Towards Automated Performance Status Assessment: Temporal Alignment of Motion Skeleton Time Series

Tanachat Nilanon, Luciano P. Nocera, Jorge J. Nieva, and Cyrus Shahabi

Abstract Patient Performance Status (PS) is used in cancer medicine to predict prognosis and prescribe treatment. Today, PS assessments rely on assessor's observation, which is susceptible to biases. A motion tracking system can be used to supplement PS assessments, by recording and analyzing patient's movement as they perform a standardized mobility task e.g. getting up from office chair to sit on examination table. A temporal alignment of the extracted motion skeleton time series is then needed to enable comparison of corresponding motions in mobility task across recordings. In this paper, we apply existing state-of-the-art temporal alignment algorithms to the extracted time series and evaluate their performance in aligning the keyframes that separate corresponding motions. We then identify key characteristics of these time series that the existing algorithms are not able to exploit correctly: task left-right invariance and vertical-horizontal relative importance. We thus propose Invariant Weighted Dynamic Time Warping (IW-DTW), which takes advantage of these key characteristics. In an evaluation against state-of-the-art algorithms, IW-DTW outperforms them in aligning the keyframes where these key characteristics are present.

Keywords Motion skeleton time series · Temporal alignment · Performance status assessment

T. Nilanon (✉) · L. P. Nocera · C. Shahabi
University of Southern California, Los Angeles, CA 90089, USA
e-mail: nilanon@usc.edu

L. P. Nocera
e-mail: nocera@usc.edu

C. Shahabi
e-mail: shahabi@usc.edu

J. J. Nieva
USC Norris Comprehensive Cancer Center, Los Angeles, CA 90033, USA
e-mail: jorge.nieva@med.usc.edu

A. Shaban-Nejad et al. (eds.), *Explainable AI in Healthcare and Medicine*, Studies in Computational Intelligence 914,
https://doi.org/10.1007/978-3-030-53352-6_32

1 Background and Motivation

Advances in sensor electronics and computer vision have led to widespread availability of single-unit color+depth (RGB+D) motion tracking systems. While these systems do not have the same level of accuracy as traditional motion capture systems, their portability and ease-of-use have made them an attractive choice for use in various medical applications, including rehabilitation and home monitoring [12]. One such system is the Microsoft Kinect, whose usability has been broadly validated for many computer vision applications, including object detection, human pose, and action recognition [6].

Performance status (PS) assessment has been used in cancer medicine to identify patients with an increased risk of complications and poor outcomes. However, widely-utilized PS scales such as the Eastern Cooperative Oncology Group (ECOG) scale [7] rely on qualitative PS descriptions, making PS assessments susceptible to biases and inter-rater disagreements [9]. These susceptibilities could potentially be addressed by supplementing PS assessments with a motion tracking system. During a clinic visit, patients could be recorded performing a set of standardized mobility tasks, then a PS score could be calculated, potentially utilizing both the observation/interview and the recordings. .

To provide a concrete example, we first describe the ChairToExamTable task used in our study (see Fig. 1). To perform this task, a patient seated in a standard office chair is asked to walk over and use a stepper to get up and sit on an examination table. This standardized task was developed based on consultation with clinicians to include movements informative to PS assessment.

As a standardized mobility task generally has broad instructions and is performed differently by different patients, a temporal alignment between the recordings is needed to avoid comparing disparate motions across recordings. A temporal alignment between two recordings can be visualized as a mapping between each of their time steps, as in Fig. 1. While these temporal alignments could be annotated manually, the process is time-consuming and would again be susceptible to biases; hence, a systematic and automated approach is preferred.

Based on a comprehensive review of our Kinect recordings, we have identified the following key characteristics of standardized-mobility-task motion skeleton time series that adversely affect state-of-the-art temporal alignment algorithms:

- **Task Left-right Invariance**: Different patients can start walking with different foot (left/right) first. This should be taken into account when generating temporal alignment. An algorithm that is not invariant to this task characteristic could generate an incorrect temporal alignment (see Fig. 1).
- **Vertical-horizontal Relative Importance**: In a typical 3-dimensional Cartesian coordinate system, two axes represent the horizontal plane and one axis represents the vertical direction. However, for a standardized mobility task, the vertical position is more relevant to PS assessments and should be given more weight than

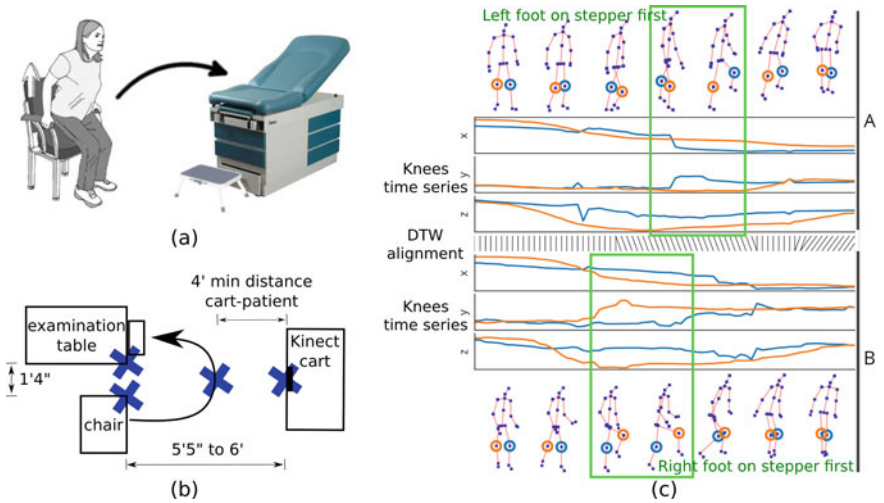

Fig. 1 **(a)** ChairToExamTable task illustration. **(b)** ChairToExamTable task set-up diagram. The small rectangle next to the examination table depicts the stepper. **(c)** Patients A and B stepping on the stepper to get up on the examination table; A steps with their left foot first while B steps with their right foot first. DTW is unable to correctly align the corresponding motions

the horizontal position, as movements in vertical direction involve changes in the subject's potential energy and require larger energy expenditure than movements in the horizontal plane.

In this paper, we propose an unsupervised temporal alignment algorithm for motion skeleton time series, designed to exploit the key characteristics above. Our method is built on DTW due to its robustness and strong baseline results.

2 Related Works

Temporal alignment of time series has rich literature, starting with Dynamic Time Warping (DTW) [8], which computes an optimal warping path whose cost between time steps is their Euclidean distance. While DTW has wide applications in various domains, its ability is limited in handling time series of different view or high dimensionality since it operates directly in the observation space.

In [14], Canonical Time Warping (CTW) was proposed for temporal alignment of human behavior, where Canonical Component Analysis (CCA) was integrated with DTW to enable linear spatial transformation of features alongside the computation of the warping path. In Generalized Canonical Time Warping (GCTW) [13], warping path basis functions were further introduced to reduce the search space for the optimal warping path, significantly reducing its time complexity.

Instead of using CCA to spatially align the features, Manifold Warping (MW) [11] integrates manifold learning with DTW to enable non-linear feature transformation. Recent methods continue to carry on this trend. In [10], Deep Canonical Time Warping (DCTW) was proposed, utilizing fully-connected layers to enable hierarchical non-linear transformation.

Departing from feature transformation, Autowarp was proposed in [1]. Autowarp uses sequence-to-sequence model to embed the time series in low-dimensional space and then utilizes the Euclidean distance in that space to guide DTW.

On the clinical application side, Wang et al. [12] proposed an algorithm to segment, align, and summarize motion skeleton time series. Their algorithm relies on the repetitiveness of the task motion as it first segments the time series by repetitive motion (e.g., walking two steps). Then all the segments are temporally-aligned for summarization. In [3], Hasnain et al. proposed to compute kinematics features such as velocity and acceleration as input for DTW distance computation; these distances are then used as features for subsequent analysis.

We note that none of these approaches can take advantage of the key characteristics of motion skeleton time series described in Sect. 1.

3 Methods

3.1 Multivariate Time Series and Temporal Alignment

The extracted motion skeletons can be represented as time series. We define a d-dimensional *time series* T of length N, $T \in \mathbb{R}^{d \times N}$: $T = (t_1, t_2, \ldots, t_N)$ where each $t_i \in \mathbb{R}^d$ is a d-dimensional vector. We then define a temporal alignment $\mathcal{A}$ between two time series $T_x \in \mathbb{R}^{d \times N_x}$ and $T_y \in \mathbb{R}^{d \times N_y}$ as a sequence of ordered pairs: $\mathcal{A}(T_x, T_y) = ((a_{x1}, a_{y1}), (a_{x2}, a_{y2}), \ldots, (a_{xL}, a_{yL}))$ where L is the length of the alignment, $a_{xi} \in \{1, 2, \ldots, N_x\}$, $a_{yi} \in \{1, 2, \ldots, N_y\}$, and the pair (a_{xi}, a_{yi}) denotes that $t_{xa_{xi}}$ is aligned with $t_{ya_{yi}}$. The alignment path is constrained to satisfy the boundary, monotonicity, and continuity conditions: $(a_{x1}, a_{y1}) = (1, 1)$, $(a_{xL}, a_{yL}) = (N_x, N_y)$, $a_{xi} \leq a_{x(i+1)} \leq a_{xi} + 1$, and $a_{yi} \leq a_{y(i+1)} \leq a_{yi} + 1$.

3.2 Dynamic Time Warping

Let $\mathcal{D}(T_x, T_y)_{i,j}$ be the distance between t_{xi} and t_{yj}. For Euclidean distance, $\mathcal{D}(T_x, T_y)_{i,j} = \sqrt{\sum_{k=1}^{d} (t_{xik} - t_{yjk})^2}$. DTW optimizes for a temporal alignment (warping) path $\mathcal{A}(T_x, T_y)$ where the total cost $\sum_{i=1}^{L} \mathcal{D}(T_x, T_y)_{a_{xi}, a_{yi}}$ is minimized. This cost can be described as a recurrence relation:

$$C(i,j) = \begin{cases} \mathcal{D}(T_x, T_y)_{i,j} & \text{if } i \text{ or } j = 1 \\ \mathcal{D}(T_x, T_y)_{i,j} + \min \begin{pmatrix} C(i-1,j) \\ C(i,j-1) \\ C(i-1,j-1) \end{pmatrix} & \text{otherwise} \end{cases}$$

where $C \in \mathbb{R}^{N_x \times N_y}$ is the accumulated cost array and the total cost is $C(N_x, N_y)$.

Our method is built on DTW due to its robustness and strong baselines. In the following sections, we explain how the key characteristics can be exploited.

3.3 Handling Task Left-Right Invariance

In a mobility task, patients are not usually instructed on which side of their body they have to move first. For example, they can start walking with their left or right foot first; or when instructed to step on a stepper, they can put their left or right foot on the stepper first. To construct a temporal alignment algorithm that is indifferent to this left vs. right variations, we extend the DTW recurrence relation with a separate cost array for each task-equivalent *feature mapping* and allow the alignment path to switch between these mappings.

In our application, the mobility task suggests that we should be indifferent to any switching between the left/right side of the lower/upper body. Let $S = \{O, A, L, AL\}$ contains these feature mappings to which our algorithm should be indifferent (see Fig. 2b for illustration). The recurrence relation $C_s, \forall s \in S$ can then be written as:

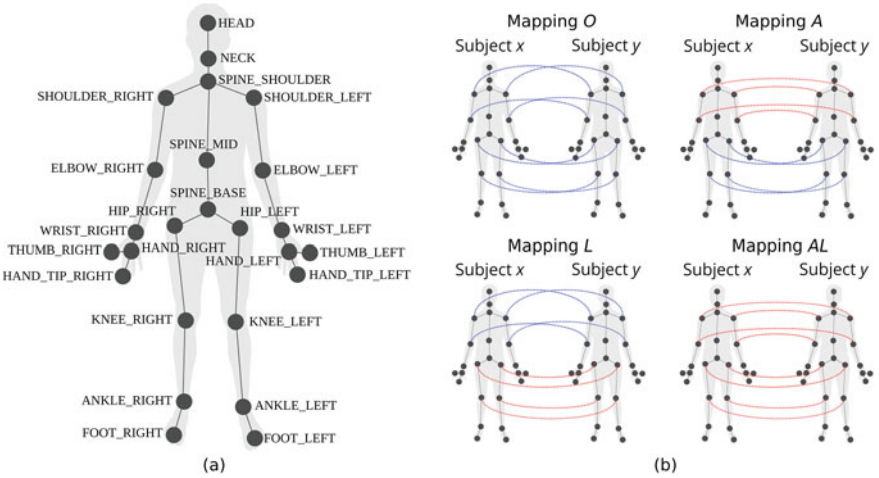

Fig. 2 **(a)** Diagram of skeleton anatomical node (n = 25) extracted from recordings by Kinect SDK v2.0. **(b)** Task-equivalent mappings between the left/right side of the lower/upper body. Some lines are omitted for clarity

$$C_s(i,j) = \begin{cases} \mathcal{D}(\mathcal{F}_s(T_x), T_y)_{i,j} & \text{if } i \text{ or } j = 1 \\ \mathcal{D}(\mathcal{F}_s(T_x), T_y)_{i,j} + \min\limits_{t \in S} \left(\begin{array}{c} \mathcal{D}(\mathcal{F}_s(T_x), \mathcal{F}_t(T_x))_{i,i} \\ + \\ \min \left(\begin{array}{c} C_t(i-1,j) \\ C_t(i,j-1) \\ C_t(i-1,j-1) \end{array} \right) \end{array} \right) & \text{otherwise} \end{cases}$$

where $\mathcal{F}_s$ denotes a transformation that maps according to s, and C_s denotes the accumulated cost array for mapping s. Note that $\mathcal{D}(\mathcal{F}_s(T_x), \mathcal{F}_t(T_x))_{i,i}$ is the cost of switching the warping path between task-equivalent mappings s and t.

To approximate $\mathcal{F}_s$, we can *reflect* the nodes between the left and the right side of the body. We define the *SPINE_BASE reference frame* to be centered at SPINE_BASE (see Fig. 2a for diagram), where its x-axis is the horizontal component of the vector pointing from HIP_RIGHT to HIP_LEFT, its y-axis is the upward vertical vector, and its z-axis is generated by the right-hand rule. After a transformation into this reference frame, a node position on one side of the body (x, y, z) can be reflected to another side of the body as $(-x, y, z)$. Then the node position can be transformed back into the global reference frame. To minimize errors, this SPINE_BASE reference frame is used for the lower body while the *SPINE_SHOULDER reference frame* (defined similarly) is used for the upper body.

3.4 Handling Vertical-Horizontal Relative Importance

In a standardized mobility task, the vertical position of the subject is more relevant to PS assessments than the horizontal position, as movements in vertical direction involve changes in potential energy and require larger energy expenditure than movements with constant elevation. We can address this in a framework of weighted DTW. For Euclidean distance, this can be written as: $\mathcal{D}(T_x, T_y)_{i,j} = \sqrt{\sum_{k=1}^{d} w_k (t_{xik} - t_{yjk})^2}$ where w_k is the weight for feature k. In our case, we set w_k to w_y if feature k is in the vertical direction and 1 otherwise.

Our method combines the two extensions to DTW described in Subsects. 3.3 and 3.4. We name it Invariant Weighted DTW (IW-DTW).

4 Experiments

4.1 Dataset

Our dataset was derived from the Analytical Tools to Objectively Measure Human Performance (ATOM-HP) clinical project, which will be partially described here.

To examine the feasibility of using a motion tracking system, a wearable activity tracker, and a set of patient-reported outcome questionnaires in PS assessment, a multi-center observational clinical study was performed with a population of cancer patients. Participants were scheduled to come in for a clinic visit twice during the study period, during which ECOG scores and Kinect recordings were collected for the ChairToExamTable task [2–5].

After exclusion of partial, noisy, and nonconforming-to-instruction recordings, 64 recordings remain. The recordings have the minimum, median, and maximum length of 110, 275, and 665 frames, respectively. The motion skeleton time series are then extracted from the recordings using Kinect SDK v2.0 and have a sampling rate of 30 Hz. Each frame originally contains 75 features ((x, y, z) of the 25 skeleton nodes); however, the extracted `WRIST`, `HAND`, `THUMB`, `HAND_TIP`, `ANKLE`, and `FOOT` nodes are found to be unreliable and therefore excluded from analysis. Thus, each frame is left with 13 skeleton nodes.

4.2 Evaluation Metrics

All Kinect recordings were annotated by annotators reviewing a visualization of the extracted motion skeleton time series. We annotated 4 keyframes important for mobility task movement analysis. The first keyframe is defined to be when the subject starts building momentum to stand. The second keyframe is defined to be when the subject starts moving one foot to walk. The third keyframe is define to be when the subject has one foot on the stepper and starts an upward motion towards the examination table. Finally, the fourth keyframe is defined to be when the subject is fully seated. Note that these 4 keyframes divide the task into 3 subtasks. Kinematics features such as accelerations and timings of these subtasks are of interest from a clinical standpoint e.g. the first subtask contains the subject's motions standing up from seated.

We posit that for any analysis performed on motion skeleton time series to be justified, it should never compare a motion in one subtask to another motion in a different subtask. Consequently, the evaluation metrics we use will reflect this. For a pair of motion skeleton time series T_x and T_y, let a computed temporal alignment between them be $\mathcal{A}(T_x, T_y)$ and let their annotated keyframes be S_{xj} and S_{yj} for $j \in \{1, 2, 3, 4\}$. We define keyframes estimated using the computed temporal alignment as $\widehat{S}_{xj} = \text{mean}(\{a \mid (a, S_{yj}) \in \mathcal{A}(T_x, T_y)\})$ and $\widehat{S}_{yj} = \text{mean}(\{a \mid (S_{xj}, a) \in \mathcal{A}(T_x, T_y)\})$. We then define *Keyframe Mean Absolute Error (MAE)* as $\text{mean}\left(|\widehat{S}_{xj} - S_{xj}|, |\widehat{S}_{yj} - S_{yj}|\right)$.

Overall MAE is then calculated as the average over all possible pairs of time series. To further observe how task left-right invariance affects the performance of a temporal alignment algorithm, we also compute the MAE averaged over pairs where subjects moved different foot first in keyframes 2 and 3.

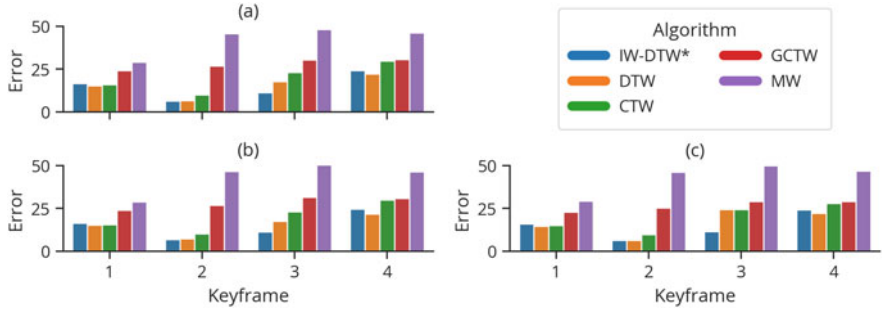

Fig. 3 Evaluation results (MAE) averaged over **(a)** all pairs of recordings and pairs where different foot was moved first in **(b)** keyframe 2 and **(c)** 3 . Task left-right invariance manifests in keyframes 2 (starts walking) and 3 (first foot on stepper)

5 Results

We evaluated IW-DTW against DTW, CTW, GCTW, and MW (see Fig. 3). Autowarp and DCTW were excluded due to poor convergence during model training.

Averaged over all possible pairs of motion skeleton time series, IW-DTW performed better than all baselines in aligning keyframes 2 and 3, where the alignment needs to handle task left-right invariance, but performed slightly worse in aligning keyframes 1 and 4, where task left-right invariance does not manifest. For keyframe 1 (starts standing up), MAE of IW-DTW (16.5) is slightly worse than the best baseline DTW (15.3). For keyframe 2 (starts walking), MAE of IW-DTW (6.4) is similar to the best baseline DTW (6.5). For keyframe 3 (uses stepper to push self up), MAE of IW-DTW (11.2) is distinctly better than the best baseline DTW (17.6). For keyframe 4 (seated on examination table), MAE of IW-DTW (24.2) is slightly worse than the best baseline DTW (22.1).

Restricting our comparison to the pairs of time series where the subjects move different foot first in keyframe 2, keyframe 2 MAE of IW-DTW (6.8) is slightly better than the best baseline DTW (7.2). As for keyframe 3, keyframe 3 MAE of IW-DTW (11.6) is better than the best baseline DTW (24.4).

6 Discussion

IW-DTW outperformed state-of-the-art algorithms in aligning keyframes when task left-right invariance clearly manifests, such as starting to walk or putting one foot on a stepper. However, when task left-right invariance does not manifest, such as getting up or being fully seated, IW-DTW performed slightly worse. We hypothesize that this is because the weighting mechanism is not expressive enough to capture local changes in relative axes importance. For example, when the subject starts getting up,

the most importance axes for alignment would be the horizontal axes (note how the body needs to bend forward before any significant vertical movement can occur). Our learned w_y (4.25) weight for the vertical direction impedes IW-DTW in correctly aligning these keyframes.

Deep learning methods hold promise in being able to extract relevant representations and learn directly from the data. However, we observed that for our use case, long sequence length still proved to be a big impediment for gradient-based optimization used in deep learning. In Autowarp [1], the authors have successfully employed their model for 4-dimensional sequences of median length 53; however, it failed to converge when training on our 39-dimensional sequences of median length 275. In DCTW [10], their experiment that is most similar to ours also has median length $<$100. Downsampling could potentially make Autowarp work for our case, but key movements such as standing up can last less than 15 frames and could be inadvertently filtered out by downsampling.

Online learning also holds promise in learning a classification task as new data comes in. However, the small number of samples ($n = 64$) make our task a poor fit for online learning. Furthermore, as our goal is to supplement traditional PS assessment, we prefer an algorithm that does not change its output with new data, so that its efficacy can be studied and validated in a randomized controlled trial.

Acknowledgements This research has been funded in part by US National Cancer Institute under award number P30CA014089, USC Integrated Media Systems Center (IMSC), and unrestricted cash gifts from Oracle, Microsoft, and Google. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the sponsors. T. Nilanon was also supported in part by DPST, IPST, Thailand.

References

1. Abid, A., Zou, J.: Autowarp: learning a warping distance from unlabeled time series using sequence autoencoders. In: Advances in Neural Information Processing Systems (NeurIPS), October 2018
2. Broderick, J.E., May, M., Schwartz, J.E., Li, M., Mejia, A., Nocera, L., Kolatkar, A., Ueno, N.T., Yennu, S., Lee, J.S.H., Hanlon, S.E., Cozzens Philips, F.A., Shahabi, C., Kuhn, P., Nieva, J.: Patient reported outcomes can improve performance status assessment: a pilot study. J. Patient-Reported Outcomes **3**(1), 41 (2019)
3. Hasnain, Z., Li, M., Dorff, T., Quinn, D., Ueno, N.T., Yennu, S., Kolatkar, A., Shahabi, C., Nocera, L., Nieva, J., Kuhn, P., Newton, P.K.: Low-dimensional dynamical characterization of human performance of cancer patients using motion data. Clinical Biomech. **56**(December 2017), 61–69 (2018)
4. Kao, J.Y., Nguyen, M., Nocera, L., Shahabi, C., Ortega, A., Winstein, C., Sorkhoh, I., Chung, Y.C., Chen, Y.A., Bacon, H.: Validation of Automated Mobility Assessment Using a Single 3D Sensor. In: Hua, G., Jégou, H. (eds.) European Conference on Computer Vision (ECCV) Workshops, vol. 3, pp. 162–177. Springer, Cham (2016)
5. Nguyen, M.N.B., Hasnain, Z., Li, M., Dorff, T., Quinn, D., Purushotham, S., Nocera, L., Newton, P.K., Kuhn, P., Nieva, J., Shahabi, C.: Mining Human Mobility to Quantify Performance Status. In: IEEE International Conference on Data Mining (ICDM) Workshop (2017)

6. Ofli, F., Chaudhry, R., Kurillo, G.: Berkeley Multimodal Human Action Database (MHAD). In: IEEE Workshop on Applications of Computer Vision (WACV), pp. 53–60 (2013)
7. Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., Carbone, P.P.: Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am. J. Clin. Oncol. **5**(6), 649–656 (1982)
8. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall Inc., Upper Saddle River (1993)
9. Schnadig, I.D., Fromme, E.K., Loprinzi, C.L., Sloan, J.A., Mori, M., Li, H., Beer, T.M.: Patient-physician disagreement regarding performance status is associated with worse survivorship in patients with advanced cancer. Cancer **113**(8), 2205–2214 (2008)
10. Trigeorgis, G., Nicolaou, M.A., Zafeiriou, S., Schuller, B.W.: Deep canonical time warping. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5110–5118 (2016)
11. Vu, H.T., Carey, C.J., Mahadevan, S.: Manifold warping: Manifold alignment over time. Proceedings of the National Conference on Artificial Intelligence **2**, 1155–1161 (2012)
12. Wang, R., Medioni, G., Winstein, C.J., Blanco, C.: Home monitoring musculo-skeletal disorders with a single 3D sensor. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 521–528 (2013)
13. Zhou, F., De La Torre, F.: Generalized canonical time warping. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 279–294 (2016)
14. Zhou, F., de la Torre, F.: Canonical time warping for alignment of human behavior. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 1–9 (2009)