creation of groups: hospital of origin, class of pathology (control, hemiplegia, diplegia, quadriplegia), laterality (left, right, bilateral), the GMFCS (I-V), botulinum toxin treatment (yes/no) and presence of orthoses (yes/no). The set of gait sessions matching the selected criteria can be further refined by an individual selection of the sessions. Kinematic and kinetic data can be visualized for each joint (pelvis, hip, knee, ankle, foot), on both sides (left and right) and along each of the three spatial dimensions (sagittal, frontal and transversal).

The comparison tool allows to compare the data from two different groups. The groups are created using the above-mentioned group management tool. The comparison can be made in terms of three state-of-the art gait indices proposed in the literature on gait analysis for CP, namely the *Gillette Gait Index* [4], the *Gait Deviation Index* [5] and the *Gait Profile Score* [6]. Histograms of the indices are generated for each group, and statistical tests allow to determine whether the distributions of indices in each of the two groups are significantly different.

The classification tool is based on a data mining algorithm called *conditional decision tree* [3], allowing to find the rules discriminating the gait patterns of two groups of subjects. The user selects the gait variables

(kinematics/kinetics, joints, sides, planes) along which the rules will operate. Gait cycles are discretized into eight phases (intial contact, loading response, mid-stance, terminal stance, preswing, initial swing, mid-swing, terminal swing). The decision tree algorithm identifies the variables whose distributions statistically differ between the two groups, and returns the sequence of tests to apply in order to determine whether a subject belongs to the first or the second group.

All the tools are implemented using open source software (R statistical language, Python, and Apache server). Data from three different hospitals, totalizing gait exams of 75, 250 and 83 subjects, respectively, are currently embedded and processed by the data mining toolbox. A total of 1828 gait sessions can be accessed with the tools.

Fig. 1: Classification tree separating subjects with GMFCS II and III.

RESULTS AND DISCUSSION

The types of data visualization offered by the proposed WEB interface provide the user with varied and rich representations of gait patterns and their relationships. For example, the visualization of gait patterns in the group management tool displays normality ranges, together with the gait data normalized along gait cycles, using as is the convention red and green curves for the left and right limbs, respectively. A second visualization tool allows to determine the gait session corresponding to a gait pattern interactively. In the comparison tool, data are reported using histograms (Gillette Gait Index and Gait Deviation Index), or barplots with confidence intervals for the Gait Profile Scores. For the gait classification tool, the result is a tree-shaped chart, whose nodes gives the tests to apply, and the leaves are stacked barplots of the proportions of gait sessions from each group falling in that leaf.

Thanks to the group management tool, the user can filter the sessions using clinically relevant criteria. Combined with the comparison and classification tools, the proposed toolbox therefore makes it possible to investigate a wide range of clinically relevant questions. Example of questions the toolbox can address are: "How do gait exams of left hemiplegia subjects visually compare to a normal group?" (visualization facility in the group management tool), "How much does Botox injection change the Gillette Gait Index of patients with diplegia?" (comparison tool) or "Which are the joints whose kinematic data allow to discriminate CP patients with GMFCS II from patients with GMFCS III?" (classification tool).

Due to space constraints, we only illustrate in more details this last example of query. Using the group management tool, a first group is created which contains all exams of patients with GMFCS II (N1=150 exams) and a second group for those with GMFCS III (N2=35 exams). The classification tree obtained from those two groups is reported in Fig. 1. Each node of the tree is a test on kinematic data, whose name gives the joint/side/plane and phase of the gait cycle. For the top node for example (first test to be applied), if the left knee flexion during the terminal stance is less than 21.9 degree, the next test is on the left branch, and otherwise on the right branch. Once in a
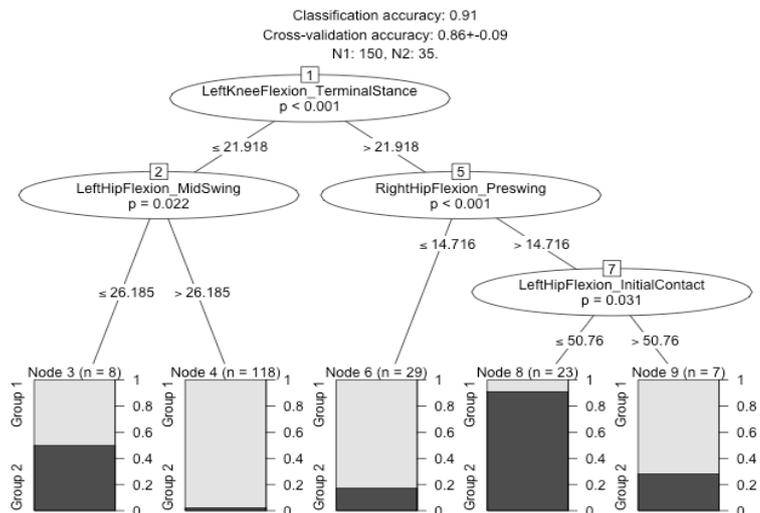
terminal node, the stacked barplot gives the probability of the subject to be in the first group (GMFCS II, light gray) or second group (GMFCS III, dark gray). Note that all tests are statistically significant with $p<0.05$, and that the tree properly classifies about 90% of the exams.

All the tools are made available online for testing at mlg.ulb.ac.be/ICT4Rehab_DM.

CONCLUSIONS

The statistical relevance of the tools depends on the amount of data available from the system, and the tools would benefit from the integration of data from other hospitals. Further developments target the design of tools for assisting in the writing of clinical reports.

ACKNOWLEDGEMENTS

REFERENCES

1. R.E. Cook, I. Schneider, M.E. Hazlewood, S.J. Hillman, and J.E. Robb. Gait analysis alters decision-making in cerebral palsy. *Journal of Pediatric Orthopaedics*, **23**(3):292– 295, 2003.
2. D.T.H. Lai, R.K. Begg, and M. Palaniswami. Computational intelligence in gait research: a perspective on current applications and future challenges. *IEEE Transactions on Information Technology in Biomedicine*, **13**(5):687–702, 2009.
3. Trevor J. Hastie, R.J. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.
4. LM Schutte, U. Narayanan, JL Stout, P. Selber, JR Gage, and MH Schwartz. *An index for quantifying deviations from normal gait*. Gait & posture, **11**(1):25–31, 2000.
5. M.H. Schwartz and A. Rozumalski. The gait deviation index: a new comprehensive index of gait pathology. *Gait & posture*, **28**(3):351–357, 2008.
6. R. Baker, J.L. McGinley, M.H. Schwartz, S. Beynon, A. Rozumalski, H.K. Graham, and O. Tirosh. The gait profile score and movement analysis profile. *Gait & posture*, **30**(3):265–269, 2009.