Generalized Time Warping for Multi-modal Alignment of Human Motion

Feng Zhou **Fernando De la Torre**
Robotics Institute, Carnegie Mellon University
www.f-zhou.com ftorre@cs.cmu.edu

Abstract

*Temporal alignment of human motion has been a topic of recent interest due to its applications in animation, tele-rehabilitation and activity recognition among others. This paper presents generalized time warping (GTW), an extension of dynamic time warping (DTW) for temporally aligning multi-modal sequences from multiple subjects performing similar activities. GTW solves three major drawbacks of existing approaches based on DTW: (1) GTW provides a feature weighting layer to adapt different modalities (*e.g.*, video and motion capture data), (2) GTW extends DTW by allowing a more flexible time warping as combination of monotonic functions, (3) unlike DTW that typically incurs in quadratic cost, GTW has linear complexity. Experimental results demonstrate that GTW can efficiently solve the multi-modal temporal alignment problem and outperforms state-of-the-art DTW methods for temporal alignment of time series within the same modality.*

1. Introduction

Alignment of time series is an important unsolved problem in many scientific disciplines. Some applications include speech recognition [23], curve matching [29], chromatographic and micro-array data analysis [18], activity recognition [14], temporal segmentation [35] and synthesis of human motion [13, 22]. In particular, alignment of human motion has recently received increasing attention in computer vision and computer graphics. Major challenges for an accurate temporal alignment of human motion include modeling the difference in subjects' physical characteristics, view point changes, motion style and speed of the action [30, 31]. Unlike existing work, this paper addresses the challenging problem of multi-modal alignment of time-series coming from different sensors where subjects are performing a similar activity. For instance, consider the problem illustrated in Fig. 1. How can we solve for the temporal correspondence between the frames of a video, the samples of motion capture data, and the accelerometer signal from different people kicking a ball?

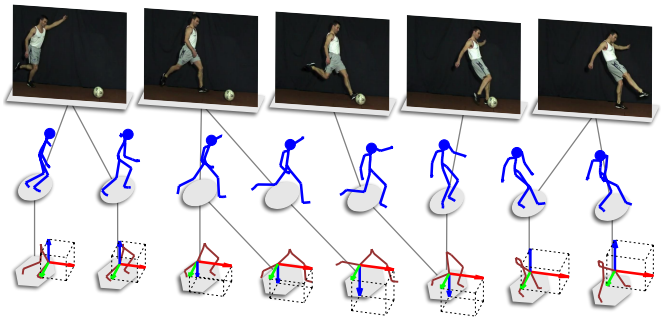

Figure 1. Temporal alignment of three sequences of different subjects kicking a ball recorded with different sensors (top row video, middle row motion capture and bottom row accelerometers).

Our work is motivated by recent success in extending dynamic time warping (DTW) for aligning human behavior. Zhou and De la Torre [34] proposed canonical time warping (CTW). CTW combines DTW with canonical correlation analysis to temporally align data of different dimensionality (*e.g.*, motion capture and video). More recently, Gong and Mendioni [7] proposed dynamic manifold warping (DMW) that extends CTW to incorporate more complex spatial transformations through manifold learning. However, CTW and DMW have three main limitations due to reliance in DTW: (1) Their computational complexity is quadratic in space and time; (2) They address the problem of aligning two sequences, and it is unclear how to extend it to the alignment of multiple sequences; (3) They compute the temporal alignment using DTW, which relies on dynamic programming to find the optimal path; however, it is unclear how to adaptively constrain the temporal warping. To overcome these limitations, this paper proposes generalized time warping (GTW), which allows an efficient and flexible alignment between two or more multi-dimensional time series of different modalities. GTW uses multi-set canonical correlation analysis to find the spatial transformations, and extends DTW by parameterizing the temporal warping as a combination of monotonic basis functions. Unlike existing DTW approaches based on dynamic pro-

gramming that usually have quadratic cost, GTW uses a Gauss-Newton algorithm that has linear complexity in the length of the sequence. Moreover, GTW allows to align several multi-modal time series.

The remaining of the paper is organized as follows. Section 2 reviews previous work on temporal alignment. Section 3 reviews previous work on DTW. Section 4 describes GTW. Section 5 illustrates the benefits of GTW on synthetic and real data.

2. Previous work

This section reviews previous work on temporal alignment of human motion. In particular, we discuss the challenges of aligning human motion from sensory data in the context of computer graphics and computer vision.

In the computer graphics literature, time warping of motion capture data has been a key component in many animation systems [32, 16]. However, most existing techniques are challenged when applied to placing stylistically different motions into correspondence. To account for the variations of human motion performed by different subjects, one popular strategy is to augument DTW with certain regression models. For instance, Hsu *et al*. [13] proposed to combine DTW with a space warping step, in which each individual degree of freedom from motion capture data can be scaled and translated. In [12], a weighted PCA algorithm [6] was used to find a low dimensional embedding such that the stylistic part of gesture sequences can be removed. Although these methods yield promising alignment results for motion capture data, they have several limitations when the extracted features in the sequences are very noisy (*e.g*., video) or come from different modalities (*e.g*., video and motion capture).

In the computer vision literature, a challenge in sequence alignment is to build view-invariant representations. In a multi-camera setting, it has been shown that both 2-D homography and 3-D epipolar geometries can form a powerful cue for alignment of two or more sequences. For instance, homography-based constraints [1, 2, 21] have been shown to be useful to align sequences in a planar scene. In addition, the fundamental matrix [25, 10] can be used to guide DTW to eliminate the distortion generated by the projection from 3D to 2D. Li and Chellappa [17] proposed a general framework for video alignment by optimizing various 2-D and 3-D constraints on a Riemannian manifold. Recent work [3] also illustrated the stability of the self-similarity matrix of actions under view changes. Built upon this observation, Junejo *et al*. [14] proposed a view-independent descriptor for video alignment using DTW. Observe that most existing works rely on certain explicit or implicit estimation of the underlying camera geometry. Unlike these works, GTW is able to efficiently align semantically similar multi-modal sequences.

3. Dynamic time warping

Given two time series[1], $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}] \in \mathbb{R}^{d \times n_x}$ and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_{n_y}] \in \mathbb{R}^{d \times n_y}$, dynamic time warping (DTW) [23] is a technique to align $\mathbf{X}$ and $\mathbf{Y}$ such that the following sum-of-square cost error is minimized [34]:

$$J_{dtw}(\mathbf{W}_x, \mathbf{W}_y) = \|\mathbf{X}\mathbf{W}_x - \mathbf{Y}\mathbf{W}_y\|_F^2, \tag{1}$$

where $\mathbf{W}_x = \mathbf{W}(\mathbf{p}_x) \in \{0,1\}^{n_x \times l}$ and $\mathbf{W}_y = \mathbf{W}(\mathbf{p}_y) \in \{0,1\}^{n_y \times l}$ are binary replication matrices (*i.e*., only replicate columns) associated with the warping paths ($\mathbf{p}_x$ and $\mathbf{p}_y$) by a non-linear mapping, $\mathbf{W}(\mathbf{p}) : \{1 : n\}^l \rightarrow \{0,1\}^{n \times l}$, which sets $w_{p_t,t} = 1$ for $t \in \{1 : l\}$ and zero otherwise. $l \geq \max(n_x, n_y)$ is the number of steps needed to align both signals. Recall that the optimal l is automatically selected by the DTW algorithm. The warping paths, $\mathbf{p}_x \in \{1 : n_x\}^l$ and $\mathbf{p}_y \in \{1 : n_y\}^l$, denote the correspondence indexes between frames. For instance, the i^{th} frame in $\mathbf{X}$ and the j^{th} frame in $\mathbf{Y}$ are aligned iff there exists $p_t^x = i$ and $p_t^y = j$ for some t.

In order to find a polynomial solution, the warping paths ($\mathbf{p}_x$ and $\mathbf{p}_y$) have to satisfy three constraints: (1) **Boundary conditions:** $[p_1^x, p_1^y] = [1, 1]$ and $[p_l^x, p_l^y] = [n_x, n_y]$. (2) **Monotonicity:** $t_1 \geq t_2 \Rightarrow p_{t_1}^x \geq p_{t_2}^x$ and $p_{t_1}^y \geq p_{t_2}^y$. (3) **Continuity:** $[p_t^x, p_t^y] - [p_{t-1}^x, p_{t-1}^y] \in \{[0,1], [1,0], [1,1]\}$. Notice that the choice of step size is not unique. For instance, replacing the step size by $\{[2,1], [1,2], [1,1]\}$ can avoid the degenerated case in which a single frame of one sequence may be assigned to many consecutive frames in the other sequence. See [23] for an extensive review on several DTW's modifications to control the warping paths.

4. Generalized time warping

Generally speaking, there are three major limitations of using DTW to align multi-modal and multi-dimensional time series: (1) DTW relies on dynamic programming (DP) to exhaustively search over all possible warping paths. This search has quadratic computational complexity ($O(n_x n_y)$) in both time and space. This might be restrictive when applying DTW to aligning long sequences. (2) A direct extension of DTW to align more than two sequences is usually infeasible due to the combinatorial explosion of possible warping paths. For instance, a DP-based alignment of m

[1] Bold capital letters denote a matrix $\mathbf{X}$, bold lower-case letters a column vector $\mathbf{x}$. $\mathbf{x}_i$ and $\mathbf{x}^{(i)}$ represent the i^{th} column and i^{th} row of the matrix $\mathbf{X}$ respectively. x_{ij} denotes the scalar in the i^{th} row and j^{th} column of the matrix $\mathbf{X}$. All non-bold letters represent scalars. $\mathbf{1}_{m \times n}, \mathbf{0}_{m \times n} \in \mathbb{R}^{m \times n}$ are matrices of ones and zeros. $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ is an identity matrix. $\|\mathbf{x}\|_p = \sqrt[p]{\sum |x_i|^p}$ denotes the p-norm. $\|\mathbf{X}\|_F^2 = \mathrm{tr}(\mathbf{X}^T\mathbf{X})$ designates the Frobenious norm. $\mathrm{vec}(\mathbf{X})$ denotes the vectorization of matrix $\mathbf{X}$. $\mathbf{X} \circ \mathbf{Y}$ is the Hadamard product of matrices. $\{i : j\}$ lists the integers, $\{i, i+1, \cdots, j-1, j\}$. $[\Rightarrow_i \mathbf{A}_i]$, $[\Downarrow_i \mathbf{A}_i]$, $[\searrow_i \mathbf{A}_i]$ are the horizontal, vertical, diagonal concatenation respectively. $\ominus$ denotes the titled minus, *e.g*., $\mathbf{A}_{6\times2} \ominus \mathbf{B}_{2\times2} = \mathbf{A}_{6\times2} - (\mathbf{1}_{3\times1} \otimes \mathbf{B}_{2\times2})$.

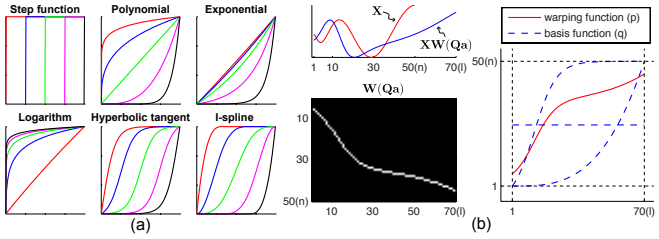

Figure 2. Temporal warping function. (a) Six common choices for monotonically increasing function $\acute{\mathbf{q}}$. (b) An example of time warping $\mathbf{XW}(\mathbf{Qa}) \in \mathbb{R}^{1\times 70}$ of 1-D time series $\mathbf{X} \in \mathbb{R}^{1\times 50}$. The warping function is a linear combination of three basis functions including a constant function and two monotonically increasing functions.

sequences incurs a complexity of $O(\prod_{i=1}^{m} n_i)$ in both time and space. (3) DTW lacks a feature weighting mechanism.

To address these issues, this section proposes generalized time warping (GTW), a technique for efficient spatio-temporal alignment of multiple time series. To accommodate for subject variability and to take into account the difference in the dimensionality of the signals, GTW uses multi-set canonical correlation analysis. To compensate for temporal changes, GTW extends DTW by incorporating a more flexible temporal warping parameterized by a set of monotonic basis functions. Unlike existing approaches based on DP with quadratic complexity, GTW efficiently optimizes the time warping function using a Gauss-Newton algorithm, which has linear complexity in the length of the sequence.

4.1. Objective function

Given a collection of m time series, $\{\mathbf{X}_1, \cdots, \mathbf{X}_m\}$, where $\mathbf{X}_i = [\mathbf{x}_1^i, \cdots, \mathbf{x}_{n_i}^i] \in \mathbb{R}^{d_i \times n_i}$. GTW finds for each $\mathbf{X}_i$, a non-linear temporal transformation $\mathbf{W}_i \in \{0,1\}^{n_i \times l}$ and a low-dimensional spatial embedding $\mathbf{V}_i \in \mathbb{R}^{d_i \times d}$, such that the resulting sequence $\mathbf{V}_i^T \mathbf{X}_i \mathbf{W}_i \in \mathbb{R}^{d \times l}$ is well aligned with the others in the least-squares sense. In a nutshell, GTW minimizes the sum of pairwise distances:

$$J_{gtw}(\{\mathbf{W}_i, \mathbf{V}_i\}) = \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{1}{2}\|\mathbf{V}_i^T\mathbf{X}_i\mathbf{W}_i - \mathbf{V}_j^T\mathbf{X}_j\mathbf{W}_j\|_F^2 + \Big(\sum_{i=1}^{m} \psi(\mathbf{W}_i) + \phi(\mathbf{V}_i)\Big), \quad (2)$$

$$\text{s.t.} \quad \mathbf{W}_i \in \Psi \text{ and } \mathbf{V}_i \in \Phi, \quad \forall i \in \{1:m\},$$

where $\psi(\cdot)$ and $\phi(\cdot)$ are regularization functions, which bias the solution in the space of temporal transformations $\mathbf{W}_i$ and the embedding for spatial transformation $\mathbf{V}_i$, respectively. Ψ and Φ represent the domains for $\mathbf{W}_i$ and $\mathbf{V}_i$. The explicit form for Ψ and Φ will be discussed in the following sections.

Generally speaking, optimizing J_{gtw} (Eq. 2) is a non-convex optimization problem with respect to the alignment ($\mathbf{W}_i$) and projection matrices ($\mathbf{V}_i$). We alternate between solving for $\mathbf{W}_i$ using a Gauss-Newton algorithm, and optimally computing $\mathbf{V}_i$ using mCCA. These steps monotonically decrease J_{gtw}, and because the function is bounded below the alternating scheme will converge to a critical point.

4.2. Parameterization of the temporal warping

To simplify the discussion, let's consider the temporal warping matrix $\mathbf{W} \in \{0,1\}^{n\times l}$ for a single sequence $\mathbf{X} \in \mathbb{R}^{d\times n}$. The DP-based approach to optimize $\mathbf{W}$ has a computational cost of $O(nl)$, which quickly becomes infeasible as the sequence length increases. In order to reduce the computational complexity and to provide a flexible way to control the warping path, GTW approximates the warping path $\mathbf{p} \in \{1:n\}^l$, which parameterizes the warping matrix $\mathbf{W}(\mathbf{p})$, as a linear combination of monotonic functions $\mathbf{q} \in [1,n]^l$, that is:

$$\mathbf{p} \approx \sum_{\bar{c}=1}^{\bar{k}} \bar{a}_{\bar{c}}\bar{\mathbf{q}}_{\bar{c}} + \sum_{\acute{c}=1}^{\acute{k}} \acute{a}_{\acute{c}}\acute{\mathbf{q}}_{\acute{c}} = \bar{\mathbf{Q}}\bar{\mathbf{a}} + \acute{\mathbf{Q}}\acute{\mathbf{a}} = \mathbf{Q}\mathbf{a},$$

where $\mathbf{a} = [\bar{\mathbf{a}}; \acute{\mathbf{a}}] \in \mathbb{R}^k, k = \bar{k} + \acute{k}$ is the weight vector and $\mathbf{Q} = [\bar{\mathbf{Q}}, \acute{\mathbf{Q}}] \in \mathbb{R}^{l\times k}$ is the basis set composed of (1) constant function $\bar{\mathbf{Q}} = [\bar{\mathbf{q}}_1, \cdots, \bar{\mathbf{q}}_{\bar{k}}] \in [1,n]^{l\times\bar{k}}$ and (2) monotonically increasing function $\acute{\mathbf{Q}} = [\acute{\mathbf{q}}_1, \cdots, \acute{\mathbf{q}}_{\acute{k}}] \in [1,n]^{l\times\acute{k}}$. Fig. 2a illustrates six common choices for $\acute{\mathbf{q}}$, including (1) step function, (2) polynomial function (ax^b), (3) exponential function ($\exp(ax+b)$), (4) logarithm function ($\log(ax+b)$), (5) I-spline, (6) hyperbolic tangent function ($\tanh(ax+b)$). Recall that [5] also used hyperbolic tangent functions as temporal basis, and the weights were optimized using a non-negative least squares algorithm. GTW differs from Fisher *et al.* [5] in three aspects: (1) GTW allows aligning multidimensional time series that have different features. Fisher *et al.* can only align one-dimensional time-series. (2) Unlike [5], we used a more efficient Eigen-decomposition to solve CCA and QP for optimizing the weights. (3) We propose to use a family of monotonic functions that allow for a more general warping (*e.g.*, subsequence matching), and constraints to regularize the solution.

As in DTW, we incorporate the following constraints on the weight $\mathbf{a}$ to constrain the warping path $\mathbf{p} = \mathbf{Qa}$.

Boundary conditions: We enforce the position of the first frame, $p_1 = \mathbf{q}^{(1)}\mathbf{a} \geq 1$, and the last frame, $p_l = \mathbf{q}^{(l)}\mathbf{a} \leq n$, where $\mathbf{q}^{(1)} \in \mathbb{R}^{1\times k}$ and $\mathbf{q}^{(l)} \in \mathbb{R}^{1\times k}$ are the first and last rows of the basis matrix $\mathbf{Q} \in \mathbb{R}^{l\times k}$ respectively. In contrast to DTW that imposes tight boundary (*i.e.*, $p_1 = 1$ and $p_l = n$), GTW relaxes the equality with inequality constraints to allow for a sub-part of $\mathbf{X}$ being indexed by $\mathbf{p}$. This relaxation can be used for sub-sequence matching.

Monotonicity: We enforce $t_1 \leq t_2 \Rightarrow p_{t_1} \leq p_{t_2}$ by constraining the sign of weight: $\acute{\mathbf{a}} \geq \mathbf{0}$. Notice that constraining the weights is only a sufficient condition to ensure

monotonicity but it is not necessary. See [24, 26, 33] for in-depth discussions on monotonic functions.

Continuity: To approximate the hard constraint on the step size (*e.g.*, $p_t - p_{t-1} \in \{0, 1\}$), we penalize the curvature of the warping path, $\sum_{t=1}^{l} \|\nabla \mathbf{q}^{(t)}\mathbf{a}\|_2^2 \approx \|\mathbf{FQa}\|_2^2$ where $\mathbf{F} \in \mathbb{R}^{l\times l}$ is the 1^{st} order differential operator.

In summary, we constrain the warping path[2] as:

$$\psi_a(\mathbf{a}) = \eta\|\mathbf{FQa}\|_2^2, \quad \Psi_a = \{\mathbf{a} \mid \mathbf{La} \le \mathbf{b}\},$$
$$\text{where } \mathbf{L} = \begin{bmatrix} \mathbf{0}_{\acute{k}\times\bar{k}} & -\mathbf{I}_{\acute{k}} \\ -\bar{\mathbf{q}}^{(1)} & -\acute{\mathbf{q}}^{(1)} \\ \bar{\mathbf{q}}^{(l)} & \acute{\mathbf{q}}^{(l)} \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} \mathbf{0}_{\acute{k}} \\ -1 \\ n \end{bmatrix}.$$

Therefore, given a basis set of k monotone functions, all feasible weights belong to a polyhedron in $\mathbb{R}^k$ parameterized by $\mathbf{L} \in \mathbb{R}^{(\acute{k}+2)\times k}$ and $\mathbf{b} \in \mathbb{R}^{\acute{k}+2}$. For instance, Fig. 2b illustrates an example of a warping function (red solid line) as a combination of three monotone functions (blue dotted lines).

4.3. Optimization of the temporal weights

Suppose that k_i basis functions, $\mathbf{Q}_i = [\mathbf{q}_1^i, \cdots, \mathbf{q}_{k_i}^i] \in \mathbb{R}^{l\times k_i}$, are associated with the i^{th} sequence $\mathbf{X}_i$, then the optimization of J_{gtw} (Eq. 2) with respect to the time warping parameter $\mathbf{W}_i$ minimizes:

$$J_{gtw}(\{\mathbf{a}_i\}) = \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{1}{2}\|\mathbf{V}_i^T\mathbf{X}_i\mathbf{W}(\mathbf{Q}_i\mathbf{a}_i) - \mathbf{V}_j^T\mathbf{X}_j\mathbf{W}(\mathbf{Q}_j\mathbf{a}_j)\|_F^2 + \sum_{i=1}^{m}\eta_i\|\mathbf{FQ}_i\mathbf{a}_i\|_2^2, \quad (3)$$
$$\text{s.t.} \quad \mathbf{L}_i\mathbf{a}_i \le \mathbf{b}_i, \ \forall i \in \{1:m\}.$$

To optimize Eq. 3, we linearize the expression and use a Gauss-Newton method similar to the Lucas-Kanade framework [19] for image alignment, where the nonlinear expression in Eq 3 is linearized by performing a first order Taylor approximation on $\mathbf{V}_i^T\mathbf{X}_i\mathbf{W}(\mathbf{Q}_i(\mathbf{a}_i+\boldsymbol{\delta}_i)) \in \mathbb{R}^{d\times l}$ given the increment $\boldsymbol{\delta}_i \in \mathbb{R}^{k_i}$, that is:

$$\text{vec}\left(\mathbf{V}_i^T\mathbf{X}_i\mathbf{W}(\mathbf{Q}_i(\mathbf{a}_i+\boldsymbol{\delta}_i))\right) \approx \mathbf{v}_i + \mathbf{G}_i\delta_i, \quad (4)$$
$$\text{where} \quad \mathbf{v}_i = \text{vec}\left(\mathbf{V}_i^T\mathbf{X}_i\mathbf{W}(\mathbf{Q}_i\mathbf{a}_i)\right) \in \mathbb{R}^{dl},$$
$$\mathbf{G}_i = [\Downarrow_t \nabla(\mathbf{V}_i^T\mathbf{x}_{\mathbf{q}_i^{(t)}\mathbf{a}_i})\mathbf{q}_i^{(t)}] \in \mathbb{R}^{dl\times k_i}.$$

Plugging Eq. 4 in Eq. 3 yields:

$$J_{gtw}(\{\mathbf{a}_i+\boldsymbol{\delta}_i\}) \approx \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{2}\|\mathbf{v}_i + \mathbf{G}_i\boldsymbol{\delta}_i - \mathbf{v}_j - \mathbf{G}_j\boldsymbol{\delta}_j\|_2^2 + \sum_{i=1}^{m}\eta_i\|\mathbf{F}_i\mathbf{Q}_i(\mathbf{a}_i+\boldsymbol{\delta}_i)\|_2^2. \quad (5)$$

[2]Notice that the constraints of ψ and Ψ in Eq. 2 associated with the warping matrix $\mathbf{W}$ are replaced by the constraints (ψ_a and Ψ_a) associated with the weight $\mathbf{a}$.

Minimizing Eq. 5 with respect to the weight increment $\boldsymbol{\delta}_i \in \mathbb{R}^{k_i}$ yields a quadratic programming problem:

$$\min_{\boldsymbol{\delta}} \quad \frac{1}{2}\boldsymbol{\delta}^T\mathbf{H}\boldsymbol{\delta} + \mathbf{f}^T\boldsymbol{\delta}, \quad \text{s.t.} \quad \mathbf{L}\boldsymbol{\delta} \le \mathbf{b} - \mathbf{La}, \quad (6)$$
$$\text{where} \quad \dot{k} = \sum_{i=1}^{m} k_i, \ \boldsymbol{\delta} = [\Downarrow_i \boldsymbol{\delta}_i] \in \mathbb{R}^{\dot{k}},$$
$$\mathbf{H} = m[\searrow_i \mathbf{G}_i^T\mathbf{G}_i] - [\Downarrow_i \mathbf{G}_i^T][\Rightarrow_j \mathbf{G}_j] + [\searrow_i \eta_i\mathbf{Q}_i^T\mathbf{F}_i^T\mathbf{F}_i\mathbf{Q}_i] \in \mathbb{R}^{\dot{k}\times\dot{k}},$$
$$\mathbf{f} = [\searrow_i \mathbf{G}_i^T]\Big(m[\Downarrow_i \mathbf{v}_i] \ominus [\Rightarrow_i \mathbf{v}_i]\mathbf{1}_m\Big) + [\Downarrow_i \eta_i\mathbf{Q}_i^T\mathbf{F}_i^T\mathbf{F}_i\mathbf{Q}_i\mathbf{a}_i] \in \mathbb{R}^{\dot{k}},$$
$$\mathbf{L} = [\searrow_j \mathbf{L}_i] \in \mathbb{R}^{(\dot{k}+m)\times\dot{k}},$$
$$\mathbf{a} = [\Downarrow_i \mathbf{a}_i] \in \mathbb{R}^{\dot{k}}, \ \mathbf{b} = [\Downarrow_i \mathbf{b}_i] \in \mathbb{R}^{(\dot{k}+m)}.$$

In all our experiments, we initialize $\mathbf{a}_i$ by uniformly aligning the sequences (the curve of GN-Init in Fig. 3b). The length of the warping path l is usually set to $l = 1.1\max_{i=1}^{m} n_i$. In practice, when the sequence length n_i is very large, an additional pre-conditioner should be used to obtain a numerically stable solution. For instance, a normalized version of Eq. 6 minimizes $\frac{1}{2}\boldsymbol{\delta}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{R}^{-1}\boldsymbol{\delta} + \mathbf{f}^T\mathbf{R}^{-1}\boldsymbol{\delta}$ subject to $\mathbf{L}\mathbf{R}^{-1}\boldsymbol{\delta} \le \mathbf{b} - \mathbf{La}$, where $\mathbf{R} = [\searrow_i n_i\mathbf{I}_{k_i}] \in \mathbb{R}^{\dot{k}\times\dot{k}}$ is the scaling matrix. After solving this new quadratic optimization problem, we need to rescale the result as $\boldsymbol{\delta} \leftarrow \mathbf{R}^{-1}\boldsymbol{\delta}$. The computational complexity of the algorithm is $O(dl\dot{k} + \dot{k}^3)$.

As discussed in [23, 28], there are various techniques that have been proposed to accelerate and improve DTW. For instance, the Sakoe-Chiba band (DTW-SC) and the Itakura Parallelogram band (DTW-IP) reduce the complexity of the original DTW algorithm to $O(\beta n^2)$ by constraining the warping path, assuming $\beta < 1$. However, using a narrow band (a small β) might cut off potential warping space, leading to a sub-optimal solution. For instance, Fig. 3a shows an example of two 1-D time series and the alignment results calculated by different algorithms. The results computed by DTW-SC and DTW-IP are less accurate than the one computed by Gauss-Newton (GN). This is because both the SC and IP bands are over-constrained (Fig. 3b).

To provide a quantitative evaluation, we synthetically generated 1-D sequences at 15 scales. For DTW-SC, we set the band width as $\beta = 0.1$. For GN, we varied k among $6, 10, 14$ to investigate the effect of the number of bases. For each scale, we randomly generated 100 pairs of sequences. The error is computed with Eq. 9 and shown in Fig. 3cd. DTW obtains the lowest error but takes the most time to compute. This is because DTW exhaustively searches the entire parameter space to find the global optima. Both DTW-SC and DTW-IP need less time than DTW because they need to search a smaller space constrained

by different bands. Empirically, DTW-IP is more accurate than DTW-SC for our synthetic dataset. This is because the global optima is more likely to lie in the IP band than the SC band. Compared to DTW, DTW-SC and DTW-IP, GN is more computationally efficient because it has linear complexity in terms of sequence length. Moreover, increasing the number of bases monotonically reduces the error.

4.4. Optimization of the spatial embedding

To optimize over $\mathbf{V}_i$ we used multi-set canonical correlation analysis (mCCA) [11], and we constrain the embedding $\mathbf{V}_i$ as:

$$\phi(\{\mathbf{V}_i\}) = \frac{m\lambda_i}{1-\lambda_i}\|\mathbf{V}_i\|_F^2, \tag{7}$$

$$\Phi = \{\mathbf{V}_i \mid \sum_{i=1}^{m} \mathbf{V}_i^T \Big((1-\lambda_i)\mathbf{X}_i\mathbf{W}_i\mathbf{W}_i^T\mathbf{X}_i^T + \lambda_i \mathbf{I}_{d_i}\Big)\mathbf{V}_i = \mathbf{I}_d\},$$

where $\lambda_i \in [0,1]$ is the regularization term. Consider the special case when $\lambda_i \to 1$, the constraint is equivalent to the one used in multi-set partial least squares (mPLS) [27]. Plugging Eq. 7 into Eq. 2 yields:

$$\max_{\mathbf{V}} \ \mathrm{tr}(\mathbf{V}^T\mathbf{C}\mathbf{V}), \quad \text{s.t.} \quad \mathbf{V}^T\mathbf{D}\mathbf{V} = \mathbf{I}_d, \tag{8}$$

$$\text{where } \dot{d} = \sum_{i=1}^{m} d_i,\ \mathbf{V} = [\Downarrow_i \mathbf{V}_i] \in \mathbb{R}^{\dot{d}\times d},$$

$$\mathbf{C} = [\Downarrow_i \mathbf{X}_i\mathbf{W}_i][\Rightarrow_j \mathbf{W}_j^T\mathbf{X}_j^T] - [\searrow_i \mathbf{X}_i\mathbf{W}_i\mathbf{W}_i^T\mathbf{X}_i^T] \in \mathbb{R}^{\dot{d}\times\dot{d}},$$

$$\mathbf{D} = [\searrow_i (1-\lambda_i)\mathbf{X}_i\mathbf{W}_i\mathbf{W}_i^T\mathbf{X}_i^T + \lambda_i\mathbf{I}_{d_i}] \in \mathbb{R}^{\dot{d}\times\dot{d}}.$$

The optimal $\mathbf{V}$ of Eq. 8 can be solved in closed form using a generalized Eigen decomposition, *i.e.*, $\mathbf{C}\mathbf{V} = \mathbf{D}\mathbf{V}\Lambda$. The dimension d is selected to preserve 90% of the total correlation.

5. Experiments

This section compares GTW against state-of-the-art DTW approaches in three experimental settings: (1) aligning time series with known ground truth to provide a quantitative comparison, (2) aligning several video sequences of different people performing similar actions using different visual features for each sequence, and (3) aligning three sequences of different subjects performing a similar action recorded with different sensors (motion capture data, accelerometers and video).

5.1. Other methods for comparison

We compared GTW against several versions of Procrustes analysis [4], which are used as baselines.

Procrustes dynamic time warping (pDTW): Procrustes analysis [4] has been extensively used for shape for alignment. We proposed a simple temporal extension pDTW, which aligns multiple time series by minimizing:

$$\begin{aligned} J_{pdtw}(\{\mathbf{W}_i\}) &= \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{2}\|\mathbf{X}_i\mathbf{W}_i - \mathbf{X}_j\mathbf{W}_j\|_F^2 \\ &= m\sum_{i=1}^{m}\|\mathbf{X}_i\mathbf{W}_i - \frac{1}{m}\sum_{j=1}^{m}\mathbf{X}_j\mathbf{W}_j\|_F^2. \end{aligned}$$

pDTW alternates between solving the warping matrix $\mathbf{W}_i \in \{0,1\}^{n_i\times l}$ by a slightly modified DTW and computing the mean sequence $\frac{1}{m}\sum_{j=1}^{m}\mathbf{X}_j\mathbf{W}_j \in \mathbb{R}^{d\times l}$.

Procrustes derivative dynamic time warping (pDDTW): In order to make DTW invariant to translation, derivative dynamic time warping (DDTW) [15] uses the derivatives of the original features. Similar to pDTW, we combined DDTW and Procrustes framework to minimize:

$$J_{pddtw}(\{\mathbf{W}_i\}) = \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{2}\|\mathbf{X}_i\mathbf{F}_i^T\mathbf{W}_i - \mathbf{X}_j\mathbf{F}_j^T\mathbf{W}_j\|_F^2,$$

where $\mathbf{F}_i \in \mathbb{R}^{n_i\times n_i}$ is the 1^{st} order differential operator.

Procrustes iterative motion warping (pIMW): Similar to GTW, iterative motion warping (IMW) [13] alternates between time warping and spatial transformation to align two sequences. In our experiment, we extended IMW to align multiple sequences by minimizing:

$$\begin{aligned} &J_{pimw}(\{\mathbf{W}_i, \mathbf{A}_i, \mathbf{B}_i\}) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{2}\|(\mathbf{X}_i\circ\mathbf{A}_i + \mathbf{B}_i)\mathbf{W}_i - (\mathbf{X}_j\circ\mathbf{A}_j + \mathbf{B}_j)\mathbf{W}_j\|_F^2 \\ &\quad + \sum_{i=1}^{m}\Big(\lambda_i^a\|\mathbf{A}_i\mathbf{F}_i^{aT}\|_F^2 + \lambda_i^b\|\mathbf{B}_i\mathbf{F}_i^{bT}\|_F^2\Big), \end{aligned}$$

where $\mathbf{A}_i, \mathbf{B}_i \in \mathbb{R}^{d\times n_i}$ are the scaling and translation parameter for the i^{th} sequence $\mathbf{X}_i$, respectively. $\mathbf{F}_i^a, \mathbf{F}_i^b \in \mathbb{R}^{n_i\times n_i}$ are 1^{st} order differential operators, enforcing a smooth change in the columns of $\mathbf{A}_i$ and $\mathbf{B}_i$.

To evaluate the time warping results, we proposed to compute the difference between the warping matrix $\mathbf{W}_i^{alg} \in \{0,1\}^{n_i\times l_{alg}}$ given by the algorithm (*e.g.*, GTW, pDTW, pDDTW, pIMW) and the ground-truth $\mathbf{W}_i^{tru} \in \{0,1\}^{n_i\times l_{tru}}$. Recall that the number of warping steps can be different, *i.e.*, $l_{alg} \neq l_{tru}$. Equivalently, we could compare the warping path $\mathbf{P}_{alg} = [\mathbf{p}_1^{alg}, \cdots, \mathbf{p}_m^{alg}] \in \mathbb{R}^{l_{alg}\times m}$ and $\mathbf{P}_{tru} = [\mathbf{p}_1^{tru}, \cdots, \mathbf{p}_m^{tru}] \in \mathbb{R}^{l_{tru}\times m}$, that is:

$$\mathrm{err}_{alg} = \frac{dist(\mathbf{P}_{alg}, \mathbf{P}_{tru}) + dist(\mathbf{P}_{tru}, \mathbf{P}_{alg})}{l_{alg} + l_{tru}}, \tag{9}$$

$$\text{where} \quad dist(\mathbf{P}_1, \mathbf{P}_2) = \sum_{i=1}^{l_1}\min_{j=1}^{l_2}\|\mathbf{p}_1^{(i)} - \mathbf{p}_2^{(j)}\|_2,$$

where $\mathbf{p}_{alg}^{(i)} \in \mathbb{R}^{1\times m}$ and $\mathbf{p}_{tru}^{(j)} \in \mathbb{R}^{1\times m}$ are the i^{th} row of $\mathbf{P}_{alg}$ and j^{th} row of $\mathbf{P}_{tru}$ respectively. Let's consider each warping path $\mathbf{P} \in \mathbb{R}^{l\times m}$ as a curve in $\mathbb{R}^m$ with l points. Thus the term of $\min_{j=1}^{l_2}\|\mathbf{p}_1^i - \mathbf{p}_2^j\|_2$ can be interpreted as the closest distance between point $\mathbf{p}_1^i$ and the curve $\mathbf{P}_2$.

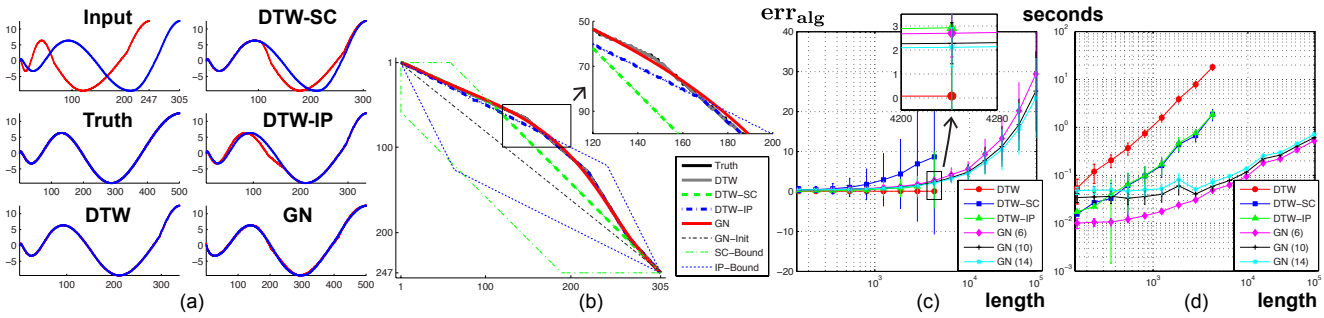

Figure 3. Comparison of temporal alignment algorithms. (a) An example of two 1-D time series (with 247 and 305 frames respectively) and the alignment results calculated using the ground truth, DTW, DTW constrained in the Sakoe-Chiba band (DTW-SC), DTW constrained in the Itakura Parallelogram band (DTW-IP) and Gauss-Newton (GN). (b) Comparison of different warping paths. GN-Init denotes the initial warping used for GN. SC-Bound and IP-Bound denote the boundaries of SC band and IP band respectively. (c) Comparison of alignment errors. (d) Time.

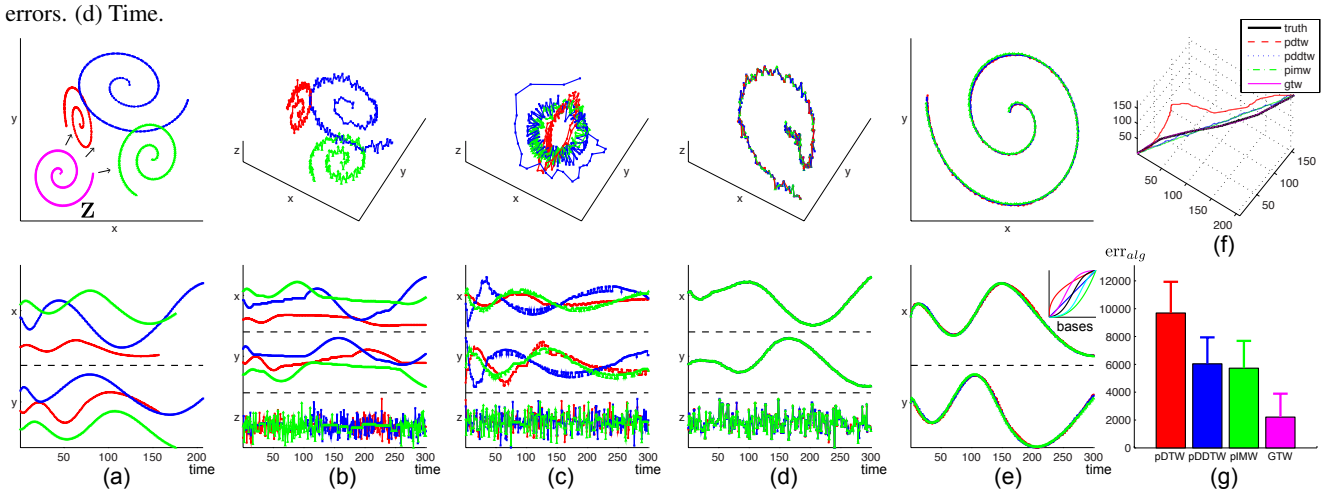

Figure 4. Synthetic data. Time series $\mathbf{X}_i, i \in \{1:3\}$ are generated by (a) spatio-temporal transformation $\mathbf{U}_i^T(\mathbf{Z} + \mathbf{b}_i\mathbf{1}_l^T)\mathbf{M}_i$ of 2-D latent sequence $\mathbf{Z}$, (b top) and adding Gaussian noise $\mathbf{e}_i$ in the 3^{rd} dimension. The spatio-temporal warping is computed by (b) pDTW, (c) pDDTW, (d) pIMW and (e) GTW, where the bases are shown in the top-right corner. (f) Comparison of different time warping techniques. (g) Mean and variance of alignment error for different methods.

5.2. Synthetic dataset

In the first experiment we synthetically generated 3-D spatio-temporal signals (2-D in space and 1-D in time) to evaluate the performance of GTW. The first two spatial dimensions and the time dimension are generated as follows:

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{U}_i^T(\mathbf{Z} + \mathbf{b}_i\mathbf{1}_l^T)\mathbf{M}_i \\ \mathbf{e}_i^T \end{bmatrix} \in \mathbb{R}^{3\times n_i}, i \in \{1:3\}$$

where $\mathbf{Z} \in \mathbb{R}^{2\times l}$ is a curve in two dimensions (Fig. 4a top). $\mathbf{U}_i \in \mathbb{R}^{2\times 2}$ and $\mathbf{b}_i \in \mathbb{R}^2$ are randomly generated projection matrix and translation vector, respectively, see Fig. 4a top. The binary matrix $\mathbf{M}_i \in \{0,1\}^{l\times n_i}$ is generated by randomly choosing $n_i \leq l$ columns from $\mathbf{I}_l$ for temporal distortion (Fig. 4a bottom). The third spatial dimension $\mathbf{e}_i \in \mathbb{R}^{n_i}$ is generated with zero-mean Gaussian noise (Fig. 4b top). Notice that in the case of synthetic data we are able to obtain the ground truth alignment matrix $\mathbf{W}_i^{tru} = \mathbf{M}_i^T$. The error between the ground truth and a given alignment $\mathbf{W}_{alg}$ is computed by Eq. 9 (Fig. 4g). We initialize all methods by uniformly aligning the three sequences, *i.e.*, $\mathbf{p}_i = round(linspace(1, n_i, l))'$, where $round(\cdot)$ and $linspace(\cdot)$ are MATLAB functions.

We set the length of the latent sequence to $l = 300$. For GTW, we set $\eta_i = \lambda_i = 0$ and selected d to preserve 90% of the total correlation. We selected three hyperbolic tangent and three polynomial functions as bases for monotonic warping function.

Fig. 4b-e show the spatial-temporal warping estimated by each algorithm. Fig. 4g shows the err_{alg} (Eq. 9) for 100 new generated time series. As can be observed in Fig. 4e GTW obtains the best performance. pDTW (Fig. 4b fails in this case since the sequences have been distorted in space. pDDTW (Fig. 4c) cannot deal with this example because the feature derivatives do not capture well the structure of the sequence. pIMW (Fig. 4d) warps sequences towards others by translating and re-scaling each frame in each dimension. Moreover, pIMW has more parameters ($\sum_{i=1}^m ln_i + 2dn_i$) than GTW ($\sum_{i=1}^m k_i + dd_i$), and hence pIMW is more prone to over-fitting. Furthermore, pIMW

tries to fit the noisy dimension (3^{rd} spatial component) biasing alignment in time, whereas GTW has a feature selection mechanism which effectively cancels the third dimension.

5.3. Aligning videos with different features

In the second experiment we applied GTW to align video sequences of different people performing a similar action. Each video is encoded using different visual features. The video sequence are taken from the Weizmann database [8], which contains 9 people performing 10 actions. To extract dynamic features from video, we extract the silhouette with background subtraction (Fig. 5a). We computed three popular shape features (Fig. 5b) for each 70-by-35 re-scaled mask image, including (1) binary image, (2) Euclidean distance transform [20], and (3) solution of Poisson equation [9]. In order to reduce the feature dimension (2450), we picked the top 123 principal components that preserve 99% of the total energy. To evaluate the performance, we randomly selected three walking sequences, each of which is manually cropped into two cycles of human walking. The ground-truth alignment was approximated by using pDTW using the same features, and it provided an accurate visual temporal alignment.

GTW was initialized with uniform alignment, and we used the parameter $\lambda = 0.1$. We used five hyperbolic tangent and five polynomial functions as the monotonic bases (Fig. 5f middle-top).

Fig. 5g shows the err_{alg} for 10 randomly generated sets of videos. Notice that neither pDTW (Fig. 5c) nor pDDTW (Fig. 5d) is able to align the videos because both of them lack the ability to solve for correspondence between signals of different nature. As observed from Fig. 5e, pIMW registers the top three components well in space; however, it overfits all the dimensions and thus obtains a biased time warping path. In contrast, GTW (Fig. 5f) warps the sequences accurately in both space and time. Fig. 5b illustrates the temporal correspondence found by GTW.

5.4. Multi-modal sequence alignment

This experiment applies GTW to align sequences of different people performing a similar activity but recorded with different sensors. We selected one motion capture sequence (Subject 12, Trial 29) from the CMU motion capture database, one video sequence (Eli, jacking) from the Weizmann database [8], and we collected the accelerometer signal of a subject performing a jacking exercise. Some instances of the multi-modal data can be seen in Fig. 6d. Observe, that to make the problem more challenging, while in the mocap (top row) and video (middle row) the two subjects are performing the same activity, in the accelerometer sequence (bottom row) the subject only moves one hand and not the legs. Even in this challenging scenario, GTW is able to solve for the temporal correspondence that maximizes the correlation between signals.

For the mocap data, we computed the quaternions for the 20 joints resulting in a 60 dimensional feature vector that describes the body configuration. In the case of the Weizmann dataset, we computed the Euclidean distance transform as described earlier. The data from the accelerometers is collected in X, Y, and Z axes by an X6-2mini USB accelerometer (Fig. 6a) at a rate of 40Hz. GTW was initialized by uniformly aligning the three sequences. We used five hyperbolic tangent and five polynomial functions as monotonic bases. Fig. 6b shows the first components of the three sequences projected separately by PCA. As shown in Fig. 6c, GTW found an accurate temporal correspondence between the three sequences. Unfortunately, we do not have ground-truth for this experiment, however visual inspection of the video suggest that results are consistent with human labeling. Fig. 6d shows several frames that have been put in correspondence by GTW.

6. Conclusions

This paper describes GTW, a technique for temporally aligning multiple multi-modal sequences. The GTW algorithm offers a more flexible and efficient framework than the state-of-art DTW algorithms because we parameterize the time warping function as a linear combination of monotonic bases.

Although GTW has shown promising preliminary results, there are still unresolved issues. First, the Gauss-Newton algorithm for time warping converges poorly in the area where the objective function J_{gtw} is non-smooth. Second, GTW is subject to local minima. A well known strategy to escape from local minima in image alignment has been to adopt a coarse-to-fine approach for optimizing GTW at different temporal scales. Third, although the experiments show admissible time warping results with fixed bases, it is more desirable to automatically learn the monotonic bases. We plan to explore these issues in future work.

Acknowledgements The first author was supported by the National Science Foundation (NSF) under Grant No. EEEC-0540865 and CPS-0931999. The second author was partially supported by the NSF grant RI-1116583. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] Y. Caspi and M. Irani. Aligning non-overlapping sequences. *Int. J. Comput. Vis.*, 48(1):39–51, 2002.

[2] Y. Caspi and M. Irani. Spatio-temporal alignment of sequences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(11):1409–1424, 2002.

[3] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):781–796, 2000.

[4] I. L. Dryden and K. V. Mardia. *Statistical shape analysis*. Wiley, 1998.

[5] B. Fischer, V. Roth, and J. Buhmann. Time-series alignment by non-negative multiple generalized canonical correlation analysis. *BMC Bioinformatics*, 8(10), 2007.

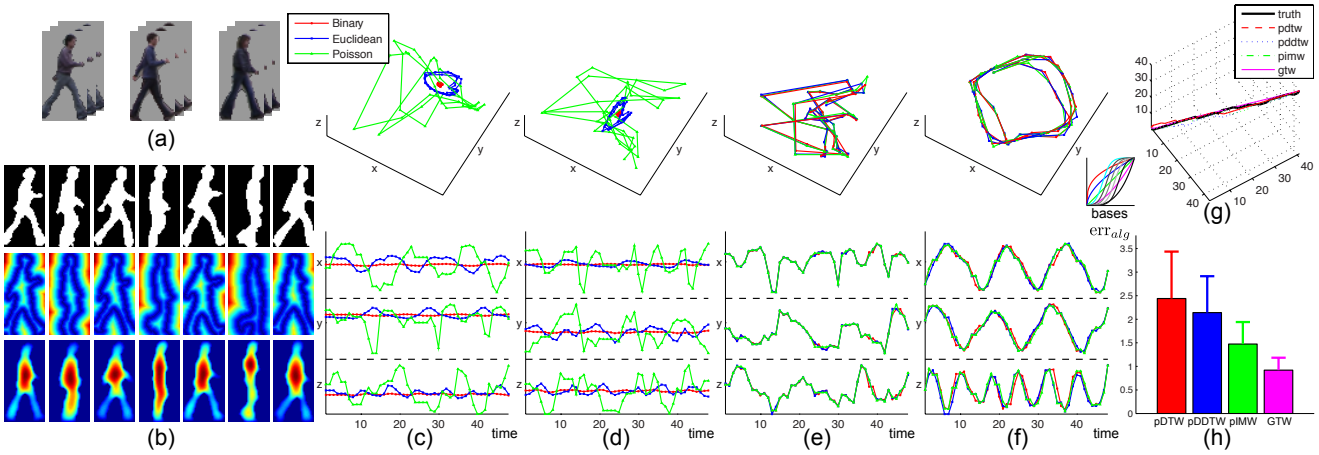

Figure 5. Example using multi-feature video data. (a) Original sequences after background subtraction. (b) Key frames aligned using GTW. The three sequences from top to bottom are represented as a binary image, Euclidean distance transform and solution of the Poisson equation respectively. (c) pDTW. (d) pDDTW. (e) pIMW. (f) GTW. (g) Comparison of different time warping techniques. (h) Mean and variance of the alignment error.

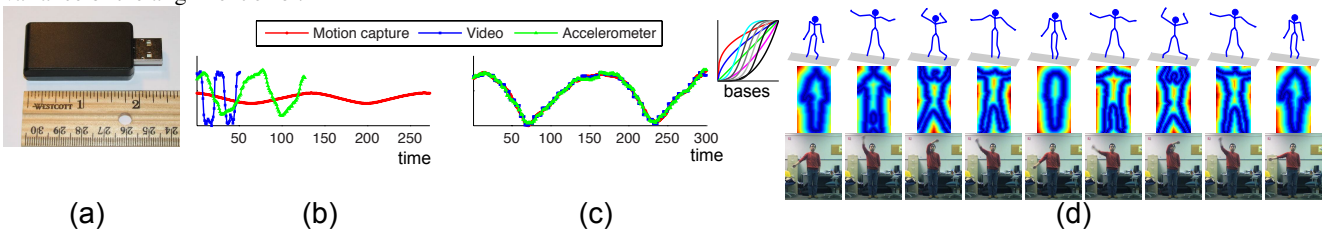

Figure 6. Example of aligning multi-modal sequences. (a) Accelerometer. (b) Projection onto the first principal component for the motion capture data, video and accelerometers respectively. (c) GTW. (d) Key frames aligned by GTW. Notice that the similar hand gestures have been aligned. On the top row we show mocap data, in the middle row video, and in the bottom the images of the accelerometer data.

[6] K. Forbes and E. Fiume. An efficient search algorithm for motion data using weighted PCA. In *ACM SIGGRAPH / Eurographics SCA*, 2005.

[7] D. Gong and G. G. Medioni. Dynamic manifold warping for view invariant action recognition. In *ICCV*, 2011.

[8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2247–2253, 2007.

[9] L. Gorelick, M. Galun, E. Sharon, R. Basri, and A. Brandt. Shape representation and classification using the Poisson equation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1991–2005, 2006.

[10] A. Gritai, Y. Sheikh, C. Rao, and M. Shah. Matching trajectories of anatomical landmarks under viewpoint, anthropometric and temporal transforms. *Int. J. Comput. Vis.*, 2009.

[11] M. A. Hasan. On multi-set canonical correlation analysis. In *IJCNN*, 2009.

[12] A. Heloir, N. Courty, S. Gibet, and F. Multon. Temporal alignment of communicative gesture sequences. *J. Visual. Comp. Animat.*, 17(3-4):347–357, 2006.

[13] E. Hsu, K. Pulli, and J. Popovic. Style translation for human motion. *ACM Trans. Graph.*, 24(3):1082–1089, 2005.

[14] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez. View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[15] E. J. Keogh and M. J. Pazzani. Derivative dynamic time warping. In *SDM*, 2001.

[16] L. Kovar and M. Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Trans. Graph.*, 23(3):559–568, 2004.

[17] R. Li and R. Chellappa. Aligning spatio-temporal signals on a special manifold. In *ECCV*, 2010.

[18] J. Listgarten, R. M. Neal, S. T. Roweis, and A. Emili. Multiple alignment of continuous time series. In *NIPS*, 2005.

[19] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.

[20] C. R. Maurer, R. Qi, and V. V. Raghavan. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(2):265–270, 2003.

[21] F. L. C. Pádua, R. L. Carceroni, G. A. M. R. Santos, and K. N. Kutulakos. Linear sequence-to-sequence alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):304–320, 2010.

[22] W. Pan and L. Torresani. Unsupervised hierarchical modeling of locomotion styles. In *ICML*, 2009.

[23] L. Rabiner and B. Juang. *Fundamentals of speech recognition*. Prentice Hall, 1993.

[24] J. O. Ramsay. Estimating smooth monotone functions. *J. R. Stat. Soc. Series B Stat. Methodol.*, 60(2):365–375, 1998.

[25] C. Rao, A. Gritai, M. Shah, and T. F. Syeda-Mahmood. View-invariant alignment and matching of video sequences. In *ICCV*, 2003.

[26] T. Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley, 1988.

[27] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *SLSFS*, pages 34–51, 2005.

[28] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, 2007.

[29] T. B. Sebastian, P. N. Klein, and B. B. Kimia. On aligning curves. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(1):116–125, 2003.

[30] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of a human action. In *ICCV*, 2005.

[31] A. Veeraraghavan, A. Srivastava, A. K. R. Chowdhury, and R. Chellappa. Rate-invariant recognition of humans and their activities. *IEEE Trans. Image Process.*, 18(6):1326–1339, 2009.

[32] A. P. Witkin and Z. Popovic. Motion warping. In *ACM SIGGRAPH*, 1995.

[33] I. W. Wright and E. J. Wegman. Isotonic, convex and related splines. *Ann. Stat.*, pages 1023–1035, 1980.

[34] F. Zhou and F. De la Torre. Canonical time warping for alignment of human behavior. In *NIPS*, 2009.

[35] F. Zhou, F. De la Torre, and J. K. Hodgins. Aligned cluster analysis for temporal segmentation of human motion. In *FG*, 2008.