Survey on Style in 3D Human Body Motion: Taxonomy, Data, Recognition and its Applications

Sarah Ribet, Hazem Wannous, Jean-Philippe Vandeborre

▶ To cite this version:

Sarah Ribet, Hazem Wannous, Jean-Philippe Vandeborre. Survey on Style in 3D Human Body Motion: Taxonomy, Data, Recognition and its Applications. IEEE Transactions on Affective Computing, 2021, 12 (4), pp.928-948. 10.1109/TAFFC.2019.2906167 . hal-02420912

HAL Id: hal-02420912
https://hal.science/hal-02420912

Submitted on 20 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

Survey on Style in 3D Human Body Motion: Taxonomy, Data, Recognition and its Applications

Sarah Ribet, Hazem Wannous, *Member, IEEE,* and Jean-Philippe Vandeborre, *Senior, IEEE*

***Abstract*—The meaning of the word *style* depends on its context. While actions have already been quite studied for a while, *style* in human body motion is a growing topic of interest. In the context of animation, *style* is crucial as it brings realism and expressiveness to the motion of a character. Even though it is undoubtedly a key element in motions, its definition and the use of the word *style* in itself, among research works, lack consensus. Achieving realistic motions is tedious. It requires either a large motion capture dataset or the considerable work of artist animators. The lack of consistent *style* data is thus a challenge. Stylistic motion generation is quite studied in order to overcome this issue. This paper focuses on the study of *style* in human body motion from 3D human body skeletal data. It establishes a taxonomy of definitions of *style*, describes the data that have been used up until now, introduces key notions about motion capture data as well as machine learning, and presents approaches about *style* recognition, person identification through their *style* and motion *style* generation.**

***Index Terms*—style, motion, character animation, motion style generation**

I. Introduction

STYLE has various meanings and refers to different fields: someone fashionable wears clothes with *style*, a painter has a specific *style*, there are different *styles* of music, etc. This paper deals with the study of *style* in human body motion. Indeed, a growing interest in body movement analysis is observed recently. It is due to the reliability of whole body sensing technologies that are now affordable and, more importantly, that are more efficient due to the increased processing power. While many approaches focus on actions, this paper focuses on *style* in 3D human body motion.

The content of this paper has to be seen in the context of animation and its multiple applications such as entertainment (feature films, video games), human computer interaction, education, health, scientific visualization and simulation, video surveillance, etc. [1]–[4]. Moreover, it focuses on approaches dealing with 3D skeletal data of human body. As a result, we focus on papers that are either recognizing *style* in human body motion or that are generating stylistic motions. We insist on the fact that the whole body has to be involved; papers working on face are out of scope of that paper. According to the taxonomy of definitions of *style* we establish in Section II, emotions are part of *style*. However, emotions are already a wide studied subject. To keep the scope of the paper reasonable and to focus solely on *style*, we take the bias to exclude papers that are dealing with emotions.

Bringing realism and expressiveness to an animated character is one of the main challenges in computer animation [5], [6]. *Style* is then an added value to the motion and this is crucial in that field [7], [8]. It is especially critical as the human eye is sharp and can easily detect unnatural motions [1], [3], [9]. Moreover, the concept of the uncanny valley [10] also applies to character animation [11]. In order to get out of the uncanny valley, where a human being experiences feelings going from uneasiness to revulsion while looking at, in this case, an animated character that is trying to resemble a human being, the animated character needs to reach a certain level of realism. *Style* can help in that purpose, surpass the uncanny valley.

Accurate realistic motions can be obtained either by the considerable work of 3D animators who manipulate details [6], [8] or by capturing motions [12]. The animators' work is time-consuming, expensive and tedious, as they do the animations from scratch by hand most of the time [6], [8], [13]. Capturing motions is also time-consuming and a burden for actors, especially when combinations of actions and styles are needed [2] and should ideally be performed several times [14]. One way of overcoming this is by generating new motions, thus reducing the amount of captures needed in datasets [8], [15] and saving animators time [6], [16]. As a result, generating new stylistic 3D human body motions is studied.

Style in 3D human body motion is per consequent studied, particularly when it comes to stylistic motion generation. We identify three types of motion style generation: synthesis of stylistic motion, editing of stylistic motion and motion style transfer. It is also studied as a classification problem and can be a tool to identify persons. However, the notion of *style* is not clearly defined and lacks a formal definition; different trends are observed. Furthermore, *style* in 3D human body motion is a growing topic of interest that has been less studied than human body motion action. Consequently, the amount and quality of 3D human body data available on *style* is not comparable to the data used in the study of motion actions.

The rest of the paper is structured as follows. Section II establishes a taxonomy of definitions of *style* in 3D human body motion, exposing the different trends that can be observed in computer literature. Data used in motion *style* approaches are described in Section III. Key concepts about motion capture data and some related machine learning techniques are exposed in Section IV. Section V highlights motion *style*

S. Ribet and J.-P. Vandeborre are with IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France. e-mail: {sarah.ribet,jean-philippe.vandeborre}@imt-lille-douai.fr

H. Wannous is with Univ. Lille, CNRS, Centrale Lille, IMT Lille Douai, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France. e-mail: hazem.wannous@univ-lille.fr

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

recognition studies. Section VI provides details about methods performing stylistic 3D human body motion generation. An overwhole discussion about motion *style*, the data related to it and its applications is conducted in Section VII.

II. Taxonomy of Style in Body Motion

In computer animation literature, the concept of motion style is vaguely defined [17]. Providing a definition of *style* in body movements is not an easy task; it leads to a lack of consensus [18]. Moreover, the words *affective state* and *body expression* also refer to *style*. We only use *style* in the rest of this paper.

This section aims at introducing a subjective proposition of a taxonomy of definitions of *style* in body motion, from how *style* is seen by research authors, and exposes relations between *emotions* and *style*.

A. Taxonomy of definitions of style in body motion

Three common trends can be observed with respect to how body motion style is seen in the literature. Fig. 1 illustrates these trends that are described in this section.

1) Style as a component of a motion: This trend gathers research works that consider *style* as an add-on to a particular motion.

Some researchers stated that *style* is one of the components of a motion. Rose *et al.* [19] introduced the notion of *verbs* and *adverbs*. They designated parameterized (stylistic) motions as *verbs* being parameterized by *adverbs*. For example, *walk* was a verb and *happy* and *sad* were some of its *adverbs*. Morawetz [20] opted to decompose an emotional motion into primary and secondary movements; an emotion was thus considered as a secondary movement in addition to a primary movement. Etemad and Arya [9], [21] also decomposed a motion into primary and secondary themes, stating that a stylistic motion amounted to a neutral motion to which a *style* was added.

2) Style as variations in a motion: Style in body motion is sometimes defined as a variation in a motion. As a result, this trend clusters the differences (variations) that can be observed between motions of the same action.

It can be a variation in the type of motion itself. For example, Brand and Hertzmann [13] claimed that "walking, running, strutting, etc., are all stylistic variations on bipedal locomotion". Chien *et al.* [18] referred to "walking, limping, running, jumping, etc." as being *styles* of human motion.

It can also be a variation within the same action, considering intuitive features such as the speed of a movement [22].

Style can also represent variations due to the undeniable fact that every human being is different which induces natural variations in our movements [23]. Indeed, each and one of us human being introduces her/his own uniqueness to motions: that is in itself a variation in motions. It is commonly admitted that someone can not perform the same motion in the exact same way twice [3], [22], [24], [25]. Indeed, Lasseter [26] mentioned that "one character would not do a particular action the same way in two different emotional states. [...] No two characters would do the same action in the same way". *Style* can thus be seen as being the *style* of a person.

Finally, *style* can also simply be seen as spatio-temporal variations in a motion [2], [7], [9], [21].

3) Style as individual-related features: Another trend, less abstract than the previously identified ones, involves individual-related features that can be shown through adjectives. They are tangible characteristics of an individual, of a human being. Table I highlights *styles* of research approaches that are introduced in further details in subsequent sections (see Sections V and VI).

We can notice the heterogeneity in adjectives used. It goes from *happy* or *proud* to *childlike* or *zombie*, including *sexy* and *old*, etc. Still, common characteristics appear such as emotions, gender, age, physical states, personality features and behaviors, that can thus be considered as subsets of *style*. It is in accordance with the definitions of *style* from Troje [29] and Abdul-Massih *et al.* [28], who respectively claimed that *style* referred to "emotions, personality or biological features, such as age or gender" and "your perceived mood, behaviour or physical properties of the motion". It was confirmed by Etemad and Arya [14] who specified that the *style*, represented by secondary themes, contained "variations caused by individual characteristics of the actor such as gender, age, emotions, energy, mood, health, and even inherited characteristics". The second column of Table I allocates the adjectives used in literature to one of the categories identified as subsets. Even though this allocation has been conducted through readings about emotions and behaviors, the reader should be aware that this proposed allocation is subjective. Fig. 2 summarizes the different categories of those adjectives depicting individual-related features that can be found up until now.

According to Table I, research approaches didn't seem to focus on a specific category of *style* and rather attempted to tackle several categories at once. Table II presents the adjectives and their occurrences, per category. The behavior category gathers more adjectives than the emotion, personal feature and physical state categories. However, most of those behavior adjectives are referred to once. They were thus not the most used. Note that both biological feature categories only contain two adjectives each, as expected, and the *no category* subset only reports the absence of *style* represented by the single adjective *neutral*. The most studied *styles* were *angry*, *happy*, *sad*, *old*, *proud*, *depressed* and *tired*. Aside from *old*, they belong to the emotion category, making it the most studied category. Furthermore, about 2/3 of the referenced papers worked with the *neutral style*, which is actually hard to define or get (either by simulation or with real data). We refer the reader to Section VII-A2 for a further analysis of this *style*.

B. Emotions and style

Emotions have been defined for quite a while now [1]. Plutchick and Kellerman [30] considered emotions as "environmental and psychological events influence brain processes that actively modulate clearly observable behaviors". Six basic – universal – emotions have been identified by Ekman [31]: *anger*, *happiness*, *sadness*, *disgust*, *fear* and *surprise*. Emotion representations have later been introduced; they identified more emotions and measured them according to dimensions. They included the model from Russel [32] for which each

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

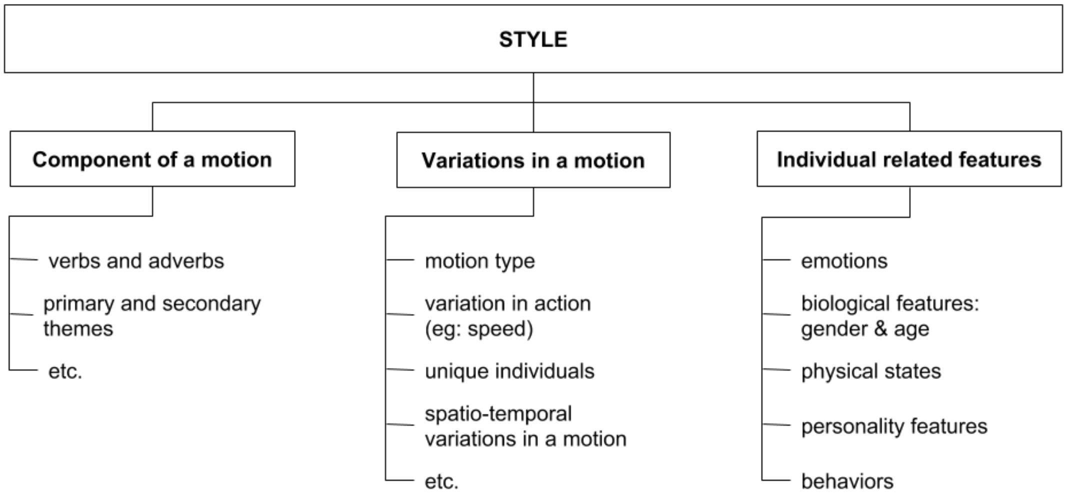

STYLE
Component of a motion
verbs and adverbs
primary and secondary themes
etc.
Variations in a motion
motion type
variation in action (eg: speed)
unique individuals
spatio-temporal variations in a motion
etc.
Individual related features
emotions
biological features: gender & age
physical states
personality features
behaviors

Fig. 1: Common trends (including examples and/or subcategories) of the definition of *style* as seen in the literature.

TABLE I: Individual-related adjective features seen as *style*.
In the category column: B= Behavior, BF:A= Biological Feature: Age, BF:G= Biological Feature: Gender, E= Emotion, PF= Personality Feature, PS= Physical State, –= no category.

Style	**Category**	[1]	[2]	[5]	[27]	[24]	[25]	[3]	[14]	[7]	[28]	[4]	[9]	[15]	[8]	[16]
afraid	E							✓								
angry	E	✓				✓	✓			✓	✓	✓			✓	
catwalk	B		✓		✓		✓									
childlike	B									✓		✓			✓	
confident	PF										✓					
cool	PF							✓								
crab walking	B						✓									
crouch	B		✓													
decided	B							✓								
depressed	E									✓		✓		✓	✓	✓
drunk	PS							✓			✓					
energetic	PS								✓				✓			
excited	E	✓														
fearful	E	✓										✓				
feminine	BF:G												✓			
flap	B		✓													
frustrated	E										✓					
goose-step	B						✓									
happy	E	✓		✓		✓	✓		✓			✓	✓			
heavy	B							✓								
in a hurry	B							✓								
injured	PS		✓											✓		✓
inverse	B		✓													
lame walking	B						✓									
limp	B		✓													
manly	PF							✓								
march	B		✓													
masculine	BF:G												✓			
mummy	B										✓					
neutral	–	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓		✓	
old	BF:A								✓	✓		✓		✓	✓	✓
proud	PF		✓				✓	✓		✓		✓			✓	
relaxed	E	✓														
sad	E	✓		✓		✓	✓	✓	✓		✓	✓	✓			
sexy	PF									✓		✓			✓	
sideways	B		✓													
sneak	B		✓				✓									
strong	PF	✓														
struggling	B														✓	
strutting	B									✓		✓				
tiptoeing	B							✓								
tired	E	✓					✓		✓				✓			
topmodel	B							✓								
weak	PS	✓														
young	BF:A								✓							
zombie	B													✓		✓

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

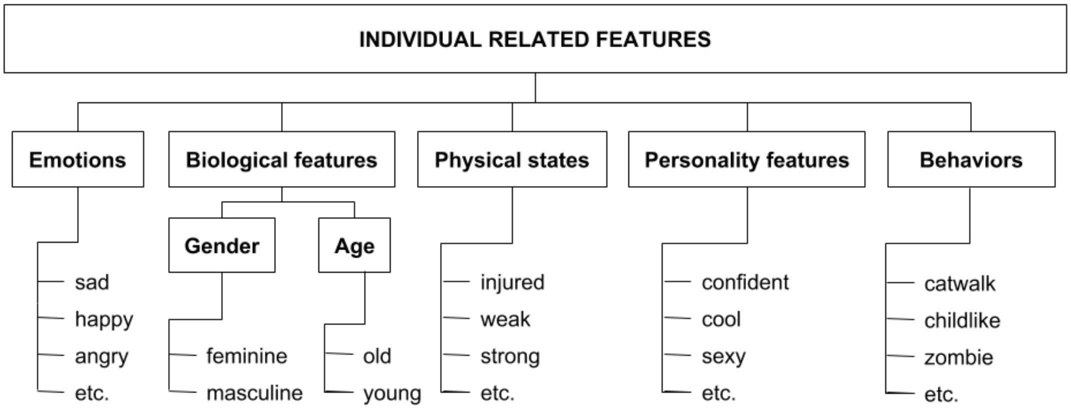


Fig. 2: Classification of *style* when seen as individual-related features.

TABLE II: Occurrences of styles as individual-related features per category.

(a) Behavior

Style	Nb. occur.	Style	Nb. occur.
catwalk	3	limp	1
childlike	3	march	1
crab walking	1	mummy	1
crouch	1	sideways	1
decided	1	sneak	2
flap	1	struggling	1
goose-step	1	strutting	2
heavy	1	tiptoeing	1
in a hurry	1	topmodel	1
inverse	1	zombie	1
lame walking	1		

(b) Biological Feature: Age

Style	Nb. occur.
old	6
young	1

(c) Biological Feature: Gender

Style	Nb. occur.
feminine	1
masculine	1

(d) Emotion

Style	Nb. occur.
afraid	1
angry	7
depressed	5
excited	1
fearful	1
frustrated	1
happy	7
relaxed	1
sad	7
tired	4

(e) Personal feature

Style	Nb. occur.
confident	1
cool	1
manly	1
proud	6
sexy	3
strong	1

(f) Physical state

Style	Nb. occur.
drunk	2
energetic	2
injured	3
weak	1

(g) No category

Style	Nb. occur.
neutral	11

emotion was a point in the space defined by valence and arousal dimensions which specified if an emotion was positive or negative and its intensity (see Fig. 3). The circumplex model of Plutchik and Conte [33] had a similar configuration. The Pleasant-Arousal-Dominant (PAD) model [34], presented in Fig. 4, proposed a representation measuring how pleasant an emotion was (pleasure dimension), its intensity (arousal dimension) and its dominant nature (dominant dimension). Surveys from Zacharatos *et al.* [35] and Karg *et al.* [36] gave more details about the different models of emotions.

Emotions are a subset of *style*. While it seems natural to include all the emotions, most state-of-the-art approaches on *style* focused only on some of the six basic emotions identified by Ekman [31], as shown in Table I. This is a restriction that induces emotions to be quite simplistic within *style*. More complex models [32]–[34] prove this wrong.

Emotions are subjective: people can have divergent interpretation of an emotion. It can for example depend on culture (see Section VII-A3) or on the fact that each human being is unique. A lot more could be said on this subject, such as the difference between *emotions*, *feelings* and *mood*. However,

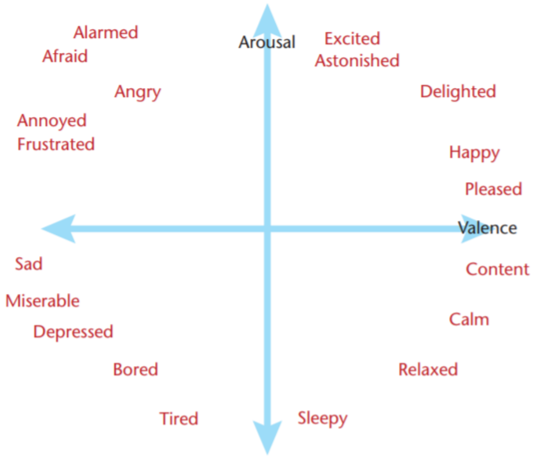


Fig. 3: The valence-arousal space. Image reproduced from [35].

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

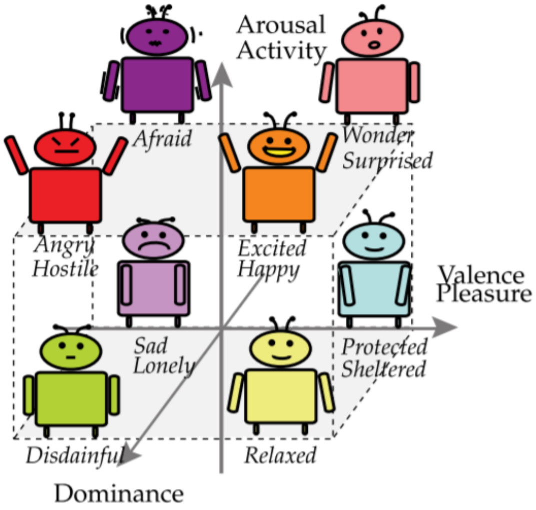


Fig. 4: The Pleasant-Arousal-Dominant (PAD) model. Image reproduced from [36].

emotions is a complex topic that can not be fully addressed in this survey.

Liu *et al.* [5] referred to *style* as a range of phenomena, such as variations due to emotional state, individual body shape, and functional activity such as walking or running. This definition gathers different trends of the definition of *style*. As a conclusion, *style* can be interpreted as representing spatio-temporal variations of a motion that add value to the motion, depending on individuals.

III. Data Used in Human Body Motion Style Approaches

As almost all of the approaches in literature chose one trend of the *style* definition presented in Section II, data that were used to study *style* were of various types. Table III presents an overview of the datasets used in motion style approaches. The data came from two main application fields: action and *style*. Moreover, most of those data have been collected via motion capture (mocap) systems. Mocap is a popular technique to represent human motion from tracked markers corresponding to different regions or joints of the human body. Further details on this technique are provided in Section IV-A.

This section only focuses on the datasets used in the reviewed papers and mentioned in Table III. Thus, this section first presents action specific datasets that were used in motion style methods. Then we introduce specific datasets for *style*. Finally, we present how they were used for motion style approaches.

A. Action datasets used in motion style approaches

Action datasets are widely used to recognize actions and gestures. We refer the reader to surveys such as the one from Presti [50] to appreciate approaches on action recognition and the data that were used to perform it. Some of them were used in motion style approaches, for the purpose of motion style recognition, motion style generation (editing/synthesis or transfer) and person identification through their style. Those datasets that were specifically sometimes used in *style* analysis are introduced.

1) Berkeley Multimodal Human Action Database (MHAD) [37]: The dataset has been made of 11 actions (*jumping in place, jumping jacks, bending – hands up all the way down, punching – boxing, waving – two hands, waving – one hand: right, clapping hands, throwing a ball, sit down then stand up, sit down, stand up*). A total of 660 action sequences were recorded by 12 subjects (seven males + five females) who performed five repetitions of each action. The optical motion capture system Impulse captured the 3D position of 43 active LED markers on a custom built tight fitting suit. Skeletons of 21 joints have been extracted using MotionBuilder software. The capturing frequency was 480 Hz.

2) Carnegie Mellon University (CMU) Graphics Lab Motion Capture Library [38]: The dataset has been recorded by 12 Vicon infrared MX-40 cameras at 120 Hz. About 144 actors performed motions divided into four categories including the locomotion one with *running*, *walking*, *jumping* and varied (*run/leap*) actions.

3) HDM05 [39]: The dataset has been made of 2,337 motion sequences divided into 130 gesture groups including *regular*, *happy* and *sad walks* performed by five non-professional actors. Motion sequences were recorded by an optical marker-based technology (made of a suit with 40-50 retro-reflective markers and 6 to 12 calibrated high resolution cameras). The frame rate was up to 240 Hz. Data were represented by 24 joints.

4) Microsoft Research (MSR) Action 3D Dataset [40]: It was designed to interact with game consoles. The dataset has been made of 20 action types: *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*, *bend*, *forward kick*, *side kick*, *jogging*, *tennis swing*, *tennis serve*, *golf swing*, *pickup & throw*. A depth camera (infra-red light) recorded depth maps of resolution 640x240 at 15 frames per second. In total, 567 sequences were captured thanks to 10 subjects who performed two or three times each action. Data were later processed to obtain skeletal data of 20 joints.

5) Microsoft Research Cambridge 12 (MSRC-12) [42]: The dataset has been made of 16 actions split into two categories: iconic gestures (*crouch or hide*, *shoot with a pistol*, *throw an object such as a grenade*, *change weapon*, *kick to attack an enemy*, *put on night vision goggles to change the game mode*) and metaphorical gestures (*start music/raise volume*, *navigate to next menu*, *wind up the music*, *take a bow to end the session*, *protest the music*, *lay down the tempo of a song*). A Microsoft Kinect camera recorded the performance of 30 participants (60% males) at 30 Hz. It resulted in 594 sequences of skeletal data of 20 joints.

6) UCF Kinect [43]: The dataset has been made of 16 actions (*balance*, *climb ladder*, *climb up*, *duck*, *hop*, *kick*, *leap*, *punch*, *run*, *step back*, *step front*, *step left*, *step right*, *twist left*, *twist right*, *vault*). A Microsoft Kinect sensor was used to capture the 1280 action samples. The OpenNI platform was

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

TABLE III: Datasets used in motion style approaches and their frequency of use.

Dataset		**Application**				
Type	**Name**	**Recognition**		**Motion generation**		
		Classification	**Person identification**	**Synthesis**	**Editing**	**Transfer**
Action	*Berkeley MHAD [37]*				[15]	
	CMU [38]			[18], [23]	[15], [22]	
	HDM05 [39]	[14]			[9], [15]	
	MSR Action 3D [40]		[41]			
	MSRC-12 [42]		[41]			
	UCF Kinect [43]		[41]			
	UTKinect [44]		[41]			
Style	*eNTERFACES'08 3D [45]*			[3]		
	Body Login Dataset [46]		[41]			
	Body Movement Library [47]	[4], [24]				
	Mockey [48]			[3]		
	UCLIC [49]	[4]				
Own data		[14]		[12]	[25], [27]	[1], [7], [17]

used to estimate skeletons of 15 joints. There were 16 subjects (13 males + 3 females).

7) UTKinect-Action3D Dataset [44]: The dataset has been made of 10 actions (*walk*, *sit down*, *stand up*, *pick up*, *carry*, *throw*, *push*, *pull*, *wave* and *clap hands*). To collect the dataset, 10 persons (nine males + one female) performed each action twice. There were 200 action samples. RGB images (resolution 480x640) and depth maps (resolution 320x240) were captured with a Kinect at 30 frames per second (fps). The final frame rate was about 15 fps though. Skeletal data (20 joints) were also recorded.

B. Motion style specific datasets

Style specific datasets used in motion style approaches (recognition, generation – editing/synthesis/transfer, person identification) are presented.

1) Body Movement Library [47]: The dataset collected 4,080 movement records split into 1356 motions. Three categories of neutral and emotional (*angry*, *happy*, *sad*, *afraid*) actions were recorded: *walking*, arm movements (*lifting*, *knocking* and *throwing*) and arm movements separated by walking. *Angry*, *happy* and *sad* styles were chosen because of their easiness of recognition and mapping into actions as well as their duration. Indeed, they contrast with reactive emotions such as *surprise* and *disgust* for example that are associated with very specific movements. An attempt was made for the *afraid* style though, with six actors. The collection of data was conducted with the eight cameras of the Falcon Analog optical motion capture system and a suit to which 35 retro-reflective markers were attached. Non professional actors (15 females + 15 males) were given non-verbal instruction through the following emotion scripts (reproduced from [47]):

> **Neutral** Imagine yourself standing by your kitchen table on a Saturday morning. You are well rested, you just had breakfast and yesterday you and your flatmates tidied the house so you are free to do whatever you want. It is a sunny day and you are thinking about what you are going to do today. There's a bit of paper on the table and you pick it up and throw it to the bin.
>
> **Angry** Today you slept in, so you had to rush to get ready. Then on the way to work, a policeman flags you down and gives you a speeding ticket, although you were just keeping up with traffic. You finally get to work where a note is waiting for you denying your request for having Friday off; now you are furious. Standing by your desk, you reach for a bit of rubbish and slam it into the bin as your temper flares.
>
> **Happy** It's Friday evening and you feel great, because earlier you handed in your final year project. Your supervisor was very pleased, he complimented you on it and hinted that you're going to get excellent marks for it. You just talked to your best friend who suggested you go out to celebrate and now you are just waiting for her to arrive. As your excitement mounts you joyously pick up a bit of rubbish on the table in front of you and throw it at the bin.
>
> **Sad** You are in your flat after just trying to finish dinner. You didn't feel like eating, your stomach is heavy and seems to be bearing all of your body's inner activity and weight. Half an hour ago, you received a telephone call that your best friend had died in a car accident. Deeply grieving, you don't know what to do. Needing to do something you reach for a bit of rubbish and throw it in the bin.

2) Body Login Dataset [46]: This dataset was specifically meant to be used for user authentication. Four Kinect cameras recorded depth image (resolution 640x480) and skeleton joint coordinates (20 joints) from 40 users (27 males + 13 females) at 30 fps. They were asked to perform 20 samples for each of two gesture types: an "S gesture" and a "user-defined gesture" *ie.* no instruction was given, each user could choose her/his own gesture. Table IV lists all the user-defined gestures.

3) eNTERFACE'08 3D [45]: This dataset was recorded in order to be able to develop 3D statistical models of human gait *styles*. It was made of 17 *walk* sequences (rectilign at different speeds and direction changes of various amplitudes) from each of the 40 subjects. The IGS 190 motion capture suit was used to record the motions at 60 fps; it had 19 inertial

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

TABLE IV: User-defined gestures per subject in Body Login Dataset [46]. Table reproduced from [51].

Subject ID	Description	Subject ID	Description
1	Double armed backstrokes	21	Shoot basketball
2	Y of Y-M-C-A	22	Parallel arms forward
3	Backwards jump-rope	23	Muscular pose
4	Backhand Tennis Swing	24	Wipe away motion
5	X-pose	25	Kamehameha
6	Upper body stretch	26	Upwards pose
7	Upper meditation	27	Salute
8	Halfway spin	28	Rockstar
9	Left arm chopping	29	Double clap
10	Long swimming frontcrawl	30	Stretches
11	Crouch forward swim	31	Taichi pose
12	Left-right body tilt	32	Taichi stretches
13	Arm-to-head stretch	33	Double wave
14	Balance-something	34	Stretch leg up
15	Chopping action	35	Jump pose
16	Dribble	36	Golf swing
17	Assorted poses	37	Shake imaginary maracas
18	Upper stretches	38	Y of Y-M-C-A (duplicate)
19	Upper flutter	39	Slow wave
20	Stretch and bend	40	Double hand stretch left right

sensors (with accelerometer, gyroscope and magnetometer). A skeleton of 20 body segments and 66 DOFs (Degrees Of Freedom) was built.

4) Mockey [48]: This dataset aimed at analyzing the expressiveness of walking sequences. A professional actor performed back and forth *walks* in 11 styles: *proud*, *decided*, *sad*, *cat-walk*, *drunk*, *cool*, *afraid*, *tiptoeing*, *heavy*, *in a hurry* and *manly*. The actor was given instructions on how to act the different styles resulting in exaggerated variations of a plain walk. These styles were chosen arbitrarily as they were recognized to have a noticeable influence on walk (see Fig. 5). A motion capture suit, IGS-190, containing 18 inertial sensors (accelerometers, gyroscopes and magnetometers) collected 247 walk cycles at 30 fps. Data were represented by 18 3D joint angles. Joint positions were discarded as they depend on the walk (displacement + orientation).

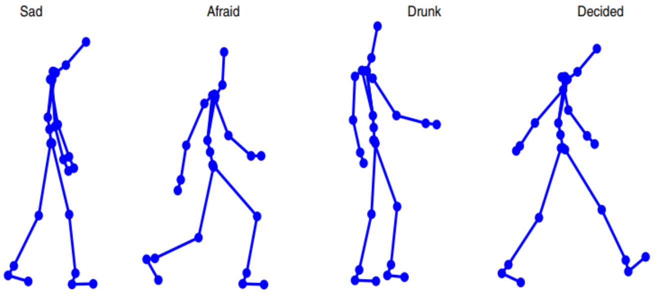


Fig. 5: Four example postures taken from the Mockey dataset. Image reproduced from [48].

5) UCLIC Affective Body Posture and Motion Database [49]: The dataset has been made of 183 body motions collected from 13 cross-cultural non professional actors (11 actors were Japanese, one was from Sri-Lanka and one was from the United States). The emotions portrayed were *anger*, *fear*, *happiness* and *sadness* (see Fig. 6). These emotions were chosen as they were part of the basic emotions identified by Ekman [31]. Data were recorded using a Vicon motion capture system of eight cameras. Actors wore a suit with 32 markers. Data were, after being captured, built into avatars. This aimed at eliminating bias, by creating non-gender nor cultural-specific avatars, and at not affecting people by facial expressions.

C. How the data are used in motion style approaches

Datasets used in motion style approaches, as well as the methods using them, are presented in Table III. Three categories of applications are identified: motion style recognition, motion style generation (editing, synthesis and transfer) and person identification through their style. Each of these categories are explored.

1) Motion style recognition: Both action and style datasets were used. Etemad and Arya [14] used a total of 48 segmented sequences from *HDM05* [39] of regular and stylistic walks (*happy* and *sad*) in addition to their own data. They indeed used a Vicon MX40 to record multiple walks from five actors in four styles (*young*, *old*, *energetic* and *tired*).

Two style datasets were also used. Bernhardt and Robinson [24] used *Body Movement Library* [47] and Crenn *et al.* used both *Body Movement Library* [47] and *UCLIC* [49].

2) Motion style generation: The use of action dataset was predominant. Indeed, in synthesis approaches, Wang *et al.* [23] used six sequences from *CMU* [38] to illustrate *walking*, *striding* and *running* sequences. Chien and Liu [18] also used *walking*, *jumping* and *running* sequences from this dataset. Only Tilmanne *et al.* [3] used *style* datasets, that they actually created: *eNTERFACES'08 3D* [45] and *Mockey* [48]. In editing approaches, Ma *et al.* [22] used *sideways*, *stepping*, *walking* and *running* sequences from *CMU* [38]. Etemad and Arya [9] segmented and extracted 16 2-step *neutral walk* sequences from HDM05 [39]. Holden *et al.* [15] used sequences of motions including *walk*, *run*, *punching* and *kicking* from *CMU* [38], *HDM05* [39] and *Berkeley MHAD* [37].

The lack of stylistic data thus led several approaches to record their own data. To perform motion style synthesis, Urtasun *et al.* [12] used a Vicon optical motion capture system to capture *walking* motions at different speeds from nine people, *running* motions at different speeds from five persons and *jumping* at different distances sequences from four people. Editing approaches also created their own data. Wang *et al.* [27] recorded two *walk* sequences (*regular walk* from an actor and *cat walk* from an actress) with a motion capture device at 33,3 fps. Min *et al.* [25] recorded *walk* sequences in 11 styles (*neutral*, *angry*, *happy*, *sad*, *tired*, *proud*, *sneaky*, *goose-step*, *cat walking*, *crab walking*, *lame waking*) from 20 persons. Holden *et al.* [15] added internal captures to the data they got from existing datasets.

The lack of stylistic data was even more present in style transfer approaches, as no specific and known dataset was used and researchers did have to create their own data. Amaya *et al.* [1] used an Optotrak system to record two actions ("pick up the glass of water, drink from it, and put it back onto the table" and "knock at the door three times") in 10 emotions (*angry*, *sad*, *happy*, *fearful*, *tired*, *strong*, *weak*, *excited*, *relaxed* as well as *neutral*). Torresani *et al.* [17] used 12 *dance* sequences recorded from professional dancers who performed motions in their own *style*. Xia *et al.* [7] decided to create a large

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

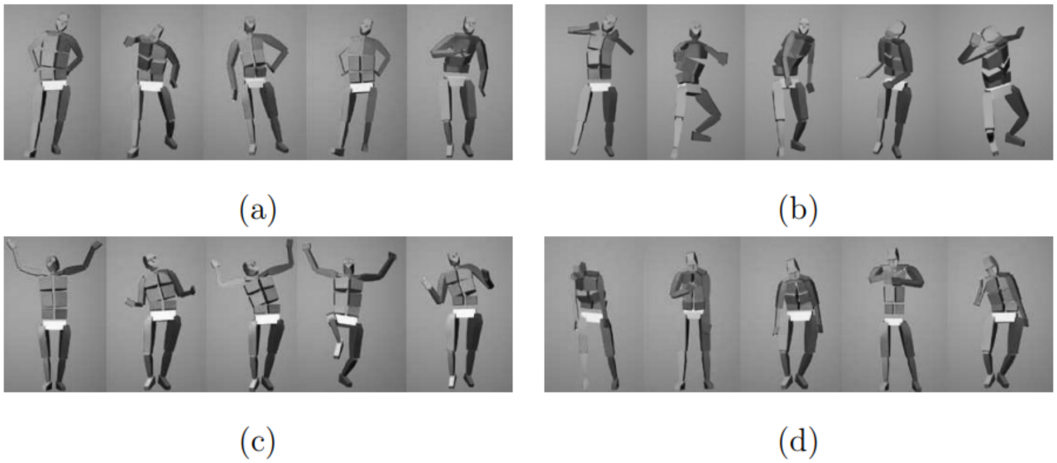


Fig. 6: Examples of the 3D affectively expressive avatars for each emotion category in the UCLIC Affective Body Posture and Motion Database. (a) Angry (b) Fear (c) Happy (d) Sad. Image reproduced from [49].

and heterogeneous human stylistic motion dataset. They thus recorded human actions (*walking*, *running*, *jumping*, *kicking*, *punching* as well as transitions between these actions) in eight *styles* (*neutral*, *proud*, *angry*, *depressed*, *strutting*, *childlike*, *old* and *sexy*). A total of 572 motions were recorded by the Vicon system with 18 cameras at 120 Hz. Their data were in fact reused in several approaches: Crenn *et al.* [4] used them for style recognition, Holden *et al.* [15] used them for editing, Yumer and Mitra [8] as well as Holden *et al.* [16] used them for style transfer.

3) Person identification through their style: Kviatkovsky *et al.* [41] used four action datasets and one style dataset to perform person authentication. Indeed, they used *UTKinect* [44], *UCF Kinect* [43], *MSRC-12* [42] and *MSRAction 3D* [40] action dataset in addition to the style dataset *Body Login Dataset* [46].

As a summary, action datasets were mostly used for motion style editing or synthesis, when *style* was considered as being variations in motions, or when neutral motions were required. The style datasets used for recognition focused mainly on emotions. Two style datasets have been recorded for editing/synthesis; however, as far as we know, only creators of the datasets have used them in motion style analysis. The datasets described in Sections III-A and III-B are publicly available online. Researchers faced a lack of style specific data, especially in style transfer; some of them created their own data. Among them, Xia *et al.* [7] recorded novel data for style transfer; their data have been used for other applications.

IV. Motion Capture Data and Machine Learning

This section presents key notions related to the representation of motion capture data and exposes key concepts of machine learning, as it is one the medium used to analyze motion capture data. We focus on dimensionality reduction and the modeling of time-series, as these two notions were used in research methods on *style* in Sections V and VI and they are part of the discussion led in Section VII.

A. Motion data representation

Motion capture (mocap) is a technique that acquires sequences of 3-dimensional joint positions at high frame rate and enables a large range of applications, from movie special effects to human machine interaction systems. There are different motion capture systems that allow to produce 3D data in real-time: magnetic, mechanic, and optical. Magnetic systems use electromagnetic sensors connected to a computer, which restricts movement due to cabling. Mechanical motion capture systems employ sensors attached to the actor's body that register the motion of articulation. Optical motion capture systems, which are the most used, employ a set of multiple synchronized cameras to capture markers placed in body parts in order to reconstruct the human body posture in movement.

These systems produce data with three degrees of freedom (DOF) for each marker, and rotational information must be inferred from their relative orientation. Human motion is therefore represented as a sequence of human poses, described through body joint positions, or through 3D-joint rotations which are then integrated via forward kinematics. The full-body pose is thus represented as the root position and orientation as well as joint relative orientations.

There are many possible ways to represent a 3D joint angle using a set of numbers, such as Euler angles [52], quaternions [53] and exponential maps [54]. Euler angles are often employed to represent 3D joint angles as rotations around x, y and z axes: three Euler angles are used for three DOFs. Quaternions are complex numbers with a real part and three imaginary parts that represent a rotation in three DOFs. Exponential maps are in general a re-parameterization in $\mathbb{R}^3$ of a quaternion. For these three types, the representation format for motion data starts with the Cartesian root joint position, followed by the first frame's corresponding value (Euler, quaternion or exponential map) of angles of each joint, then the second frame's, and so on. The storage format for such data is different according to the capture system. It usually consists of a hierarchical skeletal structure and trajectories of

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

degrees of freedom of joints.

During the last decade, markerless motion capture approaches have developed and gained an increasing interest thanks to efforts made by the computer vision community [55]. These markerless systems do not require subjects to carry special equipment or markers for tracking, so multiple streams of optical input are often analyzed to identify human forms, such as silhouettes or skeletons, to break them down into component parts for tracking.

Imaging technologies have recently shown a rapid advancement with the introduction of RGB-D sensors with real-time capabilities, such as the Microsoft Kinect, which have provided an alternative for developing markerless motion capture systems at low cost. Note that a camera such as the Kinect, with compromised accuracy and capture rate, exhibits instability when tracking bodies in fast movements, rotating, or whose parts are occluded.

B. Machine learning

The goal of machine learning generally is to understand the structure of data and fit those data into effective mathematical models, which are often statistical, and thus useful in cases that are difficult to model exactly. An important issue in learning from motion data is how to represent their content. In order to achieve a good representation, 3D configurations derived from the motion capture systems, like the relative positions of joints, their temporal difference and the normalized trajectory of the motion, can be used. Motion capture data are time-series data which often contain much noise and have high dimensionality with unique properties that make them challenging to analyze and model. Hence, there is an interest of reducing the dimensionality of time series and extracting relevant information.

1) Dimensionality reduction: Dimensionality reduction techniques can be widely classified as linear and nonlinear techniques. Principal component analysis (PCA) is a linear technique in common use, which aims at finding orthogonal principal directions of a data set by solving an eigenvalue problem, allowing to retain a small number of principal components on a linear, low-dimensional subspace. Independent Component Analysis (ICA) also produces a linear mapping. It assumes that each sample of data is a mixture of components having independent non-Gaussian distributions and it aims at finding these independent components. ICA can capture higher-order statistics in the data matrix instead of only finding correlations between its components. Instead of working on the original data, Kernel PCA, which is an extension to PCA, works on the linear feature space transformed from a nonlinear feature space using a kernel function [56] to perform a nonlinear dimension reduction. Unlike linear techniques, nonlinear ones do not estimate the parameters of mapping but margin them. Hence, Gaussian Process Latent Variable Models (GPLVM) maximize the posterior probability to reduce dimensionality given prior probability.

2) Time-series modeling: Recognition systems based on traditional learning algorithms can be designed in two ways. First, a whole motion is represented by a single feature matrix and classified by a classifier as a whole [57], [58]. Second, the motion is decomposed by a sliding window or key features to build a codebook thanks to the learning phase from the whole dataset, and then each motion can be represented as a bag or histogram of words [59]. For an effective representation of motion data, both the spatial and temporal dynamics of human motion must be modeled. The Hidden Markov Model (HMM) is a popular technique for modeling sequential data. The HMM represents the human motion as a succession of states. At each state, local statistics and state transition probabilities are determined by the training phase on the dataset. After the recent progress in deep learning techniques, many applications of computer graphic field, including motion data recognition and prediction, have shown a change of paradigm. In particular, Recurrent Neural Networks (RNNs) are capable of preserving states as they pass through a step, hence they are suitable for sequence-based problems. The Long Short Term Memory networks (LSTMs) are a special kind of RNNs, with the main difference lying in the inclusion of memory states and gates, that can learn long-term dependencies in time series problems. They notably solve the problem of vanishing gradient [60], which arises in very deep neural networks, including RNNs.

V. Motion Style Recognition

While some papers studied categories of *style* such as emotion or gender recognition [29], or even person identification [61], this section focuses on *style* in body motion. We distinguish two aspects of recognition: motion style classification and person identification through their style.

A. Motion style classification

The recognition of *style* in body motion remains a rarely studied subject. Some methods identified features and classified *styles* with machine learning techniques.

Bernhardt and Robinson [24] analyzed non-stylised motions in order to detect the implicitly communicated affect, especially in *knocking* motions. They segmented motion sequences in order to obtain basic motion primitives and periods with absence of motion. To do so, they computed the motion energy thanks to an objective function and set thresholds (see Fig. 7a). Then, extracted segments of motions were gathered into semantically meaningful clusters (see Fig. 7b). Angles from a 15-joint skeleton were considered; segments were time-normalized and their mean was subtracted to get the relative motion. The normalized segments were finally clustered with the k-means method to represent the motion primitives (see Fig. 7c). Features were afterwards computed. For the example of *knocking* motions, as the right hand and elbow exhibit the movement information, they computed eight statistical measures: maximum distance of hand from body, average hand speed, average hand acceleration, average hand jerk and similar features for the elbow. Individual movement bias, estimated by the average of all the *knocking* motions in the dataset, was removed from motions leading to non-stylised motion sequences. They used a Support Vector Machine (SVM) with a polynomial kernel for classification purposes. They conducted their experiments on the 1200 *knocking* sequences in four *styles* (*neutral*, *happy*, *sad*, *angry*) from the *Body Movement*

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

Library dataset [47]. They chose the leave one subject out cross validation protocol and obtained recognition rates of 50% and 81% respectively when the features were biased (*ie.* the individual bias was not removed) and unbiased.

Etemad and Arya [14] decomposed motion trajectories into basic functions through a linear least square method and modeled secondary features of motions, that they considered as *style*, with three Radial Basis Functions (RBF). Then they aligned the sequences with a Correlation optimized Time Warping (CoTW) method [62]. Finally, they conducted a dimension reduction of the data with a Principal Component Analysis (PCA) that showed that only six principal components were enough to represent 94% of the sequences. For classification purposes, they used RBF neural networks. The training was led using an orthogonal least square technique. They proposed an ensemble of networks to perform classification. Indeed, for each *style*, their system contained one network per DOF and a subsystem, based on majority vote, classified the entire motion sequence. The chosen *style* was determined as having the minimum distance to all classifiers. They used 48 *neutral*, *happy* and *sad walk* sequences from the *HDM05* dataset [39] as well as 75 *neutral*, *young*, *old*, *tired* and *energetic walking* sequences that they recorded themselves. They used a 15 fold cross validation protocol for their data and a 16 fold cross validation protocol for data coming from the *HDM05* dataset [39]. They reached a recognition rate of $93,5\%$ in average.

Crenn *et al.* [4] computed a total of 136 features based on visual cues. Indeed, they first computed 68 low-level descriptors based on the geometry, the motion and the frequency of eight specific body part joints (head, pelvis, elbows, shoulders and hands). Then they computed meta features on those low-level features (*ie.* mean and standard deviations for each feature, abstracting the time). All the features were scaled within the range $[-1,+1]$. They used a SVM with a RBF kernel. They tested their method on three datasets: *UCLIC* [49], the *Body Movement Library* dataset [47] and the data from Xia *et al.* [7] for which they respectively obtained recognition rates of 78%, 57% and 93%. They used a 10 fold cross validation protocol.

B. Person identification through their style

Style can be recognized but it can also, in itself, help recognize persons. It emphasizes that *style* is really related to persons as it is mentioned in Section II-A2. Thus, Kviatkovsky *et al.* [41] were interested in recognizing a person based on their style.

They used a joint position representation leading to a 57-dimensional vector. Data were temporally normalized using a combination of Dynamic Time Warping (DTW) and Fourier temporal pyramid, and spatially normalized with respect to the root joint. A dimension reduction was made with PCA. To enhance the discrimination of persons, a linear discriminant analysis was then performed. They used a probabilistic framework based on generative models (see Fig. 8). The classification of the user's identity was made using a Maximum A Posteriori (MAP) classifier. The relationship between *style* and content was represented by an application of a 1-nearest neighbor kernel density estimation. They proved that performing action recognition with a SVM ahead of the user identification was beneficial, as it was shown by the Mahalanobis distance. They conducted their experiments and obtained good results on four action recognition datasets and one person identification dataset: *MSR Action 3D* [40], *MSRC-12* [42], *UCF Kinect* [43], *UTKinect-Action3D* [44] and *Body Login Dataset* [46].

They were the first ones to work on the identification of a person from general actions, *ie.* not only on locomotion.

Only few methods proposed approaches to recognize *style* in motions. Features to describe *style* in motions were per consequent rare and not all the styles mentioned in Section II were actually studied in motion style recognition.

VI. Motion Style Generation

In animation and video games, large datasets of actions and styles are required. Capturing all the possible combinations is a burden for actors, tedious and time-consuming. One way of solving this issue is to generate motions, avoiding the capture of all combinations of actions and *styles*.

We distinguish three types of motion style generation: motion style synthesis, motion style editing and motion style transfer. We define motion style editing as a subpart of motion style synthesis that implies the user intervention. Motion style transfer was defined by Hsu *et al.* [2] as being "the process of transforming an input motion into a new style while preserving its original content".

This section describes these motion style generation methods, that are discussed in Section VII-D.

A. Motion style synthesis

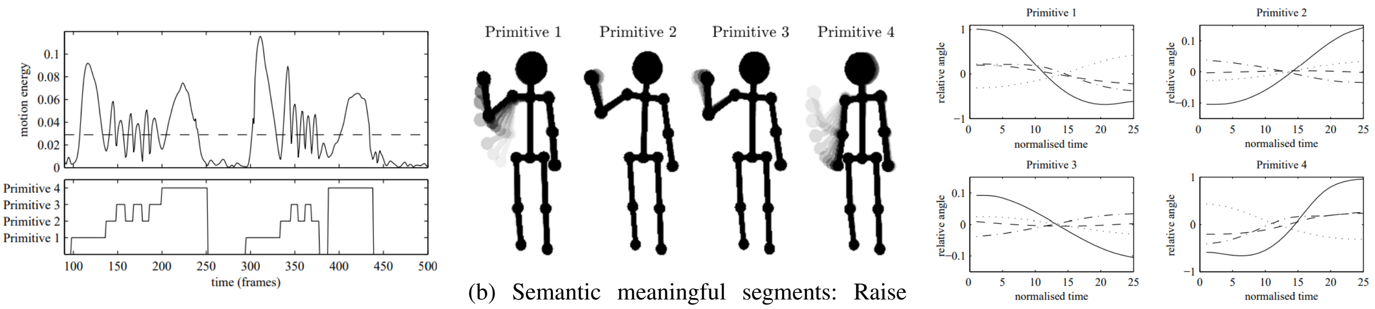
Some of the motion style synthesis methods in the literature made use of style-content separation process to be able to perform the synthesis. Others focused on motion adaptation to environmental constraints (see Fig. 9).

Grochow *et al.* [63] used Inverse Kinematics (IK), represented as a maximization of an objective function describing how desirable a pose is, to interpolate motions between *styles* and perform motion style synthesis. Poses were represented over a probability distribution function which described the likelihood function over poses. This was done using a Scaled Gaussian Process Latent Variable Model (SGPLVM). Each pose was represented by a 42 dimensional vector (joint angles in addition to position and orientation of the root of the kinematic chain where the root orientation was represented as quaternions and joint angles as exponential maps). The SGPLVM learning required the definition of a kernel function that measured the similarity between two points in the input space and the algorithm selected a subset of original poses to keep. Pose synthesis was led through the optimization of the objective function derived from SGPLVM.

Urtasun *et al.* [12] segmented their data and represented them as vectors of 78 angular DOFs. They proceeded to a PCA decomposition of motions. When a new motion came, it was projected into the PCA space and they computed the

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing



(a) Motion energy.

(b) Semantic meaningful segments: Raise arm, Knock, Retract, Lower arm.

(c) Motion primitives.

Fig. 7: Basic motion primitives extraction example for a knocking motion. Reproduced from Bernhardt and Robinson [24].

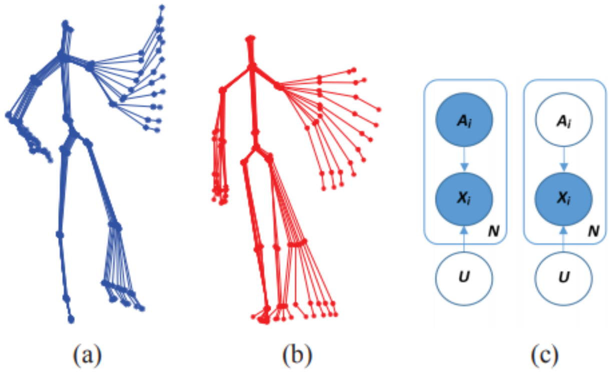


Fig. 8: Male (a) and Female (b) users performing a hand wave gesture. (c) Generative models. Image reproduced from [41].

Mahalanobis distance to the motions in their dataset. New motions could be created by extrapolation, as being a weighted average of motions with target characteristics. Given one single example, they could modify the speed, length or body size while preserving the *style* of an actor.

Tilmanne *et al.* [3] used a variation of Hidden Markov Model (HMM) called Hidden Semi-Markov Model (HSMM) that took state duration modeling into account. Neutral walk sequences were used to train an average model which was then used for automatic adaptation to a particular style. Style adaptation training was performed with constrained structural Maximum A Posteriori (MAP) linear regression transformation followed by a MAP adaptation that further transformed the models. Synthesis was performed via an algorithm which directly generated the optimal parameter sequence. They represented their data by 54 values corresponding to 3D angles (exponential maps) of 18 joints.

Early work from Tenenbaum and Freeman [64] separated *style* from content in the context of speech recognition and it inspired researchers to apply it to human body motions style analysis. Thus, Brand and Hertzmann [13] separated structure and *style* out of dance motions via a state-space representation given by a HMM to which they added a style variable. That variable could be used to vary the HMM parameters, leading to a stylistic HMM called *style machine*. The separation between structure and *style* was made through entropy minimization. They represented their motion capture data with an arrangement of 20 markers and only considered joint angles. A PCA showed that only three stylistic DOFs explained 93% of the variations between 10 models: global pose, a DOF controlling balance and gender and a DOF that could be characterized as the amount of swagger and energy in the motion. Then, they made interpolations and extrapolations within the space of HMM to make new styles.

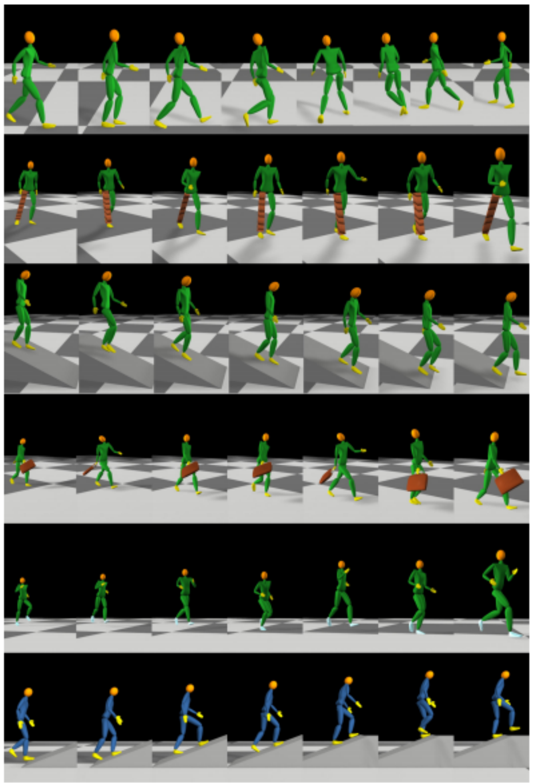


Fig. 9: Examples of synthesized motions in various walking and running styles. From top to bottom: 180-degree walking turn, limp walk, descending an incline, walking with a suitcase, running with springy shoes, ascending an incline. Image reproduced from [5].

It also inspired Wang *et al.* [23] to apply style-content separation to body motions with a probabilistic latent

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

variable model. Indeed, they modeled each pose from motion sequences as a combination of 3D vectors standing for three independent factors: the identity of the subject performing the motion denoted s, the gait of locomotion denoted g and the current state of the motion sequence denoted x. The locomotion included *walk*, *run* and *stride* motions. They considered s and g as being the *style* of the motion and x as being the content. Each pose was represented by a 89 dimensional feature vector (43 angular DOFs, velocities, transitional velocity). Joints were represented as Euler angles except for the joints with three DOFs and the global orientation which were represented as exponential maps. They applied Gaussian Process Latent Variable Model (GPLVM) with a RBF kernel to properly separate *style* from content and then generate new stylistic motions.

A few approaches focused on motion adaptation and studied how to conserve *style* and generate new stylistic motions in new environments or when constrained.

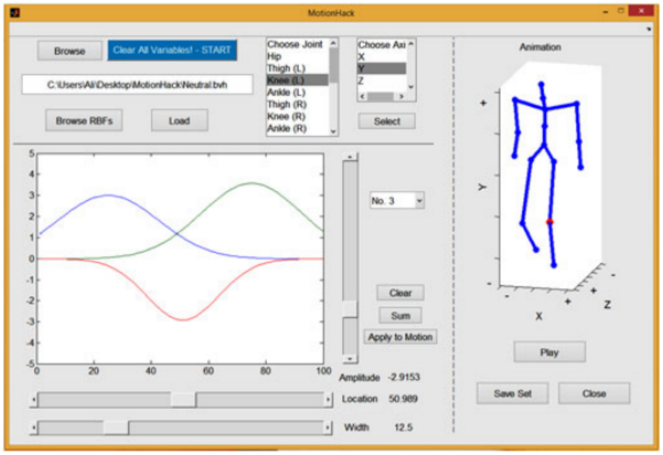
Liu *et al.* [5] represented their data with 29 joint DOFs and 6 root DOFs (rotational joints represented by exponential maps). They considered style parameters based on biomechanical studies and hypotheses on human locomotion. Thus, a person's *style* was described by the simplified set of musculoskeletal parameters and muscle preferences. These parameters were automatically estimated by nonlinear inverse optimization. A parameter of 147 dimensions represented the *style* of an actor. Motion synthesis was conducted by minimizing the muscles usage; constraints on footsteps were imposed. Then, motions of the same *style* but with actors performing different tasks, with news constraints, were generated (see Fig. 9).

Chien and Liu [18] worked on stylistically similar motion pairs. Each motion sequence was segmented into a set of limb motion segments (based on the interaction of its end effector with the environment), that produced whole body motion segments when fused. Motion data were represented by six rigid DOFs for the root joint and 60 principal components for each limb that touched the environment. Features were extracted providing a 30 dimensional feature vector for the limb segments and a 108 dimensional feature vector for whole body segments. The stylistic similarity of a new motion to an example motion was represented by constraints and objective functions. A similarity region was approximated by one class SVM and Kernel PCA.

B. Motion style editing

We define motion style editing as a subpart of motion style synthesis that implies the user intervention in the synthesis process, usually by manipulating a style parameter (see Fig. 10).

Shapiro *et al.* [65] combined, frame by frame, two motions together to automatically decompose them into style components that represented the *style* and expressiveness of a motion. Usually three to five style components were enough to represent 95% of the differences in motion data. The decomposition was performed via the unsupervised learning technique Independent Component Analysis (ICA). The user then selected the components she/he found the most interesting. Style components that highlighted the differences in posture, cadence or any other nuance that appeared in one motion but not the other should be preferred. The selected style components were then merged before being transferred in order to create a new motion. The amount of style components could be interpolated. Motion style generation was done with an alignment step DTW on one of the DOF of the character (skeleton) that the user selected according to her/his desired motion. They led their experiments on two kinds of skeleton hierarchies: either 31 joints with 62 DOFs or 26 joints with 84 DOFs represented by Euclidean coordinates.



C:\Users\Ali\Desktop\MotionHack\Neutral.bvh

Choose Joint: Knee (L)

Choose Axi: Y

No. 3

Amplitude -2.9153

Location 50.989

Width 12.5

Fig. 10: GUI developed and used to generate the features. Image reproduced from [9].

Wang *et al.* [27] extended the HMM and created a HMM/Mix-SDTG model, designed to learn a full-body motion under the control of a style variable. They trained a HMM with four hidden states on two walk sequences. Each output density contained a mixture of components Stylized Decomposable Triangular Graphs (SDTG). Each dimension of the style variable had a meaning for the user who could give it an arbitrary value to designate the desired *style* of the new motion. The motion generation was made in two steps. First, a path connecting two specified hidden states of the HMM was selected to maximize the transition probabilities along the path. Then the mean vector of the output densities along the path were calculated with a given style value and an interpolation was made between the samples. They transformed the positions into 3D rotations parameterized as exponential maps leading to a 120 dimension pose space formed by the global position, global velocity, joint rotations and angular velocities of joints.

Ikemoto *et al.* [6] developed an interactive learning method and controllable system that propagated the edits made by an animator. An artist selected an existing animation clip and edited it. The system automatically generalized and propagated those edits to a new motion clip. The artists could make additional edits, refining the system by providing feedbacks. The process was reiterated until the result was satisfying enough. The system thus incrementally built a model of edits. Their method merged estimations made by a Gaussian process regression. A set of features represented the dynamics and

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

kinematics of both source and target skeletons as they used four types of features to represent the source and target kinematics and acceleration. The method could generalize to new characters (retargeting), even with different morphologies.

Ma *et al.* [22] had the user specify control parameters (for example, stride length) and a hierarchical model was built. The skeletons were separated into four joint groups (legs, left arm, right arm, rest upper body). For each joint group, a latent variation parameter was automatically selected. A Bayesian network was constructed, automatically learnt from example motions, in order to describe the relationship between user-defined style parameters and the latent variation parameters. Motion synthesis was conducted via a Gaussian process regression (Kriging model) that amounted to a *style* and variation interpolation to generate partial motions of all group joints that were assembled to synthesize whole body motions. Thus, new motions were generated by the user who interactively controlled the style parameters. They preprocessed their data. Indeed, their example motions were normalized: each pose contained the global 3D position and orientation of the root node and the rotations of all the other joint nodes relative to their parent joint nodes. Note that all rotations were represented by quaternions. They also segmented the motions into short clips, by an automatic key frame extraction according to two body parts: lower body (legs) and upper body (rest of the body). They also established a correspondence with the key frames, both spatially, performing rotations, and temporally, with Iterative Time Warping (ITW).

Min *et al.* [25] modeled identity and *style*. They applied dimension reduction (PCA) to their data obtaining a 77 dimensional vector which could represent 99% of the geometric variations in motions. They established a temporal correspondence with warping functions. They annotated their data with environmental constraints and important key events. They applied multilinear analysis techniques to motions and constructed a generative model that interpreted *style* and identity variations thus providing parameters. They could synthesize new motions, retarget motions from one actor to another and edit stylistic motions. Editing was made by adjusting the style and identity parameters.

Etemad and Arya [9] constructed secondary features with Gaussian RBFs. Through an interface, animators provided up to three Gaussian RBFs per DOF in cartesian space that converted a neutral base motion into several stylistic variations (see Fig. 10). The RBFs edits were collected and summarized to create a feature set which was applied to neutral walking sequences. Their data were represented by 54 DOFs. They temporally aligned sequences using Correlation optimized Time-Warping (CoTW) [62].

Holden *et al.* [15] synthesized motions based on high level parameters (trajectory of the character over the terrain and the movement of the end effectors). They learnt a motion manifold training a convolutional autoencoder where the motion manifold was represented by hidden units of the autoencoder. They mapped high level parameters to motion manifold (*ie.* hidden units) with a deep forward neural network stacked on top of the trained autoencoder, that trained on only some of the data that were the most relevant to the task. A user could specify constraints (*eg.* drawing a curve over the terrain so that the character could walk along it) that led to the generation of a new motion. It could then be edited in the space of hidden units; motion editing was represented as a minimization problem. Constraints, represented as costs, were applied in hidden space: positional constraints (to fix foot sliding artifacts), bone length constraints (to preserve the rigidity of the body) and trajectory constraints. The costs were defined by two terms, relative to the content and the *style* of the output motion. They used huge amount of data coming from several sources. They retargeted them all to match to a unique skeleton represented by 3D positions.

C. Motion style transfer

Motion style transfer, or motion style translation, can be defined as the process of transforming an input motion into a new *style* while preserving its original content [2]. Fig. 11 presents an example.

Amaya *et al.* [1] decomposed their motion sequences into basic periods. They computed emotional transforms thanks to signal processing techniques. Those transforms established motion differences with respect to speed and spatial amplitude components. The speed component was calculated via a nonlinear time warping technique. The spatial amplitude component was calculated via signal amplifying methods.

Hsu *et al.* [2] used Iterative Time Warping (ITW), inspired by the motion warping of Witkin and Popovic [66], to compute a dense correspondence between motions applying spatial and temporal warps. They then described the relationship between input and output styles using a linear time invariant model. Style transfer was then applied with simple linear transformations.

Torresani *et al.* [17] based their method on the Laban Movement Analysis (LMA) and more specifically on the LMA-Effort dimension and its factors Flow, Weight and Time. Similar action fragments were matched with DTW. They referred to fragments when each motion sequence was manually segmented by a LMA human expert. They focused on the rotation DOFs, considering joint angles as exponential maps. A space-time interpolation was learnt with a SVM and a Gaussian RBF kernel. Style transfer was done by applying DTW to fragments, computing a distance between the motion blends and concatenating the best approximated fragments.

Abdul-Massih *et al.* [28] decided to manually split spatially the skeletons into groups of body parts. They extracted positional and angular amplitude features. Positional features were represented by joint-specific relative paths on a source character S and were used to transfer the motion from S to a target character T. Angular amplitude features were extracted between stylistic and neutral source motions. They were used to scale and offset angles in the motion of T. Those features created constraints that were taken into account during the transfer for which a space-time approach was used.

Xia *et al.* [7] developed an online learning algorithm which aimed at approximating the spatio-temporal transformations in frames. This was done by constructing a series of local regression models. The current pose was then translated

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

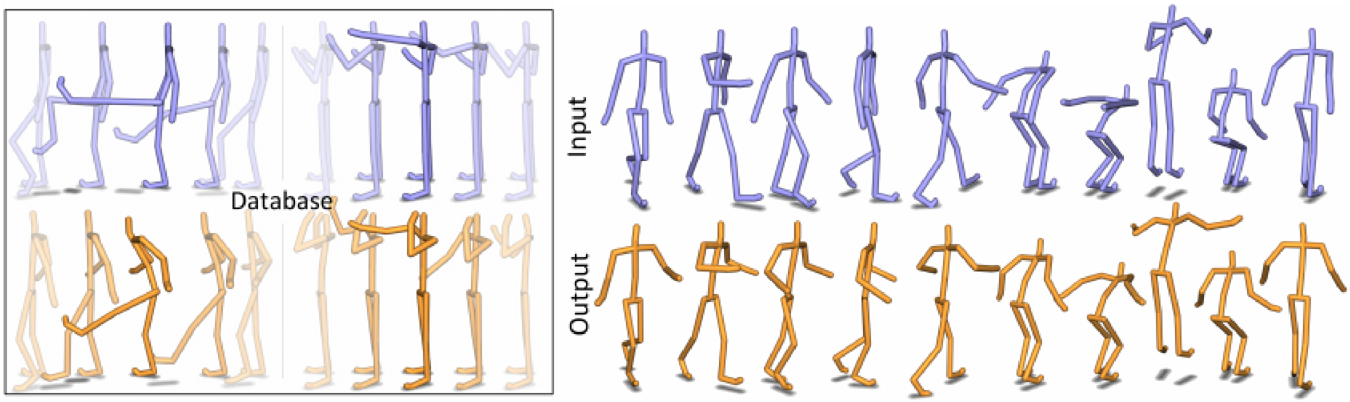


Fig. 11: A neutral heterogeneous walk ⇒ jump motion (top-right) stylized as childlike (bottom-right) using a dataset where only kick and punch actions are available (left). Image reproduced from [8].

through linear transformations. A timing prediction was done by a k-Nearest Neighbor (KNN) algorithm with a gaussian interpolation kernel. Unlabeled data were handled with a probabilistic framework mixture of local autoregressive models. They worked on heterogeneous data and converted joint angles to Cartesian parameters with an exponential map parameterization. Moreover, they needed to have their data annotated with footplant contact information, in order for their artifact correction classifier to work.

Yumer and Mitra [8] proposed a spectral approach. In a pre-processing step, in order to help differentiate actions, they computed a skeletal representation. They used the translation and rotation DOFs of all the joints except the root joint. Real-time transfer was made by applying a discrete Fourier transform. A sliding window filter was applied in time domain to handle heterogeneous data.

Etemad and Arya [14] used RBF Neural Networks (RBFNNs) to transfer *style*. They used the features that were extracted with RBFs and the warped motion data with CoTW [62] to generate a set of 5 to 10 RBFNNs for each *style*.

Holden *et al.* [16] also proposed a neural approach for motion style transfer. They focused on joint positions in 3D Euclidean space as well as the turning velocity, the forward velocity and the sideways velocity. Their pipeline was composed of two networks. A transformation network, performing the style transformation, was a feed forward convolutional autoencoder. A loss network aimed at computing the loss between motion content and *style* and correcting artifacts except foot sliding output motions. Errors calculated in hidden units were back propagated into the feedforward network.

Methods for motion style generation (synthesis, editing and transfer) have been exposed. Beside their main common goal, *ie.* motion style generation to enhance datasets and save animators and actors time, they have similarities and divergences that are discussed in Section VII.

VII. Discussion on Applications of Style in Motion

This section highlights interesting facts resulting from previous sections. We first discuss matters that arise when it comes to the definition of *style* in human body motion and what it implies. We then focus on data that are used in motion style approaches. An analysis of the methods in motion style recognition precedes another analysis of the approaches in motion style generation, where common points and divergences of the methods are discussed.

A. Motion style analysis

The taxonomy of definitions of *style* in Section II arises several questions about *style* in motion and how it is defined and seen.

1) Number of styles per motion: For the sake of simplicity, existing approaches usually focused on one *style* in motion sequences. For example, combinations of *styles* such as *feminine-sad* were discarded in the method of Etemad and Arya [9]. It was not always limited to one *style* per motion sequence though. Wang *et al.* [27] generated motions with the combinations *normal-masculine* and *catwalk-feminine*. Xia *et al.* [7] interpolated a motion between up to three *styles*.

2) Neutrality in motion style: Neutral style was also named *no style*, *normal* or even *regular*. Out of 15 paper referenced in Section II-A3, 11 considered this *style*. However, rare methods questioned the *neutral* aspect of a motion. Even though, while recording a *neutral* sequence motion, actors were asked to display minimum *style* in their motion, an absolute *neutral* motion was impossible to achieve: actors brought their own personal bias in the recorded motions. Etemad and Arya [9] proposed to average several neutral walking sequences in order to minimize if not eliminate that bias.

3) Influence of culture: Etemad and Arya [9] highlighted that *style* could be influenced by several factors, including the culture. Besides, while basic emotions can be considered as universal in facial expressions, studies showed that when body emotional expressions were taken into account, culture involved divergences in their expression and perception, especially between Americans and Japanese [67], [68]. This

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

emphasized the lack of consensus on motion style, even when focusing on the subset of emotions. Indeed, for instance, an American and a Japanese have divergent ideas on what a *happy* motion is. Furthermore, depending on the culture, some stereotypes could be reinforced: how would you define a *sexy* or *manly* motion? It could imply a normative view on gender, which is also suggested by the *feminine* and *masculine styles*. Besides, Wang *et al.* [27] had an actress perform the *catwalk* style, while a male actor performed the *neutral* walk.

4) Expressiveness of body versus face: Face, speech, body movements, etc. are means of verbal and non-verbal communication. More importance is given to facial expressions in affective computing, where it has been studied for several decades already. It can moreover benefit from knowledge in intuitive adjacent domains such as biology, as a face contains a lot of muscles, or theater plays, where actors put on masks or exaggerate their expressions to convey as explicitly as possible an emotion. Still, stylistic motions can sometimes convey more information of affective expressions than the face [69]–[71]. As an example, De Gelder [69] highlighted that in a situation of fear, while the face indicated that there was something scary, the body position indicated what was scary and how the individual intended to deal with it. Ekman and Friesen [72] even suggested that the body was more reliable than the face to convey affective expressions. They concluded that it was easier to deceive people with the face than the body. It emphasized that studying the *style* of human body motion is proved to be as relevant as studying facial expressions.

B. Style data analysis

Section III presents the data used in motion style approaches. It points out the following elements for which a particular attention should be paid to.

1) Need for data: Style datasets that have been presented in Section III-B were designed either for emotion recognition (*Body Movement Library* [47] and *UCLIC* [49]), for motion style synthesis (*eNTERFACE'08 3D* [48] and *Mockey* [48]) or for person identification (*Body Login Dataset* [46]). As a result, no publicly available motion style dataset has been created yet to perform motion style recognition specifically, and, for motion style generation purposes, only two datasets were recorded to perform motion style synthesis.

Furthermore, Table III highlights that none of the action as well as style datasets has been used to perform both style recognition, person identification included, and motion style generation (synthesis, editing and transfer). Data from Xia *et al.* [7] paved the way though. Recorded to perform motion style transfer, their data have been used for motion style recognition and motion style generation. Furthermore, their data were the only ones to consider different categories of *style* as well as various heterogeneous actions.

Besides, deep learning techniques arise. A lot of data need to be considered, such as in the method of Holden *et al.* [15]. Moreover, Etemad and Arya [14] pointed out that data used for neural networks needed to be very consistent. Such data are, to the best of our knowledge, missing.

The lack of consistent, publicly available and specifically recorded datasets to perform motion style recognition as well as motion style generation is consequently obvious. It was more striking when *style* was considered as being individual-related features than when *style* was considered as being variations in a motion, where action datasets could suffice. It induces a difficulty in the comparison of results from the different methods that tackled motion style analysis.

2) Motion capture technologies: Data sequences were mostly recorded with motion capture (mocap) systems. Most of the systems used in the mentioned papers were optical infrared systems with markers: Impulse, Vicon, and Falcon Analog. The Microsoft Kinect camera is a markerless optical solution.

Motion capture technologies are affordable now and enable users to have readily realistic data [1], [3], [12], [65]. However, they require specific equipment. They are constrained to a limited space [1], [3], impairing the naturalness of actors'motions. IGS-190 is an exception as it relies on inertial sensors on a suit and has no space constraints. Motion capture technologies also provide highly dimensional data that are usually reduced in the process [3].

3) Motion types: Most of the used datasets were composed of homogeneous motions (*ie.* one kind of motion per sequence – eg. walking, running, etc. – was considered). Data from Xia *et al.* [7] contained heterogeneous motions (*ie.* several kinds of motions were considered per sequence) and it has led to a new kind of motion study. *CMU* dataset also included heterogeneous motions.

Besides, the used datasets were made of actions of different kinds: locomotion like *walking*, *running*, *jogging*, etc. but also dynamic gestures like *jumping*, *kicking*, *punching*, etc. or even specific actions like *draw X*. The most frequent action was *walking*, especially in stylistic data, then running. This makes sense knowing that locomotion is central to human movement [5].

4) Acted and naturalistic motions: Recorded data can be acted or non-acted (*ie.* naturalistic). Data are "acted" if the person performing the motion expressed the motion purposely. When a motion is expressed naturally, data are "non-acted". The style specific datasets presented in Section III-B all used acted data. The "user-defined gesture" of the *Body Login Dataset* [46] could be considered as non-acted though (the "S-gesture" making the dataset acted), as each user could have their own gesture. It is hard to record naturalistic data. Acted data have the advantage of being clearly defined; however, they introduce a bias.

The actors also play a determinant role in the naturalness of the recorded data. Actors can be professional or not. In the datasets presented in Section III-B, most of the actors were non professional. The *Mockey* dataset [48] recording requested professional actors to perform motions. Motions were exaggerated there. Even though the difference in motion is more visible (acted motions), the naturalness of the motions is impaired. One could expect non professional actors to provide more naturalistic motions than professional actors. However, the same observation could be made for the *UCLIC* dataset [49]: even though actors were non professional, emotions were exaggeratedly portrayed (see Fig. 6). This emphasizes that

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

recording naturalistic data is not easy, whether motions are performed by professional or non professional actors.

C. Analysis of motion style recognition

A few approaches tackled the motion style recognition challenge. It poses two main questions: what styles can be recognized and which features are extracted?

1) Styles to choose to perform recognition: Researchers mentioned that sometimes two styles could be very similar. Indeed, some styles can be mistaken with others, even by human beings. For example, the difference between an *old walk* and *an injured walk* is not always straightforward. Etemad and Arya [14] highlighted that the confusion rates for *happy*, *energetic* and *young styles* were quite high, as well as for *sad*, *tired* and *old styles*. They decided, for this reason, to evaluate their methods on opposite *styles*, by evaluating *happy* and *sad*, separately from *young* and *old* as well as *energetic* and *tired*.

Furthermore, when recording their dataset, Ma *et al.* [22] emphasized the fact that emotions such as *surprise* and *disgust* could hardly be studied in motions as they were immediate and usually short reactions to an event and resulted in very specific movements. As a result, sets of *styles* should be carefully chosen when it comes to perform motion style recognition.

2) Features to represent a style: Methods presented in Section V highlighted that many kind of features have been studied. Bernhardt and Robinson [24] chose them accordingly to the very specific motion (*knocking*) they were studying. Moreover, they removed what they considered as being an individual bias, considered by Crenn *et al.* [4] as being part of the *styles* to classify. Crenn *et al.* [4] computed quite a lot of different geometry, frequency and motion-based descriptors. Etemad and Arya [14] let RBFs and PCA estimate the best descriptors. In another method, Etemad and Arya [9], [21] mentioned that *style*, in the field of biology, was composed of two general types of spatio-temporal features: posture (no change throughout time) and movement (dynamics - variation throughout the motion).

Kleinsmith and Bianchi-Berthouze [68] gathered information about the features of affective states or dimensions from body expressions studied in the literature. Discriminative features have thus actually been identified. However, most of the *styles* reported in their survey belonged to the emotion subcategory of *style*. Etemad and Arya [9] worked on six *styles* (*happy*, *sad*, *energetic*, *tired*, *feminine* and *masculine*) and identified common features for each of them. In both studies, those features relied heavily on visual cues. For example, a happy person would tend to perform movements with high amplitudes, while a sad person would be more prone to lower her/his arms, etc.

Identifying descriptors of motions styles thus remains a field to investigate more deeply, as no agreement has yet been made on non visual features describing motion style. Moreover, the features that were extracted to recognize motion style are not the same as the ones that characterized motion style when it came to motion style generation approaches, as suggested by Etemad and Arya [9] and as it can be seen in Section VI.

D. Analysis of stylistic motion generation methods

Section VI presents methods on motion style generation, either by performing motion style synthesis, editing or transfer. Common characteristics that appear for those three types of motion style generation are presented, as well as characteristics specific of methods of one motion style generation type.

1) Trend of style definition: Most of the approaches presented in this paper opted for the trend exposed in Section II-A3, namely considering *style* as individual-related features characterized by adjectives. There was indeed a quasi unanimity in motion style transfer approaches, as well as in motion style recognition approaches.

2) Motion types: Most of the approaches studied *walking* motion sequences, as stated previously (Section VII-B3), but also *running* sequences and *kicking* or *punching* sequences. Some of these approaches stated that their method, applied to *walking* sequences, should theoretically apply to *running* [25], *boxing* [22] or *kicking* [22] but that it would not apply to dance motions as they are too complex. A few approaches though, managed to deal with dance motions [13], [17]. The same observation could be made with heterogeneous data as only a few recent approaches decided to deal with them [7], [8]. Data that have been used up until now were thus mostly related to human locomotion.

3) Data representation: In methods performing motion style synthesis, most of the data were represented by joint angles [12], [13]. Some were represented by Euler angles [23] but exponential maps were undoubtedly more used [3], [5], [23], [63].

In methods performing motion style editing, data representations were more various: DOFs were represented either in Cartesian space [9], in Euclidean space [65] or by exponential maps [6], [27].

As for motion style transfer approaches, data representations were also various. When joint angles were considered, they were represented as exponential maps [7], [17]. Positions in euclidean space were also considered [16].

However, even if common characteristics appeared such as the use of exponential maps to represent angular DOFs, the number of DOFs used was never the same, even though data were usually skeletons that were recorded with motion capture techniques and thus had a similar skeleton structure.

4) Motion decomposition: Motion sequences needed to be segmented only when style transfer was applied. Indeed, no approach studying motion style editing decomposed its data. Among motion style synthesis approaches, only Chien and Liu [18] decomposed spatially their data. On the contrary, motion decomposition was much more frequent in style transfer studies. It could be spatially [28] but it was mostly temporally [1], [8], [17].

5) Establishment of the difference between two motions: Motion style transfer approaches were the only ones that did so. Indeed, transferring motion style implied most of the time to establish a representation of the difference between input and output *styles*. It could be the difference between a *neutral* and stylistic motions such as in the approaches of Amaya *et al.* [1] – and their emotional transforms, Abdul-Massih *et al.* [28] or Yumer and Mitra [8] – in the frequency domain, or even

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

Etemad and Arya [14] with neural networks. It could also be the difference between two stylistic motions such as for Xia *et al.* [7], where the relationship between input and output *styles* was explained by a time-varying autoregressive model whose parameters were estimated via a linear least square regression, or for Hsu *et al.* [2].

6) Motion alignment: In motion style synthesis approaches, Tilmanne *et al.* [3] aligned the directions of all their sequences.

In motion style editing approaches, alignment has been a key step, especially temporal alignment [25] with DTW [65] and its variants ITW [22] or CoTW [9]. Some approaches also used spatial alignment [22]. Moreover, in the case of methods relying on neural networks, the amount of data that was used is huge, which implied a normalization step [15], [22].

Motion alignment has also been widely used in motion style transfer studies. Abdul-Massih *et al.* [28] decided to achieve it manually when groups of body parts were set by users. DTW [17] or its variants – ITW [2], CoTW [14] – was often used. Xia *et al.* [7] opted for a KNN search they applied to the dataset to find motions that were close to the current input frame thanks to a distance computation. If the input sequence had a different structure than the sequences in the dataset, a retargeting step was applied.

This step has been central to the motion generation pipelines. Holden *et al.* [16] highlighted that it was difficult for artists to intervene in post processing step, so pre-alignment should be preferred. Other motion alignments exist such as Canonical Time Warping (CTW) [73] and Generalized Time Warping (GTW) [74].

7) Dimension reduction: Lots of motion style synthesis approaches used a dimension reduction algorithm [12], [13], [18]. Among motion style editing methods, only Min *et al.* [25] used PCA. Shapiro *et al.* [65] did a similar dimension reduction, as they stated that only a few of their style components could cover for 99% of the geometric variations of the data. No approaches from motion style transfer methods applied dimension reduction.

This steps seems rational as motion capture data, that are used here, are highly dimensional [3]. This is due to the motion capture technologies but also to the skeleton structure [41] and the fact that human motions are repetitive [6], [65].

8) Learning methods: In Section VI, methods for motion style generation are gathered by subcategories of motion style generation (synthesis, editing and transfer). These methods can also be gathered according to their common points.

Some of the methods used unsupervised learning techniques with a HMM [3], [13], [27] or the ICA [65]. Others used supervised learning with a SVM [17] or a KNN search [7]. Statistical models, namely Gaussian processes, have been quite investigated: Grochow *et al.* [63] used a SGPLVM, Wang *et al.* [23] used a GPLVM, both Ikemoto *et al.* [6] and Ma *et al.* [22] used a Gaussian process regression and Etemad *et al.* [9] used Gaussian RBFs. Neural networks have been used as well, mostly in the last few years: Ma *et al.* [22] used Bayesian networks, Etemad and Arya [14] used neural networks based on RBFs and Holden *et al.* [15] used a deep forward neural network combined with a convolutional autoencoder. Note that in action recognition, Long Short Term Memory (LSTM) networks are widely used [75], whereas they were not used at all in the methods on 3D human body motion style presented in this paper.

9) Post-processing step to clean up foot sliding artifacts: In computer animation, some visual artifacts, such as the foot sliding or foot skating artifact, can appear. A foot sliding artifact appears when the contact between the feet and the floor is not correctly enforced [76].

No researcher used a post-processing step for cleaning foot sliding artifacts in motion style synthesis approaches.

On the opposite, motion style editing approaches dealt with it. Only Shapiro *et al.* [65] explicitly mentioned that they required a post processing step to clean up the foot skating artifacts: the global translation DOF, removed before the ICA decomposition, was added again to synthesized motion. This is nonetheless a persistent problematic. Ikemoto *et al.* [6] did not directly handle environmental contacts like footstrikes as they cleaned them up with an automated postprocessing step but stated that it was a lead for future work. Ma *et al.* [22] created transitions if long motion sequences were wanted and eliminated foot skates. Min *et al.* [25] took care of foot sliding artifacts by enforcing the environmental constraints (foot contact) with an inverse kinematics process. Holden *et al.* [15] detected whether the toe or heel of the skeleton went beyond a certain height or velocity thus getting foot contact labels. Then they added positional constraints during editing to act against artifacts.

Motion style transfer approaches also considered foot sliding artifacts. Classifiers were used to correct them. Hsu *et al.* [2] built a footplant classifier with a multi discriminant analysis and univariate Gaussians modelisation. An heuristic generalisation, blending the input to the transferred motion, gave priority to the content over *style*, in case a conflict appeared. Xia *et al.* [7] developed a KNN classification algorithm to automatically process footplant artifacts: they annotated in a pre-processing step contact information and detected them with the KNN; noise reduction was applied with Gaussian filters. Yumer and Mitra [8] also proceeded to a foot-plant nearest neighbor search before cleaning the artifacts in a post processing step. Holden *et al.* [16] did not do any artifact correction, but they mentioned that it should have been done.

10) Bone length: In motion style editing approaches, in addition to the foot sliding artifact, a particular attention should be paid to the bone lengths. Etemad and Arya [9], who used cartesian data representation, stated that the bone lengths were not necessarily preserved after edits were applied. Shapiro *et al.* [65] highlighted that the change of limb length impacted foot plants and also created occasional foot skating or violation of floor constraints, which was why if the data were represented by marker positions instead of joint angles, they applied a filter to restore the correct limb lengths (and a low-pass filtering was automatically done to eliminate high-frequency motions). Holden *et al.* [15] applied bone length constraints.

11) Evaluation metrics: The most frequent way to evaluate motion style generation was to proceed by conducting a visual evaluation, for example presenting the synthesized motion and the original one. In motion style synthesis approaches, that was

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

the most common evaluation mean [3], [5], [12], [18]. This was not the only one though: Urtasun *et al.* [12] also measured statistical values (interpolation error, intra-variability and inter-variability, that are specific to their data).

In motion style editing approaches, evaluations were conducted by the perception of the user. Indeed, the specificity of motion style editing is that it implies a visual tool [9], [25], [65]. This leads to the absence of other mean of evaluation than the perception of the user [6], [9], as these methods are based on perceptual approaches and opinions of users (animators). Still, Ma *et al.* [22] conducted a leave-one-out cross validation between the validation clips and the predicted motion clips. Min *et al.* [25] evaluated their motion synthesis process using cross validation techniques to compare synthesized motion with ground truth motion data. Etemad and Arya [9] conducted a survey.

In motion style transfer approaches, some researchers used vision and perception to assess the correctness of their method [14], [28]. A few methods used measures. Thus Xia *et al.* [7] measured temporal and pose errors. Most of the approaches actually proposed user surveys [14] to assess the naturalness [28] or realism [8] of the generated motions.

Broadly speaking, most studies lacked qualitative assessments.

12) Applications and further investigations: Motion style generation became a key to save time during motion capture sessions. Moreover, it aims at helping animators by automating their work. Focusing on *style* appears crucial in the generation of crowds, as it brings realism. However, only motion style editing methods using neural networks explored this application [15], [22]. Brand and Hertzmann [13] talked about a potential use of their method for a cast of thousands though, perhaps implying that their technique could be used for crowd simulation. This application is thus to be further investigated.

A common trend is to study other *style* features. Indeed, Amaya *et al.* [1] highlighted that it could be desirable to work on personalities, culture and gender *ie.* other subsets of *style*. Torresani *et al.* [17] supported the idea of learning person-specific styles. Kviatkovsky *et al.* [41] also suggested that gender classification could be investigated.

Amaya *et al.* [1] suggested to transfer a human emotional motion to an animal. Abdul-Massih *et al.* [28] explored this path and worked on style transfer applied to morphologically different characters: human, T-Rex, dragon, three-headed creature and snake. Note that Wampler [77] focused only on animals.

VIII. Conclusion

This paper exposes approaches on the study of *style* in 3D human body motion based on skeletal data. It includes style motion analysis, recognition and generation. *Style* is critical as it brings realism and expressiveness to different applications such as character animation.

The first key element is to set a definition of motion style in 3D human body motion, where it lacks consensus. Each author of the approaches that have been presented in this paper had their own definition and worked on their own *styles*. A taxonomy of definitions of *style* highlights that human body motion *style* is sometimes seen as variations in motions, or as a component of a motion, or as individual-related features. Among the approaches presented in this paper, the most frequent trend was to consider *style* as individual-related features. Even when opting for this trend, the features defined with adjectives are diverse. Subcategories are identified: emotions, biological features (age and gender), physical states, personality features and behaviors. The most studied *styles* were *angry*, *happy*, *sad*, *old*, *proud*, *depressed* and *tired*; most of them being emotions. Furthermore, $^2/_3$ of the referenced papers worked with the *neutral style*, even though it is difficult to represent it, in particular as the actors add an individual bias.

A second key point lies in the data that are used in motion style approaches. The lack of publicly available datasets on *stylistic* motions is highlighted, as most of the approaches either used action and mostly emotional datasets or recorded their own data. This makes any comparison between approaches difficult. Action datasets were mostly used when *style* in human body motion was seen as being variations in a motion, or when *neutral* sequence motions were required. When *style* was seen as individual-related features, only two publicly available datasets (*eNTERFACE'08 3D* [45] and *Mockey* [48]) truly focused on *style* with several of its subcategories (other datasets focus on emotions), explaining why many researchers decided to record their own data.

Only a few methods tackled 3D human body motion style recognition. Several approaches have already been conducted on emotions though. Finding features that describe *style* without relying on visual cues (and hence being specific to the *styles* that are studied) remains a challenge. Person identification through their *style* is one of the applications of *style* in human body motion. It highlights that it is related to persons.

Stylistic motion generation has been more broadly investigated. We distinguish three types of motion style generation: motion style synthesis, motion style editing and motion style transfer. These types of motion style generation have common characteristics and challenges, such as the way to perform a relevant evaluation of the generated stylistic motions as there is a lack of qualitative assessment. They also have specific characteristics or steps, as the establishment of the difference between two motions that has only been performed in motion style transfer. *Stylistic* motion generation is a solution to the lack of data and it can save time for actors when recording datasets as well as for 3D animators when creating a character animation. Animators are not that much involved in the process though: only motion style editing approaches inquired their opinion.

Some challenges have been highlighted in this survey. There is definitely a need for data, especially databases that would relate to the different trends of the definition of *style* presented in Section II. The *neutral* style also needs to be clarified. Some methods already paved the way to work on stylistic motion generation. More investigations would need to be done in body motion style recognition though, especially in determining what could be the *style* features. An evaluation metric of the performance of the methods that wouldn't be relying on visual

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

cues should also be investigated. Analysis of body motion *style* and stylistic motion generation could be valuable in character animation. Indeed, it would for instance be helpful to surpass the uncanny valley, or to make improvements in crowd simulation.

References

[1] K. Amaya, A. Bruderlin, and T. Calvert, “Emotion from motion,” in *Graphics Interface*, vol. 96, May 1996, pp. 222–229.

[2] E. Hsu, K. Pulli, and J. Popović, “Style translation for human motion,” *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 1082–1089, July 2005.

[3] J. Tilmanne, A. Moinet, and T. Dutoit, “Stylistic gait synthesis based on hidden Markov models,” *European Association for Signal Processing (EURASIP) Journal on advances in signal processing*, vol. 2012, no. 1, p. 72, 2012.

[4] A. Crenn, R. A. Khan, A. Meyer, and S. Bouakaz, “Body expression recognition from animated 3D skeleton,” in *3D Imaging (IC3D), 2016 International Conference on*. IEEE, 2016, pp. 1–7.

[5] C. K. Liu, A. Hertzmann, and Z. Popović, “Learning physics-based motion style with nonlinear inverse optimization,” *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 1071–1081, 2005.

[6] L. Ikemoto, O. Arikan, and D. Forsyth, “Generalizing motion edits with gaussian processes,” *ACM Transactions on Graphics (TOG)*, vol. 28, no. 1, pp. 1–12, January 2009.

[7] S. Xia, C. Wang, J. Chai, and J. K. Hodgins, “Realtime style transfer for unlabeled heterogeneous human motion,” *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 119–129, 2015.

[8] M. E. Yumer and N. J. Mitra, “Spectral style transfer for human motion between independent actions,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 137–144, 2016.

[9] S. A. Etemad and A. Arya, “Expert-driven perceptual features for modeling style and affect in human motion,” *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 534–545, 2016.

[10] M. Mori, “The uncanny valley,” *Energy*, vol. 7, no. 4, pp. 33–35, 1970.

[11] L. Louis and M. Azaini, “Evaluating the uncanny valley theory based on human attitudes,” *Archives of Design Research*, vol. 28, no. 2, pp. 27–41, 2015.

[12] R. Urtasun, P. Glardon, R. Boulic, D. Thalmann, and P. Fua, “Style-based motion synthesis,” in *Computer Graphics Forum*, vol. 23, no. 4. Blackwell Publishing Ltd., 2004, pp. 799–812.

[13] M. Brand and A. Hertzmann, “Style Machines,” in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., July 2000, pp. 183–192.

[14] S. A. Etemad and A. Arya, “Classification and translation of style and affect in human motion using RBF neural networks,” *Neurocomputing*, vol. 129, pp. 585–595, 2014.

[15] D. Holden, J. Saito, and T. Komura, “A deep learning framework for character motion synthesis and editing,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 138–149, 2016.

[16] D. Holden, I. Habibie, I. Kusajima, and T. Komura, “Fast neural style transfer for motion data,” *IEEE Computer Graphics and Applications*, vol. 37, no. 4, pp. 42–49, 2017.

[17] L. Torresani, P. Hackney, and C. Bregler, “Learning motion style synthesis from perceptual observations,” in *Neural Information Processing Systems (NIPS)*, vol. 6, 2006, pp. 1393–1400.

[18] Y.-R. Chien and J.-S. Liu, “Learning the stylistic similarity between human motions,” *Advances in Visual Computing*, pp. 170–179, 2006.

[19] C. Rose, M. F. Cohen, and B. Bodenheimer, “Verbs and adverbs: Multidimensional motion interpolation,” *IEEE Computer Graphics and Applications*, vol. 18, no. 5, pp. 32–40, 1998.

[20] C. L. Morawetz, “A high-level approach to the animation of human secondary movement,” Ph.D. dissertation, Theses (School of Computing Science)/Simon Fraser University, 1989.

[21] S. A. Etemad and A. Arya, “Extracting movement, posture, and temporal style features from human motion,” *Biologically Inspired Cognitive Architectures*, vol. 7, pp. 15–25, 2014.

[22] W. Ma, S. Xia, J. K. Hodgins, X. Yang, C. Li, and Z. Wang, “Modeling style and variation in human motion,” in *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association, July 2010, pp. 21–30.

[23] J. M. Wang, D. J. Fleet, and A. Hertzmann, “Multifactor gaussian process models for style-content separation,” in *Proceedings of the 24th international conference on Machine learning*, vol. 3. ACM, June 2007, pp. 975–982.

[24] D. Bernhardt and P. Robinson, “Detecting affect from non-stylised body motions,” in *International conference on affective computing and intelligent interaction*. Springer, 2007, pp. 59–70.

[25] J. Min, H. Liu, and J. Chai, “Synthesis and editing of personalized stylistic human motion,” in *Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games*. ACM, February 2010, pp. 39–46.

[26] J. Lasseter, “Principles of traditional animation applied to 3d computer animation,” in *ACM Siggraph Computer Graphics*, vol. 21, no. 4. ACM, 1987, pp. 35–44.

[27] Y. Wang, Z.-Q. Liu, and L.-Z. Zhou, “Learning style-directed dynamics of human motion for automatic motion synthesis,” in *Systems, Man and Cybernetics, 2006. SMC’06. IEEE International Conference on*, vol. 5. IEEE, 2006, pp. 4428–4433.

[28] M. Abdul-Massih, I. Yoo, and B. Benes, “Motion style retargeting to characters with different morphologies,” in *Computer Graphics Forum*. Wiley Online Library, March 2016.

[29] N. F. Troje, “Decomposing biological motion: A framework for analysis and synthesis of human gait patterns,” *Journal of vision*, vol. 2, no. 5, pp. 371–387, 2002.

[30] R. Plutchik and H. Kellerman, *Emotion, Theory, Research, and Experience: Theory, Research and Experience*. Academic press, 1980.

[31] P. Ekman, “Are there basic emotions?” 1992.

[32] J. Russel, “A circumplex model of affect,” vol. 39, no. 6, pp. 1161–1178, 1980.

[33] R. E. Plutchik and H. R. Conte, *Circumplex models of personality and emotions.* American Psychological Association, 1997.

[34] A. Mehrabian, “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[35] H. Zacharatos, C. Gatzoulis, and Y. L. Chrysanthou, “Automatic emotion recognition based on body movement analysis: a survey,” *IEEE computer graphics and applications*, vol. 34, no. 6, pp. 35–45, 2014.

[36] M. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, “Body movements for affective expression: A survey of automatic recognition and generation,” *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 341–359, 2013.

[37] F. Ofli, R. Chaudry, G. Kurillo, R. Vidal, and R. Bajscy, “Berkeley MHAD: A comprehensive multimodal human action database,” *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pp. 53–60, 2013.

[38] J. K. Hodgins, “CMU graphics lab motion capture database,” 2002, http://mocap.cs.cmu.edu.

[39] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, “Documentation mocap database hdm05,” 2007.

[40] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3d points,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, June 2010, pp. 9–14.

[41] I. Kviatkovsky, I. Shimshoni, and E. Rivlin, “Person identification from action styles,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 84–92.

[42] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, “Instructing people for training gestural interactive systems,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, May 2012, pp. 1737–1746.

[43] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. LaViola, and R. Sukthankar, “Exploring the trade-off between accuracy and observational latency in action recognition,” *International Journal of Computer Vision*, vol. 101, no. 3, pp. 420–436, 2013.

[44] L. Xia, C.-C. Chen, and J. K. Aggarwal, “View invariant human action recognition using histograms of 3d joints,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, June 2012, pp. 20–27.

[45] J. Tilmanne, R. Sebbe, and T. Dutoit, “A database for stylistic human gait modeling and synthesis,” in *Proceedings of the eNTER-FACE’08 Workshop on Multimodal Interfaces*. Paris, 2008, pp. 91–94.

[46] J. Wu, J. Konrad, and P. Ishwar, “The value of multiple viewpoints in gesture-based user authentication,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 90–97.

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2019.2906167, IEEE Transactions on Affective Computing

[47] Y. Ma, H. M. Paterson, and F. E. Pollick, “A motion capture library for the study of identity, gender, and emotion perception from biological motion,” *Behavior research methods*, vol. 38, no. 1, pp. 134–141, 2006.

[48] J. Tilmanne and T. Dutoit, “Expressive gait synthesis using pca and gaussian modeling,” *Motion in Games*, pp. 363–374, 2010.

[49] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, “Cross-cultural differences in recognizing affect from body posture,” *Interacting with Computers*, vol. 18, no. 6, pp. 1371–1389, 2006.

[50] L. L. Presti and M. La Cascia, “3d skeleton-based human action classification: A survey,” *Pattern Recognition*, vol. 53, pp. 130–147, 2016.

[51] V. I. P. V. Laboratory, “BodyLogin Dataset: Multiview,” 2017, https://vip.bu.edu/projects/hcis/body-login/datasets/multiview/#dl.

[52] F. Han, B. Reily, W. Hoff, and H. Zhang, “Space-time representation of people based on 3d skeletal data: A review,” *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017.

[53] L. Herda, R. Urtasun, A. Hanson, and P. Fua, “Automatic determination of shoulder joint limits using experimentally determined quaternion field boundaries,” *International Journal of Robotics Research*, vol. 22, no. 6, 2003.

[54] F. S. Grassia, “Practical parameterization of rotations using the exponential map,” *Journal of graphics tools*, vol. 3, no. 3, pp. 29–48, 1998.

[55] B. Rosenhahn, C. Schmaltz, T. Brox, J. Weickert, D. Cremers, and H. Seidel, “Markerless motion capture of man-machine interaction,” in *Computer Vision and Pattern Recognition (CVPR), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.

[56] N. F. Samatova, W. Hendrix, J. Jenkins, K. Padmanabhan, and A. Chakraborty, *Practical graph mining with R*. CRC Press, 2013.

[57] M. Raptis, D. Kirovski, and H. Hoppe, “Real-time classification of dance gestures from skeleton animation,” in *Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation*. ACM, 2011, pp. 147–156.

[58] M. Müller and T. Röder, “Motion templates for automatic classification and retrieval of motion capture data,” in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2006, pp. 137–146.

[59] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[60] S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[61] N. F. Troje, C. Westhoff, and M. Lavrov, “Person identification from biological motion: Effects of structural and kinematic cues,” *Attention, Perception, & Psychophysics*, vol. 67, no. 4, pp. 667–675, 2005.

[62] S. A. Etemad and A. Arya, “Correlation-optimized time warping for motion,” *The Visual Computer*, vol. 31, no. 12, pp. 1569–1586, 2015.

[63] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popović, “Style-based inverse kinematics,” in *ACM transactions on graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 522–531.

[64] J. B. Tenenbaum and W. T. Freeman, “Separating style and content,” *Advances in neural information processing systems*, pp. 662–668, 1997.

[65] A. Shapiro, Y. Cao, and P. Faloutsos, “Style components,” in *Proceedings of Graphics Interface 2006*. Canadian Information Processing Society, June 2006, pp. 33–39.

[66] A. Witkin and Z. Popović, “Motion warping,” in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. ACM, 1995, pp. 105–108.

[67] S. Sogon and M. Masutani, “Identification of emotion from body movements: A cross-cultural study of americans and japanese,” *Psychological Reports*, vol. 65, no. 1, pp. 35–46E, 1989.

[68] A. Kleinsmith and N. Bianchi-Berthouze, “Affective body expression perception and recognition: A survey,” *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[69] B. De Gelder, “Towards the neurobiology of emotional body language,” *Nature reviews. Neuroscience*, vol. 7, no. 3, p. 242, 2006.

[70] M. Argyle, *Bodily communication*. Routledge, 2013.

[71] P. E. Bull, *Posture & gesture*. Elsevier, 2016, vol. 16.

[72] P. Ekman and W. V. Friesen, “Nonverbal leakage and clues to deception,” *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[73] F. Zhou and F. Torre, “Canonical time warping for alignment of human behavior,” in *Advances in neural information processing systems*, 2009, pp. 2286–2294.

[74] F. Zhou and F. De la Torre, “Generalized time warping for multimodal alignment of human motion,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1282–1289.

[75] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[76] L. Kovar, J. Schreiner, and M. Gleicher, “Footskate cleanup for motion capture editing,” in *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*. ACM, 2002, pp. 97–104.

[77] K. Wampler, Z. Popović, and J. Popović, “Generalizing locomotion style to new animals with inverse optimal regression,” *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 49–59, 2014.

Sarah Ribet received two engineering degrees in Computer Science in 2016 from Télécom Lille, France, and École de technologie supérieure, Canada. She is currently a PhD candidate at the Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL - UMR CNRS 9189) at the University of Lille. Her current research interests are mainly focused on the study and the analysis of 3D human motions and their applications in computer vision and computer graphics.

Hazem Wannous is an Associate Professor at the University of Lille and IMT Lille Douai. He is also a member of the Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL - UMR CNRS 9189). He received his Ph.D. degree in Computer Science from the University of Orléans, France in 2008. His current research interests are mainly focused on 3D action recognition, machine learning, pattern recognition, video indexing, and geometric vision. He serves as a regular reviewer for a number of top journals in the field (IEEE Trans. on PAMI, IEEE Trans. on Image Processing, IEEE Trans. on Medical Imaging, IEEE Trans. on Multimedia), and he served also as Area Chair for the IAPR ICPR 2014 conference. He is a PC member for a number of international conferences and workshops. He is co-author of several papers in refereed journals and proceedings of international conferences.

Jean-Philippe Vandeborre is a Full Professor of Computer Science at IMT Lille Douai (Institut Mines-Télécom Lille Douai), France. He received the M.S. degree in 1997, and the Ph.D. degree in Computer Science in 2002, both from the University of Lille, France. He also holds the French degree *Habilitation à Diriger des Recherches (HDR)* from the University of Lille, France, since 2012. He is also a researcher of CRIStAL (UMR CNRS 9189), the Computer Science Department of the University of Lille linked to the CNRS (Centre National de la Recherche Scientifique). His current research interests are mainly focused on 3D-model analysis, and include 3D-mesh segmentation, multimedia indexing and retrieval from content and their applications. He is also investigating the topic of human gesture analysis with the help of depth-cameras. He is the co-author of several publications in high-impact journals and conferences, and serves as program committee member or reviewer for international conferences and workshops in the 3D-model processing and computer vision fields.

1949-3045 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.