4D Model-based Spatiotemporal Alignment of Scripted Taiji Quan Sequences

Jesse Scott, Robert Collins, Christopher Funk, and Yanxi Liu
School of Electrical Engineering and Computer Science
The Pennsylvania State University. University Park, PA. 16802, USA
{jescott, rcollins, funk, yanxi}@cse.psu.edu

Abstract

We develop a computational tool that aligns motion capture (mocap) data to videos of 24-form simplified Taiji (TaiChi) Quan, a scripted motion sequence about 5 minutes long. With only prior knowledge that the subjects in video and mocap perform a similar pose sequence, we establish inter-subject temporal synchronization and spatial alignment of mocap and video based on body joint correspondences. Through time alignment and matching the viewpoint and orientation of the video camera, the 3D body joints from mocap data of subject A can be correctly projected onto the video performance of subject B. Initial quantitative evaluation of this alignment method shows promise in offering the first validated algorithmic treatment for cross-subject comparison of Taiji Quan performances. This work opens the door to subject-specific quantified comparison of long motion sequences beyond Taiji.

1. Introduction

In the realms of health and sports, digital recording and analysis of human movement provide rich content for performance characterization and training. In this paper we address some basic challenges involving spatiotemporal warping between 4D (3D+time) mocap data and 3D (2D+time) video data. We focus on cross-modality alignment between mocap and video data for Taiji routines performed by different subjects at different times.

Taiji Quan is a form of Chinese martial arts practiced for competition and health purposes by millions of people worldwide. Simplified *24-form Taiji Quan* is comprised of a scripted routine of 24 movement forms performed sequentially, usually taking 4-5 minutes to complete (Figure 1 Top). A distinctive feature of Taiji is its slow and seamless transitions between poses, which introduces difficulties for motion-segmentation based activity recognition methods.

We propose to leverage a prerecorded mocap routine performed by an instrumented lab subject for analysis of video sequences of other performers. The first and most crucial step of this analysis is to align the mocap with the video both temporally, in terms of time synchronization of different forms and movements, and spatially, by determining camera viewpoint. A correct spatiotemporal alignment allows for frame-by-frame pairing of joints detected in video with labeled motion capture points, allowing the body joints and limb segments of the performer to be projected and overlaid onto the image frames. Using our method, two video performers can be aligned through an underlying mocap reference model. Figure 1 illustrates that our nonlinear model-based method can achieve much better pose correspondences than linear time warping.

The spatiotemporal alignment problem addressed here is difficult because the recording modalities capture very different types of data (3D marker locations vs pixel intensities), recording parameters differ (resolution; frame rate), and video capture conditions vary (viewpoint; lighting). There are also significant technical hurdles raised by inter-subject variation of body characteristics, performance pace, style, and skill levels. In general, we consider a progression of mocap to video matchings in difficulty as follows:

1. Simultaneous Data Mapping: both mocap and video are recorded at the same time of the same subject;
2. Intra-Subject Mapping: mocap and video are of the same person but recorded at different times;
3. Inter-Subject Mapping: mocap and video come from two different performers, and the video may come from an unknown source camera (e.g. YouTube video).

Our work demonstrates that it is feasible to achieve sustained spatiotemporal alignment of a prerecorded mocap sequence with video of a lengthy, complex human action performed by different subjects. Other contributions include

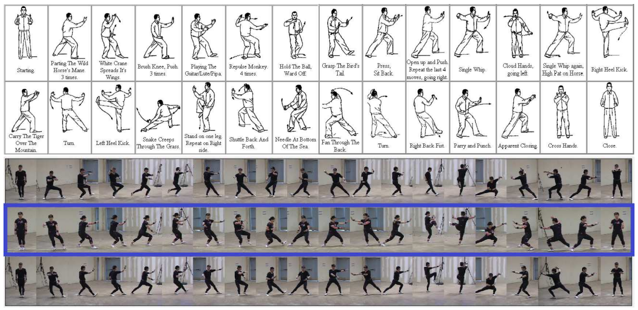
Starting.
Parting The Wild Horse's Mane. 3 times.
White Crane Spreads It's Wings.
Brush Knee, Push. 3 times.
Playing The Guitar/Lute/Pipa.
Repulse Monkey. 4 times.
Hold The Ball, Ward Off.
Grasp The Bird's Tail.
Press, Sit Back.
Open up and Push. Repeat the last 4 moves, going right.
Single Whip.
Cloud Hands, going left.
Single Whip again, High Pat on Horse.
Right Heel Kick.
Carry The Tiger Over The Mountain.
Turn.
Left Heel Kick.
Snake Creeps Through The Grass.
Stand on one leg. Repeat on Right side.
Shuttle Back And Forth.
Needle At Bottom Of The Sea.
Fan Through The Back.
Turn.
Right Back Fist.
Parry and Punch.
Apparent Closing.
Cross Hands.
Close.

Figure 1: Top: 24-form simplified Taiji is a scripted action sequence [11]. Bottom: Inter-subject alignment results. A reference subject (in box) is aligned with a second performer using linear time warping (above box) and nonlinear warping computed by our algorithm (below box).

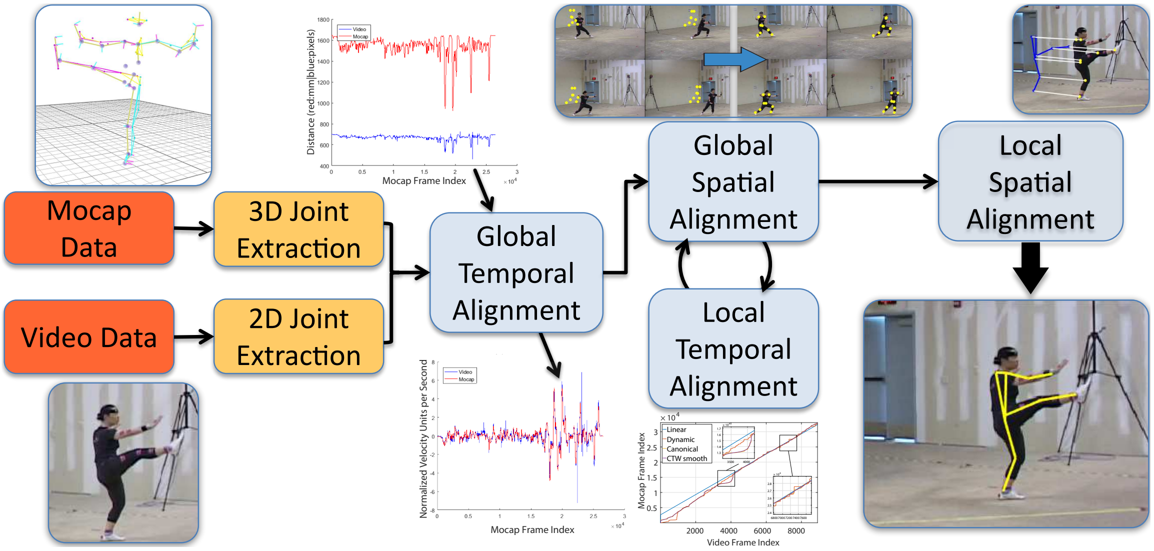

Figure 2: Overview of our method for matching motion capture data to video through temporal and spatial alignment.

the use of MCMC to search for camera parameters yielding spatial alignment and a dataset with measured ground truth for pose estimation evaluation.

2. Related Work

2.1. Understanding Movement, Activity, and Action

Motion perception has been described in [1] as a hierarchical model similar to the building blocks of language, with *movements* (letters) like jump, sit, and throw being the fundamental unit, *activities* (words) like load boxes, shoot free throws, and chop wood being composed of multiple *movements*, and *actions* (sentences) like make breakfast, swing dance, and repair a car being a series of *activities*. In this paper we take a global to local top-down approach to correlate *actions* so that *activities* and then *movements* can be more easily matched.

2.2. Current Commercial Methods

The use of video to analyze movements and simple activities has developed significantly over the last 25 years. Gait analysis as a tool for medical diagnostics [20] is a mature field. Activity analysis is used commercially in professional sports to perfect everything from baseball pitches [12] to golf swings [19]. Current commercial methods are limited to *movements* and simple or repetitive *activities*.

2.3. Finding Joints by Leveraging Motion Capture

Our work employs alignment and warping procedures to leverage motion capture data for 3D pose analysis of video. A recent work related to this is Zhou and De La Torre [30]. They track feature points using dense optical flow on segments of *movements* and assign trajectories to parts detected independently on each frame. The mocap data is split into equal sized segments, clustered using Procrustes analysis, then spatiotemporally matched with the video trajectories. While this approach is promising, it suffers from a high computational cost, is designed for *movements*, and no code was available for quantitative comparison when requested. Our approach differs in that we separate temporal and spatial matching into different, interleaved steps, and we align an entire sequence of video with a complete motion capture sequence at the *action* level.

2.4. Convolutional Pose Machines

To align an articulated human body model to a person in video, we must overcome the twin problems of pose estimation and body part segmentation. From 2D data alone there are large variations in image appearance caused by pose, viewpoint, clothing, illumination, and clutter.

For localizing 2D body joint positions, Convolutional Neural Networks (CNNs) are currently the state-of-the-art [25, 14, 22, 6, 21, 2]. The algorithms either directly output the joint coordinates [25], regress into a heat map for each joint [14, 22], or use the output of the CNN as a feature descriptor [6]. There have also been pose detection algorithms specifically designed for video that capitalize on optical flow [22, 9, 33, 23] or that use spatial and temporal tracking [5]. Many of these algorithms are trained and tested on images of people in mostly upright, standing positions, facing the camera. Convolutional Pose Machines (CPM) [27] leverages deep learned features for human pose detection, and passes the image feature maps back into the network multiple times to help localize the joints. Cao *et al.* [3] extends this work by adding pairwise dependencies between each connected set of joints to improve the accuracy of the pose and to aid in multi-person tracking.

One persistent issue with all single image pose detectors is missing parts due to self-occlusion. Attempts to eliminate those errors either work with depth information [7], without depth information [9, 18], or by estimating occurrences between parts [28]. Another common difficulty is distinguishing between Observer-Centric (OC) versus Person-Centric (PC) body labels [10]. Determining whether an arm or leg is a left or right limb requires knowing whether the subject is facing towards or away from the camera. Both issues, occlusion and left/right ambiguity, are easily solved in our framework by leveraging 3D motion capture data to guide interpretation of 2D image detections.

2.5. Temporal Alignment

Time alignment presents three levels of difficulty. The first is correlating video recorded simultaneously with the mocap data (Simultaneous). The second is correlating video recorded at a different time but of the same subject (Intra-subject). The third and most difficult is correlation of video and mocap from two independent subjects (Inter-subject).

A 2 DoF linear transformation allows for time scaling and offset with as little as one pair of time series. Our linear transform approach leverages the head position as the cross-modal feature. The change in height over time of a subject's head is relatively insensitive to viewing angle for upright cameras, and is also easily computed from the mocap data. For the special case of simultaneous mocap-video recording, linear mapping is capable of providing optimal temporal correlation. Intra-subject pairing requires more than slope/intercept transformations because, no matter the skill level, each subject has temporal variations in performance that cause nonlinear temporal differences between capture sessions. Inter-subject pairings face additional challenges due to variable skill level, physical ability, and body attributes (height, weight, gait) of the different subjects. Both intra- and inter-subject pairings require a higher degree of freedom nonlinear transformation for temporal alignment, but a linear mapping can still provide a reasonable initial time alignment to then be refined.

The current state-of-the-art method for temporal mapping comes from Zhou and De La Torre, who propose three different methods for aligning temporal data: Dynamic Time Warping (DTW); Canonical Time Warping (CTW)[32]; and Generalized Canonical Time Warping (GCTW) [29, 31]. DTW is a dynamic programming approach based on aligning sequences to minimize total ℓ_2 distance. DTW has the limitation that it cannot weight the feature vectors or be used with multi-modal data since both datasets need to have the same dimensionality. CTW combines the DTW algorithm and Canonical Correlation Analysis to perform feature selection and dimensionality reduction while aligning signals of different dimensions. GCTW extends CTW by allowing multi-set analysis.

2.6. Other Datasets

There are NO existing multi-modal datasets comparable to our Taiji dataset. Berkeley MHAD dataset [26] used by [30] contains a small collection of atomic scripted *movements*, but the data have little relationship to our *action* performance focus and contain few repeated performances. The Human3.6M dataset [13, 4] is a mix of unscripted performances with a variety of takes per performer. However, it only contains subjects performing unrepeatable, free-form simple *activities*, which are also not relevant to this research. The CMU Graphics Lab Motion Capture Database [16] contains a broad range of performances and performers. Most of the data are *movements* or multiple *activities* but almost none of the *activities* are repeated and none are scripted. As far as the authors are aware, aside from the dataset we have collected, there are no multi-modal datasets (mocap with video) containing scripted and repeated *actions* available for experimental testing and evaluation.

3. Our Approach

Figure 2 presents an overview of the proposed approach. The core is an interleaving of temporal and spatial alignment routines, aimed at bringing the 3D mocap model data into alignment with the 2D video. Temporal and spatial alignments run in two sequential stages, *global* and *local* alignment. Loosely speaking, global alignment performs a linear estimation to coarsely fit the mocap data to the video data, while local alignment estimates a nonlinear fit.

3.1. Preparing Input Datasets

3.1.1 Joints Extracted from Mocap

The reference mocap model contains a time series of 12 labeled 3D joint positions providing information about position of body segments and movement of each labeled joint with a joint position accuracy of less than 0.5mm. The mocap model thus provides prior information in the form of a complete, 4D "script" of what poses and motions to expect at each stage of a 24-form Taiji performance. The raw motion capture data provided by Vicon Nexus software can suffer several deficiencies that need to be corrected to provide a clean reference model, and Vicon's provided plug-in gait model does not sufficiently address them. We have therefore developed a homemade algorithm that performs the following steps: automated marker labeling; missing data replacement (gap filling); high frequency noise filtering; and marker to joint modeling. These cleaning steps result in every frame of a mocap sequence having a 3D joint measured with sub-millimeter accuracy for each of our 12 joint locations. This post-processing requires one hour per minute of raw mocap data, averaging 5 hours of cleaning per capture. Cleaning mocap data is a task required for any mocap data capture and is implemented in an automatic process integrated as part of the Vicon system.

3.1.2 Joints Extracted from Video

Given video of a subject performing 24-form Taiji, we apply 2D joint detection by Convolutional Pose Machines (CPM) [27, 3] to extract the 12 joints of a performer in each video frame. This is an image appearance-based method and quality of extracted joints is much lower than the mocap measurement data. Because the CPM joint data is detected independently for each frame, the detections are noisy and inconsistent, especially for highly mobile body parts like wrists, elbows, and ankles. The method can not find self-occluded joints, and may yield duplicate joint detections. We have developed an automated, video-based method for post processing the raw CPM joint detections to "clean" many of these errors. A constant velocity Kalman filter model is applied to track detected joints of a single 2D subject, verify/correct joint spatial consistency, fill in short periods of missing data, correct left/right label swaps, detect/remove outliers, improve the temporal consistency of detected joints, and smooth high frequency localization noise. After this completely automated cleaning process, some self-occluded joints may still be missing (red curve in Figure 4A,H) but there are improvements to consistency and the periods of missing detections are shortened. Video joint cleaning takes approximately 10 minutes and is required independent of the application of our method.

Additionally, we have observed that the visually defined joint locations marked by CPM [3] differ from the biomechanical points of rotation provided by the Vicon software. Thus, when comparing CPM joints with ground truth projected mocap joints, there can be a large spatial difference, as highlighted in Figure 3. The most accurate CPM joint detections are the hips (purple curve) with an average of 80% of detected hip joints being located within 10 pixels of the ground truth mocap joint location. The accuracy is significantly worse for wrists, with an average of 20% of detected video joints being more than 23 pixels from the corresponding mocap joint. These errors are not constant 2D offsets and retraining the network is not a viable option due to the expense of capturing mocap data. Figure 3 is interpreted in this paper as showing the radius of uncertainty for the true position of a detected joint. For example, hips detected by CPM have an 80% likelihood to be within 10 pixels of the ground truth mocap location. This inherent noise in the video joint detection process creates a minimum threshold (noise floor) for detection accuracy.

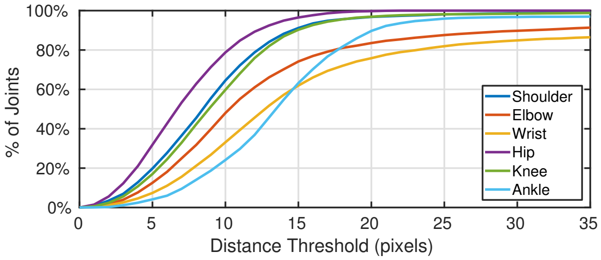

Figure 3: Accuracy of detected video joints generated by [3] as compared to measured ground truth mocap joints, shown as the percentage of detected joints within a given distance of ground truth. 80% detection occurs within 10-23 pixels depending on the joint type.

3.1.3 Video Mosaicing

Our spatial alignment method assumes a stationary camera view. However, it can be adapted for panning and zooming video by first performing video mosaicing to align each frame within a panoramic field of view. We generate a panorama by iteratively aligning a series of video frames using 3 DoF (translation+scale) transformations computed by RANSAC from noisy correspondence matches between sparse, corner intensity patches. More sophisticated methods could be used; video mosaicing is a well-developed area [24] and is not a focus of this work.

3.2. Global Temporal Alignment

As an initial step, a global temporal alignment estimates a linear time warp providing scale and offset between the mocap and video time series data. Since this is the first step in spatiotemporal alignment, we do not assume a known viewing direction. However, assuming the very common case of upright camera view, we use the position of the subject's head over time. In the mocap data this is simple to extract by following the bottom of the 3D head (neck joint) of the subject relative to the floor (Z=0 plane). In the video data, subject's head height is provided by taking the maximum of the vertical distance between the detected head and the ankles of both feet, since one foot is always on the ground in 24-form Taiji.

Global temporal alignment is a two parameter linear mapping between the mocap and video sampling time indices: $t_{video} = a * t_{mocap} + b$. The temporal offset b is determined by correlating normalized derivatives of the two height signals. Time scale a is computed as the ratio between known video and mocap sampling rates.

3.3. Global Spatial Alignment

A correct alignment between camera coordinates and video allows 3D motion capture joints to be projected into spatial registration with 2D joints of a performer in video. Since the video subject is not necessarily wearing visible markers, we use approximate 2D joint locations determined by the CPM algorithm[3].

We use multiple frames spread out across the sequence to maximize the spread of observed points used to compute the spatial alignment. Although off-the-shelf perspective-n-point (PnP) registration [17] works well for Simultaneous recordings, it fails for intra- and inter-subject alignment. We therefore adopt a stochastic search method, Markov Chain Monte Carlo (MCMC), which has been shown previously to be a viable general purpose method for robustly estimating camera pose [15].

Let the camera projection model that projects 3D point P_i into 2D point p_i be $p_i = \Phi(P_i|R, c, \theta)$ for known internal camera parameters θ and external camera parameters (R, c) that we want to solve for. To assess quality of a hypothesized R and c, we define a Gibbs likelihood function:

$$\mathcal{L}(R,c) \propto \exp\{-\sum_{i=1}^{N} D(p_i, \Phi(P_i|R,c,\theta))\} \tag{1}$$

where D measures distance between 2D image points. We use a robust distance function that clamps the Euclidean distance at a max threshold (e.g. 100 pixels). Using MCMC to explore modes of this likelihood function is equivalent to exploring camera poses having high likelihood.

The key to an efficient MCMC sampler is to define proposal "moves" that map current parameter values into new values that are likely to score as well or better than the current state. We define two proposal moves: a diffusion move on camera location, and a novel modified diffusion move over camera rotation. With these two move proposals, searching for camera location and orientation proceeds iteratively using the standard Metropolis-Hastings algorithm.

Location diffusion: Let (R, c) be the current state. Sample a 3D offset vector from an isotropic zero-mean Gaussian $\delta \sim N(0, \sigma^2 I)$ and propose a new state $(R, c+\delta)$. We currently use a standard deviation σ of 500mm.

Rotation diffusion: Let (R, c) be the current state. Perturb azimuth, elevation and roll angles by zero mean Gaussian noise (standard deviation 1 degree) to form a proposed rotation matrix S. Although it is tempting to make the new state proposal be (S, c), this yields a bad move because small changes in camera orientation can lead to large offsets of 2D projected points in the image, resulting in a high rejection rate. To fix this problem, when proposing a new orientation matrix S we also solve for a location d so that the camera-centered coordinates of all 3D marker points stay roughly the same. This is set up as a least squares problem minimizing the following equation over d:

$$E(d) = \sum_{i=1}^{N} \|S(p_i - d) - R(p_i - c)\ .\| \tag{2}$$

The location that minimizes E is

$$\hat{d} = (I - S^T R)\bar{p} + S^T R\, c \tag{3}$$

where $\bar{p} = \sum_i p_i / N$ is the center of mass of the 3D marker points across all images being used. The new proposed state becomes $(S, \hat{d}\,)$.

3.4. Local Temporal Alignment

Local temporal alignment is achieved with canonical time warping (CTW) [32], which nonlinearly maps one time sequence to another to allow short-duration localized time expansions or dilations. This nonlinear warping is needed because there are fluctuations in performance speed even for the same subject trying to repeat the same routine.

To achieve more accurate and flexible temporal alignment per frame, we adapt the techniques of [32] to use the joints of the 3D mocap model projected into 2D using the previously calculated global spatial alignment. These projected 2D mocap joints are then paired with the corresponding 2D joints detected from the video by CPM. All joint pairs are used to determine a frame by frame time warping constrained by the restriction of maintaining causality with a monotonically increasing warp. CTW is used to seek a nonlinear time warping function that minimizes

$$J(V_x, V_y, W_x, W_y) = \|V_x^T X W_x - V_y^T Y W_y\|_F^2 \tag{4}$$

where X and Y are the two multidimensional joint signals to be aligned, W_x and W_Y specify the time warping, and V_x and V_y project the multi-dimensional signals into a canonical coordinate system that maximizes their correlation. Subject to a set of monotonicity and boundary constraints, the time warping components W_x and W_y are computed optimally using dynamic programming.

3.5. Local Spatial Alignment

After nonlinear time warping, spatial alignment is performed again to re-estimate camera pose, this time using the updated temporal mapping of 3D mocap frames to 2D image frames. We use the same MCMC-based camera pose estimation algorithm described in Section 3.3. This loop of refinement, using time warp T_i to estimate camera pose (R_i, c_i), which is then used to estimate a new time warp T_{i+1}, may be iterated multiple times until convergence. In our experiments to follow, we require four iterations or less to achieve steady state of spatiotemporal alignment.

Even after multiple refinement iterations, mocap data may not project perfectly onto the corresponding video frames due to subject and performance variability. To address this remaining error, an additional 2D translational offset for each frame is calculated to bring the projected mocap points more closely into alignment with detected 2D CPM joint locations, and is computed as the median of the remaining point-to-point residual difference errors. This is a form of non-rigid spatial alignment, since the resulting camera projection across the sequence can no longer be described by a single center of perspective.

4. Experimental Results

We have collected a dataset of 24-form Taiji performances by recording both motion capture and video using a Vicon Nexus motion capture system. We evaluate three types of matching between motion capture data and video:

1. Simultaneous (same time, same subject);
2. Intra-Subject (different time, same subject);
3. Inter-Subject (different time, different subject).

4.1. Data Collection

Our Vicon data collection system consists of the Nexus software and 12 IR Cameras providing sub-millimeter mocap accuracy in a 20x30x14 foot capture volume at 100Hz sampling rate. We captured 5 subjects performing 24-form Taiji multiple times. Subject 5 was also captured using two integrated Vicon Vue 1080HD cameras at 50FPS. Video from the Vue cameras is temporally synchronized and spatially calibrated with the motion capture cameras, with precise calculation of position, orientation, and optical distortion all integrated as part of the Nexus software, yielding a video to motion capture alignment error of less than 2 pixels. All other lab subjects (1,2,3,4) were recorded using independent Sony XV2100 720x480 DV cameras recording at 29.97FPS. The Sony cameras are not synchronized nor calibrated for position and orientation with respect to the mocap system. They have been calibrated to remove optical lens distortion, however. An additional video performer, subject 6 in Section 4.4, is from a downloaded public YouTube video captured by an unknown camera with 1280x720 resolution.

4.2. Simultaneous Alignment Results

Our data collection system is a hardware/software system that provides temporal synchronization and calibrated spatial alignment of video and mocap. For the special case of calibrated, synchronously recorded datasets, we therefore know the ground truth temporal alignment, spatial alignment, and camera viewpoint as determined by the highly accurate Vicon system. This spatiotemporal ground truth provides a foundation for quantitative evaluation of our algorithm as well as CPM joints detected by [3].

Figure 4A shows the spatiotemporal error of our method compared to ground truth for mocap of subject5-session5 (5-5). Our algorithm has a mean temporal error of 5ms with a standard deviation of 159ms, a median temporal error of 2 ms, and a peak error of 860 ms. A large peak in temporal error (orange curve) of 670ms during video frames 8800-9000 (Figure 4E) is caused by a larger than usual percentage of missed CPM joint detections (only 66% detected). There are also temporal errors at the start (530ms at frame 50) and end (860ms at frame 13160) of the performance when the subject is standing still – these have minimal effect on spatial error or visual alignment. Figure 4 shows qualitative examples of video frames with both good (B-D) and poor (E-G) spatial alignment results. (A complete video with overlaid results is included in supplemental material.) Simultaneous results requires iterating spatiotemporal warping twice to achieve a steady state temporal warp.

Simultaneous	CPM [3]			Ours		
Joint	$\mu\pm\sigma$	Median	% Det	$\mu\pm\sigma$	Median	%Det
Right Shoulder	8.6±5.8	7.7	100.0	**8.0±4.1**	**7.5**	100
Right Elbow	14.6±18.0	10.0	100.0	**8.4±5.7**	**7.2**	100
Right Wrist	19.6±28.4	11.3	97.6	**11.5±8.4**	**9.7**	100
Left Shoulder	9.9±**5.1**	9.5	100.0	**8.0**±5.2	**7.5**	100
Left Elbow	18.1±22.0	10.8	99.6	**8.8±6.0**	**7.5**	100
Left Wrist	25.6±33.1	13.8	94.9	**11.7±8.1**	**9.6**	100
Right Hip	7.5±4.0	6.9	100.0	**4.8±3.2**	**4.3**	100
Right Knee	11.6±8.6	10.9	99.9	**6.1±4.1**	**5.3**	100
Right Ankle	22.3±58.8	13.6	99.9	**10.5±16.8**	**9.4**	100
Left Hip	6.9±3.7	6.5	100.0	**5.7±3.2**	**5.0**	100
Left Knee	8.3±11.4	**6.6**	100.0	**7.8±4.2**	7.0	100
Left Ankle	16.2±16.2	13.7	99.9	**10.5±5.0**	**9.5**	100
All Joints	14.1±24.4	9.8	99.3	**8.5±7.4**	**7.4**	100

Intra-subject	CPM [3]			Ours		
Joint	$\mu\pm\sigma$	Median	% Det	$\mu\pm\sigma$	Median	% Det
Right Shoulder	12.5±8.4	10.8	100.0	**9.4±6.8**	**7.8**	100
Right Elbow	18.4±18.7	13.3	100.0	**12.1±8.9**	**10.0**	100
Right Wrist	26.8±30.3	17.3	97.9	**17.4±15.5**	**13.8**	100
Left Shoulder	13.5±8.1	9.4	100.0	**10.0±6.8**	**8.5**	100
Left Elbow	21.6±21.4	15.4	99.7	**12.4±8.8**	**10.2**	100
Left Wrist	29.2±33.1	17.7	95.0	**17.6±14.2**	**13.9**	100
Right Hip	8.8±**6.3**	7.2	100.0	**7.6**±19.3	**6.0**	100
Right Knee	9.9±**8.4**	8.1	100.0	**9.6**±18.6	**7.9**	100
Right Ankle	21.5±60.9	11.5	100.0	**12.7±23.6**	**10.5**	100
Left Hip	8.5±**5.2**	7.5	100.0	**7.1**±18.2	**5.6**	100
Left Knee	9.1±**10.5**	**6.7**	100.0	**8.9**±17.3	7.8	100
Left Ankle	15.1±**14.4**	12.1	99.9	**11.5**±17.6	**9.6**	100
All Joints	14.4±23.8	9.5	99.4	**11.4±15.9**	**8.7**	100

Table 1: Quantitative analysis of spatial alignment relative to ground truth for CPM [3] and our method with errors measured as ℓ_2 pixel distances (lower is better) for one subject, a beginner. Best values are shown in **bold**. Also shown are percentages of joints detected. Our method locates all joints all the time because it is based on 3D mocap data.

Table 1 presents a statistical evaluation of spatial alignment both by our algorithm and the CPM-detected joints [3] compared to ground truth. The table shows μ, σ, and median of offset errors for each of the 12 joints and all joints combined. This can be done only for simultaneous and intra-subject pairings where the ground truth is measured spatially and temporally. The table shows our method produces better localization overall and lower standard deviations for most joints. Our method has 100% detection, while [3] does not because it can not detect occluded joints.

4.3. Intra-subject Alignment Results

For mocap subject5-session5 and video subject5-session4 (5-4), linear time warping is no longer sufficient due to variations in pacing and movement between the two performances. Time warping is difficult to evaluate quantitatively in non-simultaneous captures because of the inability to capture temporal ground truth. However, we can still evaluate the alignment of projected mocap data overlaid on the video, and consider residual spatial errors as due to the time alignment's secondary impact. Figure 4H shows that our method when applied to intra-subject data produces

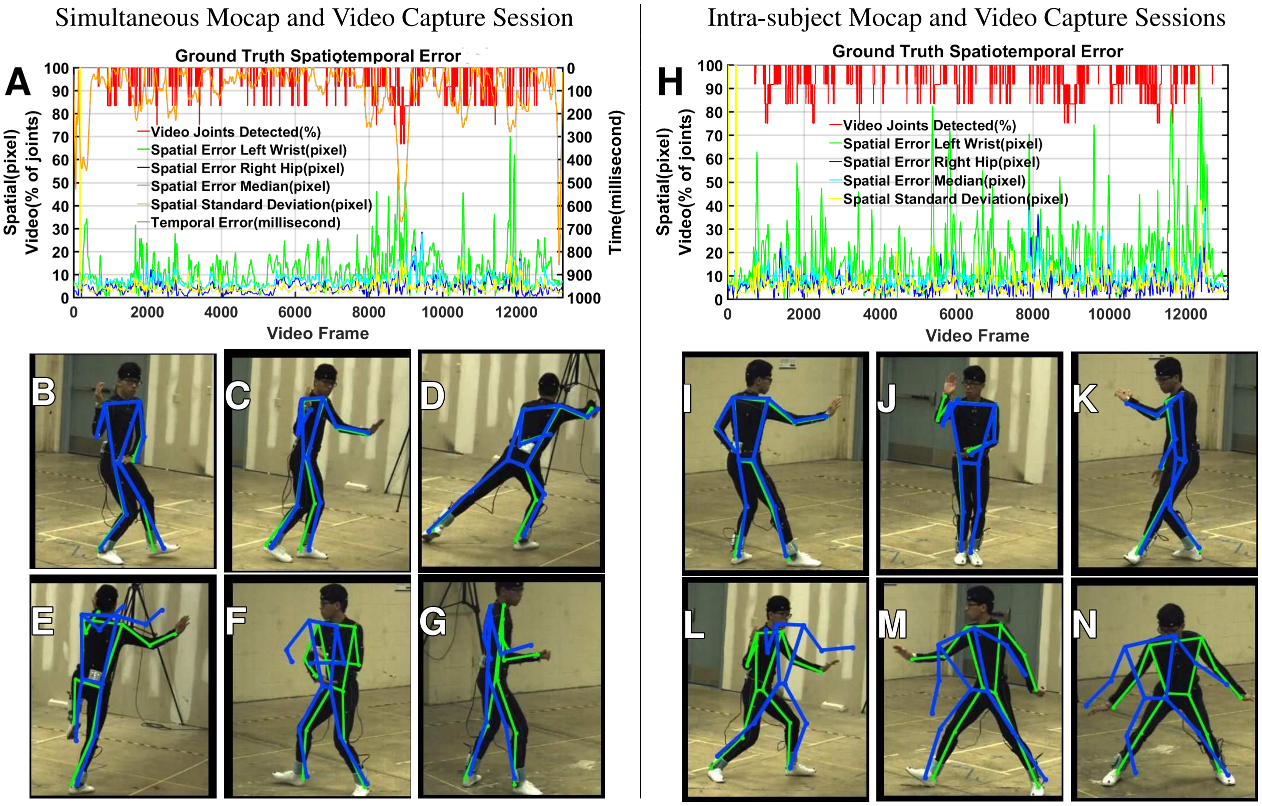

Figure 4: Spatiotemporal error of synchronously recorded mocap and video of subject 5, session 5 (A-G). Intra-subject pairing with mocap subject 5, session 5 and video subject 5, session 4 (H-N). A,H: Graphs demonstrating the accuracy of the matching where average spatial errors are 8.5 and 11.4 pixels even with up to 33% and 25% of joints not detected by [3] (Figure 3) for simultaneous and intra-subject video recordings, respectively. The maximum error in mocap data collection is 2 pixels based on the Vicon calibration data. B-D and I-K are examples of low error frames while E-G and L-N are examples of high error frames. The frame numbers for each image are B:2561, C:7885, D:9108, E:9000, F:10526, G:11920, I:3888, J:7556 K:10275, L:7900, M:9822, N:12509.

comparable spatial alignment results to simultaneous mappings. The spatial error mean is 11.4 pixels, below the 16 pixel average offset needed for joints from [3] as shown in Figure 3. (A video with overlaid results is included in the supplemental material.) Qualitative overlays are shown for both good (I-K) and poor (L-M) spatial alignments. This intra-subject pairing requires iterating spatiotemporal warping three times to achieve a steady state temporal warp.

4.4. Inter-subject Alignment Results

A large number of historical videos of Taiji performances exist. We wish to leverage this rich set of video resources to study Taiji masters' skills, analyze diversity of performance styles, and preserve cultural heritage. We take a first step towards this goal by aligning mocap data from subject1-session1 recorded in our lab against a public YouTube video of Master performer Jiamin Gao (subject 6) of China [8]. We did not record the Gao video, and had no control over acquisition conditions or camera parameters.

Neither temporal nor spatial ground truth is available for this inter-subject experiment. However, qualitative visualization of the alignment shows strong temporal correlation as evidenced by synchronized movements (a video with overlaid results is included in the supplemental material). As can be seen from Figure 5, the pose of the subject in key frames matches reasonably well with the pose of the overlaid mocap data, indicating good temporal alignment. This inter-subject pairing requires iterating spatiotemporal warping five times to achieve a steady state temporal warping.

In previous experiments, we quantified spatial error with ground truth mocap joint locations in the video and then compared our methods spatial error compared to that ground truth. We also did the same for joints detected by CPM [3] as a comparable method. However, since the video was selected from YouTube there are no ground truth error locations in the video as there is no mocap data recorded. As a result, spatial error can not be determined for either our method or CPM joints and therefore Figure 5 only includes qualitative examples.

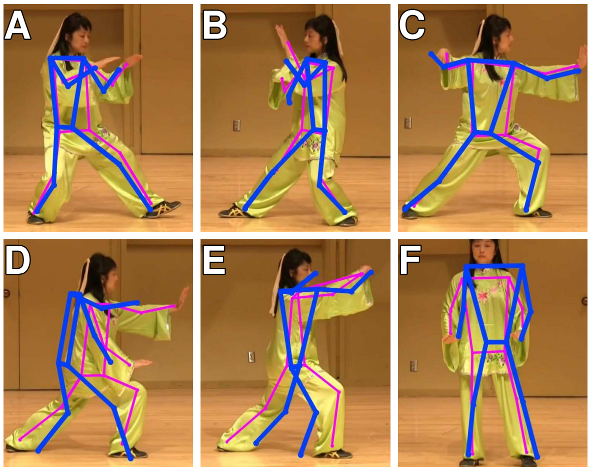

Figure 5: Spatiotemporal comparison of an inter-subject pair, mocap 1-1, video 6-3. A-F visualize sample frames overlaid with detected CPM video joint skeletons (Pink) and those produced by our algorithm (Blue). A, B, C are low error frames (3341, 3894, 4439). D, E, F are high error frames (1804, 5819, 8215). Complete video is included in the supplemental material.

The key technical hurdle of inter-subject mappings is subject variability. Each subject has a different body shape and size, and varying skill level differences involving range of motion, balance, flexibility and stride. For example, a novice may not be able to lift their leg as high or lunge as low as an advanced or master performer. All of these subject and performance differences negatively impact the quality of the resulting spatial joint alignment. Another contributing factor to the diferences between our method and CPM [3] is presented in Figure 3, where CPM joints are measuring slightly different body locations than Vicon mocap joints, and can be tens of pixels off in the image.

Figure 6 visualizes a small section of the calculated time warping between mocap of Subject5-Session2 and video of Subject1-Session1. The vertical "cliff" in the mapping is a time period where roughly 2.4 seconds of mocap data is compressed to align with merely 1/3 of a second of video data. This happens because the mocap subject incorrectly performs four copies of a "cloud hands" movement instead of the usual three. As a result, mocap time needs to be accelerated to catch back up with the video performer, who was doing the correct sequence of movements. (A video clip of our method adapting to a performer mistake is included in the supplemental material.)

5. Conclusions and Future Work

We have presented an approach to temporally and spatially align motion capture data of a subject performing 24-form Taiji with video data of either the same or different performer. This preliminary work is a stepping stone for accurate localization of joints and body segments in 2D video, a precursor for comparison and analysis of the quality of two independent video performances.

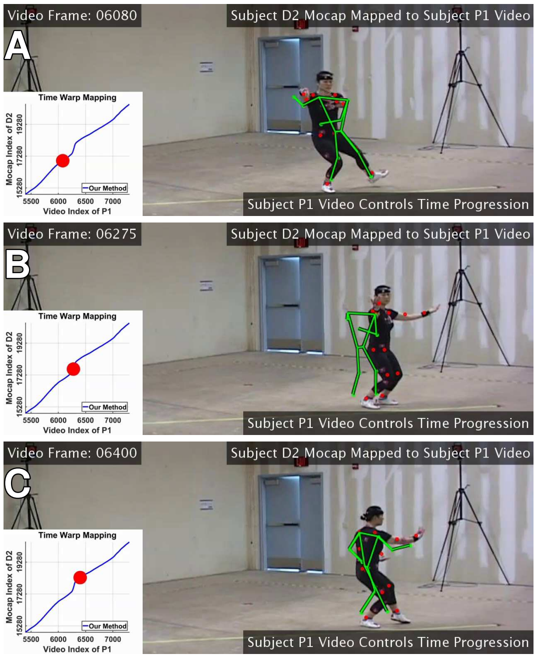

Figure 6: Inter-subject spatiotemporal alignment: when the mocap subject (green skeleton) mistakenly performs an additional unscripted action, our non-linear time warping method is capable to align the two sequences by speeding up the mocap subject motion. A-C demonstrates the alignments before, during and after the mistake. Example video is included in the supplemental material.

Inter-subject alignment is challenging due to large variation in body shape and performance skill/style, requiring far more spatial transformation flexibility than is available by choosing a 6-DoF camera viewpoint. Our future work will explore adding physically-meaningful parameters into the MCMC spatial alignment search to adjust for different body sizes and limb rotation angles.

While the CPM method [3] used for detecting joints in video is state of the art, quantitative evaluation of its joint localization accuracy has not been done before, and our results (Figure 3; Table 1) indicate that CPM joints are offset from the biomechanical rotational joints measured by the Vicon mocap system. Improvements in 2D image based joint detection, such as reducing missed detections and increasing localization accuracy, could yield the largest improvement to both temporal and spatial alignment across all pairing types.

Acknowledgments. This work was supported in part by NSF grant IIS-1218729. The authors thank Professor Pingping Xie, Dr. Wenlin Fan and Mr. David Nguyen for their support of this work.

References

[1] A. F. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Royal Society Workshop on Knowledge-based Vision in Man and Machine*, 352:1257–1265, 1997. 2

[2] A. Bulat and G. Tzimiropoulos. *Human Pose Estimation via Convolutional Part Heatmap Regression*, pages 717–732. Springer International Publishing, Cham, 2016. 2

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4, 6, 7, 8

[4] C. S. Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *International Conference on Computer Vision*, 2011. 3

[5] J. Charles, T. Pfister, D. Magee, and A. Hogg, D. Zisserman. Personalizing human video pose estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[6] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2

[7] X. Chen and A. Yuille. Parsing occluded people by flexible compositions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3945–3954, 2015. 3

[8] Chinese Wushu Academy. 24 form Yang style taichi chuan by "queen of taichi " master Jiamin Gao of US wushu center. 7

[9] N.-G. Cho, A. L. Yuille, and S.-W. Lee. Adaptive occlusion state estimation for human pose tracking under self-occlusions. *Pattern Recognition*, 46(3):649–661, 2013. 2, 3

[10] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *Computer Vision–ACCV 2012*, pages 138–151. Springer, 2012. 3

[11] M. P. Garofalo. Taijiquan 24 form Yang style. [Online; accessed August 6, 2017]. 1

[12] H. Ghasemzadeh and R. Jafari. Coordination analysis of human movements with body sensor networks: A signal processing model to evaluate baseball swings. *IEEE Sensors Journal*, 11(3):603–610, March 2011. 2

[13] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, Jul 2014. 3

[14] A. Jain, J. Tompson, Y. LeCun, and C. Bregler. Modeep: A deep learning framework using motion features for human pose estimation. In *Asian Conference on Computer Vision, Singapore*, pages 302–315, 2014. 2

[15] T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. K. Mansinghka. Picture: A probabilistic programming language for scene perception. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4390–4399, 2015. 5

[16] C. M. G. Lab. CMU graphics lab motion capture database. [Online; accessed August 6, 2017]. 3

[17] V. Lepetit, M. Moreno-Noguer, and P. Fua. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision*, 81(2):155–166, 2009. 5

[18] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, 2017. 3

[19] J. Mengoli. Method and apparatus for automating motion analysis, Feb. 4 2003. US Patent 6,514,081. 2

[20] A. Muro-de-la Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla. Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors*, 14(2):3362–3394, 2014. 2

[21] A. Newell, K. Yang, and J. Deng. *Stacked Hourglass Networks for Human Pose Estimation*, pages 483–499. Springer International Publishing, Cham, 2016. 2

[22] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *IEEE International Conference on Computer Vision, Santiago, Chile*, pages 1913–1921, 2015. 2

[23] J. Romero, M. Loper, and M. J. Black. FlowCap: 2D human pose from optical flow. In *Proc. 37th German Conference on Pattern Recognition (GCPR)*, volume LNCS 9358, pages 412–423. Springer, 2015. 2

[24] H.-Y. Shum and R. Szeliski. Systems and experiment paper: Construction of panoramic image mosaics with global and local alignment. *International Journal of Computer Vision*, 36(2):101–130, Feb 2000. 4

[25] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA,*, pages 1653–1660, 2014. 2

[26] R. Vidal, R. Bajcsy, F. Ofli, R. Chaudhry, and G. Kurillo. Berkeley mhad: A comprehensive multimodal human action database. In *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision*, WACV '13, pages 53–60, Washington, DC, USA, 2013. IEEE Computer Society. 3

[27] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 2, 4

[28] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2878–2890, 2013. 3

[29] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1282–1289. IEEE, 2012. 3

[30] F. Zhou and F. De la Torre. Spatio-temporal matching for human detection in video. In *Computer Vision–ECCV 2014*, pages 62–77. Springer, 2014. 2, 3

[31] F. Zhou and F. D. la Torre. Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):279–294, Feb 2016. 3

[32] F. Zhou and F. Torre. Canonical time warping for alignment of human behavior. In *Advances in neural information processing systems*, pages 2286–2294, 2009. 3, 5

[33] S. Zuffi, J. Romero, C. Schmid, and M. J. Black. Estimating human pose with flowing puppets. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3312–3319, 2013. 2