Human gait data mining by symbol based descriptive features

V. Ergovic[1], S. Tonkovic[2] and V. Medved[3]

[1] IBM Croatia / Software Group, Zagreb, Croatia
[2] Faculty of Electrical Engineering and Computing / ZESOI, Zagreb, Croatia
[3] Faculty of Kinesiology / Biomechanics, Zagreb, Croatia

Abstract— **Most medical applications which include large dataset require search by example capability. Such similarity-based retrieval has attracted a great deal of attention in recent years. Although several different approaches have appeared, most are not specialized for problems of time series signals, typically found in gait analysis or ECG. This paper proposes an approach for efficient processing of human gait by using symbolization of different features by means of feature description. We evaluated this approach against database that holds limited set of kinematics, kinetic and EMG data, describing simple step (a subset of standard gait variables). The measurement was performed in the Biomechanics Laboratory at the Faculty of Kinesiology, University of Zagreb, using 2002-ELITE system with Kistler platform (40 cm by 60 cm).**

Keywords— **descriptive features, data mining, gait analysis.**

I. Introduction

Gait analysis and diagnosis are mostly based on limited observations and measurements from parts of the complicated movement system. Inexperienced therapist often does not have a clear understanding of the integrated movement system and, therefore, cannot properly set diagnosis. The learning process takes significant amount of time; it is based on facts and rules. This leads to generation of statements and provides an input which can be used by an expert system. Based on our observation of typical process and deductions performed by the therapist we propose novel approach for times series data mining which imitates therapist approach. We propose a human readable text description of signal just as one used by the therapist. We than map this description to symbols. Each signal is expressed by symbolic word where symbols are alphabetically ordered. This approach greatly improves search across large databases, dimension reduction and interpretation of the results.

A. Background work

Humans have vast amount of cultural and domain knowledge and experience to draw upon, as well as a remarkable ability to recognize similar situation when making decision about new information, while machine learning methods can only generalize based on the data that has already been seen and even than is in a limited manner. Many machine-learning algorithms rely heavily on mathematics and statistics. There are many different machine-learning algorithms all with different strengths and weaknesses and they are suited to different types of problems. Some, such as decision trees, are so easy to interpret and are among the most widely used methods in medical decision-making. Others, such as neural networks are of a black box type. They produce an answer but it is very difficult to reproduce and understand reasoning behind the mechanism [1]. Decision trees and rule based reasoning are typically used in medical signal analysis especially in ECG analysis [2]. However, in domain of gait analysis there are only few papers which have similar approach. Table 1 shows an excerpt from Winter's table, originally titled Elaborated Strategy of the Clinical Application of Gait Analysis [3].

Table 1 Gait Diagnostic Chart [3]

Observed Abnormality	Possible Causes	Diagnostic Evidence
Foot slap at heel contact approximation	Below normal dorsiflexor activity at heel contact	Below normal tibialis anterior EMG or dorsiflexor moment at heel contact
Forefoot or flatfoot initial contact	(a) Hyperactive plantar-flexor activity in late swing	(a) above normal plantar-flexor EMG in late swing
	(b) Structural limitation in ankle range	(b) decreased dorsiflexion range of motion
	(c) Short step-length	(c) See a,b,c,d bellow
Short step-length	(a) weak hip push-off prior swing	(a) Below normal plantar-flexor moment or power generation or EMG during push-off
	(b) weak hip flexors at toe off and early swing	(b) Below normal hip flexor moment or power generation or EMG during late push-off and early swing
	(c) excessive deceleration of leg in late swing	(c) Above normal hamstring EMG or knee flexor moment or power absorption late in swing
	(d) Above normal contralateral hip extensor activity during contralateral stance	(d)Hyperactivity in EMG of contralateral hip extensors

Selected statements can be easily stored in a database and used for comparison with other gait signals by leveraging text mining. Text mining, roughly equivalent to text analytics, refers generally to the process of deriving high-quality information from text. High-quality information is typically derived through the dividing of patterns and trends through the means such as statistical pattern learning. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output [1]. In order to leverage text mining, time series signals are converted to text and stored in database table(s). Basically, input time series datasets are described by words like in a table. However, the area of gait analysis is not so well researched and structured as for example ECG. It is more complex and it can hold much more data (default setup of our system gives over 20 different variables which hold time series data – kinetics and kinematics). Figure 1 shows abnormal pelvic obliquity and tilt. We can apply descriptive approach on the signal level and describe signals with statements (right pelvic obliquity curve is shifted down; left pelvic obliquity curve is shifted up, pelvic tilt curve is shifted down; pelvic tilt has large extreme values; pelvic obliquity degrees are completely out of normal ranges and completely shifted up, multiple deviations on the curves, phase shift, etc.). Those statements, together with selected statements from table can be stored in database for purpose of text mining or construction of decision trees.

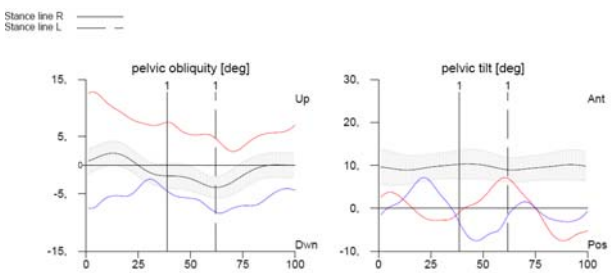

Fig. 1 Normal range and discords - pelvic

For subjects who have walking problems, it would be expected that parts of some of the graphs would be outside the expected range. A skilled clinician requires a considerable amount of time to analyze this data. In addition to this, simply browsing through all possible graphs for a single trial is time consuming [4, 5].

B. Symbolization

Symbolization is the method of representing time series dataset as a set of symbols in the form of one or more symbolic words which brings significant dimension reduction. The most popular symbolic algorithm in the field of research in recent years is SAX which is an extension to PAA (**P**iecewise **A**ggregate **A**pproximation). Real valued time series converted to the SAX word baabccbc [6] are shown with Figure 2. All three possible symbols are approximately equally frequent.

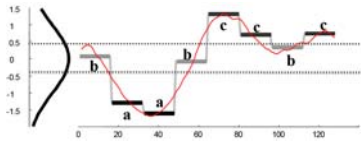

Fig. 2 Time series signal converted to symbolic word [6]

In PAA, each sequence of time series data is divided into k segments with equal length and the average value of each segment is used as a coordinate of a k-dimensional feature vector. PAA is fast and simple for implementation and provides linear time indexing. SAX uses an intermediate representation between the raw time series and the symbolic strings because the first step is transformation of the data into PAA representation. Second step is to symbolize the PAA representation into a discrete string. Time series is discretized by first obtaining a PAA and then, using predetermined breakpoints, PAA coefficients are mapped into SAX symbols. SAX uses MINDIST function that returns the minimum distance between the original time series of two words like shown with formula 1.

$$MINDIST(\hat{Q},\hat{C}) \equiv \sqrt{\frac{n}{w}}\sqrt{\sum_{i=1}^{w}(dist(\hat{q}_i,\hat{c}_i))^2} \quad (1)$$

Where n is the length of the string to be reduced to a string of arbitrary length w, ($w < n$, typically $w << n$). The alphabet size is also an arbitrary integer a, where $a > 2$. Q and C are words for distance calculation and q_i and c_i are individual symbols in the words. This function is used to calculate distance between two symbolic words and it is used for classification purposes. Typically this function is implemented via lookup table.

In processing human gait it is very common to process very long time series. Disadvantages of PAA and SAX include inability to deal with significant phase shifts and predetermined alphabet without the possibility for runtime extension. A distance function brings computational overhead during queries when larger alphabet is used. Although there are few initiatives that are dealing with these problems they can not avoid some limitations such as strict order of symbols in the word, compression limitations, and problem with runtime alphabet extension [8, 9].

II. SYMBOL BASED DESCRIPTIVE FEATURE APPROACH

A. Proposed approaches

Our approach introduces a concept of descriptive feature of time series data set such as a signal.

Definition. Descriptive feature is text representation of observed facts stated in human readable and understandable format.

Our approach does not include text mining method which would normally calculate word statistics and relevance in statement. Each descriptive feature can be mapped to a symbol. Symbol does not represent any segment in the curve like SAX does and therefore position in a word is irrelevant (it can be sorted alphabetically). This feature provides fast retrieval by exact match, almost as a hash function and enables additional extensions to the alphabet during the runtime without additional computation costs. Each description can be placed in separate table to provide mapping between symbols which are represented by primary key in description table to descriptive feature. This helps to improve search and classification across complex time series dataset such as human gait variables. The disadvantage is a descriptive feature is that is based on heuristics. Therefore someone should identify and describe the feature before it is stored in repository.

B. Comparison with previous work

Figure 3 represents logical model of the data warehouse with hierarchical lookup tables (summary tables). In this model query is first performed on the first level of the lookup hierarchy table and then based on the similarity, queries are performed on the sub levels until sufficient similarity has been reached or query has not returned a result at all (not found). In ETL (**E**xtract **T**ransform and **L**oad) stage each signal is summarized at a different level of SAX granularity. SAX word generation is performed more than ones with different alphabet size and different word lengths. This idea of hierarchy has been taken from Hierarchical clustering algorithm. Drawback of this solution is a number of queries that have to be performed before results are obtained and for each fact table additional lookup table is required [9]. SAX is also heavily dependant on the alphabet size, word length, segmentation of the input signal, and additional computation during queries when larger alphabet and larger words are used.

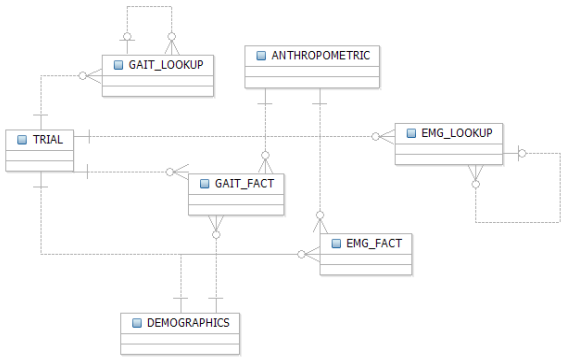

Fig. 3 Logical warehouse model for gait analysis with lookup tables [9]

C. Improvement of the model

Descriptive feature approach requires a table with description (code table which maps symbols to statements). This also enables easier signal interpretation and human understandable query construction. Modified lookup tables point to correct trial which is shown with the Figure 4.

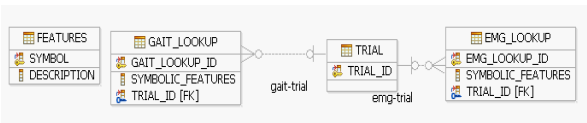

Fig. 4 Feature based lookup tables

In each row of the lookup table symbolic features are inserted as one word, with symbols sorted in ascending order (e.g. abfhlxz). We enforce ordering since position of our symbols is irrelevant. If we look into this model even more we will notice that EMG and gait are only some features of the trial. With that approach we can reduce this model even more by eliminating gait and lookup table, merging them together as shown with Figure 5.

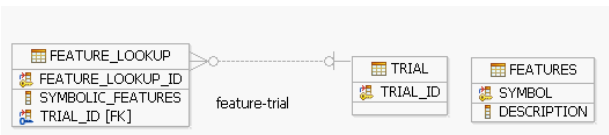

Fig. 5 Final model

Elimination of lookup table increases alphabet, there are more feature descriptions and it is appropriate for alphabets under 200 symbols, not recommended for larger alphabets.

III. RESULTS

We have worked with ground reaction force components (X,Y,Z), sensor marker readings (knee extension and rotation) and EMG. Created dataset tends to be very large. In our case average reading included 10 different parameters sampled at 10 ms for 20 seconds period of movement that produced more than 20000 samples for one movement pattern including five to six steps. Presented results are referring to GAIT_FACT table. Gait fact table was loaded with 7.201.050 rows, each row had 10 columns. Fourier approximation used 44 elements to approximate the signal (22 for magnitude and 22 for phase) as stated by formula 2.

$$f(t) = a_0 + a_1 \cos(w_0 + \varphi_1) + .. + a_n \cos(nw_0 + \varphi_n) \quad (2)$$

Based on previous work, lengths of symbolic words were set to 4 and 9, with alphabet size of 3 and 6. Each signal was aggregated in 8 words. Average dimensional reduction was 384.61, but reduction in number of rows was 10000 times as summarized in Table 2. Gait fact table populated with 7.201.050 rows was described with less than 720 rows (the worst case scenario - 360 on the level one of the hierarchy and 360 on the level two). Since original data is contained in GAIT_FACT table there were no limitations on the use of additional algorithms after lookups were considered [9]. Similar results were with EMG. We used 10 selected statements regarding our dataset.

Table 2 Dimension reduction and data base rows

Method	Average reduction	Number of rows
Fourier approximation	476.16	171352
SAX hierarchy (2 level)	384.61	360-720
Descriptive features	1996	359

IV. CONCLUSIONS

Like any other approach in expert system diagnostics, this approach should be considered just to offer recommendations and possible indications about the problem. We have identified two major benefits of our approach. In the segment of dimension reduction, descriptive features can be stored in database alone, however in that case we original data are lost and additional algorithms such as clustering support vector machines or regression can not be utilized since original data is lost. The other segment that is affected is the speed of the retrieval during searches performed on the database – in that case our descriptions are stored in lookup tables to enable fast retrieval. Usage of additional algorithms for drilling down specific subset in that case is supported. Since expert systems in biomechanics tend to contain huge amount of time series data, structural approach and aggregation mechanism are needed. Further research should be oriented towards evaluation of data mining chaining possibilities and domain specific ontologies which should describe diagnostic process and standardize vocabulary which describes signals on description level.

ACKNOWLEDGMENT

The results presented are the product of scientific projects „Noninvasive measurements and procedures in bio-medicine“, „Automated motion capture and expert evaluation in the study of locomotion“ and „Real-life data measurement and characterization“, realized with the support by The Ministry of Science, Education and Sports, Republic of Croatia.

REFERENCES

1. Segaran T, (2007) Programming Collective Intelligence. O'Reilly, Sebastopol, 2007, US
2. Yanowitz F, ECG Conduction Abnormalities at http://library.med.utah.edu/kw/ecg/ecg_outline/Lesson6/index.html
3. Winter D (1991), Gangbildanalyse-Stand der Messtechnik und Bedeutung fur die Orthopädie-Technik. Duderstadt: Mecke Druck. pp 266-277
4. The Robert Gordon University (School of Comptuing) at http://www.comp.rgu.ac.uk/docs/info/project/ce-humangait-rn.htm
5. Noble R.A, White R (2005), Visualisation of gait analysis data, Information Visualization, Proceedings 9th International Conference on Information Visualization, London, UK, 2005 pp 247 - 252
6. Keogh E, Lonardi S, Ratanamahatana C (2004) Towards Parameter-Free Data Mining, 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, US, 2004, pp 206-215
7. Korn F, Jagadish H, Faloutsos C (1997) Efficiently supporting ad hoc queries in large datasets of time sequences, In Proceedings of SIGMOD, pp 289-300
8. Ergovic V (2008) Multi Stage Symbolic-Based Shape Indexing and Retrieval, In proceedings of ELMAR 2008, Zadar, Croatia, 2008, pp 87-90
9. Medved V, Ergovic V, Tonkovic S (2008), Towards a High Performance Expert System for Gait Analysis, IFMBE Proceedings of 4th European Congress for Medical and Biological Engineering 2008, Antwerp, Belgium, 2008, pp 2105-2108

- Author: Vladimir Ergovic
- Institute: IBM
- Street: Miramarska 23
- City: Zagreb
- Country: Croatia
- Email: vladimir.ergovic@hr.ibm.com