Spectral Style Transfer for Human Motion between Independent Actions

M. Ersin Yumer
Adobe Research

Niloy J. Mitra
University College London

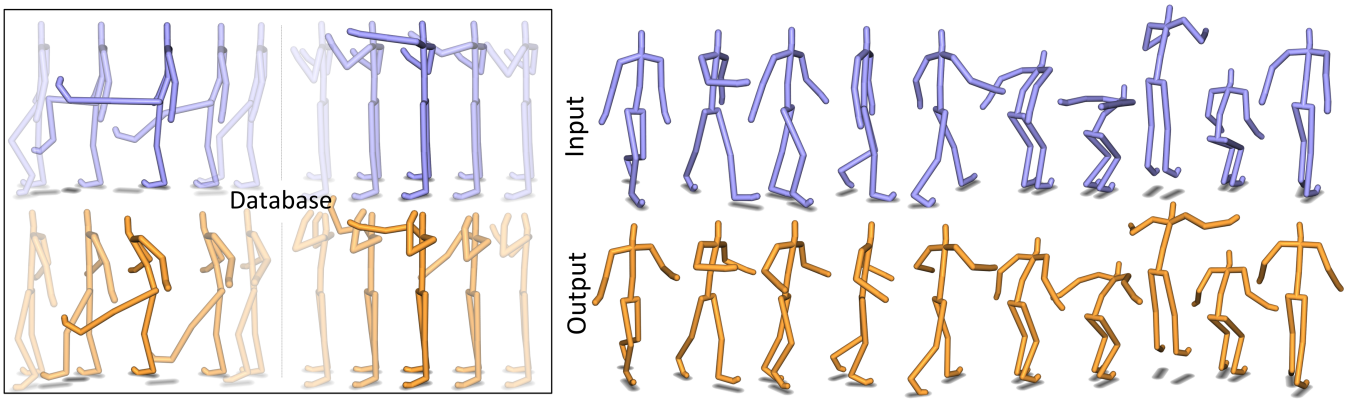

Figure 1: *Spectral style transfer between independent actions. A **neutral** style heterogeneous **walk** ⇒ **jump** motion (top-right) stylized as **childlike** with our method (bottom-right) using a database where only **kick** and **punch** actions are available (left). Note the high energy upper body and low energy lower body distinguishing the **childlike** style from the **neutral** style both in the database and the stylized motion.*

Abstract

Human motion is complex and difficult to synthesize realistically. Automatic style transfer to transform the mood or identity of a character's motion is a key technology for increasing the value of already synthesized or captured motion data. Typically, state-of-the-art methods require all independent actions observed in the input to be present in a given style database to perform realistic style transfer. We introduce a spectral style transfer method for human motion between independent actions, thereby greatly reducing the required effort and cost of creating such databases. We leverage a spectral domain representation of the human motion to formulate a spatial correspondence free approach. We extract spectral intensity representations of reference and source styles for an arbitrary action, and transfer their difference to a novel motion which may contain previously unseen actions. Building on this core method, we introduce a temporally sliding window filter to perform the same analysis locally in time for heterogeneous motion processing. This immediately allows our approach to serve as a style database enhancement technique to fill-in non-existent actions in order to increase previous style transfer method's performance. We evaluate our method both via quantitative experiments, and through administering controlled user studies with respect to previous work, where significant improvement is observed with our approach.

Keywords: motion style transfer, spectral motion synthesis

Concepts: •**Computing methodologies** → **Motion processing;**

SIGGRAPH '16 Technical Paper, July 24-28, 2016, Anaheim, CA
ISBN: 978-1-4503-4279-7/16/07
DOI: http://dx.doi.org/10.1145/2897824.2925955

1 Introduction

While human motion primarily represents the action it embodies, the style of the action is also a significant component of a natural human act. Accurate stylization of human motion to portray mood, character, or identity is critical for realistic humanoid animation. Style transfer is an important tool to achieve this required realism for virtual characters in games and movies, without exhaustively capturing the combination of all possible actions and styles.

Realistic motion synthesis is notoriously difficult [Min and Chai 2012]. Hence, high quality animations are still created by retargeting captured real motion data. This results in an expensive, time consuming, and tedious pipeline for the animator because all steps (motion capture, retargeting, and cleanup) need to be repeated for each different action and style of motion required for the character. Therefore, multiplying the value of captured data by stylizing it to resemble various moods and identities have been extensively studied.

Previous approaches were introduced both for homogeneous human motion stylization, typically using linear methods [Hsu et al. 2005], and recently, for heterogeneous motion (*e.g.,* walk ⇒ jump ⇒ run, and transitions in between) through temporally local nearest neighbor blending [Xia et al. 2015] in spatial-temporal space. However, these approaches typically require the training database to include the type of actions that the target motion sequence consists of. This consequently results in the requirement of significantly large style database curation for transfer.

Our goal is to circumvent the costly and time consuming requirement of all inclusive style databases, and be able to perform style transfer from limited databases where available actions may be independent from those in the input motion sequence. To this extent, we present a novel spectral human motion style transfer method, which not only handles heterogeneous motion sequences but also transfers style between independent actions (Figure 1). The key idea of our approach is performing the style transfer in spectral space. This allows us to transfer style *without* establishing a frame-by-frame spatial correspondence between the target motion

sequence and the source motion database. Our approach immediately allows database enhancement, which can increase previous methods performance significantly.

We demonstrate the effectiveness of our method on a variety of actions and styles. Through user studies, we show that our method performs significantly better than the previous state-of-the-art including the online learning method [Xia et al. 2015], especially when style database lacks certain action types contained in the target motion sequence. Our qualitative and quantitative evaluations show that our method gracefully transfers style even when there is a significant discrepancy between target and source action types (*e.g.,* transfer of *childlike* style from kick action to punch action), where alternative methods may fail.

Our main contributions are

- Formulation of a spectral style transfer for human motion that enables transfer of style independent of frame-by-frame spatial similarity between source and target motions, thereby facilitating style transfer between independent actions.
- An extension to the spectral style transfer, which exploits sliding window filters to temporally localize the style transfer, thereby facilitating style transfer between independent actions even in the presence of heterogeneous motion sequences.
- Demonstration of our approach as a style database enhancement method to increase other style transfer approaches' performance.

2 Background

Data-Driven Motion Synthesis. Arikan *et al.* [2003] used an optimization method to generate a continuous concatenation of training sequences under path and action constraints, whereas Kovar *et al.* [2004] introduced a directed graph approach to build combinations of motion sequences from a given motion database and follow user imposed constraints (*e.g.,* follow a specific path). There is a significant body of work in controlled statistical human motion synthesis [Chai and Hodgins 2007; Min and Chai 2012; Lau et al. 2009]. Once the parameters are learned, statistical models can be used to generate novel motion that statistically satisfies the properties of the data it is trained on. Although the motion can be guided through constraints, the variation in the generated motion is limited with the variation in the training dataset. Hence, an action type that was missing in the training dataset cannot be gracefully synthesized.

Motion Stylization. In an early work Amaya *et al.* [1996] introduced a method where they extracted full periodic cycles, termed 'basic periods', and they considered style transfer as changing the speed and the peak magnitude of these cycles. This operation in the spatial-temporal space, is similar to a limited version of our core method; instead of considering the entire frequency spectrum, Amaya *et al.* [1996] is equivalent to considering only the dominant frequency component. The algorithm is limited to clearly labeled full cycles and because of operating with time signals it requires manual correspondence between the cycles of target and source styles. In a similar direction, analyzing two motions with the goal of transferring properties from one to the other, Shapiro *et al.* [2006] proposed using independent component analysis [Hyvärinen et al. 2004] to separate motion into different components and transfer.

Hsu *et al.* [2005] introduced a style transfer method where they used a linear time-invariant system to model the difference between aligned homogeneous motion sequences of the same action stylized differently. They then used this system to transfer the identified style to new instances of similar homogeneous motions. Analysis and synthesis of motion patterns in images were also studied with linear models [Giese and Poggio 2000]. Urtasun *et al.* [2004] proposed decomposing motion data into PCA components to encode style or identity differences. A new motion sequence's component weights are then adjusted to the ones learned for synthesis of stylized data. Ikemoto *et al.* [2009] introduced a model based on Gaussian processes of kinematics and dynamics to edit long sequences of motion based on a small sample provided by the user. [Urtasun et al. 2004; Hsu et al. 2005; Ikemoto et al. 2009] focus on learning a parametric or non-parametric model based on a specific example and applying the model to a novel motion. Although they are very fast, they suffer from the generalization imposed by the specific model they fit. Hence, their performance significantly diminish when applied to motion sequences that incorporate heterogeneous action.

Generative statistical models have also been used to synthesize stylistic motion. Brand and Hertzmann [2000] introduced a statistical motion model where a set of Hidden Markov Models were trained to interpret motion variations caused by style. Similarly, Wang *et al.* [2007] utilized Gaussian Process Latent models to capture stylistic motion variation caused by gait and identity. Min *et al.* [2010] extended the method to capturing style and identity in the same action.

Data-Driven Style Transfer. Most relevant to our work, Xia *et al.* [2015] recently introduced a data-driven method for style transfer using a database of captured motion sequences labeled in various styles and actions. Their method builds on a local mixtures of autoregressive models that builds temporally local nearest neighbor mixtures from the source style database to transform each frame successively. This enables handling of heterogeneous motion more gracefully compared to global methods (*e.g.,* [2005]). However, because of the frame by frame spatial similarity matching, the method works well when the style database includes the type of actions in the target motion sequence.

In contrast, we introduce a method where similarity and style transfer is performed through operations in the frequency domain independent of direct spatial correspondences. This allows us to transfer style between significantly different actions (*e.g.,* from kick and punch database to walk and jump (Figure 1)).

3 Method

Human motion, when threated as time domain signals (Figure 2), is complex and it is difficult to establish correspondences to perform meaningful style transfer, especially between two independent actions. However, we observe that when separated from the time domain synchronization dependency in the spectral domain, the frequency component magnitudes yield useful information. Figure 3 shows an example of this idea: the difference between the stylized motions and neutral states are highly correlated even though the actions are different. We build on this idea, and exploit the compact spectral domain representation.

3.1 Core Method

Discrete Fourier Transform (DFT). For the sake of completeness, we first re-cap the well known DFT formulation here. Let $f[t]$ be a discrete time domain signal of one of the degrees of freedom (DOF) of a human motion data. Consequently, the Discrete Fourier Transform $F[k]$ of $f[t]$ is given by

$$F[k] = \sum_{t=0}^{N-1} f[t] e^{-i \frac{2*\pi}{N} kt} \quad k = 0, \ldots, N-1 \qquad (1)$$

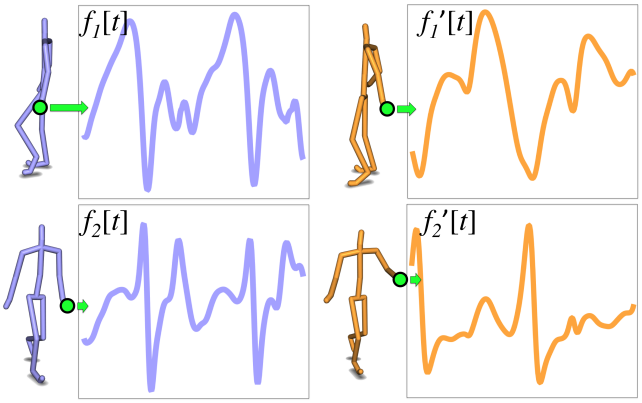

Figure 2: *Difference between neutral (left) and stylized (right) motions is not trivial to correlate when the actions are different (top: kick, bottom: walk).*

where N is the length of the signal and $i^2 = -1$. The single-sided spectrum $F[\omega]$ is given by

$$F[\omega] = \frac{2}{N} F[k] \quad k = 0, \ldots, N/2 \tag{2}$$

where $\omega = (f_s/N)k$ is the frequency transform from samples k in the spectral space. Here, f_s is the sampling frequency of the original time domain signal $f[t]$. Let $R[\omega]$ and $A[\omega]$ represent the magnitude and phase of the spectra $F[\omega]$ given by:

$$\begin{aligned} R[\omega] &= |F[\omega]| \\ A[\omega] &= \angle F[\omega] \end{aligned} \tag{3}$$

We only use the single-sided spectrum in the positive frequency range ($\omega = \{0 : f_s/2\}$) since we know that our time domain signal is real which always yields a symmetric spectrum.

It is trivial to reconstruct $f[t]$ from $F[k]$ (hence, from $F[\omega]$) using the inverse discrete Fourier transform:

$$f'[t] = \frac{1}{N} \sum_{k=0}^{N-1} F'[k] e^{i \frac{2*\pi}{N} kt} \quad t = 0, \ldots, N-1 \tag{4}$$

The magnitude, $R[\omega]$, defines the existence and intensity of a motion at ω frequency whereas the phase, $A[\omega]$ describes relative timing.

Style Transfer in Spectral Space. Let $f^s[t]$ be a time domain signal of the source style, and a different action than the target motion $f[t]$ (*e.g.,* walk vs. kick). Let $f^r[t]$ be a time domain signal of the reference style, that represents the same action as $f^s[t]$ and the same style as $f[t]$ (Later in this section, we introduce how we find the best candidates for $f^s[t]$ and $f^r[t]$, in the spectral space). Our goal is to extract the difference between $f^s[t]$ and $f^r[t]$, and apply that difference to $f[t]$. However, this is not trivial in time domain, since the length of the three signals, as well as their synchronization, and spatial correspondences can be arbitrarily different. We therefore resort to the spectral domain and formulate the style transfer by computing a new magnitude, $R'[\omega]$ for the entire signal, solving the following optimization problem:

$$\begin{aligned} \underset{R'[\omega]}{\text{minimize}} \quad & \sum_{\omega} R'[\omega] - (R[\omega] + s[\omega]\,(R^s[\omega] - R^r[\omega])) \\ \text{subject to} \quad & A'[\omega] = A[\omega] \\ & \text{Im}\left\{f'[t]\right\} = 0. \end{aligned} \tag{5}$$

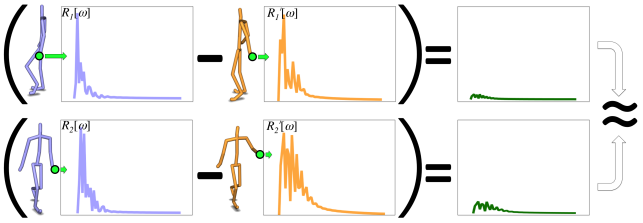

Figure 3: *Difference of style (left vs. right) is highly correlated in the frequency domain magnitude component even when the actions are different (top: kick, bottom: walk). Compare to Figure 2.*

where $R^s[\omega]$, and $R^r[\omega]$ are the source and reference style spectral magnitudes. $s[\omega] = \frac{R[\omega]}{\max_{\omega}(R[\omega])}$ and $f'[t]$ is the inverse discrete Fourier transform given by Equation 4 of the new spectrum whose magnitude is given by $R'[\omega]$. We solve for each DOF in the skeleton independent from the others. Although each DOF is treated independently, keeping the phase of the signals constant ensures the synchronization between the joints are preserved because of the fact that we preserve the timing with the first constraint. Note that the second constraint is also very important because the resulting signal should be a real valued signal representing the stylized motion in spatial-temporal space. We illustrate this process schematically in Figure 4.

Style Database and Spectral Similarity. Above, we assumed that $f^s[t]$ and $f^r[t]$ were given. Here, we explain how we find the best candidates automatically from an available database of available motion data. In order to efficiently measure similarity in the spectral domain, we define a skeletal power representation P, which is a vector of discrete cumulative root mean square (DRMS) powers of all DOF of the animated skeleton, where DRMS power of i^{th} DOF in spectral domain is given by

$$P_i = \sqrt{\frac{1}{N} \sum_{\omega} ||R_i[\omega]||^2} \tag{6}$$

We define the similarity between two motions as the square norm of their skeletal powers. Table 1 shows classification of specific actions (rows) with databases of different availability (columns). We can observe that the skeletal power representation helps differentiate between action similarity whereas it is less sensitive to style. This is expected since style might be more prominent in individual components of the the spectra, but the total energy use in different parts of the body is generally dominated by the type of action performed. This helps us choose the best candidate action to transfer from, in the case of limited action availability in the database as shown in Table 1.

The core approach introduced above performs well if $f[t]$ is a dominantly cyclic, homogeneous action (*e.g.,* walk). However, there are two limitations to it: (1) its performance diminishes for heterogeneous actions, or for abrupt transient-like actions. (2) If the phase differences between different degrees of freedoms of the skeleton play a role in style, they will be missed. We address these two drawbacks in Sections 3.2 and 3.3, respectively.

3.2 Windowing

We generalize the core spectral transfer approach introduced in Section 3.1 to handle arbitrarily heterogeneous actions through a sliding window filter in time domain, and compute the spectral transfer for fixed-length windows around each time step. The intuition behind windowing is to decrease the coupling of transfer computation between parts of the clip distant in the time domain.

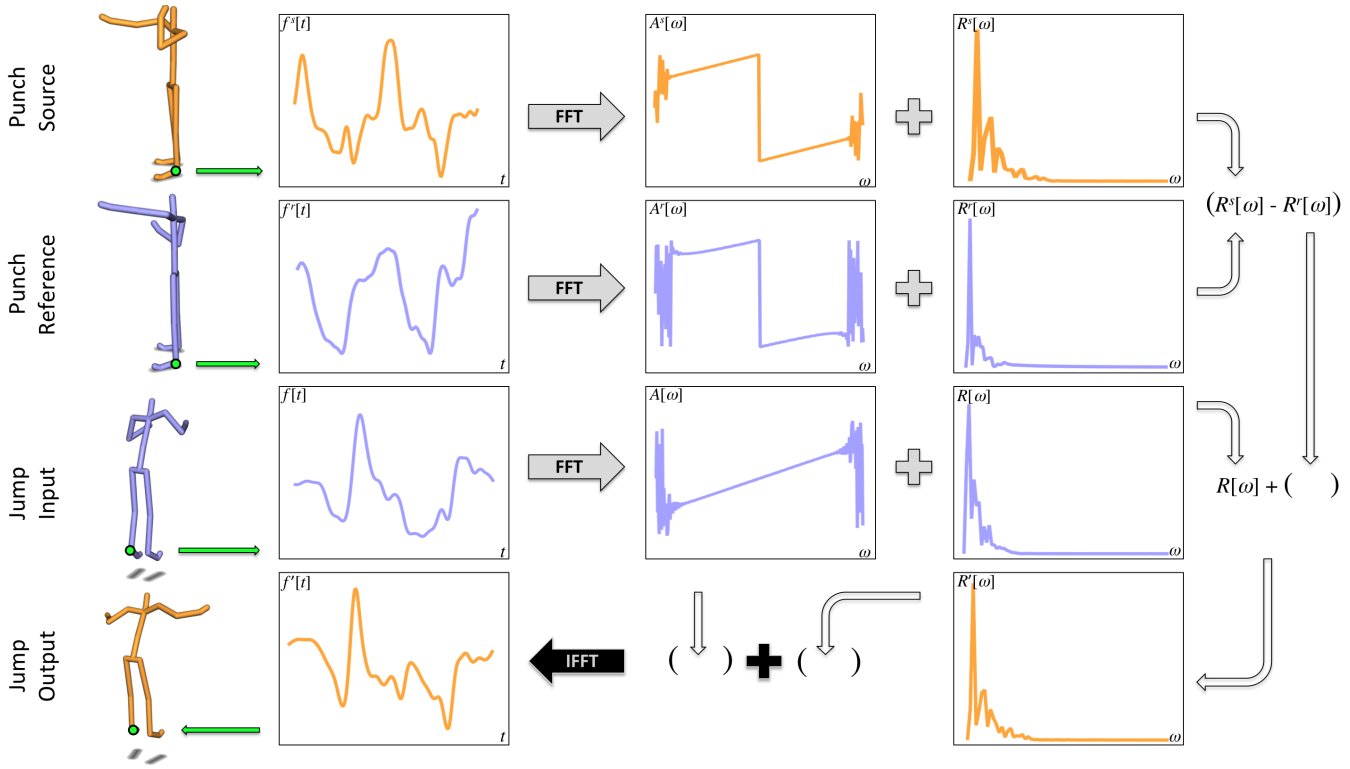

Figure 4: *Time domain signals: target $f[t]$, source $f^s[t]$, and reference $f^r[t]$.Spectral domain processing: we keep $A[\omega]$ constant, and apply the difference of $R^s[\omega]$ and $R^r[\omega]$ to $R[\omega]$ under real-only time-domain signal constraint to compute $R'[\omega]$. Stylized magnitude $R'[\omega]$ and constant $A[\omega]$ result in the stylized time domain data.*

Table 1: *Skeletal power metric action classification performance. Columns show available action in the database. A random selection of 100 actions for each row is tested against all databases. Each cell shows the top classified action from the database and total percentage of the actions classified same as the top classified action. W: walk, R: run, P: punch, K: kick, J: jump.*

	WRPKJ	RPK	WP	PK
W	W 93%	R 85%	W 99%	K 100%
R	R 94%	R 97%	W 93%	K 100%
P	P 93%	P 95%	P 95%	P 100%
K	K 85%	K 89%	W 81%	K 100%
J	J 85%	R 75%	W 74%	K 100%

Let t_f be the current time step we are calculating the window for, and $w[t]$ be a window function. To minimize the spectral leakage, we choose $w[t]$ as a sliding Hanning window [Oppenheim and Schafer 2009] as follows:

$$w[t] = \begin{cases} 0.5\left(1 - \cos\left(\frac{2\pi\left(t - t_f + \frac{N}{2}\right)}{N-1}\right)\right) & \text{if } -\frac{N}{2} \leqslant t \leqslant \frac{N}{2} - 1 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Let $g[t]$ be the discrete time domain motion of one of the DOF of a heterogeneous human motion. For the sake of simplicity, without loss of generality, we overload $f[t]$ (Section 3.1) as follows: $f[t] = g[t] \cdot w[t]$. We can then utilize the approach introduced in Section 3.1 to compute $f'[t]$ of the windowed signal for frame, and consequently $g'[t_f] = f'[t_f] \cdot 1/w[t_f]$ leading to the stylized time step for t_f.

Windowing enables us to solve for each time step by using a constant and temporally local signal. It is well suited for style transfer when $g[t]$ is an arbitrarily long and has heterogeneous action content, because of temporal localization.

3.3 Generalized Spectral Style Transfer

Solving for shorter temporally local windows for each time step of motion data enables us to handle heterogeneous motions gracefully. To ensure continuity between the stylized time steps, we introduce a smoothness term to Equation 5. Moreover, as mentioned earlier in Section 3.1, we need to take into account the relative timing dependency between different DOF of the motion. Hence, we refine our optimization problem as follows:

$$\begin{aligned} \underset{R'[\omega], A'[\omega]}{\text{minimize}} \quad & \sum_{\omega} R'[\omega] - \left(R[\omega] + s[\omega]\left(R^s[\omega] - R^r[\omega]\right)\right) \\ & + \lambda_s \sum_{\omega} \left(R'[\omega] - R'^p[\omega]\right) \\ & + \lambda_a \sum_{\omega} \left(A'[\omega] - A[\omega]\right) \\ & + \lambda_b \sum_{j \in \mathbb{N}} \sum_{\omega} \left(\left(A'[\omega] - A'_j[\omega]\right) - \left(A^s[\omega] - A^s_j[\omega]\right)\right) \\ \text{subject to} \quad & \operatorname{Im}\left\{f'[t]\right\} = 0. \end{aligned} \tag{8}$$

where $R'^p[\omega]$ in the second unary term is the spectral magnitude computed for the current time step in the previous time step's window. This serves as the smoothness term between the time steps. The third unary term is the relaxed phase constraint from Equation 5, this constraint is relaxed to balance the binary term. In the binary term, A^s is the phase of the source style data, and $\mathbb{N}$ is the set of all DOF in the current DOF's joint, the immediate parent and

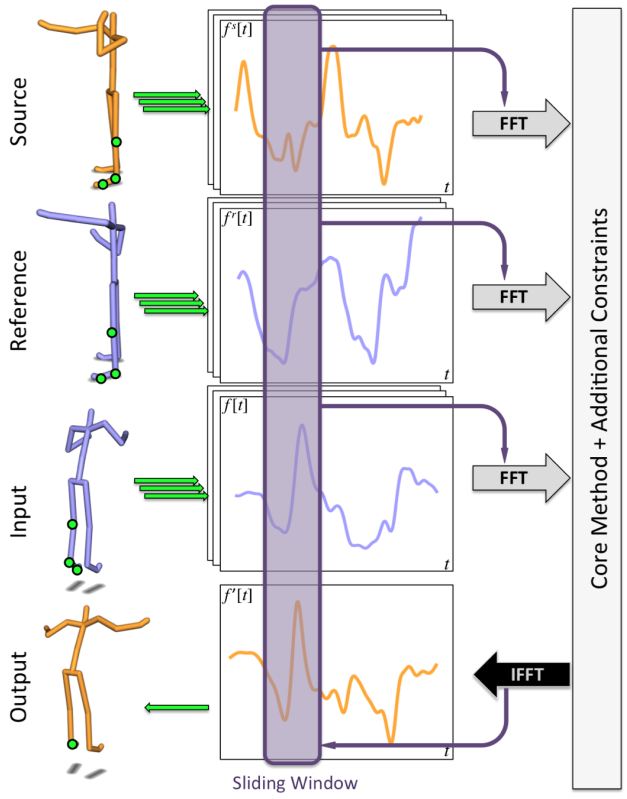

Figure 5: *We generalize our core method both to handle arbitrarily long heterogeneous motion sequences by processing through a sliding window filter and to take into account the phase relationship between joints by introducing additional constraints based on parent and children in the skeleton hierarchy (See text for details).*

immediate children joints. This term accounts for relative timing dependency between different DOF of the motion. λ_s, λ_a and λ_b are normalization factors[1].

4 Implementation

Style database. We use a variety of subsets of the captured human motion database that was introduced by Xia *et al.* [2015]. We curate controlled subsets that include only a limited number of actions, which does not fully describe the test motion (*e.g.,* stylizing *jump* action with a database of only *walk* and *kick* actions) in order to demonstrate our methods capability in transferring style between different actions. The complete database includes eight styles (neutral, angry, childlike, depressed, proud, sexy, struggling, old), and six actions (walk, run, jump, kick, punch, transitions). The database is also annotated with foot-plant (right foot on/off, left foot on/off) information.

Preprocessing. We compute skeletal RMS power features (Equation 6) and frequency domain representations ($R[w]$ and $A[w]$ given by Equation 3) for multi-scale windows (with $N = 17, 19, 21$ for 120 *fps*) centered around each time-step. We cache a fast nearest neighbor search database from power feature - frequency domain data pairs. Including a set of different size windows in our database helps accounting for slight variations of the general pace difference between motion examples. We include all DOF of the skeleton except for translation and rotation of the root node in the ground plane. We also build a foot-plant detection table using a concatenated feature vector of $R[w]$ and $A[w]$ for all lower body DOF vs. foot-plant.

[1] In our experiments, $\lambda_s = 1$, $\lambda_a = \lambda_b = \text{mean}\,\{R[\omega]/(\pi/2)\}$.

Real-time style transfer. At runtime, we solve Equation 8 for each time-step successively, in the frequency domain, and reconstruct time-domain data. We use gradient descent and it performs well in for the core formulation provided that the sequence is dominated by a homogeneous motion. Since the initial condition plays an important role, for the generalized formulation we initialize the optimization from the previous frame's solution. For the beginning frames of the signal, we follow a forward-backward pass approach to get a good initial condition: we randomly initialize and do a forward pass on a number of frames as many as the length of the window, followed by a backward pass reversing the order of the frames. This results in a reliable initial condition for the first frame. Finally, as a post-processing step we detect the foot-plant using the precomputed foot-plant nearest neighbor search database, and clear foot sliding artifacts trivially with the detected foot-plant information [Kovar and Gleicher 2004].

Performance. Our system achieves an average of 50 fps processing speed (with a $\sim$0.1 second buffer before the first frame due to the forward-backward initialization process) on an Intel i7-3.5GHz eight core PC, when the full database is included in the search tables, with parallelized lookup.

5 Evaluation and Discussion

5.1 Experiments

Homogeneous Action. We evaluate our method quantitatively by generating controlled input data from known stylized data (100 randomly selected continuous motion sequences) available in the database: We convert all stylized data to neutral by using our method, the method introduced by Xia *et al.* [2015], and the method of by [Hsu et al. 2005] (Three neutral counter-parts for each stylized motion: resulting from translating 100 clips from their stylized version to neutral by all methods (ours, [Hsu et al. 2005] and [Xia et al. 2015]), resulting in 300 [neutral]-[stylized] pairs. Each method translates all these 300 pairs back to stylized, and error is reported on the aggregate of these 300 pairs. So, each method is stylizing the other methods neutralized sequences). We then stylize the synthetically created neutral motions sequences and compare with the original. Figure 6 shows the average error of leave-one-out experiments with different database combinations for our method (Section 3.3), [Xia et al. 2015] and [Hsu et al. 2005]. As the disparity between the synthesized motion and the actions available in the database increase, our method significantly out performs the others.

Database Enhancement. One of the immediate applications of our method is to utilize it for enhancing existing databases by populating data for non-existent actions in a database. This leads to an enhanced database, which could then be utilized in order to perform style transfer with previous methods. To show the effectiveness, we test stylized motion error of [Xia et al. 2015] with limited database, full-database, and enhanced data-base for various actions. We generate the enhanced database by using the stylized limited database to generate stylized versions of neutral states of missing actions. As seen in Figure 7, average joint position leave-one-out error reduces considerably with the enhanced database that is generated using our method. For this example, we used 50 randomly selected neutral motions for each missing action and generated stylized versions with our method to complement the limited database. Since style transfer using a previous method with the enhanced database vs. limited database show significant improvement, this establishes that our style transfer between independent actions actually transfer meaningful components.

Visual Comparison and Results. Figure 8 shows an example motion sequence with heterogeneous action from real motion capture

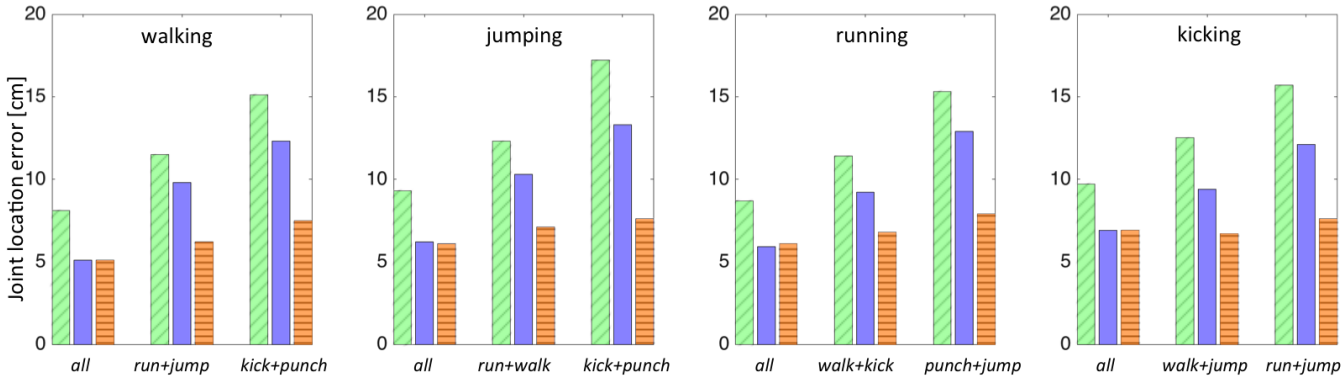

Figure 6: *Leave-one-out homogeneous action experiments. Joint location error of [Hsu et al. 2005]:, [Xia et al. 2015]:, and Our Method:. Name of the stylized action is on top of each graph. X-axis of each group shows the available actions in the database during stylization. (Note: maximum standard deviation among all methods were less than 2.0 cm. with insignificant variation between methods, hence, are excluded from the charts.)*

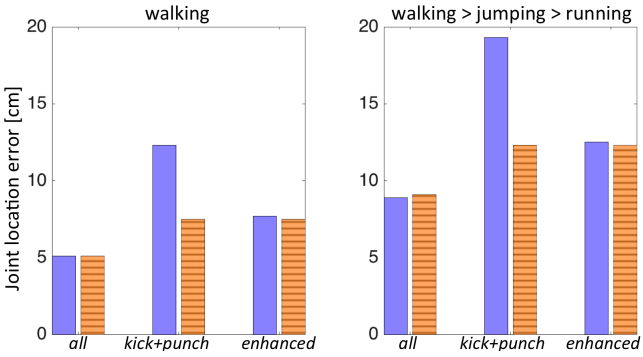

Figure 7: *Leave-one-out experiments with different databases: all actions available, only kick and punch available, and enhanced database where the missing actions are filled-in with our method. Joint location error of [Xia et al. 2015]:, Our Method:. Stylized motion sequence on top of each graph.*

(Figure 8a) stylized with our method with the 'sexy' style (Figure 8b), and [Xia et al. 2015] (Figure 8c). We use a reduced database of only walk and punch actions. Note the stylization of the jump motion in our result, where the previous work stays closer to the original 'neutral' content because the action is missing from the database. Please refer to our supplementary video for more visual results in continuous video format.

5.2 User Studies

We conducted a user study on Amazon Mechanical Turk (AMT) to both assess how well our system performs stylization, and to compare with previous methods. We designed and administered four different question types (Please refer to our supplementary material for representative screen shots of each question type). We rejected individual responses where the user recorded their response without waiting the videos to finish run. We rejected all responses of a user if they have exhibited this behavior for more than two questions (in 20 questions). In total, we collected more than 250 unique user responses excluding 30% of rejected data. This resulted in at least 600 responses for each question type (and their repetitions if mentioned) explained in detail below:

(1) Recognition. The user is presented with a video of a stylized motion and is asked to select the best style that is present in the action. We ask the same questions with a motion capture as well, to assess the performance against the confusion of style in motion capture data. We repeated this user study with real mo-cap data as well.

Table 2: *Ratio of users who confused sexy style with proud style (User marked 'proud' although the motion was labeled as 'style').*

	Walk	Run	Kick	Jump	Punch
Our method	87%	22%	12%	5%	0%
Real mo-cap	85%	23%	14%	4%	0%

Table 3: *Ratio of users responses to the realism questions.*

	Ours	Mo-Cap	Both	None
Angry	22%	25%	**48%**	5%
Childlike	35%	37%	24%	4%
Depressed	32%	30%	32%	6%
Proud	14%	16%	**67%**	3%
Sexy	26%	24%	15%	**35%**
Struggling	32%	30%	14%	**24%**
Old	21%	23%	**54%**	2%

Figure 9 shows the confusion matrix of our stylized motions and the real motion capture data. Note that we achieve recognition rates similar to the real capture data. An interesting outcome of this user study is the fact that *sexy* style being significantly confused with *proud* style. The results shown in Figure 9 are averaged over all action types, however we breakdown the confusion between sexy and proud in Table 2. Note that most of the confusion is concentrated in the walk motion, with run and kick also being significantly confused.

(2) Realism. The user is presented with two simultaneous videos: one corresponding to a real stylized motion capture from the database, and the other corresponding to a stylized motion in the same style with our method. We ask the user to choose the video(s) that represent real motion capture data. Table 3 shows the results for each style. The users were given the options to demarcate 'both' or 'none' of the motions as 'real capture'. The results show that our algorithm is easily confused with real motion capture (Column 1 and 2 in Table 3 are similar for all styles). Note that the least convincing results were 'Sexy' and 'Struggling' for both our method and the motion capture since both our results and motion capture were not convincing to the user.

(3) Heterogeneous Action Comparison vs. [Xia et al. 2015]. The user is presented with two simultaneous videos corresponding to a heterogeneous motion that is stylized with our method, and [Xia et al. 2015]. The user is given the style name asked to choose the more realistic video that represents the given style. We utilized a number of different heterogeneous human capture data made publicly available by Ohio State University Motion Capture Lab[2], and

[2]http://accad.osu.edu/research/mocap/

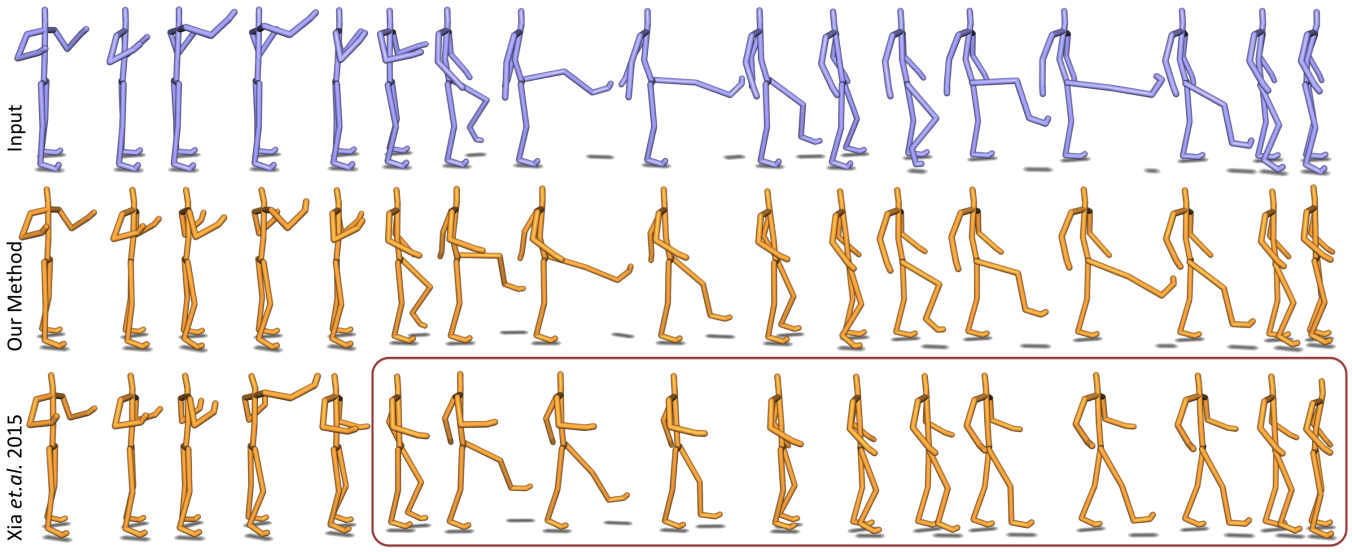

Figure 8: *Key frame visual comparison. Real motion capture input in 'neutral' style (top). 'Sexy' stylization with a limited database of 'walk' and 'punch' motions: our method (middle),[Xia et al. 2015] (bottom). Note the 'kick's being represented similar to 'walk' by [Xia et al. 2015] because of the missing 'kick' data in the database.*

CORRECT STYLE LABEL

USER SELECTED STYLE LABEL	Neutral	Angry	Childlike	Depressed	Proud	Sexy	Struggling	Old
Neutral	96 % 94 %	- 6%	- -	- -	4% 0%	- -	- -	- -
Angry	- -	93 % 90 %	- -	- -	7% 10%	- -	- -	- -
Childlike	- -	- -	92 % 94 %	- -	- -	- -	8% 6%	- -
Depressed	- -	- -	- -	89 % 86 %	- -	- -	- -	11% 14%
Proud	- -	8% 10%	- -	- -	92 % 88 %	- 2%	- -	- -
Sexy	- -	- -	10% 8%	- -	32 % 35 %	58 % 57 %	- -	- -
Struggling	- -	- -	- -	- -	- -	- -	92 % 91 %	8% 9%
Old	- -	- -	- -	- -	- -	- -	6% 5%	94 % 95 %

Figure 9: *Recognition user-study confusion matrices. In each cell: real motion capture (top), and our stylized motion (bottom) percentage values are given. Note that we achieve recognition rates similar to the real capture data. There is significant confusion in recognizing sexy style, which is thought as proud by the users both for the real motion capture and our transfer results.*

the CMU Graphics Lab[3], pre-processing for skeleton compatibility [Monzani et al. 2000]. We repeated this question type with multiple subsets of the database, with different heterogeneous motions as shown in Figure 10.

(4) Database Enhancement Comparison. The user is presented with two simultaneous videos corresponding to a heterogeneous motion that are both stylized with [Xia et al. 2015]. However, one of the videos is stylized using a limited database, whereas the other one is stylized using a database which was enhanced with our method starting from the limited database. The results (Figure 11) show that the users found the enhanced motion significantly more convincing than the motion synthesized from the limited database. The results of [Xia et al. 2015] on the enhanced database is indistinguishably similar to ours which shows that if our method is used to fill-in motions in a database with our method, then existing methods can also be used to achieve similar results.

[3]http://mocap.cs.cmu.edu/

6 Conclusion

We introduced a spectral style transfer method which takes advantage of the spectral space to compute similarity and synthesize stylized motion enabling transfer style between independent actions. We also showed that our method can be used to enhance a style transfer database by creating stylized motions of missing actions from their neutral counterparts. This in turn results in significantly increased performance for methods tailored to use full action representation in the database.

Limitations and Future Work. A limitation of our work arises from the fact that style and action are significantly correlated for certain actions. A particular move or behavior can be unique to both a particular combination of style and action, and might not be observable in other action types. Our method will not be able to transfer style in such cases, since there is no observed data for the unique style-action correlation. In addition, if the motion is very different from what is seen in the database, the method is likely fail but gracefully – there will be minimal style transfer. Since our transfer is energy-based, if the database has motion where energy is concentrated only on the upper body (eg. punch), and the input sequence has motion only in the lower body (eg. jump), the transfer will be minimal and may not be noticeable. A promising future direction is building a stylized motion synthesis system using our approach. However, since we operate in the spectral domain, imposing direct spatial constraints (*e.g.,* hand has to follow a certain path) for motion editing requires further research.

Acknowledgements

We thank the authors of [Xia et al. 2015] for sharing their data. This work was partly supported by ERC Starting Grant SmartGeometry (StG-2013-335373), Marie Curie CIG 303541, and Adobe Research fundings.

References

AMAYA, K., BRUDERLIN, A., AND CALVERT, T. 1996. Emotion from motion. In *Graphics interface*, vol. 96, Toronto, Canada, 222–229.

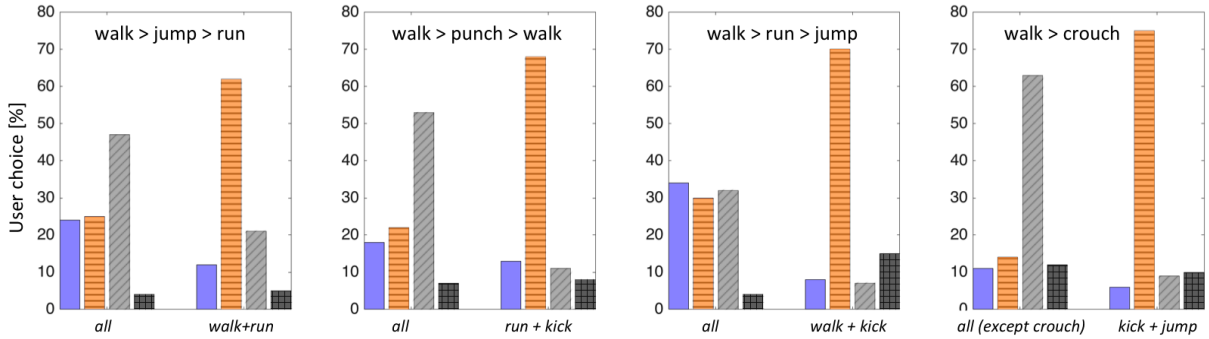

Figure 10: *Heterogeneous stylization user-study results: Users choose from:* [Xia et al. 2015]:■, Our Method:■, Both:■, *and* **None:**■. *Name of the stylized action sequence is on top of each graph. The x-axis marks the type of actions that were available in the database used for stylization by both methods: Left hand side marked as all, means that all action types were available in the database when stylizing the motion, Right hand side marks the type of actions that were available in the database when stylizing the motion.*

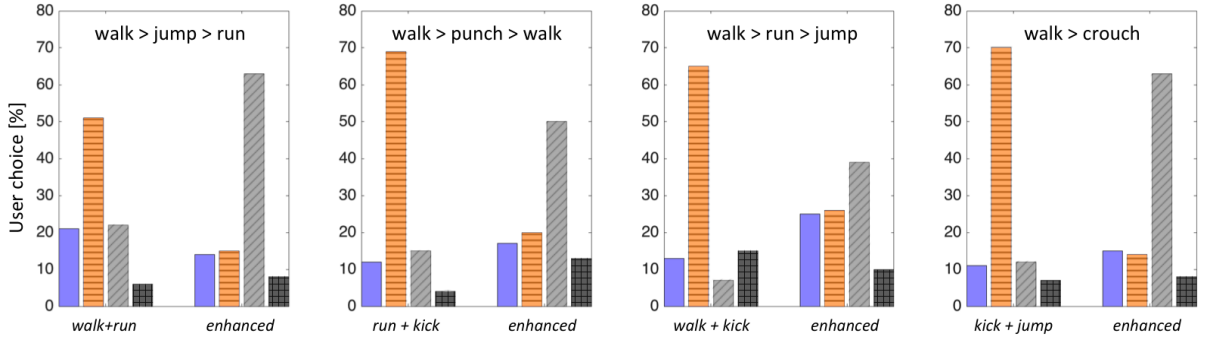

Figure 11: *Database enhancement user-study results: Users choose from:* [Xia et al. 2015]:■, Our Method:■, Both:■, *and* **None:**■. *Name of the stylized action sequence is on top of each graph. X-axis of each group shows the available actions in the database for stylization similar to Figure 10 (missing action's in enhanced database is filled-in with our method).*

ARIKAN, O., FORSYTH, D. A., AND O'BRIEN, J. F. 2003. Motion synthesis from annotations. *ACM Transactions on Graphics (TOG) 22*, 3, 402–408.

BRAND, M., AND HERTZMANN, A. 2000. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., 183–192.

CHAI, J., AND HODGINS, J. K. 2007. Constraint-based motion optimization using a statistical dynamic model. *ACM Transactions on Graphics (TOG) 26*, 3, 8.

GIESE, M. A., AND POGGIO, T. 2000. Morphable models for the analysis and synthesis of complex motion patterns. *International Journal of Computer Vision 38*, 1, 59–73.

HSU, E., PULLI, K., AND POPOVIĆ, J. 2005. Style translation for human motion. *ACM Transactions on Graphics (TOG) 24*, 3, 1082–1089.

HYVÄRINEN, A., KARHUNEN, J., AND OJA, E. 2004. *Independent component analysis*, vol. 46. John Wiley & Sons.

IKEMOTO, L., ARIKAN, O., AND FORSYTH, D. 2009. Generalizing motion edits with gaussian processes. *ACM Transactions on Graphics (TOG) 28*, 1, 1.

KOVAR, L., AND GLEICHER, M. 2004. Automated extraction and parameterization of motions in large data sets. In *ACM Transactions on Graphics (TOG)*, vol. 23, ACM, 559–568.

LAU, M., BAR-JOSEPH, Z., AND KUFFNER, J. 2009. Modeling spatial and temporal variation in motion data. In *ACM Transactions on Graphics (TOG)*, vol. 28, ACM, 171.

MIN, J., AND CHAI, J. 2012. Motion graphs++: a compact generative model for semantic motion analysis and synthesis. *ACM Transactions on Graphics (TOG) 31*, 6, 153.

MIN, J., LIU, H., AND CHAI, J. 2010. Synthesis and editing of personalized stylistic human motion. In *Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games*, ACM, 39–46.

MONZANI, J.-S., BAERLOCHER, P., BOULIC, R., AND THALMANN, D. 2000. Using an intermediate skeleton and inverse kinematics for motion retargeting. In *Computer Graphics Forum*, vol. 19, Wiley Online Library, 11–19.

OPPENHEIM, A. V., AND SCHAFER, R. W. 2009. *Discrete-time signal processing*, vol. 3. Prentice-Hall.

SHAPIRO, A., CAO, Y., AND FALOUTSOS, P. 2006. Style components. In *Proceedings of Graphics Interface 2006*, Canadian Information Processing Society, 33–39.

URTASUN, R., GLARDON, P., BOULIC, R., THALMANN, D., AND FUA, P. 2004. Style-based motion synthesis. In *Computer Graphics Forum*, vol. 23, Wiley Online Library, 799–812.

WANG, J. M., FLEET, D. J., AND HERTZMANN, A. 2007. Multifactor gaussian process models for style-content separation. In *Proceedings of the 24th international conference on Machine learning*, ACM, 975–982.

XIA, S., WANG, C., CHAI, J., AND HODGINS, J. 2015. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG) 34*, 4, 119.