Biologically Inspired Cognitive Architectures (2014) 7, 15–25

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bica

ELSEVIER

RESEARCH ARTICLE

Extracting movement, posture, and temporal style features from human motion

S. Ali Etemad [a,*], Ali Arya [b]

[a] *Department of Systems and Computer Engineering, Carleton University, Canada*
[b] *School of Information Technology, Carleton University, Canada*

Received 5 August 2013; received in revised form 2 October 2013; accepted 3 October 2013

KEYWORDS
Motion;
Gaits;
Secondary features;
Style;
Affect;
Optimization

Abstract
Small variations in biological motion responsible for perception of characteristics, styles, or affects of the person performing the actions, are referred to as secondary features. This paper presents a novel method for separating and extracting spatiotemporal sets of secondary features from human motion data. The technique employs a dataset of sequences and identifies a corresponding neutral sequence through maximizing a similarity index based on correlation. Specific control points or temporal cues are then distributed through the input sequence. Distribution is carried out with the goal of maximizing an objective function successive to time warping. The optimized set of cues are used to reconstruct the neutral component of the signal using cubic splines. Accordingly, both spatial (movement and posture) and temporal secondary features are extracted from the stylistic input sequence. To illustrate one of the possible applications of the proposed technique, style translation is carried out. We illustrate that our proposed system can be used to extract various classes of secondary features from different actions such as walking, jumping, and running.

Introduction

Recently, computational human motion studies have drawn much deserved attention. This is in part due to advancements in computing capabilities, which in turn have resulted in a growth in applications such as animation, interactive games, and virtual worlds.

Analysis of style and affect in human motion is often aimed at extracting, classifying, interpreting, and synthesizing affective and stylistic features. These features can be used in a wide range of applications especially in the field of human–computer interaction (HCI), ranging from games and animation (Ma et al., 2010) to biomedical engineering (Picard, 2009). The complex nature of human motion, as well

Corresponding author.
E-mail addresses: ali.etemad@carleton.ca (S.A. Etemad), arya@carleton.ca (A. Arya).

2212-683X/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bica.2013.10.001

as the wide range of stylistic features including actor attributes and characteristics such as gender, age, energy, mood, emotion, health, and inherited traits, all add to the difficulty of the task. Furthermore, behavior and cognition depend on a variety of parameters (Kleinsmith, De Silva, & Bianchi-Berthouze, 2006), making the task even more difficult.

Previously, a model was proposed for describing the relationship between basic action classes and styles (Etemad & Arya, 2010). In this model, inspired by the work of Laban (Guest, 2005), the main action and related styles were named primary themes and secondary themes respectively. Accordingly, we name the stylistic features of human motion signals, secondary features (SF). These features have shown to exist in the form of spatiotemporal add-ons (Rose, Cohen, & Bodenheimer, 1998; Unuma, Anjyo, & Takeuchi, 1995) and appear in three different forms, namely movement, posture, and temporal variations (Amaya, Bruderlin, & Calvert, 1996; Normoyle, Liu, Kapadia, Badler, & Jörg, 2013; Thrasher, Van der Zwaag, Bianchi-Berthouze, & Westerink, 2011; Troje, 2002; Troje, Cord, & Lavrov, 2005). Psychophysics experiments and statistical analysis of recorded motion data are often used to characterize these features (Blake & Shiffrar, 2007) while computational techniques are developed for motion control (Coros, Beaudoin, & van de Panne, 2010) and style translation (Hsu, Pulli, & Popovic, 2005). These computational approaches, some of which are reviewed in Related work section, all provide effective and valuable solutions for the problem. However, we believe it is essential to be able to systematically extract the SF sets in spatiotemporal domain. This will allow researchers in both psychology and multimedia fields to employ the features for related studies. This paper provides a solution for this problem.

In this paper, we propose a technique for separating and extracting secondary features from primary actions. This approach is valuable as the extracted features can be analyzed towards behavioral and perceptual studies or directly used in animation and multimedia systems. Our proposed system utilizes a dataset of motion sequences and through maximizing a similarity index, selects a neutral sequence that best corresponds to the inputs stylistic motion. Correlation optimized time warping (CoTW) (Etemad & Arya, 2013; Nielsen, Carstensen, & Smedsgaard, 1998) is used in the process as the sequences need to be accurately aligned. A set of temporal cues are then distributed through the sequence while maximizing a correlation-based objective function. Spatiotemporal cubic splines are used to approximate the neutral component of the input. Movement and posture feature components are subsequently computed and extracted. The temporal feature is calculated as the third component of the SF set. The proposed algorithm is tested on various examples and the results are provided. As an application of the algorithm, style translation is carried out, through which, the SF set of stylistic inputs are transferred onto neutral ones. The overall schematic of the system is illustrated in Fig. 1.

Related work

Inverse kinematics (IK) and control-based methods are popular approaches for modeling motion and stylistic variations. Grochow, Martin, Hertzmann, and Popovic (2004) used IK for synthesizing natural and stylistic human motion. Coros et al. (2010) employed controllers for synthesizing stylistic gait sequences. Liu, Yin, van de Panne, and Guo (2012) used linear feedback control to learn a variety of motion skills through dynamic simulation. Inverted pendulum-based models have also been explored in this area (Tsai, Lin, Cheng, Lee, & Lee, 2010).

Dimensionality reduction is a common tool for mapping motion sequences (De La Torre & Black, 2001). Shapiro, Cao, and Faloutsos (2006) employed independent component analysis for extracting motion style features. Learning-based techniques have also been proposed and utilized. Brand and Hertzmann (2000) employed probabilistic models for interpolation and extrapolation of different styles for synthesis of new stylistic dance sequences using a cross-entropy optimization structure which enables their style machine to learn from various style examples. A neural network setting was used in (Etemad & Arya, 2009) which learned neutral-to-stylistic mappings and used it for style translation. Hsu et al. (2005) proposed LTI system identification for style translation successive to a novel time and space warp technique. Finally, Ma et al. (2010) used Bayesian networks along with kriging for modeling style and variation in animation of human motion. Kriging process uses a set of control points, similar to the approach proposed in this paper.

Relative modeling or editing of motion signals are another popular approach. In one of the earlier works in the field, Amaya et al. (1996) developed a method to alter speed and range of motion to achieve emotional actions from neutral ones. Interpolation/extrapolation systems are common among relative editing techniques. Rose et al. (1998) used this type of editing to blend motion styles successive to time warping. Bruderlin and Williams (1995) employed several methods for altering motion data. These methods included blending through interpolation (successive to alignment), waveshaping, displacement mapping, and filtering. The authors concluded that when filtering is employed, the increase in lower frequency gains are visualized as decreases in intensity of performed actions; an increase in middle band frequencies results in exaggerated movements; and finally by increasing higher frequencies, nervous twitches will be produced. Pullen and Bregler (2002) used motion capture data to add texture to keyframed animation. Their proposed model utilizes

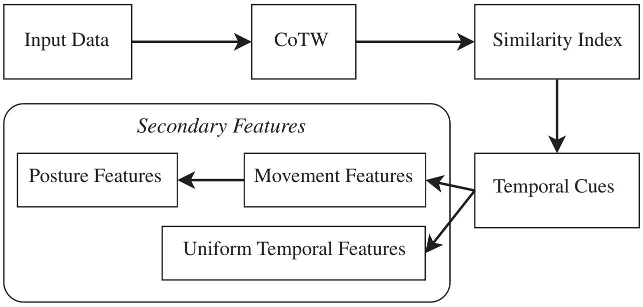


Fig. 1 Overall process of the proposed system for extraction of SF. CoTW is carried out and similarity index is maximized followed by optimization of temporal cues and extraction of the three components of the SF set.

correlations between distinct body parts and adds mid and high-level frequency alterations to keyframed or synthesized signals through a process referred to as texturing. Successful outcome of interpolation/extrapolation methods and filtering/frequency-based techniques presented in these works, confirm that stylistic features are most likely in the form of add-ons that contain frequency differences with respect to the main action. These properties form the basis of our proposed approach.

Linear model and data

Linear model

It has been illustrated that human motion sequences can be represented by:

$$\mathbf{Y} = \mathbf{P} + \sum_{i=1}^{r} \mathbf{w}_i \cdot \mathbf{S}_i + \mathbf{e} \tag{1}$$

where **Y** is the action as perceived, **P** and **S** are the primary and secondary themes (or features) respectively, r represents the number of secondary themes, **w** is the weight associated with each secondary theme, and **e** is the noise present in the data (Etemad & Arya, 2010).

In order to simplify the problem, it is often assumed that $r = 1$, and therefore combinational secondary themes such as young-tired or energetic—feminine are not taken into consideration. Therefore Eq. (1) can be re-written as $\mathbf{Y} = \mathbf{P} + \mathbf{w} \bullet \mathbf{S} + \mathbf{e}$. Having only one secondary theme in the motion reduces the need for a weight factor in the model which is mostly for relative emphasis on particular SFs. Thus, we can conclude $\mathbf{Y} = \mathbf{P} + \mathbf{S} + \mathbf{e}$. The main goal of this research is to extract **S** from **Y** as precisely as possible, and without affecting **P**. Doing so will facilitate the study of spatiotemporal properties of actor characteristics, styles, and affects, regardless of the actions being performed. It will also allow the creation of new sequences with the same primary and different secondary themes (style translation).

Given $\mathbf{Y}_1 = \mathbf{P}_1 + \mathbf{S}_1 + \mathbf{e}_1$, $\mathbf{Y}_2 = \mathbf{P}_2 + \mathbf{S}_2 + \mathbf{e}_2$, and $\mathbf{P}_1 = \mathbf{P}_2$, we have $\Delta\mathbf{Y} = (\mathbf{S}_2 - \mathbf{S}_1) + (\mathbf{e}_2 - \mathbf{e}_1)$. By selecting one of the sequences to be stylistically neutral ($\mathbf{S}_1 = \varnothing$) and minimizing the noise in the data, the SF set of the second sequence can be extracted. This forms the basis for separating the stylistic features from the primary actions using the proposed method. The notion of $\mathbf{S}_1 = \varnothing$ is revisited and further discussed in the following sections.

Data

Motion capture data can be represented by a number of consecutive multidimensional postures variable with time. This model represents each posture with a finite number of markers corresponding to different regions or joints of the body. The motion matrix can be characterized using time series of joint locations or joint angles. Accordingly, the motion matrix D can be represented by $\mathbf{D} = [\mathbf{R}_1\ \mathbf{R}_2\ \cdots\ \mathbf{R}_n]$ where $\mathbf{R}_j = \{R_j^i, i = 1 \cdots m,\ R \in \mathbb{R}\}$ represents the jth DOF (also referred to as trajectory or signal) of the motion sequence for m frames or time instances.

The Carnegie Mellon motion capture dataset (http://mocap.cs.cmu.edu/) and a dataset that we recorded using a Vicon MX40 motion capture system (located at the School of Information Technology, Carleton University) are the two different resources used in this research. The CMU motion capture data have a 96 DOF structure while ours have 60 DOFs. The extra DOFs in the CMU data mostly belong to minor limbs such as fingers and toes, which are often not studied in gait sequences. Also, both models contain *zero-columns* (DOFs which are constant with respect to their parent DOF) which we remove prior to use in our experiments. Ethics approval is secured for our data and experiments. More details regarding the exact number and classes of actions and SF are provided in Result and discussion section.

The proposed technique and dataset include stylistically neutral actions. While it is almost impossible to perform and capture perfectly neutral actions, we asked the actors to display minimal stylistic behavior. The notion of neutral actions has also been employed in other literature such as (Amaya et al., 1996; Heloir, Courty, Gibet, & Multon, 2006; Hsu et al., 2005). Furthermore, the confusion matrix presented in the results section indicates that in general, people have a correct understanding of a *neutral action sequence*.

Correspondence

Many of the proposed techniques for feature extraction and editing, or synthesis of motion sequences are dependent on examples or datasets. For example, (Arikan & Forsyth, 2002; Arikan, Forsyth, & O'brien, 2003; Kovar, Schreiner, & Gleicher, 2002; Ma et al., 2010; Pullen & Bregler, 2002; Wu, Tournier, & Reveret, 2011) are all as such. Similarly, the method proposed in this paper utilizes a dataset of motion sequences. To extract the SF from a stylistic input sequence, a neutral sequence that best corresponds to the input needs to be selected. Therefore, the input is compared to the neutral sequences in our dataset(s) successive to time warping, and the neutral sequence that maximizes a similarity index is selected. Sequences under investigation must be aligned to achieve more reliable results. Therefore, we apply time warping prior to calculation of the similarity index. A warped signal is basically the same signal after nonlinear stretching and compressions at different intervals to maximize alignment with another signal. Here, we first describe the utilized time warping technique followed by the similarity index.

Correlation optimized time warping

Corresponding motion signals, even for similar actions performed by the same actor, contain temporal misalignments. Processing the signals to achieve correct alignment is therefore a vital step towards motion editing. Especially for content-based methods, warping is often a critical step which ensures that correct posture mapping is accomplished. Many have studied and proposed different techniques for spatio-temporal alignment of signals, among which (Hsu, da Silva, & Popović, 2007; Hsu et al., 2005; Sakoe & Chiba, 1978; Witkin & Popovic, 1995; Zhou & Torre, 2009) can be mentioned.

In this paper, we use CoTW (Etemad & Arya, 2013; Nielsen et al., 1998).

CoTW warps an input sequence with respect to a reference with the objective of maximizing a weighted sum of Pearson's correlation coefficient for different DOFs as a representation of similarity between the two sequences. The correlation coefficient in general is defined by:

$$\rho(x,y)=\frac{\sum_{i=1}^{m}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{m}(x_i-\mu_x)^2\sum_{i=1}^{m}(y_i-\mu_y)^2}} \tag{2}$$

where x and y are the two signals for which the coefficient is being calculated, m represents the temporal length of the signals, and μ denotes the mean.

Through CoTW, the input sequence is first divided into a number of segments, λ, each of which is permitted to linearly warp using uniform time warping (UTW) (Fu, Keogh, Lau, Ratanamahatana, & Wong, 2008) in either temporal direction. The room to warp for each segment is called slack and denoted by σ. The optimum λ and σ are calculated using dynamic programming and subsequently used to warp the input sequence while maintaining maximum signal to noise ratio for minimizing the distortion caused by the warping procedure.

CoTW shows several advantages over other warping techniques such as (Hsu et al., 2005; Sakoe & Chiba, 1978; Zhou & Torre, 2009) in terms of alignment, distortion, and smooth warping of signals. Moreover, compared to typically used distance-based cost functions, correlation shows better performance in characterizing similarity between motion sequences (Etemad & Arya, 2013; Nielsen et al., 1998).

Similarity index

In order to select a neutral sequence that best corresponds to the input we define the objective function:

$$G=\sum_{i=1}^{n}\boldsymbol{u}_i\cdot\rho(D_{\text{input},i},\widetilde{D}_{\text{neutral},i}) \tag{3}$$

where D represents the motion data, n is the number of DOFs, and $\boldsymbol{u}$ is a set of weights used in the process. The $\sim$ sign denotes the sequence after warping. Subsequently, the neutral sequence that maximizes G is selected as the corresponding sequence.

In motion, each joint or limb has a different impact and contribution towards execution and perception of a displayed sequence. For example, in (Etemad, Arya, & Parush, 2011), it was illustrated that legs possess the most weight in perception of energy-related features (tired walk vs. energetic walk). Therefore, in Eq. (3), different weights are assigned to different DOFs. The weights may vary for different action classes (for example walk vs. jump). Ad-hoc means are employed for assigning this parameter. For regular gait sequences, the weight of the shoulder and thigh joints are set to almost twice the weight of hand and foot joints. The weights for fingers and toes can be set to zero due to their insignificance. As mentioned earlier, the CoTW process also utilizes a weight set based on which the significance of different DOFs in the warping process is decided. We assign the warping weight set and u to be equal since the aim of both parameters is to customize the processes based on the relative importance of different joints.

Extraction of features

Several key components have been identified for features that compose SF in human motion (Amaya et al., 1996; Normoyle et al., 2013; Thrasher et al., 2011; Troje, 2002; Troje et al., 2005). These features relate to: (a) posture (also referred to as structure), (b) movement (also referred to as dynamics), and (c) timing (also referred to as speed or uniform temporal features). Posture features are those that often stay unchanged through the sequence. Rather, they are changes to the initial pose of the body with which the motion is carried out. Movement features are the changes to the motion signals and vary throughout the sequence. The uniform temporal feature relates to the speed with which the sequence is performed. This feature is often correlated with the secondary theme in the sequence, meaning different styles are displayed with different speeds (Mather & Murdoch, 1994).

Based on the above, given a stylistic (input) sequence and a selected corresponding neutral sequence, we first model a neutral version of the stylistic action over the input. By subtracting the modeled approximation from the input, movement SFs are extracted. Subsequently, we utilize the modeled approximation and the corresponding neutral sequence to extract the posture SFs. Finally, the input and the corresponding neutral sequence are used to calculate the uniform temporal feature.

Movement features

We employ spatiotemporal cubic splines for modeling the neutral component of the input sequence. The non-linear nature and controllability of splines make them very suitable means for many applications where complex data are being modeled. Cubic splines are piecewise third-order polynomials that pass through a number of control points. The control points when projected onto human motion trajectories can be seen as temporal cues for segments of the motion sequence. These models have been widely used in geometric modeling, computer graphics, and other such applications (Bartels, Beatty, & Barsky, 1998; Etemad & Arya, 2011). Generally, a spline function is in the form of:

$$F(i)=\begin{cases} f_1(i) & \text{if } t_1\leqslant i<t_2\\ f_2(i) & \text{if } t_2\leqslant i<t_3\\ & \vdots\\ f_{n-1}(i) & \text{if } t_{n-1}\leqslant i<t_n \end{cases} \tag{4}$$

where $f: [a,b]\to\mathbb{R}$ and $t_1,t_2,\ldots,t_n$ are the control points, meaning $F(i)$ must pass through all n cues. In the cubic form, $f_j(i)$, is defined by:

$$f_j(i)=\sum_{k=0}^{3}c_{j,k}(i-t_j)^k \quad \text{for } j=1,2,\ldots,n-1 \tag{5}$$

Besides the necessity for **F** to pass through all t (which entails Eq. (7)), Eqs. (6) and (8) must also hold true to achieve a continuous and smooth approximation. Moreover, $f_j(i)$ must be continuous for all subintervals of j.

$$f_j(t_j) = f_{j+1}(t_j) \quad \text{for } j = 1, 2, \ldots, n-1 \tag{6}$$

$$f'_j(t_j) = f'_{j+1}(t_j) \quad \text{for } j = 1, 2, \ldots, n-1 \tag{7}$$

$$f'_j(t_j) = f''_{j+1}(t_j) \quad \text{for } j = 1, 2, \ldots, n-1 \tag{8}$$

Solving Eqs. (6)—(8) and including the boundary conditions for the starting and ending points yields $n+1$ equations and $n+1$ unknowns which can be solved for acquiring F.

Generally, when modeling a signal (or in the case of this study, joint angle curve) using cubic splines, using more control points leads to an approximation which follows the actual signal more precisely and accurately, reconstructing the higher frequency curves. Fewer and more widely distributed control points, on the other hand, result in a more loose and general approximation of the signal, leaving out higher frequency components. While the control points can be distributed non-uniformly, in order to measure a frequency value, uniform and evenly spaced cues are used in this paper. Henceforth we define a measure for the frequency rate of the temporal cues named ω, (Eq. (9)) where n_{cues} is the number of temporal cues, m is the number of frames in the sequence, and f_s is the original sampling frequency of the signal during recording of the signal. In this research a constant sampling rate of f_s = 60 fps is used.

$$\omega = \frac{n_{cues} f_s}{m} \tag{9}$$

In order to approximate the neutral component of a stylistic input sequence, **Y**, we calculate ω with the aim of maximizing similarity of the approximated sequence, $\mathbf{Y}_{approx}$, with respect to the warped version of the corresponding neutral sequence $\widetilde{\mathbf{N}}$. Hence, the objective function $J = \rho(\mathbf{Y}_{approx}, \widetilde{\mathbf{N}})$. Accordingly ω is computed using $\arg\max_\omega J$. The operator ρ calculates the correlation coefficients of corresponding DOFs of the sequence and the frequency value is calculated with the aim of maximizing similarity between corresponding joint angle curves. This approach calculates a separate frequency value for each DOF of the sequence. Another approach is to calculate a single frequency for the sequence (as a whole) through $J_{seq} = \sum_{i=1}^{n} \boldsymbol{u}_i \rho(\mathbf{Y}_{approx,i}, \widetilde{\mathbf{N}}_i)$ where $\boldsymbol{u}$ is the weight parameter and i represents the DOF of the sequence. The same weight vector used for time warping or correspondence measurement can be used. As a result, arg-max$_\omega$ J_{seq} calculates a frequency value that maximizes a weighted sum of correlations of all DOFs of the two sequences.

Fig. 2 illustrates the correlation values for several joint angle curves of a stylistic sequence. It is shown that in all cases, for maximizing correlation, ω < 10 Hz is required. The correlation values converge towards an asymptote which implies increasing the number of cues beyond a certain point does not result in significant changes in the model. This is because as the number of cues increases, the modeled approximation tends towards the actual stylistic signal rather than the neutral one. In other words, the goal is to stop increasing ω as soon as the modeled signal resembles the neutral signal.

Successive to calculating the optimal frequency for the temporal cues and approximating the neutral component ($\mathbf{Y}_{approx}$) of the sequence with splines, a feature set is extracted using:

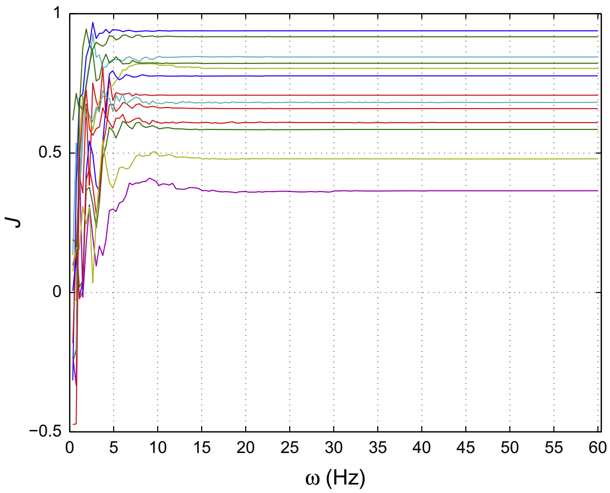


Fig. 2 J vs. ω for several joint angle curves of a stylistic walk. The maximum of each curve (usually occurring between 3 and 8 Hz) is employed as the optimum frequency for cues.

$$\mathbf{\Phi}_{movement} = \mathbf{Y} - \mathbf{Y}_{approx} \tag{10}$$

where $\mathbf{\Phi}_{movement}$ is the movement SF. The reason that this feature set is associated with *movement* is its varying nature throughout the sequence. This reasoning becomes more clear as we extract the posture feature in the following section. Fig. 3 illustrates a signal being modeled with the optimal ω value (~3 Hz) and the movement SF set being extracted. It is observed that the approximated neutral signal using splines is very similar to the corresponding neutral signal. However, posture features, in the form of spatial offsets, remain to be extracted.

Posture features

It was illustrated in Fig. 3 that the temporal cues used to approximate the neutral component are spatiotemporally located on the stylistic input signal. Therefore, the extracted movement SF set is always positioned on the input

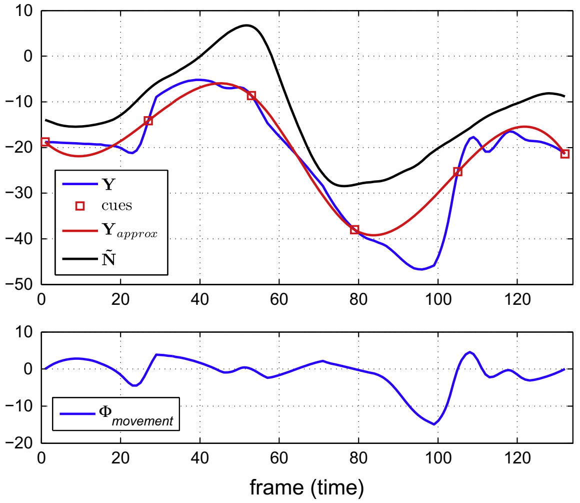


Fig. 3 Extraction of movement SF from a motion signal with ω = ~3 Hz.

and excludes any spatial offsets with respect to the corresponding neutral sequence. However, as illustrated in Fig. 3, spatial offsets do exist as segments of the approximated signal need to be vertically shifted in order to fall on top of the neutral signal. Due to its unvarying nature, this spatial difference between the modeled sequence and the neutral sequence is highly resembling of the posture features described earlier. In other words, when put in the context, this feature set describes how the body structure of the neutral sequence should be spatially modified to become similar to the stylistic body structure. Subsequently, this feature set can be calculated by:

$$\mathbf{\Phi}_{posture,avg} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{Y}^{i}_{approx} - \widetilde{\mathbf{N}}^{i}) \tag{11}$$

which results in a constant average posture feature. Here, m represents the length of the sequence. Fig. 4 illustrates the average posture feature extracted for the joint angle curve used in Fig. 3.

This feature can also be modeled and extracted using the same calculated set of temporal cues, increasing the resolution of this component of the SF. While this modification may seem to contradict the very definition of posture features, it makes computational sense and improves the results. Accordingly, for each segment p (the segment of the signal between two consecutive cues), the mean shift is calculated using:

$$\mathbf{\Phi}_{posture,segment} = \frac{1}{m_p}\sum_{i=1}^{m_p}(\mathbf{Y}^{i}_{approx} - \widetilde{\mathbf{N}}^{i}|\mathrm{p}) \qquad for\ \mathrm{p} = 1 : \mathrm{n}_{\mathrm{cues}} - 1 \tag{12}$$

where m_p represents the number of frames in that segment. A sample extracted posture set is illustrated in Fig. 4. This quantized version of the posture feature will cause discontinuities in the SF and any signal to use the feature. To remedy this, we approximate the piece-wise feature using cubic smoothing splines. The smoothed $\mathbf{\Phi}_{posture,approx}$ for the quantized signal $\mathbf{\Phi}_{posture,segment}$, is calculated through Eq. (13) based on (De Boor, 1978):

$$\arg\min_{\mathbf{\Phi}_{posture,approx}} \left\{ r\sum_{i}\beta_i\|\mathbf{\Phi}^{i}_{posture,approx} - \mathbf{\Phi}^{i}_{posture,segment}\|^2 + (1-r)\sum_{i}\gamma_i\left\|\frac{\Delta^2\mathbf{\Phi}^{i}_{posture,approx}}{\Delta i^2}\right\|^2 \right\} \tag{13}$$

Here, γ is the roughness measure, for our application set to 1, $r \in [0, 1]$ is the smoothing parameter, for our application set to 0.01, and β is the weight equal to 1. Fig. 4 illustrates the extraction and approximation process for the posture SF set. It is important to note that the posture features should be extracted only successive to extraction of the movement features.

Uniform temporal feature

When different segments of a motion sequence are performed with different speeds, the temporal changes are manifested as spatiotemporal curves. However, there are instances where the timing and speed of the entire sequence is one of the influential components of the SF set (Amaya et al., 1996). Consequently, in addition to movement and posture SFs, we introduce a separate SF component, $\Phi_{temporal}$, which describes the relationship between the temporal lengths of stylistic sequences with respect to neutral ones. We calculate this feature using:

$$\Phi_{temporal} = \frac{m_{\mathrm{Y}}}{m_{\mathrm{N}}} \tag{14}$$

where m_{Y} is the temporal length of the input and m_{N} is the length of the corresponding neutral sequence. An important

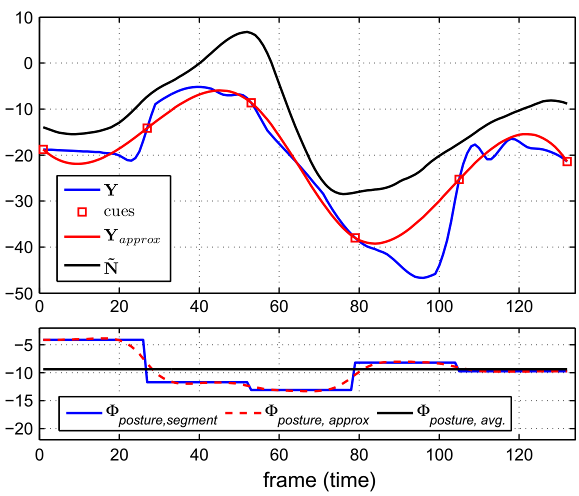


Fig. 4 Extraction of posture SF from the input signal. Smoothing splines are used to approximate the feature.

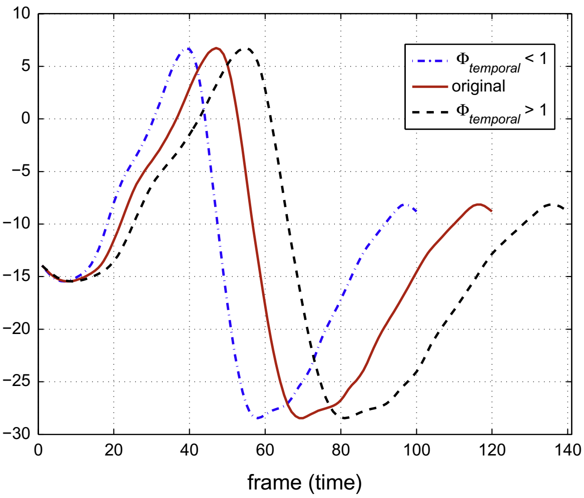


Fig. 5 $\Phi_{temporal}$ for versions of the signal with different speeds.

Table 1 The actions and datasets used in this study.

	CMU dataset					Carleton dataset								
Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Action	Walk	Walk	Walk	Walk	Walk	Walk	Walk	Walk	Jump	Jump	Jump	Run	Run	Run
SF type	Neutral	Sad	Macho	Drunk	Lavish	Neutral	Tired	Energetic	Neutral	Tired	Energetic	Neutral	Tired	Energetic

point to consider is that for this SF, m_Y and m_N need to correspond to sequences that contain similar actions (primary themes). For example, if **Y** is a walking sequence with two strides, **N** must similarly contain walking with only two strides. This feature can also be characterized or described by the speed with which a particular stylistic action in performed with respect to the neutral version of that action.

Successive to calculating $\Phi_{temporal}$, UTW can be used to speed-match a neutral sequence with the stylistic input, hence transferring this feature. UTW is basically a linear stretching and compressing technique that uses interpolations for uniformly changing the temporal length of a signal or sequence. If $\Phi_{temporal} = 1$, the neutral sequence requires no temporal modification as it is equal in length to the stylistic sequence. For $\Phi_{temporal} < 1$, the neutral sequence needs to be compressed, and when $\Phi_{temporal} > 1$, it needs to be stretched. Examples of $\Phi_{temporal}$ are illustrated in Fig. 5.

Generally, the temporal features and properties of the input and corresponding neutral sequence might not be uniformly distributed. In other words, it is not simple to distinguish whether $\Phi_{temporal}$ has resulted from a uniformly faster/slower sequence in all segments or whether the sequence is only performed faster in some segments and slower in others. As an example, it is possible that $\Phi_{temporal} = 1$, with the first half of the input being faster than neutral, followed by a slower second half, resulting in equal overall lengths, and thus $\Phi_{temporal} = 1$. In such cases, the non-uniformities in the sequence is manifested as spatiotemporal features and are most likely extracted as movement or posture features, even though the source is temporal variations. It is therefore, not necessary to extract non-uniform temporal features. Moreover, it can be argued that one of the significant roles of CoTW is to warp the sequences non-uniformly, thus resolving the issue of non-uniform temporal features and allowing them to be extracted as movement and posture SFs.

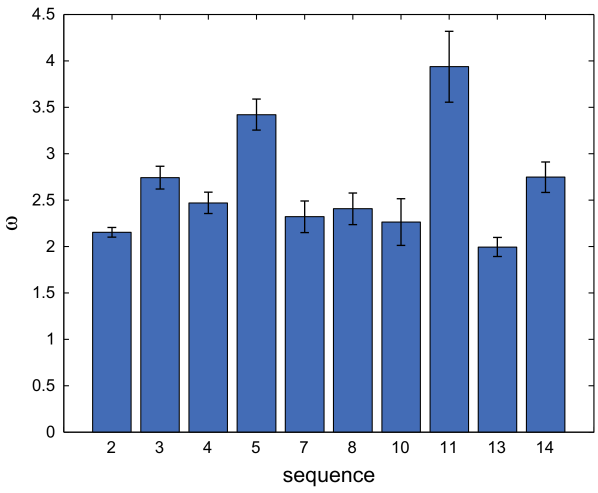


Fig. 6 Average ω measured for different input sequences. Error bars represent standard errors over DOFs.

Results and discussion

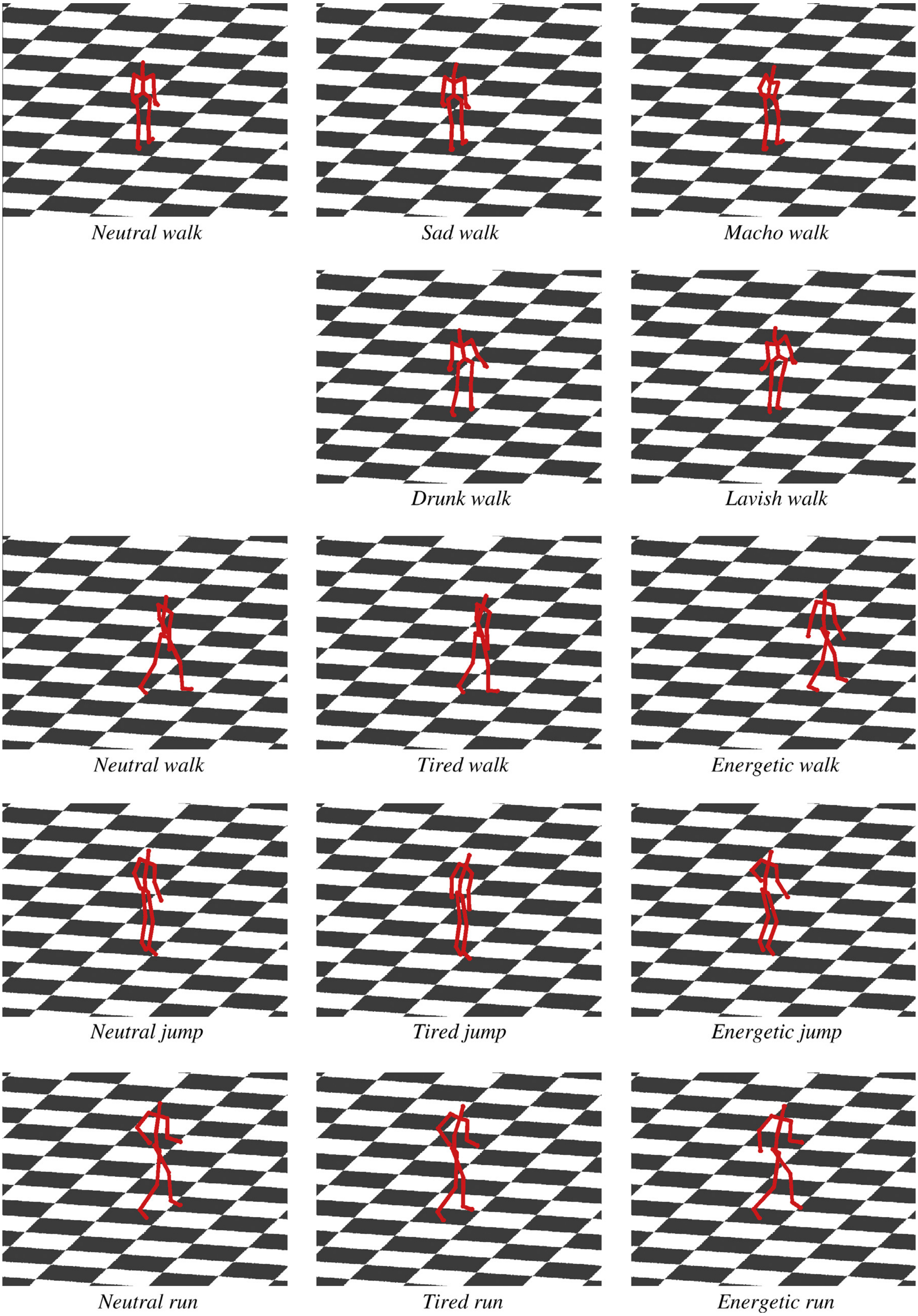
As described earlier, we utilize the CMU dataset as well as our own recorded data. From the CMU dataset, sad, macho, drunk, and lavish walks are employed. Also, 5 neutral walks are used from which the best corresponding sequence is selected for each stylistic input walk. From our own dataset, tired and energetic walks, jumps, and runs are used as inputs, along with 5 neutral walks, 5 neutral jumps, and 5 neutral runs. An actor from the School of Information Technology, Carleton University, with prior motion capture experience was asked to perform the actions. Initially, at a time when the actor indicated an energetic state, energetic sequences were recorded. After few minutes of walking around the capture room and indication of neutral energy by the actor, the neutral actions were recorded. Finally, after 10 more minutes of exercise in the capture room and confirmation of fatigue by the actor, tired actions were recorded. They were also asked to exaggerate in displaying the energy levels. Table 1 presents the types of primary actions and SFs used in this study. The 5 neutral actions are represented only once in the table. Following, the movement, posture, and uniform temporal features are extracted in the presented order. Furthermore, to illustrate the performance of the proposed method, style translation is carried out.

Extraction of features

Fig. 6 presents the average and standard errors of ω over DOFs of the sequences presented in Table 1. One-way analysis of variances, ANOVA, for the effect of different DOFs on

Table 2 Uniform temporal features for the test data.

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\Phi_{temporal}$	1.00	1.23	1.19	1.65	0.98	1.00	1.09	0.75	1.00	1.02	0.95	1.00	1.05	0.84



Neutral walk

Sad walk

Macho walk

Drunk walk

Lavish walk

Neutral walk

Tired walk

Energetic walk

Neutral jump

Tired jump

Energetic jump

Neutral run

Tired run

Energetic run

Fig. 7 Snapshots from style translation outputs. The complete sequences are available in the accompanying video submitted with this manuscript.

the calculated ω, shows no significant effect at the $p < 0.05$ level despite the existing variations. Specifically, for the CMU samples, $F(75, 228) = 1.2$, $p = 0.158$ and for Carleton samples, $F(51, 260) = 1.31$, $p = 0.091$. One-way ANOVA for the effect of actions on calculated ω shows significance with $F(3, 300) = 19.6$, $p < 0.0001$ for the CMU samples and $F(5, 306) = 9.49$, $p < 0.0001$ for Carleton samples. This can be clearly observed in Fig. 6 where, for example, energetic jump requires a significantly higher average ω compared to tired jump. This analysis indicates that while within a particular action, the DOF does not significantly impact ω, the action from which the features are being extracted, and the feature types themselves, do have a significant influence on the optimum frequency values.

Table 2 presents the results for the calculated $\Phi_{temporal}$ values. Energetic sequences along with lavish walk, which are shorter (faster) than neutral, have resulted in $\Phi_{temporal} < 1$. For the rest of the samples, $\Phi_{temporal} > 1$, indicating longer lengths (slower) with respect to neutral. Generally, $\Phi_{temporal}$ is quite intuitive and simple to speculate without computation. Sequences that have SFs associated with low energy are often longer than neutral, thus $\Phi_{temporal} > 1$, while those attributed to higher energy are faster, and so $\Phi_{temporal} < 1$. However, for instances where exact ratios are required, Eq. (14) is utilized.

Style translation

Once the three SF components have been extracted from a stylistic input sequence, the features can be applied to a neutral sequence to produce a stylistic theme and change the action from neutral to stylistic. This process is called style translation (Hsu et al., 2005). Here, we utilize style translation to visualize the features extracted using our approach.

When manipulating motion data, synchronization is often lost (Hsu et al., 2005). As a result, out of tune motion of the feet causes an artifact in which the character seems to be skating rather than taking firm and solid steps. This is referred to as footskating (Lyard & Magnenat-Thalmann, 2007). By using J_{seq} which results in a constant ω for all DOFs, footskating can be reduced. This approach, however, introduces a trade-off as sub-optimal frequencies are utilized for some DOFs. As a result, some extracted SFs will be inaccurate. In this paper, for style translation, we use the individually calculated ω values for each DOF. Footskating can be remedied through post-processing, for example, by using footplant-based techniques (Kovar et al., 2002). In this paper, we first map the joint angle curves of the outputs onto the Cartesian space. Subsequently, the skate trajectory of the stance foot is calculated for the duration of that particular foot-stance, and subtracted from the motion, eliminating the skating artifact.

The accompanying video submitted with this manuscript shows the style translation results, where successful conversion of neutral to designated SFs point to the accuracy of our method for extraction of style features. Snapshots of the results are shown in Fig. 7, where the neutral actions are the inputs and the stylistic ones are the translation outputs.

To further evaluate the outputs, 10 participants, were asked to watch and provide feedback on the SFs that are perceived from the outputs. The average age of participants was 31.3, the standard deviation was 11.5, 6 were males, and 4 were females. A force choice questionnaire was used for this purpose. Sequences from the two datasets were presented separately and choices were divided based on the dataset. This was done for two reasons: (a) the skeleton structure for the two datasets was different which could influence the results had they been investigated together, (b) sad (from the CMU dataset) and tired (from the Carleton dataset) actions are very prone to misperception by participants. Such sequences are difficult to distinguish in real life as both contain features such as slower motion, decreased sway, and downward tilt in shoulders and head. Fig. 8 presents the confusion matrix where most sequences have been correctly perceived for the translated SFs.

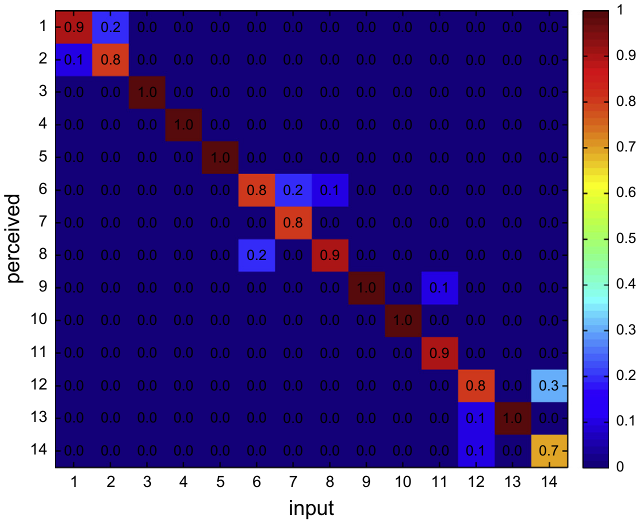


Fig. 8 Confusion matrix for style translation outputs.

In general, our style translation outputs are in par with methods such as (Hsu et al., 2005) or (Rose et al., 1998) where high quality stylistic sequences were generated. Our system performs well for different actions such as walking, jumping and running and a variety of styles such as sad, macho, drunk, lavish, tired, and energetic. However, compared to linear interpolation/extrapolation approaches such as (Rose et al., 1998), our method illustrates higher generalization capabilities where using a frequency range of ω = 2—8 Hz for temporal cues, we can extract the movement SFs without a neutral reference. To extract the posture features, however, our system also needs a neutral reference. In addition, the uniform temporal feature can be speculated.

Unlike most existing techniques for stylistic motion control which are purely computational, our proposed method draws inspiration from the way biological motion is executed and perceived. In reality, variations in biological motion originate from the same three sources which our system successfully differentiates and separately extracts. As a result, we believe our system can be used to study biological style features and contribute to psychophysics. The system can also be used for multimedia and HCI applications such as animation and gesture-based interaction systems.

Conclusion

In this paper, the problem of extracting stylistic features from motion sequences was investigated. Inspired by the way in which different actions are physically performed, three types of spatiotemporal features were defined: movement, posture, and temporal. For a stylistic input sequence, a correlation-based similarity index was calculated with respect to a dataset of neutral sequences. Correlation optimized time warping (CoTW) was carried out to align the training and input motion sequences. Temporal cues were optimally distributed throughout the stylistic input sequence and cubic splines were used to model the neutral component of the sequence. Movement features were then extracted followed by posture features, which were smoothed to prevent discontinuities when employed. Finally, uniform temporal features were calculated. Style translation was carried out on samples from two different datasets to show the performance of the method. The samples consisted of different actions such as walking, running, and jumping as well as different styles such as sad, macho, drunk, lavish, tired, and energetic. The style translation videos and subjective evaluations illustrated the accuracy and significance of the proposed technique.

Acknowledgments

This work was supported in part by The Natural Sciences and Engineering Research Council of Canada (NSERC) and Ontario Centres of Excellence (OCE).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bica.2013.10.001.

References

Amaya, K., Bruderlin, A., & Calvert, T. (1996). Emotion from motion. *Graphics Interface, 96*, 222–229.

Arikan, O., & Forsyth, D. A. (2002). Synthesizing constrained motions from examples. *ACM Transactions on Graphics, 21*(3), 483–490.

Arikan, O., Forsyth, D. A., & O'brien, J. F. (2003). Motion synthesis from annotations. *ACM Transactions on Graphics, 22*(3), 402–408.

Bartels, R. H., Beatty, J. C., & Barsky, B. A. (1998). *An introduction to splines for use in computer graphics and geometric modeling*. San Francisco, CA: Morgan Kaufmann.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology, 58*, 47–73.

Brand, M., & Hertzmann, A. (2000). Style machines. In *Proceedings of the 27th annual conference on computer graphics and interactive techniques* (pp. 183–192). ACM Press/Addison-Wesley Publishing Co..

Bruderlin, A., & Williams, L. (1995). Motion signal processing. In *Proceedings of the 22nd annual conference on computer graphics and interactive techniques* (pp. 97–104). ACM.

Coros, S., Beaudoin, P., & van de Panne, M. (2010). Generalized biped walking control. *ACM Transactions on Graphics, 29*(4), 130.

De Boor, C. (1978). *A practical guide to splines* (1st ed.). Springer.

De La Torre, F., & Black, M. (2001). Dynamic coupled component analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 643–650).

Etemad, S. A., & Arya, A. (2009). 3d human action recognition and style transformation using resilient backpropagation neural networks. In *Proceedings of the IEEE international conference on intelligent computing and intelligent systems* (Vol. 4, pp. 296–301).

Etemad, S. A., & Arya, A. (2010). Modeling and transformation of 3D human motion. In *Proceedings of the 5th international conference on computer graphics theory and applications* (pp. 307–315).

Etemad, S. A., & Arya, A. (2011). Separation and extraction of energy variants from human motion using temporal minimization. In *Proceedings of the IEEE international conference on virtual environments human–computer interfaces and measurement systems* (pp. 1–5).

Etemad, S. A., & Arya, A. (2013). A customizable time warping method for motion alignment. In *Proceedings of the 7th IEEE international conference on semantic computing* (pp. 387–388).

Etemad, S. A., Arya, A., & Parush, A. (2011). Spatial perceptual weights of energy-related features in animation of human motion. In *Proceedings of computer graphics, international* (p. S15).

Fu, A. W. C., Keogh, E., Lau, L. Y., Ratanamahatana, C. A., & Wong, R. C. W. (2008). Scaling and time warping in time series querying. *The VLDB Journal—The International Journal on Very Large Data Bases, 17*(4), 899–921.

Grochow, K., Martin, S. L., Hertzmann, A., & Popovic, Z. (2004). Style-based inverse kinematics. *ACM Transactions on Graphics, 23*(3), 522–531.

Guest, A. H. (2005). *Labanotation: The system of analyzing and recording movement*. Psychology Press.

Heloir, A., Courty, N., Gibet, S., & Multon, F. (2006). Temporal alignment of communicative gesture sequences. *Computer Animation and Virtual Worlds, 17*(3–4), 347–357.

Hsu, E., da Silva, M., & Popović, J. (2007). Guided time warping for motion editing. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on computer animation* (pp. 45–52). Eurographics Association.

Hsu, E., Pulli, K., & Popovic, J. (2005). Style translation for human motion. *ACM Transactions on Graphics, 24*(3), 1082–1089.

Kleinsmith, A., De Silva, P. R., & Bianchi-Berthouze, N. (2006). Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers, 18*(6), 1371–1389.

Kovar, L., Schreiner, J., & Gleicher, M. (2002). Footskate cleanup for motion capture editing. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on computer animation* (pp. 97–104). ACM.

Liu, L., Yin, K., van de Panne, M., & Guo, B. (2012). Terrain runner: Control, parameterization, composition, and planning for highly dynamic motions. *ACM Transactions on Graphics, 31*(6), 154.

Lyard, E., & Magnenat-Thalmann, N. (2007). A simple footskate removal method for virtual reality applications. *The Visual Computer, 23*(9–11), 689–695.

Ma, W., Xia, S., Hodgins, J. K, Yang, X., Li, C., & Wang, Z. (2010). Modeling style and variation in human motion. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics symposium on computer animation* (pp. 21–30). Eurographics Association.

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 258*(1353), 273–279.

Nielsen, N. P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A, 805*(1), 17—35.

Normoyle, A., Liu, F., Kapadia, M., Badler, N. I, & Jörg, S. (2013). The effect of posture and dynamics on the perception of emotion. In *Proceedings of the ACM symposium on applied perception* (pp. 91—98). ACM.

Picard, R. W. (2009). Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), 3575—3584.

Pullen, K., & Bregler, C. (2002). Motion capture assisted animation: Texturing and synthesis. *ACM Transactions on Graphics, 21*(3), 501—508.

Rose, C., Cohen, M. F., & Bodenheimer, B. (1998). Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications, 18*(5), 32—40.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing, 26*(1), 43—49.

Shapiro, A., Cao, Y., & Faloutsos, P. (2006). Style components. In *Proceedings of graphics interface* (pp. 33—39). Canadian Information Processing Society.

Thrasher, M., Van der Zwaag, M. D., Bianchi-Berthouze, N., & Westerink, J. H. (2011). Mood recognition based on upper body posture and movement features. In *Affective computing and intelligent interaction* (pp. 377—386). Berlin Heidelberg: Springer.

Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision, 2*(5), 371—387.

Troje, N. F., Cord, W., & Lavrov, M. (2005). Person identification from biological motion: Effects of structural and kinematic cues. *Perception and Psychophysics, 67*(4), 667—675.

Tsai, Y.-Y., Lin, W.-C., Cheng, K. B., Lee, J., & Lee, T.-Y. (2010). Real-time physics-based 3D biped character animation using an inverted pendulum model. *IEEE Transactions on Visualization and Computer Graphics, 16*(2), 325—337.

Unuma, M., Anjyo, K., & Takeuchi, R. (1995). Fourier principles for emotion-based human figure animation. In *Proceedings of the 22nd annual conference on computer graphics and interactive techniques* (pp. 91—96). ACM.

Witkin, A., & Popovic, Z. (1995). Motion warping. In *Proceedings of the 22nd annual conference on computer graphics and interactive techniques* (pp. 105—108). ACM.

Wu, X., Tournier, M., & Reveret, L. (2011). Natural character posing from a large motion database. *IEEE Computer Graphics and Applications, 31*(3), 69—77.

Zhou, F., & Torre, F. (2009). Canonical time warping for alignment of human behavior. In *Advances in neural information processing systems* (pp. 2286—2294).