International Journal of
Physical Medicine & Rehabilitation

Vakanski et al., Int J Phys Med Rehabil 2017, 5:3
DOI: 10.4172/2329-9096.1000403

Research Article Open Access

Metrics for Performance Evaluation of Patient Exercises during Physical Therapy

Aleksandar Vakanski[1]*, Jake M. Ferguson[2] and Stephen Lee[3]

[1]*Industrial Technology, University of Idaho, Idaho Falls, ID, USA*
[2]*Center for Modeling Complex Interactions, University of Idaho, Moscow, ID, USA*
[3]*Department of Statistical Science, University of Idaho, Moscow, ID, USA*

Abstract

Objective: The article proposes a set of metrics for evaluation of patient performance in physical therapy exercises.

Methods: Taxonomy is employed that classifies the metrics into quantitative and qualitative categories, based on the level of abstraction of the captured motion sequences. Further, the quantitative metrics are classified into model-less and model-based metrics, in reference to whether the evaluation employs the raw measurements of patient performed motions, or whether the evaluation is based on a mathematical model of the motions. The reviewed metrics include root-mean square distance, Kullback Leibler divergence, log-likelihood, heuristic consistency, Fugl-Meyer Assessment, and similar.

Results: The metrics are evaluated for a set of five human motions captured with a Kinect sensor.

Conclusion: The metrics can potentially be integrated into a system that employs machine learning for modelling and assessment of the consistency of patient performance in home-based therapy setting. Automated performance evaluation can overcome the inherent subjectivity in human performed therapy assessment, and it can increase the adherence to prescribed therapy plans, and reduce healthcare costs.

Keywords: Physical therapy; Performance metric; Log-likelihood; KL divergence; RMS distance

Introduction

Functional recovery from neuromotor disabilities, various surgical procedures, or musculoskeletal trauma is strongly dependent on patient participation in a physical therapy program. While a large portion of all therapy exercises is performed by patients in a home-based setting, the lack of supervision and motivation for continued involvement in the therapy program in outpatient environment conduce low adherence to prescribed treatment regimens [1]. The presented work in this article was motivated by our belief that the latest progress in machine learning furnishes a potential to be harnessed for analysis and monitoring of patient progress toward recovery during in home physical rehabilitation, and accordingly, can greatly benefit both patients and healthcare providers.

The recent rapid advancements in artificial intelligence (AI), driven predominantly by its sub field machine learning, have been reflected by ubiquitous deployment across a wide spectrum of application domains, ranging from miscellaneous image-, text-, and voice-processing apps in smart phones and computers to autonomous cars and personalized recommender systems.

It is expected that as the field further evolves in the years to come, AI-enabled systems will have even more pronounced and transformative impact on society as a whole and on all aspects of our lives as individuals.

In the medical field, the number of machine learning applications has proliferated recently due to the demonstrated capacity for discovering complex patterns by analysing large numbers of electronic medical records. Not surprisingly, the most notable medical AI success has been in the domain of medical image processing. For example, the medical team at Deep Mind have applied deep artificial neural networks (ANNs) for analysis of digital scans of the eye in diagnosis of age-related macular degeneration and diabetic retinopathy [2], and for analysis of radiotherapy scans for detection of oral and neck cancer [3]. Other exemplary AI applications include image processing of skin lesions in screening and detection of melanoma cancer [4], and image processing of scans for detection of invasive brain cancer cells [5]. Machine learning approaches have also been implemented in a variety of other biomedical research problems [6], such as analysis of genomics sequences [7], drug discovery and repurposing [8], and robotic healthcare assistants [9].

The benefits of applying machine learning algorithms to medical data analytics are numerous, and encompass customized and personalized diagnosis and treatment, faster screening and early detection of conditions, which can potentially lead to improved healthcare quality and patient satisfaction, reduced healthcare costs, reduced need for hospital stay, and similar.

As more archived traditional medical records are transferred to digital form, and as the personal wearable devices and mobile apps unobtrusively collect massive amounts of information about our bodily functions and activities, more training data will become available, which will improve the outcomes of the machine learning algorithms and leverage the extraction of subtle health related and behavioural patterns. For instance, one creative solution employing images taken from a regular cell phone camera is the mobile app AiCure [10], which uses AI-supported image processing for monitoring users' habits in taking prescription medications, with an objective to increase the adherence rates, as well as to update the respective physician on patient habits related to taking the prescribed medications.

***Corresponding author:** Aleksandar Vakanski, Industrial Technology, University of Idaho, 1776 Science Center Drive, TAB 309, Idaho Falls, ID, 83402, USA, Tel: 208-757-5422; E-mail: vakanski@uidaho.edu

Received April 01, 2017; **Accepted** April 17, 2017; **Published** April 20, 2017

Citation: Vakanski A, Ferguson JM, Lee S (2017) Metrics for Performance Evaluation of Patient Exercises during Physical Therapy. Int J Phys Med Rehabil 5: 403. doi: 10.4172/2329-9096.1000403

Likewise, application of machine learning algorithms for monitoring and evaluation of patient compliance with a prescribed physical therapy program can improve the adherence rates, reduce the required time for functional recovery, and consequently, reduce treatment cost. The development of such systems requires hardware components, i.e., a dedicated computer for data processing, and a sensory system for capturing patient exercises during rehabilitation sessions. Among the different sensory systems for motion capturing, the vision-range sensors of the type of Microsoft Kinect are currently an excellent option for the task at hand, considering their affordability (price in the range of $150), reliability for different research and industrial applications, and availability of open source libraries for program development with a broad range of capabilities. Two such existing systems KiRes (Kinect Rehabilitation System) [11] and VERA (Virtual Exercise Rehabilitation Assistant) [12] utilize the motion capturing feature of Kinect to present an avatar on a computer display that reproduces patient motions in real time, and simultaneously displays the desired motions. The visualization of the performance provides an instantaneous feedback to the patient, helps in recognizing any needs for correcting the exercises, as well as motivates the patient to comply with the prescribed treatment. A comprehensive review of the technical and clinical merits of the application of Microsoft Kinect for motion capturing of patient exercises in physical rehabilitation is presented by Hondori and Khademi [13].

Equally important to the requirement for adequate hardware components is the development of a methodology for computer-driven analysis of patient therapy efforts, related to evaluating the consistency of the performance with the PT-prescribed exercises, the day-to-day patient progress, and the level of compliance with the prescribed treatment plan. Such methodology is predicated upon the provision of: (i) efficient mathematical models for representation of bodily movements undertaken during physical therapy exercises [14], and (ii) efficient metrics for quantifying the patient executed motions and collating the performance to the prescribed motions by the PT.

The objective of this article is to present a survey of the current literature in reference to the metrics for evaluation of patient performance in physical therapy. The existing practice for evaluation of physical rehabilitation has exclusively relied on assessment by a PT. For instance, a common test for evaluation of motor recovery after stroke is Fugl-Meyer Assessment [15], where a PT evaluates a patient's performance on a set of pre-defined movements and assigns a numerical score on a scale of 0 to 2 for each of the movements. Related tests for evaluation of the level of recovery after stroke include the Motor Assessment Scale [16] and the motricity index [17]. Another test for assessing the ability of upper motor movements is the Wolf Motor Function Test [18], which is a timed test consisting of several functional tasks, scored on a scale of 0 to 5. These and several other tests for assessment of patient performance and the corresponding level of functional recovery that are currently performed by a trained PT are suitable candidates for automation, since they rely on a set of standard pre-defined movements. Accordingly, drawbacks of this type of assessment include: it is time consuming, and it produces subjective scores where different PTs can provide different assessment scores due to human inability to accurately measure and quantify body trajectories. Automated performance evaluation can overcome these limitations by providing more accurate and quantified assessment, also can be involved in daily monitoring of the therapy sessions, and can provide instantaneous corrective feedback and send the performance data to the respective PT on a daily basis.

With regards to the proposed metrics for automated performance evaluation in the published literature, to the best of our knowledge only the work by Komatireddy et al. [12] has partially addressed this topic. The authors proposed a quantitative metric, related to the number of correctly performed repetitions of an exercise, and a qualitative metric, related to ratio of optimal **vs.** sub-optimal repetitions of the exercise. The study does not provide a clear explanation of which discriminative approach was applied for distinguishing between optimal and suboptimal repetitions.

This article reviews metrics that have been used, or that can be potentially used, for evaluation of patient therapy motions. Motivated by the work in Komatireddy et al. [12], we employ a taxonomy that classifies the metrics as quantitative and qualitative. Further, quantitative metrics are categorized into model-less and model-based metrics. Model-less metrics perform the assessment based on the raw time series of the motions as acquired by a sensory system. Metrics is this category are: root-mean square distance, and norm of jerk. Model-based metrics calculate the consistency of patient exercises in comparison to a mathematical model of the motion as prescribed by a PT. Metrics in this category include: log-likelihood, Kullback Leibler divergence, heuristic consistency, and prediction intervals. Other related metrics not explored in this work are the Hellinger distance and the Bhattacharyya distance. While the quantitative metrics evaluate the motions at a low level of abstraction, i.e., at a level of individual measurement points in a sequence, the qualitative metrics evaluate the motions at a high level of abstraction, i.e., at a motion sequence level. Metrics in this category involve: number of optimal attempts, Fugl-Meyer Assessment, and Wolf Motor Function Test.

The article is organized as follow. The next section introduces the used mathematical notation for the human motions. Afterwards the metrics for patient performance evaluation are described. The reviewed metrics are next compared for evaluation of five human motions. The last section summarizes the presented study.

Notation

In a physical rehabilitation setting, a PT will prescribe a collection of desired therapy motions to a patient, by either performing the motions in front of the patient, or by physically moving the body parts of the patient along the required paths. It is assumed here that the PT will provide several demonstrations of each motion in order to reinforce the perception of the motion by the patient, which may be related to required range, speed of movement, and other respective constraints in the execution of the motion. The set of reference examples of a motion prescribed by the PT is denoted $\mathcal{O}=\{O_m\}_{m=1}^{M}$ where m is used for indexing the individual examples of the motion, and M is the total number of examples of the motion $\mathcal{O}$ demonstrated by the PT. It is also assumed that a sensory system is used for capturing the prescribed therapy exercises, where each motion $\boldsymbol{O}_m$ is acquired by the sensor as a temporal sequence of measurements $\boldsymbol{O}_m=\left(\mathbf{o}_m^{(1)},\mathbf{o}_m^{(2)},\ldots,\mathbf{o}_m^{(T_m)}\right)$. Each measurement $\mathbf{o}_m^{(k)}$ in the motion sequence represents a D-dimensional vector where the subscript m denotes again the example index, and the superscript k denotes the temporal position of the measurement in the motion sequence $\boldsymbol{O}_m$. The total number of temporal measurements in the demonstration $\boldsymbol{O}_m$ is denoted T_m. In general, the length of the motion examples T_m in the set $\{\boldsymbol{O}_1, \boldsymbol{O}_2,.., \boldsymbol{O}_M\}$ will be similar but different, due to the inherent variability of human movements.

Analogously, let's assume that the patient is attempting to perform the prescribed motion $\mathcal{O}$ in a home-based rehabilitation program in front of a sensory system for motion capturing. The patient is presumably asked to repeat the motion a predefined number of times at a predefined time period (e.g., 10 times daily). The measured motion examples performed by the patient are denoted $\mathcal{R}=\{\boldsymbol{R}_n\}_{n=1}^{N}$ where by analogy n represents a motion index, and N is the number of performed motion examples. Each motion example is a temporal sequence $\boldsymbol{R}_n=\left(\mathbf{r}_n^{(1)},\mathbf{r}_n^{(2)},\ldots,\mathbf{r}_n^{(T_n)}\right)$, consisting of T_n D-dimensional vectors denoted in this work $\mathbf{r}_n^{(k)}=\left[r_n^{(k,1)} \quad r_n^{(k,2)} \quad \cdots \quad r_n^{(k,D)}\right]^T$.

The metrics for performance evaluation are to describe in a quantitative or a qualitative manner, or both, the consistency of the patient performed examples of the motion $\mathcal{R}$ with the PT prescribed examples of the motion $\mathcal{O}$. Due to musculoskeletal constraints, pain, or other conditions, the patient may not be able to correctly perform the motion at the beginning of the therapy program, which may, or may not, improve as the therapy program progresses.

Metrics

The reviewed metrics for performance evaluation are classified in this work into two main categories: quantitative and qualitative metrics. Accordingly, quantitative metrics assign a numerical score for the consistency of the patient performance, whereas qualitative metrics assign either a non-numerical evaluation (e.g., correct versus incorrect performance) or a discrete numerical score from a finite and limited range of values or states.

Quantitative metrics

Quantitative metrics can be also referred to as low-level metrics, since they evaluate the consistency of each measurement with regards to the prescribed sequence of measurements, or with respect to a model of the motion in the form of a probability distribution. The quantitative metrics are further classified into model-less and model-based metrics.

Model-less metrics

The model-less metrics compare the motions captured during a physical therapy exercise by a patient, with the motions captured when prescribing the therapy exercise by the PT. These metrics compare the measured raw trajectories of the body parts as acquired by the sensory system.

The following metrics are classified in this group:

a) *Root-mean square (RMS) distance*-obtained as a sum of differences between the points of a captured trajectory $\boldsymbol{R}_n$ and a set of prescribed trajectories $\boldsymbol{R}_n$ and a set of prescribed trajectories $\mathcal{O}=\{\boldsymbol{O}_m\}_{m=1}^{M}$

$$\mathcal{L}_1(\boldsymbol{R}_n,\mathcal{O})=\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{T_m}\left\|\mathbf{r}_n^{(k)}-\mathbf{o}_m^{(k)}\right\|=$$

$$=\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{T_m}\sqrt{\left(r_n^{(k,1)}-o_m^{(k,1)}\right)^2+\cdots+\left(r_n^{(k,D)}-o_m^{(k,D)}\right)^2} \tag{1}$$

One constraint of the RMS distance is the requirement that the trajectories have the same length, i.e., the same number of observations T_m. Therefore, the observed trajectories need to be scaled to a same length before the RMS distance is calculated. For the case when the trajectories are linearly scaled to a same length, if there are great spatial differences along their temporal dimension, that will result in a large RMS distance between the trajectories. This limitation is typically mitigated by employing approaches for temporal alignment of the trajectories, such as Dynamic Temporal Warping (DTW) [19].

Another metric that can be derived from the RMS distance for a single motion example $\boldsymbol{R}_n$ is the mean of the RMS distances for all motion sequences in the set $\mathcal{R}=\{\boldsymbol{R}_n\}_{n=1}^{N}$, i.e.,

$$\mathcal{L}_{1,mean}(\mathcal{R},\mathcal{O})=\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_1(\boldsymbol{R}_n,\mathcal{O}) \tag{2}$$

b) *Norm of jerk*-where the term jerk is related to the time derivative of the acceleration, i.e., the third derivative of the position. The metrics calculates the norm of the jerk for each trajectory point as:

$$\mathcal{L}_2(\boldsymbol{R}_n)=\frac{1}{T_n}\sum_{k=1}^{T_n}\left\|\frac{d}{dt^3}\mathbf{r}_n^{(k)}\right\|=$$

$$=\frac{1}{T_n}\sum_{k=1}^{T_n}\sqrt{\left(\frac{d^3r_n^{(k,1)}}{dt^3}\right)^2+\cdots+\left(\frac{d^3r_n^{(k,D)}}{dt^3}\right)^2} \tag{3}$$

This metric quantifies the level of smoothness of the movement [20], and high value of jerk can be indicative of shaky patient movements during the physical exercises. In certain rehabilitations exercises and conditions, it is expected that the patients will produce high level of jerks at the beginning of the treatment, which will gradually reduce as the recovery improves. Although this metric evaluates only one aspect of the movements, when combined with other metrics it can provide valuable information regarding the level of progress toward functional recovery.

Model-based metrics

These metrics rely on a model of the prescribed motions and/or a model of the patient motions. Common methods used for modeling human motions include probabilistic approaches, such as Gaussian mixture models [21] and hidden Markov models [22]. These approaches model the sequences through a set of latent states that describe a statistical distribution of the motion dynamics. Other common approach for modeling human movements is by employing a set of deterministic latent states connected by weights, such as the artificial neural networks [21].

The metrics in this category include:

a) *Log-likelihood*-expresses the probability $\mathcal{P}$ that a performed motion example by the patient is drawn from a model of the motions as prescribed by the PT. For a model described with a set of parameters λ, the log-likelihood of a motion example $\boldsymbol{R}_n$ is calculated as a natural logarithm of the likelihood for all data points given the model parameters λ [21], that is,

$$\mathcal{L}_3(\boldsymbol{R}_n)=\frac{1}{T_n}\log\mathcal{P}(\boldsymbol{R}_n)|(\lambda)=$$

$$=\frac{1}{T_n}\log\prod_{k=1}^{T_n}\mathcal{P}\left(\mathbf{r}_n^{(k)}|\lambda\right)=\frac{1}{T_n}\sum_{k=1}^{T_n}\log\mathcal{P}\left(\mathbf{r}_n^{(k)}|\lambda\right) \tag{4}$$

Similar to (2), the mean of the log-likelihood for all sequences in the set $\mathcal{R}=\{\boldsymbol{R}_n\}_{n=1}^{N}$ can be employed as a measure of consistency of the repetitions of a single motion in reference to a model λ of the prescribed set $\mathcal{O}$.

$$\mathcal{L}_{3,mean}(\mathcal{R},\mathcal{O})=\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_3(\boldsymbol{R}_n) \tag{5}$$

b) *Kullback Leibler (KL) divergence*-is a measure of the similarity between two probability distributions [23]. One of the distributions is considered to represent the true theoretical distribution of the data, in this case that is the empirical distribution of the prescribed movements by the PT, i.e., $\mathcal{P}(\mathcal{O})$. The other distribution represents an approximation of the true distribution, which in this case is the distribution of the executed movements by the patient, i.e., $\mathcal{P}(\mathcal{R})$. The KL divergence between $\mathcal{P}(\mathcal{O})$ and $\mathcal{P}(\mathcal{R})$ is defined as:

$$\mathcal{L}_4(\mathcal{R},\mathcal{O})=\mathcal{P}(\mathcal{O})\log\frac{\mathcal{P}(\mathcal{O})}{\mathcal{P}(\mathcal{R})} \tag{6}$$

If the probability distributions of the motions are modelled with a parameter set, the KL divergence can be found by calculating the mean probability of the data points in the motion sequences as

$$\mathcal{L}_{4,mean}\left(\mathcal{R},\mathcal{O}\middle|\lambda\right)=\frac{1}{MT_m}\sum_{m=1}^{M}\prod_{k=1}^{T_m}\mathcal{P}\left(\mathbf{o}_m^{(k)}\middle|\lambda\right)\cdot$$
$$\cdot\left(\frac{1}{MT_m}\sum_{m=1}^{M}\sum_{k=1}^{T_m}\log\mathcal{P}\left(\mathbf{o}_m^{(k)}\middle|\lambda\right)-\frac{1}{NT_n}\sum_{n=1}^{N}\sum_{k=1}^{T_n}\log\mathcal{P}\left(\mathbf{r}_n^{(k)}\middle|\lambda\right)\right)\quad(7)$$

This metric is also known as relative entropy, and is a measure of the lost information when the probability distribution $\mathcal{P}(\mathcal{R})$ is used to approximate the probability distribution $\mathcal{P}(\mathcal{O})$.

Other alternative metrics to the KL divergence that have been used to quantify the difference between two probability distribution and can be as well considered for evaluation of human motion consistency are the Hellinger distance and Bhattacharyya distance.

c) *Heuristic consistency*-is a simple qualitative measure that determines the proportion of patient movements that are contained within the extremums of the demonstrated movements $\mathcal{O}$. The measure is defined as:

$$\mathcal{L}_5(\mathcal{R},\mathcal{O})=\frac{1}{NT_n}\sum_{k=1}^{T_n}\left(1-\mathbf{1}_{\left\{\min\left(\mathbf{o}^{(k)}\right),\max\left(\mathbf{o}^{(k)}\right)\right\}}\left(\mathbf{r}_n^{(k)}\right)\right)\quad(8)$$

The indicator function $\mathbf{1}_{\left\{\min\left(\mathbf{o}^{(k)}\right),\max\left(\mathbf{o}^{(k)}\right)\right\}}\left(\mathbf{r}_n^{(k)}\right)$ evaluates to 1 if the captured trajectory data at time step k, $\left(\mathbf{r}_n^{(k)}\right)$ and otherwise the indicator function evaluates to 0. Higher values of the measure indicate increased consistency between the patient performed, and the prescribed movement examples. This metric may require a larger number of movement examples.

d) *Prediction intervals*-can be used to determine if the estimated means of the patient movements are consistent with the fitted model of PT's demonstrated movements. For this purpose, 95% confidence intervals from the relative likelihood are constructed, determined by the bounds

$$\ln\left(\mathcal{P}\left(\mathbf{o}^{(k)}\middle|\hat{\lambda}\right)\right)-\ln\left(\mathcal{P}\left(\mathbf{o}^{(k)}\middle|\lambda\right)\right)\le 1.92\,.\quad(9)$$

Next, the proportion of estimated means from the captured patient trajectory that is contained within the confidence interval is calculated, and averaged over all captured trajectories to obtain the metric $\mathcal{L}_6(\mathcal{R},\mathcal{O})$. If the captured trajectories are consistent with the demonstrated movements then $\mathcal{L}_6(\mathcal{R},\mathcal{O})$ should have a value of approximately 5%.

Qualitative metrics

Qualitative metrics can be referred to as high level metrics because they evaluate each patient's performed motion example as an individual repetition with respect to the prescribed motion examples, as opposed to evaluating the individual sequential measurements at the trajectory level.

The following metrics have been used for qualitative assessment of therapy exercises in previous works in the literature:

a) Number of optimal attempts-is used in the work of Komatireddy et al. [12] to assess patient performance. As stated before, it is not clear what type of approach the authors applied in labeling the motions as either optimal or suboptimal.

On the other hand, it is possible to use any of the quantitative approaches listed above to calculate a numerical score for each repetition of a motion, and then to label it as optimal if the score is greater than a predefined threshold value.

b) Fugl-Meyer assessment (FMA)-introduces a series of standardized exercises intended to evaluate the development of motor functions and balance in patients recovering from stroke [15]. The FMA test encompasses five principle domains for assessment: motor function, sensory function, balance, joint range of motion, and joint pain. Each domain involves several assessment steps related to the performance of respective movements. The movements are evaluated by a PT on a scale with 3 grades, with 0 as minimum and 2 as maximum grade. The assessment produces a cumulative numerical score representing the progression toward functional recovery of the stroke patient.

This assessment method can be employed in the development of metrics for automated performance evaluation, by either drawing insights from the PT evaluator's way of scoring the movements, or by training a machine learning algorithm to score in a similar manner by using PT's scores as inputs.

In addition, the FMA test has been reported to be complex and time consuming [16]. Consequently, an automated version of the test based on machine learning methodology could be a valuable contribution to the domain of physical rehabilitation. Another potential advantage of automated assessment is the provision of more precise evaluation than the three grades scale.

Several faster alternative tests to the FMA have been introduced, including the Motor Assessment Scale [16] and the motricity index [17]. These tests have been frequently used in practice, and can also be exploited in the development of an automated performance metric.

c) Wolf motor function test (WMFT)-is a timed test of functional tasks used to assess the ability of upper motor movements [18]. The test relies on using a number of objects as props, such as a chair, table, weights. The required motions are performed by using the props. The tasks are timed, with each motion given a maximum time of 2 minutes. The performance of each task is scored on a scale from 0 to 5. Summary scores are calculated based on the medians of the timings of the motions, and on the means of the ratings for the functional abilities.

Similar to the observation regarding the FMA test, WMFT is also suitable for automation and can provide understanding into the development of automated performance metrics.

Evaluation

Dataset

The proposed metrics were evaluated on the publically available dataset of human motion UTD-MHAD (University of Texas at Dallas – Multimodal Human Action Dataset) [24]. The dataset includes 27 actions, each performed by 8 subjects 4 times. A Kinect sensor and a wearable inertial sensor were used for collecting the data.

The following 5 actions were employed here for evaluation purposes: two hands front clap, right arm throw, draw circle clockwise, draw triangle, and tennis serve. Sample images for the actions are presented in Figure 1.

Evaluation Results

The following metrics were evaluated for the five actions: rot-mean square distance, log-likelihood, KL divergence, heuristic consistency, and prediction intervals. The results are presented in Table 1.

The data for the five actions was divided into 2 sets: a training

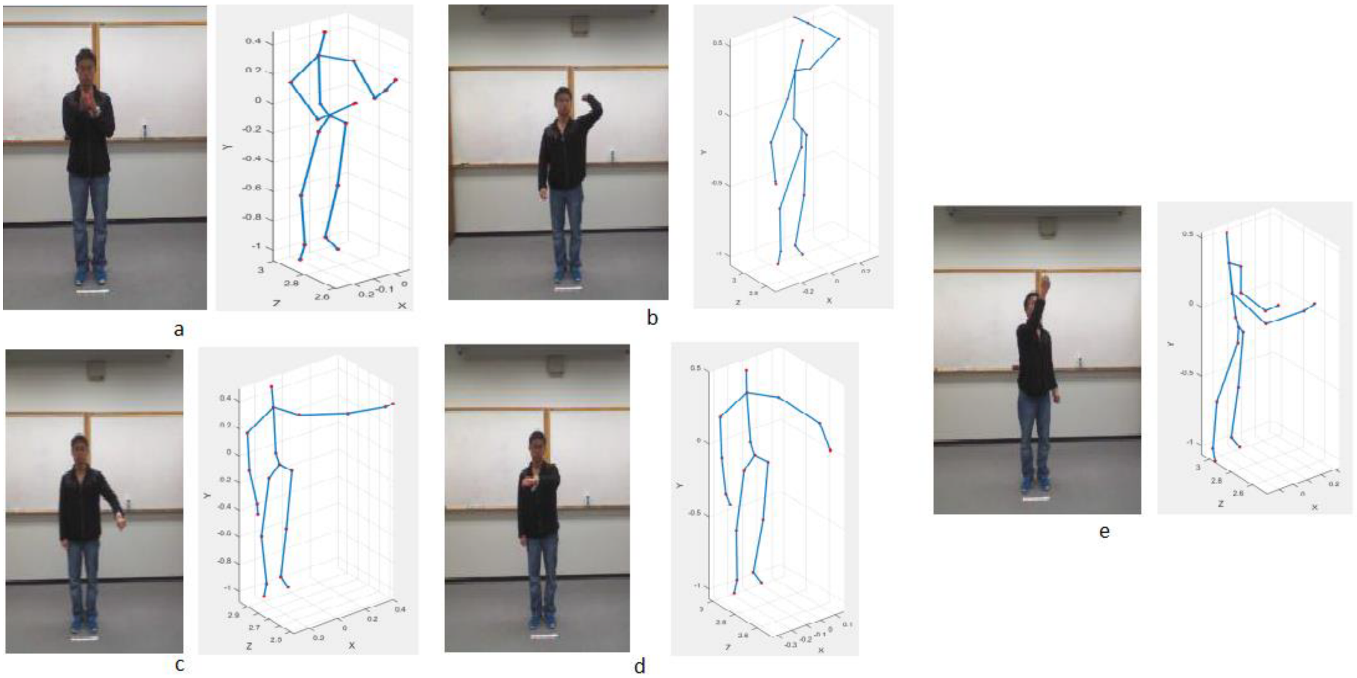

Figure 1: Sample images and skeletal representations for the selected actions in the UTD-MHAD dataset: (a) Two hands front clap; (b) Right arm throw; (c) Draw circle clockwise; (d) Draw triangle; and (e) Tennis serve.

	Action 4 - Two Hand Front Clap	Action 5 – Right Arm Throw	Action 9 – Draw Circle Clockwise	Action 11 – Draw Triangle	Action 17 – Tennis Serve
RMS	5.616 (0.139)	6.508 (0.172)	6.580 (0.144)	6.098 (0.230)	8.195 (0.208)
Log-likelihood: Autoenconder+Mixture Density Network (GMM)	1.707 (0.569)	0.968 (0.834)	0.488 (0.897)	0.788 (0.467)	0.823 (0.626)
Log-likelihood: Autoenconder+Expectation Maximization (GMM)	1.808 (3.526) 8 states	0.557 (3.481) 8 states	0.429 (5.189) 10 states	0.199 (10.567) 7 states	0.575 (2.999) 7 states
Log-likelihood: Autoenconder+Hidden Markov Model	-0.954 (0.066) 23 states	-0.731 (0.041) 28 states	-0.905 (0.017) 26 states	-1.459 (3.025) 30 states	-0.730 (0.015) 25 states
KL Divergence: Autoenconder+Mixture Density Network (GMM)	Train *vs.* test data: 1.232	Train *vs.* test data: 0.685	Train *vs.* test data: 2.617	Train *vs.* test data: 1.626	Train *vs.* test data: 0.709
Heuristic Consistency	0.1002	0.0972	0.0728	0.1053	0.0724
Prediction Intervals	0.0759	0.0966	0.0788	0.0837	0.0559

Table 1: Performance metrics for 5 action movements from the UTD-MHAD Dataset.

set consisting of 21 sequences for each action, and a testing dataset consisting of 7 sequences of each action. Both the training and the testing set correspond to actions performed by the same group of subjects. One may note that it is preferred the motions to correspond to therapy exercises, and the testing set to include suboptimal examples of the motions. As part of the future work, we have plans to create a dedicated dataset related to motions performed in physical therapy.

The root-mean square distance was calculated for the recorded trajectories. The motion capture feature of Kinect provides a skeletal data, where the human skeleton (shown in Figure 1) consists of 20 joints. The temporal measurements for each joint are spatial 3-dimensional coordinates. Hence, the data comprises 60 dimensional data sequences. The recorded motion sequences were scaled to a same number of measurements by using the DTW algorithm. The provided results in Table 1 present the mean values for the root-mean square distance for the 7 motion sequences in the testing dataset.

Log-likelihood of the testing data was calculated for several different mathematical models of the training data. The dimensionality of the raw observation data was first reduced from 60 to 3-dimensions, by employing an autoencoder neural network [25]. Afterwards, the 3-dimensional sequences were modeled using a mixture density network [14], Gaussian mixture model by employing expectation maximization, and a hidden Markov model [26]. The mean log-likelihood of the testing dataset is shown in Table 1.

The mean KL divergence of the testing data is also presented in Table 1. Similar to the log-likelihood metric, an autoencoder is employed to reduce the dimensionality of the observed data, and a mixture density network is afterwards used to model the data.

The last two columns in the table present the heuristic consistency and prediction intervals metrics.

Conclusion

The article presents a survey on the current literature on the metrics for evaluation of patient performance in physical therapy. The metrics are classified into quantitative and qualitative metrics. The quantitative metrics assign a numerical score for the patient performance, and are categorized into model-less and model-based metrics, based on whether a mathematical model of the motions is employed for performance evaluation.

The existing practice in physical therapy predominantly relies on assessment by a physical therapist. The studies related to automated assessment of therapy motions are scarce in the published literature, and consequently little attention has been paid to the development and definition of metrics for performance evaluation. This article reviews some of the reported metrics in the literature. In addition, the article reviews metrics that have been used for evaluation of human motions in other fields. Examples are root-mean square distance and norm of jerk, which have been used in the domain of robotic learning from human demonstrations. Other metrics, such as Kullback Leibler divergence, heuristic consistency, have been used in general for comparison of probability distributions.

The presented metrics in this article can be used for evaluation of human motions in other application domains, or also for assessment of sequential data in other fields, if applicable.

Acknowledgement

This work was supported by the Center for Modeling Complex Interactions through NIH Award #P20GM104420 with additional support from the University of Idaho.

References

1. Jack K, McLean SM, Moffett JK, Gardiner E (2010) Barriers to treatment adherence in physiotherapy outpatient clinics: a systematic review. Manual Therapy 15: 220–228.
2. De Fauw J, Keane P, Tomasev N, Visentin D, van der Driessche G, et al. (2016) Automated analysis of retinal imaging using machine learning techniques for computer vision. F1000 Res 5: 1573.
3. Chu C, De Fauw J, Tomasev N, Paredes BR, Hughes C, et al. (2016) Applying machine learning to automated segmentation of head and neck tumour volumes and organs at risk on radiotherapy planning CT and MRI scans. F1000 Res 5: 2104.
4. Jafari MH, Nasr-Esfahani E, Karimi N, Reza Soroushmehr SM, Samavi S, et al. (2016) Extraction of skin lesions from non-dermoscopic images using deep learning. arXiv: 1609.
5. Jermyn M, Desroches J, Mercier J, Tremblay MA, St-Arnaud K, et al. (2016) Neural networks improve brain cancer detection with Raman spectroscopy in the presence of operating room light artifacts. J Biomed Opt 21: 094002.
6. Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. Mol Pharmac 13: 1445-1454.
7. Ditzler G, Polikar R, Rosen G (2015) Multi Layer and recursive neural networks for metagenomic classification. IEEE Transactions on NanoBioscience 14: 608-616.
8. Hughes TB, Miller GP, Swamidass SJ (2015) Modeling epoxidation of drug-like molecules with a deep machine learning network. ACS Central Science 1: 168-180.
9. Shademan A, Decker RS, Opfermann JD, Leonard S, Krieger A, et al. (2016) Supervised autonomous robotic soft tissue surgery. Science Translational Medicine 8: 337-364.
10. AiCure – Advanced Medication Adherence.
11. Anton D, Goni A, Illarramendi A, Torres-Unda JJ, Seco J (2013) KiReS: A Kinect based telerehabilitation system. Int. Conf. on e-Health Networking, Applications and Services: 456-460.
12. Komatireddy R, Chokshi A, Basnett J, Casale M, Goble D, et al. (2016) Quality and quantity of rehabilitation exercises delivered by a 3-D motion controlled camera: a pilot study. Int J Phys Med Rehab 2: 1–14.
13. Hondori HM, Khademi M (2014) A review on technical and clinical impact of Microsoft Kinect on physical therapy and rehabilitation. J Med Eng: 1–16.
14. Vakanski A, Ferguson JM, Lee S (2017) Mathematical modeling and evaluation of human motions in physical therapy using mixture density neural networks. J Physiother Phys Rehabil 1: 1-10.
15. Fugl-Meyer AR, Jääskö L, Leyman I, Olsson S, Steglind S (1975) The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. Scandinavian J Rehab Med 7: 13–31.
16. Carr JH, Shepherd RB, Nordholm L, Lynne D (1985) Investigation of a new motor assessment scale for stroke patients. Phys Ther 65: 175-180.
17. Poole JL, Whitney SL (1988) Motor assessment scale for stroke patients: concurrent validity and interrater reliability. Arch Phys Med Rehabil 69: 195–197.
18. Wolf SL, Lecraw DE, Barton LA, Jann BB (1989) Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. Exp Neurol 104: 125-132.
19. Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-26: 43-49.
20. Calinon S, D'halluin F, Sauser EL, Caldwell DG, Billard AG (2010) Learning and reproduction of gestures by imitation: an approach based on hidden Markov model and Gaussian mixture regression. IEEE Robotics and Automation Magazine 17: 44-54.
21. Bishop CM (2006) Pattern Recognition and Machine Learning. New York, USA: Springer.
22. Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc of the IEEE 77: 257-286.
23. Kullback S, Leibler RA (1951) On information and sufficiency. Annals Math Stat 22: 79-86.
24. University of Texas at Dallas – Multimodal Human Action Dataset.
25. Bourlard H, Kamp Y (1998) Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics 59: 291-294.
26. Vakanski A, Mantegh I, Irish A, Janabi-Sharifi F (2012) Trajectory learning for robot programming by demonstration using Hidden Markov Model and Dynamic Time Warping. IEEE Transactions on Systems, Man, and Cybernetics 44: 1039-1052.