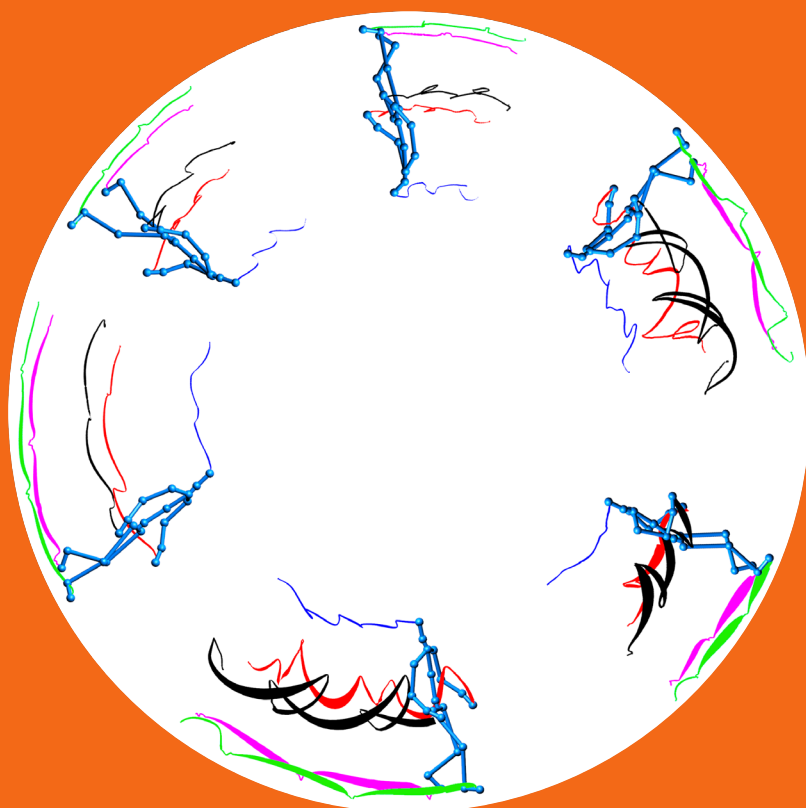


From motion capture to performance synthesis: A data based approach on full-body animation

Klaus Förger



From motion capture to performance synthesis: A data based approach on full-body animation

Klaus Förger (born Lehtonen)

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall AS1 of the school on 9 October 2015 at 12.

**Aalto University
School of Science
Department of Computer Science**

Supervising professor

Prof. Tapio Takala

Thesis advisor

Prof. Tapio Takala

Preliminary examiners

Assoc. Prof. Ronald Poppe,
Utrecht University,
Netherlands

Dr. Kari Pulli,
Light (<https://light.co/>),
USA

Opponents

Assoc. Prof. Hannes Högni Vilhjálmsson,
Reykjavík University,
Iceland

Aalto University publication series

DOCTORAL DISSERTATIONS 90/2015

© Klaus Förger

ISBN 978-952-60-6350-8 (printed)

ISBN 978-952-60-6351-5 (pdf)

ISSN-L 1799-4934

ISSN 1799-4934 (printed)

ISSN 1799-4942 (pdf)

<http://urn.fi/URN:ISBN:978-952-60-6351-5>

Unigrafia Oy
Helsinki 2015

Finland



Author

Klaus Föhrer

Name of the doctoral dissertation

From motion capture to performance synthesis: A data based approach on full-body animation

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 90/2015

Field of research Media Technology

Manuscript submitted 15 June 2015

Date of the defence 9 October 2015

Permission to publish granted (date) 18 August 2015

Language English

☐ **Monograph**

☒ **Article dissertation (summary + original articles)**

Abstract

Human motions such as walking or waving a hand can be performed in many different styles. The way the perceived styles are interpreted can vary depending on the context of the motions. The styles can be described as emotional states such as aggressiveness or sadness, or as physical attributes such as being tense or slow. This thesis studies synthesis of expressive styles and real-time interaction between autonomous characters in order to enable controllable performance synthesis.

Presented research relies on motion capture as it enables reproduction of realistic human motion in off-line animations, and recording expressive performances with talented actors. The captured motions can then be used as inputs for several motion synthesis methods that enable real-time animations with actions that can adapt to changing surroundings.

While the main field of this thesis is computer animation, building an understanding of motion style is also related to fields of perception, psychology and semantics. Furthermore, to recognize and to enable control of created styles, methodology from the field of pattern recognition has been used.

In practice, the research includes implementations and evaluations of proof-of-concept systems, and questionnaires where varying motion styles have been rated and described. Both quantitative analysis of answers of the questionnaires, and visualizations of the data have been made to form a qualitative understanding of motion style.

In the context of single character motion, the main result is in enabling accurate verbal control of motion styles. This was found to be possible when the styles are modeled as continuous attributes that are allowed to vary independently, and when individual styles are numerically defined through comparisons between motions. In the context of expressive interaction between characters, the research builds on the observation that motions can be interpreted as expressive behaviors when portrayed as reactions to an action. The main contribution here is a new method for authoring expressive interaction through recorded actions and reactions.

The results of the dissertation are useful for development of virtual characters as many existing systems do not take full advantage of bodily motions as an expressive medium. More specifically, the presented methods enable creating characters that can interact fluidly while still allowing the expressiveness to be controlled.

Keywords computer animation, motion capture, human motion, motion style

ISBN (printed) 978-952-60-6350-8

ISBN (pdf) 978-952-60-6351-5

ISSN-L 1799-4934

ISSN (printed) 1799-4934

ISSN (pdf) 1799-4942

Location of publisher Helsinki

Location of printing Helsinki

Year 2015

Pages 139

urn <http://urn.fi/URN:ISBN:978-952-60-6351-5>

Tekijä

Klaus Förger

Väitöskirjan nimi

Ilmaisuvoimaisten koko kehon animaatioiden tuottaminen liikekaappauksen avulla

Julkaisija Perustieteiden korkeakoulu**Yksikkö** Tietotekniikan laitos**Sarja** Aalto University publication series DOCTORAL DISSERTATIONS 90/2015**Tutkimusala** Mediatekniikka**Käsikirjoituksen pvm** 15.06.2015**Väitöspäivä** 09.10.2015**Julkaisuluvan myöntämispäivä** 18.08.2015**Kieli** Englanti☐ **Monografia**☒ **Yhdistelmäväitöskirja (yhteenvedo-osa + erillisartikkelit)****Tiivistelmä**

Ihmisen liikkeitä kuten kävelyä tai käden heilutusta voi esittää monella tyyllillä, joiden tulkinta voi vaihdella riippuen siitä millaisessa tilanteessa liikkeet esitetään. Liikkeiden tyyli voidaan tulkita tunnetiloina, kuten aggressiivisuutena tai surullisuutena, tai fyysisinä ominaisuuksina, kuten jäykkyytenä tai hitautena. Tässä väitöksessä on tutkittu liikkeiden tuottamista ja reaaliaikaista vuorovaikutusta autonomisten hahmojen välillä, jotta voidaan luoda ilmeikkäitä kehollisia esityksiä.

Esitetty tutkimus hyödyntää liikkeenkaappausta, koska se mahdollistaa todenmukaisten liikkeiden toistamisen animaatioelokuvissa, ja ilmeikkäiden esitysten tallentamisen. Kaapattuja liikkeitä voidaan käyttää lähtömateriaalina liikesynteesimetoille, jotka mahdollistavat reaaliaikaisen animaation ja liikkeiden mukautumisen muuttuvaan ympäristöön.

Väitöksen päätutkimusala on tietokoneanimaatio, mutta tyylien ymmärtämiseksi on käytetty menetelmiä myös havaintotutkimuksen, psykologian ja semantiikan aloilta. Tyylien tunnistamiseksi ja säädeltävyyden mahdollistamiseksi käytössä on myös menetelmiä, jotka liittyvät hahmontunnistukseen.

Käytännössä tutkimus sisältää esimerkkijärjestelmien toteuttamista ja arviointia. Lisäksi on tehty kyselyitä, joissa osallistujat ovat arvioineet tyylejä ja kuvaileet niitä omin sanoin. Kyselyiden vastauksia on analysoitu määrällisesti, ja dataa on visualisoitu laadullisen kuvan luomiseksi tyyleistä.

Yksittäisen hahmon tapauksessa päätulos on, että on mahdollista säätää tyylejä kielellisten kuvausten perusteella, kun eri tyylejä käsitellään jatkuvina ominaisuuksina, joiden annetaan vaihdella toisistaan riippumattomasti, ja kun yksittäiset tyyli määritellään numeerisesti liikkeiden vertailuihin pohjautuen. Hahmojen vuorovaikutuksen tapauksessa liike voidaan tulkita ilmeikkääksi, jos se esitetään vastauksena tiettyyn toimintaan. Päätulos tässä yhteydessä on uusi menetelmä luoda ilmeikstä vuorovaikutusta näyteltäjen liikkeiden ja niiden herättämien reaktioiden pohjalta.

Väitöksen tuloksia voidaan käyttää hyväksi kehitettäessä animaatiohahmoja ilmaisuvoimaisemmiksi kehollisten tyylien avulla, mihin useat olemassa olevat järjestelmät eivät anna tukea. Esitetyillä menetelmillä voi tuottaa hahmoja, jotka voivat olla sujuvassa vuorovaikutuksessa samalla kun niiden ilmeikkyyttä säädellään.

Avainsanat tietokoneanimaatio, liikekaappaus, ihmisen liike, liikkeen tyyli**ISBN (painettu)** 978-952-60-6350-8**ISBN (pdf)** 978-952-60-6351-5**ISSN-L** 1799-4934**ISSN (painettu)** 1799-4934**ISSN (pdf)** 1799-4942**Julkaisupaikka** Helsinki**Painopaikka** Helsinki**Vuosi** 2015**Sivumäärä** 139**urn** <http://urn.fi/URN:ISBN:978-952-60-6351-5>

Preface

The work for this dissertation has been funded over the years by the Department of Media Technology (currently part of the Department of Computer Science) of Aalto University, the Hecse doctoral program, aivoAALTO project of Aalto University, and Academy of Finland projects Enactive Media (128132) and Multimodally grounded language technology (254104). Thank you for your financial support as without it this dissertation would not have been possible.

For guiding the research, I thank Prof. Tapio Takala and Prof. Timo Honkela. Thanks also go to my co-workers Roberto, Meeri, Jari, Tuukka, and Päivi for your comments, suggestions and practical assistance that I have relied on. During the writing process constructive comments by Jussi Hakala and Jussi Tarvainen were very helpful. I also have gratitude for everyone who have put on the mocap suit and performed motions that I have used as raw data. I would mention you by names if I had not promised to keep your identities secret. I am also grateful for the comments from the pre-examiners, Prof. Ronald Poppe and Dr. Kari Pulli. I also thank family, friends and especially you Vilja, for listening to my rambling talks about the research.

Finally, the most important thing for the research has been good will among people as that allows spending less time on fighting, and more time on building the human civilization and related stick figure centric activities.

Espoo, August 24, 2015,

Klaus Förger (born Lehtonen)

Contents

Preface	1
Contents	3
List of Publications	5
Author's Contribution	7
1. Introduction	9
1.1 Motivation and scope	9
1.2 Research objectives, questions and methods	12
1.3 Structure of the thesis	13
2. Motion style in related fields of research	15
2.1 Virtual characters and expressive behavior	15
2.2 Affects and emotions	17
2.3 Historical view on animation techniques related to styles . .	18
2.4 Capture and representation of human motion	20
2.5 Example-based synthesis of human motion style	22
2.6 Recognition of motion styles	25
2.7 Semantics of motion styles	27
3. Styles in motion of a single character	31
3.1 Background	31
3.2 Perception of styles	32
3.3 Semantics of human motion	35
3.4 Controlling styles	38
3.4.1 Implementation of relative style control	39
3.4.2 Evaluation of relative style control	42
3.5 Discussion	45

4. Styles in interaction between characters	49
4.1 Background	49
4.2 Experiment on continuous bodily interaction	49
4.3 Authoring expressive interaction	52
4.4 Discussion	55
5. Conclusions	57
Bibliography	59
Errata	67
Publications	69

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

- I** Klaus Lehtonen and Tapio Takala. Evaluating Emotional Content of Acted and Algorithmically Modified Motions. In *24th International Conference on Computer Animation and Social Agents (CASA 2011)*, Chengdu, China, Transactions on Edutainment VI, Lecture Notes in Computer Science, Volume 6758, pages 144-153, May 2011.

- II** Roberto Pugliese and Klaus Lehtonen. A Framework for Motion Based Bodily Enaction with Virtual Characters. In *11th International Conference on Intelligent Virtual Agents (IVA 2011)*, Reykjavik, Iceland, Lecture Notes in Computer Science, Volume 6895, pages 162-168, September 2011.

- III** Klaus Förger, Tapio Takala and Roberto Pugliese. Authoring Rules for Bodily Interaction: From Example Clips to Continuous Motions. In *12th International Conference on Intelligent Virtual Agents (IVA 2012)*, Santa Cruz, USA, Lecture Notes in Computer Science, Volume 7502, pages 341-354, September 2012.

- IV** Klaus Förger, Timo Honkela and Tapio Takala. Impact of Varying Vocabularies on Controlling Motion of a Virtual Actor. In *13th International Conference on Intelligent Virtual Agents (IVA 2013)*, Edinburgh, UK, Lecture Notes in Computer Science, Volume 8108, pages 239-248, August 2013.

- V** Klaus Förger and Tapio Takala. Animating with Style: Defining Expressive Semantics of Motion. *The Visual Computer*, Online First Articles, February 2015.

Author's Contribution

Publication I: "Evaluating Emotional Content of Acted and Algorithmically Modified Motions"

The author was the main writer of the publication, and performed the practical work related to programming and conducting the experiments.

Publication II: "A Framework for Motion Based Bodily Enaction with Virtual Characters"

The work was an equal contribution between the writers. The author of this dissertation had a major role in creating the animation-related components of the system, and also contributed in conducting the experiments and writing the paper.

Publication III: "Authoring Rules for Bodily Interaction: From Example Clips to Continuous Motions"

The author was the main writer of the publication, and performed the practical work related to programming and testing the system.

Publication IV: "Impact of Varying Vocabularies on Controlling Motion of a Virtual Actor"

The author was the main writer of the publication, and performed the practical work related to programming and conducting the experiments.

Publication V: "Animating with Style: Defining Expressive Semantics of Motion"

The author was the main writer of the publication, and performed the practical work related to programming and conducting the experiments.

1. Introduction

1.1 Motivation and scope

This thesis studies style of human motion in animations. Possible applications of expressive styles include entertainment uses such as video games [54], and educational systems [26] that can, for example, teach how to cope in situations where other people show strong emotions. The motivation behind the research is that while motion style allows displaying many expressive variations of actions (see Fig. 1.1 for examples), it is not always used to its full potential in real-time and interactive animations. Practical limitations on taking advantage of motion style have been recently lowered as high quality motion capture has become more available and cheaper [54]. Even consumer level sensors can capture full body motions [102]. The research in this dissertation suggests that the main difficulty related to motion styles is not in synthesis of style variations, but in accurately defining individual styles. The main contributions of this thesis are methods for defining and controlling synthesized style seen from a single character and emerging from interaction between characters.

Motion styles are considered as tools for displaying hidden attributes ranging from predetermined physical and emotional states of an animated character to fluctuating attitudes and feelings towards other virtual characters or real humans. The styles are studied in the context of real-time animation in which it is not possible to use a lot of time on synthesizing motions or on evaluating appearance of motions.

It can be asked, why motion style should be used as a medium for relaying emotions and attitudes when other modalities such as facial expressions, symbolic gestures and speech could produce the same impressions. In practice, the most expressive results will likely come from combined

use of all modalities. Combination of modalities is important as all of them are not perceivable all the time. For example, a character could spend long durations without saying anything. Motion style is advantageous in this respect as style is present always when an action is performed. Also, consistency between the modalities can be important for example when emotions need to be communicated unambiguously [6, 16]. Therefore, some attention to motion style can be useful even when it is not the main modality of expression.

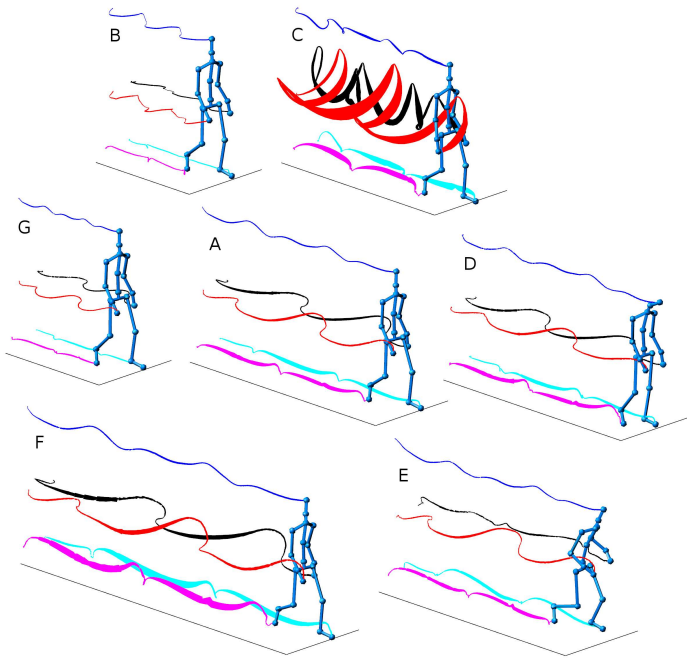


Figure 1.1. Examples of motion styles showing the end pose of the walk, trajectories of head, hands and feet, and instantaneous velocities as thickness of the trajectories. Motion A is an acted regular walk, B and C have changes in the shape of the motion trajectories, D and E have changes in overall posture, and F and G have different velocities and traveled distances.

This dissertation presents developments of methods that allow taking advantage of motion capture, performance capture, and motion synthesis to enable synthesis of performances controlled in the same way as one could instruct a human actor. Here, motion capture is referred as the technical means of recording motion. Performance capture means recording motions performed by talented actors, and enables for example animating expressive characters in movies. Example-based motion synthesis in turn extends motion capture towards flexible reuse of motions in new surroundings or as new variations. Performance synthesis aims to combine

all the three aspects as illustrated in Figure 1.2.

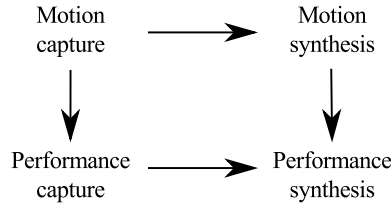


Figure 1.2. Dependencies between techniques related to motion styles. The arrows indicate that the target requires and extends the source.

The main field of the research is computer animation, but methodology related to the fields of pattern recognition, semantics, perception, and psychology have been applied as elaborated in Table 1.1. Methods have been included when they support the goal of producing expressive behaviors. Exclusions to keep the scope of the research manageable are explained in Chapter 2 under the sections concerning the respective fields. The included methods related to control of motion style aim to allow control of semantically meaningful styles. In the case of single character motion, this is realized with verbal commands.

Table 1.1. Scope of the research

	Pattern recognition	Computer animation	Semantics / Perception / Psychology
Core topics	<ul style="list-style-type: none"> • Style of animated human motion • Relative definitions for motion styles • Semantic control of motion style 		
Included topics	<ul style="list-style-type: none"> • Feature extraction from human motion • Automatic recognition of motion styles 	<ul style="list-style-type: none"> • Motion capture based synthesis methods • Real-time motion synthesis • Interaction in animations • Numerical representation of human motion 	<ul style="list-style-type: none"> • Perceived emotional content of motions • Verbal description of actions (verbs) • Verbal description of motion style (adverbs)
Excluded topics	<ul style="list-style-type: none"> • Action recognition 	<ul style="list-style-type: none"> • Manual key-frame editing • Physics-based animation • Combining motion with other modalities 	<ul style="list-style-type: none"> • Symbolic gestures with culturally varying meanings • Emotions felt by a performer of a motion

1.2 Research objectives, questions and methods

This dissertation is driven by two research objectives: Enabling verbal control of motion styles of a single character, and authoring of expressive behaviors that emerge from interactions between characters. To satisfy these objectives, answers were needed for the questions: 1. How people perceive and describe animated motion styles? 2. How the human motion styles can be modeled numerically in a compact and generalizable way?

The posed research questions are of exploratory nature, and this is reflected in the scientific methods applied in the research. To find out what kind of a phenomenon perceived motion style is, an analysis by synthesis approach has been adopted. In practice, questionnaires containing animations of acted and synthetic motions have been made, and people have been asked to rate (Publication I) or to describe (Publication IV) the styles they perceive. The answers of the questionnaires have then been analyzed with help of data visualizations to reach a qualitative understanding of perceived motion styles. Also, people interacting with a virtual character displaying varying behaviors have been interviewed (Publication II). Findings from these publications support the view that motion style is a multidimensional phenomenon, and that several styles may be simultaneously perceived. Also, the results imply that defining motion styles through comparisons of motions could be more precise than describing individual motions.

Numerical modeling of styles is present in all the publications. Its role is especially large in the Publications III and V as the practical implementations that allow control of styles depend on the numerical models.

Ways to fulfill the objectives have been demonstrated with two proof-of-concept systems. In the context of single character motion, accurate verbal control of motion style is made possible by considering styles as continuous attributes that are allowed to vary independently, and defining individual styles numerically through comparisons between motions (Publication V). A quantitative validation of accuracy of controlling a motion style is also presented as relying only on example cases was not considered sufficient. In the context of expressive interaction between characters, control is enabled by representing behaviors with recorded action and reaction pairs, and letting an expert select what are the most relevant features in the examples (Publication III).

1.3 Structure of the thesis

The next chapter presents applications of motion style and ways the term is used in the related fields of research. The third and fourth chapters present the research done for this dissertation from the points of view of style in motion of a single character (Publications I, IV and V) and style in interaction between characters (Publications II and III), respectively. The interactive case can be seen as a direct extension of the single character case, and this is the reason for the order of the chapters. However, the research started with the interactive case, which raised questions to be studied with the single characters. In practise, this means that part of the lessons learned from the single character case do not appear in the presented interaction framework as discussed in more detail in Section 4.4.

In the text, the general approaches and main results of the publications are elaborated, and the impact of the results is discussed. Details such as exact formulas used in practical implementations can be read from the individual publications. Finally, overall conclusions, suggestions for design of expressive virtual characters, and possible future directions for research are presented.

2. Motion style in related fields of research

2.1 Virtual characters and expressive behavior

A major application area for recorded human motion is in animation of human-like virtual characters similar to the one in Figure 2.1. Moving virtual characters can appear for example in games and animated films [54], they can be museum guides [90], simulated persons when planning physical cooperation [18], a part of a large crowd [70], ballroom dance teachers [36], virtual actors visualizing a script [85], pedagogical agents simulating a real-life situation [26], virtual dancers responding interactively to music [86], and conversation partners in an interview situation [37]. Synthesized human motion is a useful expressive modality in all these cases, but the requirements set by the cases can be very different. A pedagogical agent might have to express many negative emotions to make a training scenario believable, while a ballroom dance teacher could benefit from infinite patience even when a student is troubled. For a virtual dancer the motion might be the main modality of expression, but bodily motions could be used only for supporting spoken communication in case of a conversational agent.

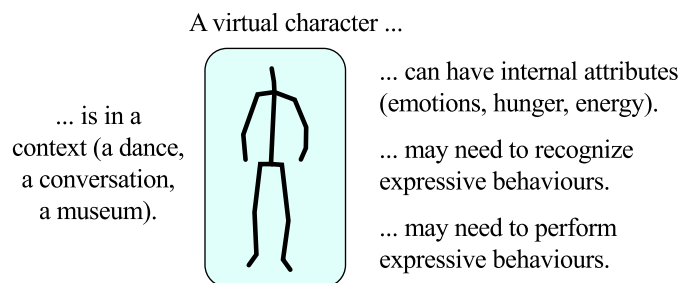


Figure 2.1. Properties of virtual characters

To participate in expressive interaction, a virtual character must be able to behave expressively, recognize behaviors, and be able to decide how to react to perceived behaviors. Early systems allowed characters to decide actions based on internal variables such as courage, hunger, intelligence and charisma [9, 71]. Simplifications in the systems included directly reading internal attributes of other characters instead of observing them from appearance of the characters. Also, the reactions tended to be mostly discretely scripted actions. The assumption of a small selection of allowed commands when planning behaviors is also built into more recent control schemes of virtual characters such as the Behavior Markup Language (BML) [95]. It has been proposed that allowing use of natural language should help authoring behaviors of virtual characters [71]. This view has been supported by a comparison of natural language and markup languages in the context of describing behaviors of characters in play scripts [85].

Bodily actions can be fluid, thus allowing continuous interaction that cannot be realized well by systems designed for turn-based dialog [105]. It has also been observed that if reactions of a virtual character are well synchronized with actions of a human, the virtual character is perceived more pleasant [37]. In practice, a fluid model of interaction has been created with a probabilistic method that uses pairs of recorded actions and reactions to learn how to react to human movements [39].

Continuous interaction with bodily motions can be seen as a form of enaction which means participatory sense-making where two interaction entities affect each others' efforts to understand their surroundings and each other [17]. The paradigm of enaction has been used in the context of interactive movies [87] and facial expressions of virtual characters [42] to turn users and observers into participants and co-authors of the content. In these contexts, the aim is to allow emergence of meaning from the interaction that would not exist if the same actions would take place separately.

In this dissertation, real-time interaction with a virtual character is an important theme. This shows also in the context of styles perceived from a single character as the used methods are primarily restricted to those that could be applied in real-time animation.

2.2 Affects and emotions

Style of bodily motions can communicate many emotions. In psychological research, several sets of basic emotions have been proposed that differ depending whether they are based on facial expression, bodily involvement, readiness to perform actions, or by being hardwired in the human brain [64]. When varying intensity levels of emotions are considered, basic emotions can be split into, for example, cold and hot anger or happiness and elated joy [96]. Another model for emotions are affective dimensions such as pleasure and arousal that have been validated to work when applying them on words related to emotions, facial expressions and felt moods [81]. Psychological studies on perception and recognition of affects in bodily motion mostly assume discrete non-overlapping classes of emotions or use abstract affective dimensions [44].

Important considerations for psychologists are whether acted or authentic emotions should be studied and if the ground truth comes from actors or observers [44]. As this dissertation views human motion through animation, the observers' perception is what counts. Also, the question of authenticity of emotions is more straightforward as virtual characters do not have real intentions or feelings.

When expressive interaction between a virtual character and a real human is desired, recognizing affects from the human becomes important. This essentially turns the virtual character into an affective computing system [66]. Challenges in affective computing include defining what is an affect, combining information from several modalities, interpretation of observed expressive behaviors, and taking context of the behaviors into account [66]. The same challenges are also relevant when the goal is to synthesize expressive behaviors that should be recognizable to humans.

It has been proposed that to solve problems of affective computing, researchers should not adopt a single theory of emotions, but rather concentrate on pragmatic models of expressive behavior learned from the users of affective systems [66]. This advice can be relevant in the context of human motion style as for example masculinity, femininity or a limping style are not part of the basic emotions nor do they fit easily into the affective dimensions. In this dissertation, emotional styles and those related to physical characteristics of human motion are considered equally important.

2.3 Historical view on animation techniques related to styles

Computer animation with human-like characters can be said to originate from the tradition of hand-drawn animation [47]. Principles such as anticipation of movements, squash and stretch during a motion, and slow in and out, can be used as guidelines in the creation of hand-drawn animations with appealing style [47]. Technically, hand-drawn animation is based on drawing key-frames that define poses, and in-between frames that allow fluent transition between the key-frames [47]. Early computer animation techniques, such as animating rotations with quaternions, were presented as ways to automate the creation of the in-between frames [84].

The ability to capture motions using optical, magnetic or mechanical systems created an alternative way to produce animations with human-like characters [54, pp. 14-24]. Motion capture usually refers to the technical process of recording motions [54, pp. 1-2]. When it is also recognized that the way the motions are performed is important, the term performance capture is often used [54, pp. 1-2]. While in traditional key-frame animation the style in the motions is created by an animation artist, in performance capture the style is created by an actor.

In early days of motion capture, it was often advertised as a cheap replacement for the work of key-frame animators, but this was quickly found not to be true [54, pp. 37-42]. The trend is also evident in the published animation research. Papers about methods for editing captured motions with techniques similar to key-frame animation have been presented such as motion warping [99] and motion path editing [23]. Also, motion captured material needs to be carefully retargeted to any character models that are differently sized than the original actor to avoid unphysical-looking results [22]. Even then the style of captured motions may be unfit to some purposes. For example, motions that should look like a giant lizard may end up looking more like a guy in lizard suit [54, p. 64].

In addition to extending traditional animation, digital human motion has opened doors for completely new kinds of animation techniques. Idea of treating human motion as a set of signals has produced motion interpolation that can create continuous ranges of motions styles between captured motions [79]. Other techniques allow editing existing styles by filtering motion signals [11] and by editing the signals in the frequency

domain [93].

Growing computing power allowed creating three dimensional animations that could be rendered in real-time. These new possibilities in turn enabled creation of interactive 3D video games. As character animation with key-frame techniques was expensive, video game industry quickly embraced motion capture as a cost-effective alternative [54, p. 34]. This development appears in animation research as techniques that allow artists and programmers to cooperate in building animations where individual clips can be smoothly concatenated [57]. Further development resulted in motion graphs that can be used for creating animations where a character ends up in a desired location [46]. While the original motion graphs allowed roughly determining what a character should do (walk, jump, etc.) and where the action should be done, the control of motion was not continuous. More fluid control of produced motion was achieved with motion graphs that allowed both concatenation of motions and interpolation of similar motions [83, 32].

Style of motions was not a high priority in early video games that used captured motions for character animation [54, p. 34]. A reason behind the low priority was that the used rendering techniques did not allow displaying all the subtle details of captured motions [54, p. 34]. As rendering techniques have developed, style has become a more important issue. This is visible in animation research as techniques that extend motion graphs with metadata related to dance styles [101] or with other information related to functional or stylistic variations [55]. A proposed alternative way to add style to real-time animation is to first produce neutral-looking motions and then use a linear time-invariant (LTI) model to add style variations [35]. In this case, the LTI model is a digital filter containing multipliers that can transform a given input teaching sample into a desired output teaching sample.

Parallel to motion capture based animation, there have been developments in animation based on kinematic modeling and physics simulations. Inverse kinematics (IK) satisfies constraints such as a character reaching for an object [98]. Usually the constraints can be satisfied by many alternative movements thus allowing style variations. The naturalness of the produced motions can be controlled to an extent, for example, by limiting used kinetic energy or distance to a default pose [27]. More flexible control of styles can be achieved by learning a style from recorded motion and making the IK prefer poses close to the learned style [27].

Similarly to IK-based methods, physics simulations allow generation of human motion without captured examples. Physics simulations enable realistic-looking contacts between objects, but require much more computation to animate a virtual human than motion capture or IK-based methods [20, 29]. Developing more natural-looking physics-based motion has been a concern, whereas production of expressive styles has had less attention [20]. A notable exception is recreation of acted styles using space-time optimization with physically derived rules [51]. This technique was used to create realistic-looking motions, but was also several magnitudes slower than real-time [51].

In practice, many animation systems are not strictly kinematics, physics or motion capture based, but combine several methods. For example, synthesis based on motion interpolation can be combined with IK to preserve contacts between the ground and the feet [79]. More elaborate systems can have a small amount of key frames, IK for handling external parameters, and physics simulation for secondary motions [80].

In this dissertation, example-based synthesis relying on motion capture is used for creation of style variations. While motion can be synthesized without examples with IK and physics-based methods, synthesis of expressive styles with those methods is often based on learning the styles from recorded examples [27, 51].

2.4 Capture and representation of human motion

The use of motion capture has been justified by its ability to record small details that make motions look natural, because reproducing the details with manual animation methods could be hard [3]. In turn, to be able to store, edit, and display human motion, the motion must be modeled numerically. The type of a system that is used for capturing motions determines what raw data is available. Optical motion tracking systems can record coordinates of markers attached to an actor [67, pp. 187-188]. Mechanical systems record distances between points and joint angles with an exo-skeleton that an actor must wear [54, pp. 24-24]. Magnetic systems can give both coordinates and orientations [67, pp. 187-188]. Modern depth cameras can record a three dimensional surface from a moving actor [102]. As the raw data can be clumsy in animating characters, motion capture systems often process the raw data and transform it to more refined formats [67, pp. 195-196]. In this dissertation, optical motion

tracking with markers (see Fig. 2.2) is the main method used for recording human motion. A detailed treatment of motion capture methods is excluded from this work.



Figure 2.2. On the left is an infrared camera and on the right is a suit with reflective markers that were used for capturing motions.

A common way to represent human motion in an animation context is to treat the human body as a transformation hierarchy. In practice, offsets between levels of the hierarchy approximate bones and rotations model the joints. The number of included bones and joints can vary depending on the capture system used and the desired level of realism. While this kind of representation is only an approximation of real bodily structures, it is usually sufficiently accurate for animations, and it offers three technical advantages in the context of animation. First is that a hierarchy is a compact presentation for the degrees of freedom (DOF) allowed by a human body. The second is that the constraints such as the distances between joints are implicitly satisfied by the hierarchy. The third is that graphics APIs such as OpenGL can process and traverse hierarchies efficiently. [25]

Further variation in representations comes from the format of the rotations. Possible alternatives include rotation matrices, Euler angles, exponential maps and quaternions [24]. Matrix transformations are commonly used in computer animation as they allow several three dimensional transformations in addition to rotations [67, pp. 133-34], but they do not enable interpolation of rotations as such [67, p. 53]. Euler angles can be considered the most intuitive to humans [67, p. 60]. However, Euler angles suffer from gimbal lock which is a mathematical singularity and can prevent editing one degree of freedom in some combinations of rotations [24]. Exponential maps are less prone to gimbal lock than Euler angles, while still having only three parameters for the three degrees of freedom [24]. This makes exponential maps an attractive format for

machine learning based motion synthesis methods [88]. Limited use of exponential maps in animation is probably related to the lack of a simple way to concatenate rotations with them [24]. Quaternions allow rotations to be concatenated and interpolated with relative ease, and they do not suffer from gimbal lock [84]. This makes them suitable for animation systems. A weak point in quaternions is that they have four parameters that model three degrees of freedom [24]. This makes them less intuitive to use than Euler angles [67, p. 60] and they need to be normalized to unit length after any changes [24].

While a hierarchical skeleton is the *de facto* standard for representing a human body in animation, it is not always the best format. For example, in decomposition of motions to style components, a representation based purely on coordinates has been found to work better than any formats based on rotations [82]. Another case where joint rotations can be sub-optimal is detection of similarities between motions [46]. In that case, joints closer to the root level of a hierarchy can have more impact on overall pose than the ones further away. Also, the impact of a joint is not always the same, but can vary depending on the current pose.

In practice, low-level motion signals specifying coordinates and rotations are often transformed to other formats to better fit specific purposes. Compression of motion data can employ methods such as approximation of data with Bezier curves, wavelet transformation or per-frame Principal Component Analysis (PCA) [4]. An opposite direction can be taken in recognition of human motion where motion may be represented with more signals than in an original format as is elaborated later in Section 2.6. Synthesis of styles can employ motion formats that isolate style-related aspects to part of the signals, while retaining the ability to return to the original format as is presented in more detail in Section 2.5.

2.5 Example-based synthesis of human motion style

Example-based methods that enable creating style variations are useful, because doing new motion capture for each variation would require time and money and is not practical in all situations [54]. In this dissertation, the main attention is given to high-level methods, in which styles can be adjusted with a few parameters, and that can work in real-time. Motion captured examples are used also in off-line methods that extend manual key-framing by, for example, propagating edits from one body part to oth-

ers [63].

As many style oriented editing operations use more than one input motion, compatibility between the motions must be considered. The methods that swap parts of motions, such as one limb, require that the input motions are equally long. The methods that are based on differences between motion signals also require that corresponding events happen at same time, in other words that the motions are in the same phase. Both requirements can be satisfied with time warping [99, 45] if the input motions contain the same action. Copying timing from one motion to another has also been suggested as a way to edit stylistic content [3].

Motion blending by linear interpolation is the standard approach for creating ranges of styles [69]. The interpolation generally requires that timings of the input motions are aligned. After that the poses in the corresponding frames of animation can be interpolated using several alternative numerical methods. Methods that perform linear interpolation can also be extended to linear extrapolation as illustrated in Figure 2.3. The use of interpolation and extrapolation of style differences relies on the finding that the changes they produce seem to correspond well to perception-based rating of styles [92]. When styles between two motions are needed, spherical linear interpolation (slerp) of quaternions gives high quality results as it guarantees a constant rotational velocity over the parameter range [84]. Slerp is useful for example in creating transitions between different actions [46]. However, as synthesis of style variations is usually restricted to motions containing the same action with different styles, simpler linear interpolation and extrapolation allowing multiple input motions is possible. Linear interpolation has been used successfully with motions encoded as spline parameters [79], components from Principal Component Analysis (PCA) [94], components from Independent Component Analysis (ICA) [58], frequency bands [11, 93], parameters of Hidden Markov Models (HMMs) [89] or as quaternions with renormalization to unit length [84].

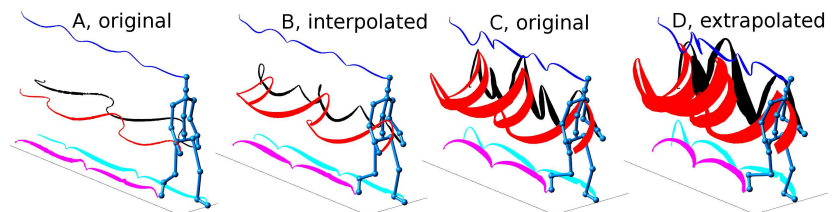


Figure 2.3. A and C are original captured motions. B is produced by interpolation of A and C. D is produced by extrapolating the difference between A and C.

A problem with interpolation and extrapolation of style differences is that the number of parameters can be impractical with a large set of motions as the weight of each motion is one adjustable parameter. This has been remedied by mapping several motions to manually labeled ranges using Radial Basis Functions (RBFs) [79]. The RBFs are commonly used in scattered data interpolation. An RBF restricts the impact of a data point in the interpolation space to a value that decreases with distance. Another alternative for reducing the number of parameters is to perform dimensionality reduction between motions with Principal Component Analysis (PCA) [21, 91, 92, 94]. The PCA is a numerical method that combines correlated variables, and allows removing dimensions with the least amount of numerical variance.

In addition to creating styles between motions and exaggerating style differences, it is possible to transfer style between motions [35, 3, 56]. Style transfer, which is based on differences between motions, requires that the example motions and the target motion contain the same action [35, 56]. In style transfer that is limited to, for example, only per-segment retiming and amplitude scaling, the target motion can be from a different action category [3]. Style transfer can be done continuously in real-time by training a linear time-invariant (LTI) model to reproduce the differences between example motions [35]. The perceptual content of the motion differences is completely dependent on used example data and can reflect, for example, styles given as actor instructions such as ‘neutral’, ‘angry’ or ‘crab walking’, or identity-related styles determined by differences between people [56].

The term ‘style transfer’ has also been used when referring to substituting a part of motion signals with ones from another motion [82]. However, this can be considered to be in a different class than the other transfer methods as it requires that the style is localized in only a part of the channels of a motion. Swapping signals of joint rotations is a straightforward operation that has been found practical in creating new style variations [38]. However, as not all swaps create natural-looking results, rule-based classifiers were used to prune the results [38]. Creating partial blends in the joints that are connected to the swapped joints can reduce the potential unnaturalness [65]. Swapping can also be applied to frequency bands [11] or ICA components [82] calculated from original channels of motions. ICA decomposition allows swaps with a reduced dimensionality, but may require post-processing to remove artifacts such as foot sliding [82].

Decomposition of motion signals to frequency bands enables a style editing operation that requires only one input motion [11, 93]. The frequency bands can be useful as they split motions to overall movements (low frequencies) and to more detailed motion textures (high frequencies) [76] as illustrated in Figure 2.4. Thus, scaling the bands can create new style variation without changing the action of the motion [11].

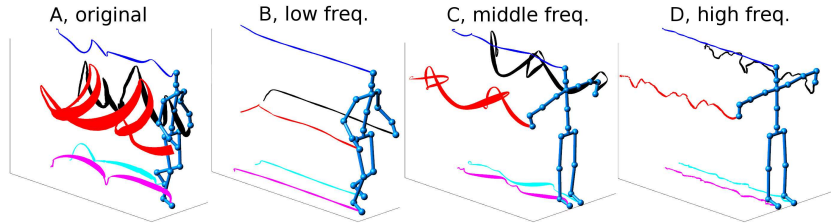


Figure 2.4. A is an original captured motion. B, C and D have been created from A by preserving global translation and low, middle and high frequencies of joint rotations, respectively.

This dissertation builds on example-based motion synthesis methods to enable performance synthesis. In this context, motion synthesis is considered as creation of style variations, while performance synthesis is seen as creation of specific styles that an animator could desire. The methods presented in this section can be considered to be on the side of motion synthesis as they either create unnamed variations or the descriptions of the produced styles come directly from labels of input motions. This dissertation presents two performance synthesis methods in which similar styles can be produced even if the used input motions would be randomly reordered or if the styles would be differently mixed in individual input motions.

2.6 Recognition of motion styles

In this dissertation, automatic recognition of motion styles is considered as a necessary step in controlling behavior of an expressive virtual character. Other possible uses for the recognition include detection of affects from a human in affective computing systems [66], and retrieval of motions from a large database [59]. For many uses, it can be beneficial if the automatic recognition would be similar to human recognition of styles.

Human recognition of affects has been studied, for example, from arm movements [73]. Ten acted affects (afraid, angry, excited, happy, neutral, relaxed, sad, strong, tired and weak) were used in the study. The result

of the study was that overall recognition rate of 30% was reached when the chance value was 10%. The low recognition rate was partly caused by confusions between the acted categories. For example, acted weakness was identified as weak, sad or tired. It was also speculated that arm movement might not be an optimal way to express all the affects. When studying affects in full-body motion, it has been found that people can recognize affects robustly if the animated body model gives at least some hints of structure and is not only a cloud of point-lights [53].

Other studies suggest that bodily expressions can have unique properties that are not present in other modalities. For example, when classifying bodily expressions, movements showing terror and happiness can be confused with each other even though as felt emotions they could be considered almost opposites [96]. Furthermore, emotions perceived from human motion have been found to affect the perceived gender [41]. For example, sad motion style caused motions acted by males to be often classified by observers as being performed by a female [41]. Also, other modalities may modulate or divert attention away from bodily motion [6, 16].

Naturalness and realism of motions can affect the way styles are perceived. In this dissertation, the aim has been to keep the considered motions in a natural range, for example by not using motion extrapolation. However, the appearance of 3D models shown in animations of the motions also have an affect on human perception. In a study comparing the recognition of emotions seen in still poses displayed with two levels of realism, it was concluded that neither the realistic nor the more simplified 3D model was better than the other [68]. In the context of facial expressions, it has been found that less realistic facial models can be exaggerated more than realistic ones before the expressions turn strange-looking [60]. A stick figure model has been selected for the animations of this dissertation. It is a good compromise showing sufficient amount of details to support recognition of motion styles while lacking diversions such as facial expressions.

On the side of automatic recognition, bodily expressions have been studied by coding them with features such as posture of the upper body, amount of movement activity, spatial extent of motions and movement dynamics [96]. These features were found to be effective in classifying acted affects (54% correctly classified versus 7% chance level). Other possible low level features include maximum distance between body parts, and speed, acceleration and jerk of a single body part [7]. More refined features include

computationally defined versions of Effort and Shape components of Laban Movement Analysis that have been used in recognition of styles from dance [15, 28]. Camera-based features used in affect recognition include Contraction Index (measures extend of a silhouette), Quantity of Motion, and Motion Fluency [13]. Styles can also be modeled numerically with features that are derived from captured motions with Fourier transform and Principal Component Analysis (PCA) [92].

Recognition of a styles can be viewed as an extension of action recognition [74] if styles are represented by groups of individual motions. In that case, methods such as Support Vector Machines (SVMs) can be used for the recognition [5]. An SVM is a machine learning method that is trained with an example classification, and can then be used to classify new data points.

None of the publications of this dissertation are solely about automatic recognition of motion styles. Instead, it appears as part of several publications where the recognition is used for enabling control over motion styles (Publication V) or in forming reactions to the recognizable behaviors (Publications II and III). Furthermore, recognition of styles is not considered as a question whether a motion has a style or not, but rather as amount of a style or as relative differences between two motions.

2.7 Semantics of motion styles

In previous sections, research related to motion style has been introduced from several points of view. However, the cited publications do not have a common definition of the term ‘motion style’.

One approach to motion style is to treat it mainly as a way to communicate emotions while giving little attention to other types of styles as is often done in psychological research [44]. Style can be also be discussed in the context of natural and artificial-looking motions [34]. Expert definitions of styles such as Laban notation for dance movements are another possibility [15, 28]. Yet another possibility is to consider human motion to consist of content (actions such as walking or running), identity (including age and sex), and style (emotions and attitudes) [31]. Style can be also viewed as all types of variations that are possible for an action without further distinctions [48]. Some consider styles to be hard to define quantitatively [51] while others [100] give definitions such as: "We define the style of motion as statistic properties of mean and standard variance

of joint quaternions in 4D unit sphere space." Style can also be considered to emerge from physical properties of motion and the human body [51]. Alternatively, from the point of view of key-frame animation, aspects such as posture, transitions between poses, simplification and exaggeration can be considered to be important for styles [62]. Further mix-ups can be caused by the division of motions to non-stylized (meaning everyday actions) and stylized (meaning artistic movements such as dance) [7].

While the overall classifications of motion styles can be interesting, verbal descriptions for individual styles can be more useful in practical situations where styles are synthesized. The descriptions can be built by experts on top of example motions [79], be based on actor instructions [56], or individual users may be allowed to define their own styles by annotating examples [92]. Properties that can be defined as exact numbers such as velocity or traveled distance may also be used [94]. It is also possible to rely completely on visual inspection while editing unnamed components [82]. In some publications styles are named ad hoc for example as "goosestep" [55] or "catty" [104].

While many publications claim to enable synthesis of styles, surprisingly few of the publications encountered during writing of this dissertation try to validate how large ranges of styles can be produced and how reliably they can be recognized by human observers. The majority of the publications rely only on showing a few examples of synthesized styles and giving opinions of the authors [1, 10, 11, 27, 31, 40, 55, 58, 61, 52, 63, 76, 79, 82, 91, 92, 93, 97, 100, 104]. A smaller number of publications compare produced styles with acted examples taken as ground truths [3, 34, 35, 56, 94]. In one study, satisfaction of users trying to create styles was tested [43]. Only four studies had systematical perceptual evaluations of produced styles done by people who did not create the styles [14, 33, 49, 89]. In all of these four studies, at least a few cases were encountered where a synthesized motion was recognized worse than an acted version or a feature that was hypothesized to predict style did not actually do that. This highlights the need for further perceptual evaluations especially in cases where claims are large such as separation of synthesized styles and identities [31] or "it can convert novice ballet motions into the more graceful modern dance of an expert" [10].

The varying and sometimes vague definitions for motion style and the absence of systematic evaluation of synthesized styles can be seen as different sides of the same phenomenon. In this dissertation, styles are con-

sidered primarily as perceivable variations of human motion that can be described with natural language. The descriptions are not considered only as words but also as symbols that can be given numerical descriptions by grounding them with physical measurements [30]. A way to perform the grounding is modeling styles as convex regions in a conceptual space [19] that may contain hierarchical relations such as ‘limping is a type of walking’ [50]. In this case a style could be represented with a group of similar motions. Alternatively, styles can be modeled as differences between example motions [104]. In this dissertation, the essence of motion style is explored with perceptual experiments, and a definition of styles as relations between motions is embraced and shown to be beneficial in practice.

3. Styles in motion of a single character

3.1 Background

This chapter presents research done in Publications I, IV and V that study motion style visible from a single character. Two underlying points of view are shared by these publications. The first is viewing motion style as a continuous phenomenon as is illustrated in Figure 3.1. The second is viewing motion styles as aspects of motion that may occur simultaneously as is demonstrated in Figure 3.2. While the two views may seem self-evident in the context of low-level motion synthesis, they are less often taken into account in published systems that allow high-level control of motion style with natural language labels for styles, for example. In the next two sections, human perception and descriptions for styles are explored. Then a method for controlling motion style with relative commands such as ‘do the same, but more sadly’ is presented.

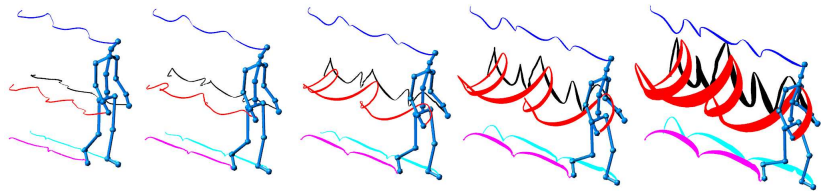


Figure 3.1. A continuum between depressed and aggressive styles

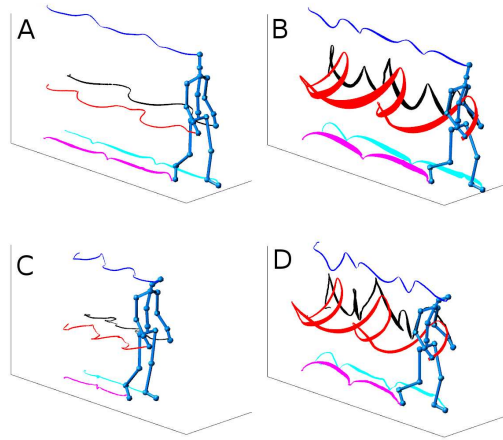


Figure 3.2. Example of style dimensions where A is a neutral starting point, B is more aggressive, C is more depressed, and D is both aggressive and depressed.

3.2 Perception of styles

Publication I presents an evaluation of emotional and stylistic content of acted and algorithmically modified motions. The idea of this work was to explore how people perceive motion styles in everyday movements. The approach was to allow the motions to be evaluated on multiple simultaneous Likert scales with each scale described with a commonly used word instead of more abstract affective dimensions [81]. Also, the intention was not to limit the scope to basic emotions [64], but to also consider motion specific attributes such as masculinity and relaxedness.

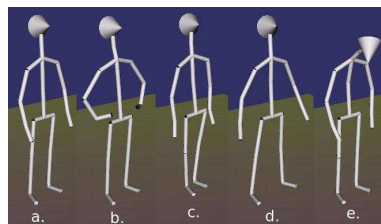


Figure 3.3. Still poses used in changing posture of animated characters in Publication I. Pose (a) is the neutral pose of the actor.

The stimulus material of the study consisted of short walks followed by a knocking motion performed by a male and a female actor. The motions were asked to be acted in styles afraid, angry, excited, happy, neutral, relaxed, sad, strong, and weak, which had been used in earlier by Pollick et al. [73]. To complement the acted styles, similar motions were produced

by an animator by applying algorithmic modifications on the neutrally acted examples. The modifications were combinations of adjustments to the pose of the characters (Fig. 3.3), scaling of frequency bands created from the motion signals [11] (Fig. 3.4), and modifications to the timings of the motions (Fig. 3.5).

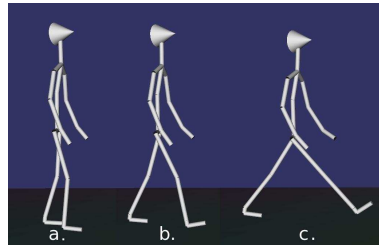


Figure 3.4. Effect of the frequency-based modification used in Publication I when a walking motion (b) is modified to create a shorter motion (a) and a longer motion (c).

The three research questions of the study were: 1. Can acted styles and emotions be distinguished by viewing motions animated with a stick figure? 2. Do the three implemented modifications change emotions seen in the motions? 3. What are suitable dimensions to be rated when evaluating motions? Answers to the questions were sought with a questionnaire containing videos of the stimulus material shown as a stick figure character. The videos were evaluated using five point scales between sad-happy, tired-excited, angry-relaxed, weak-strong, afraid-confident and masculine-feminine. The questionnaire was answered by 28 volunteers.

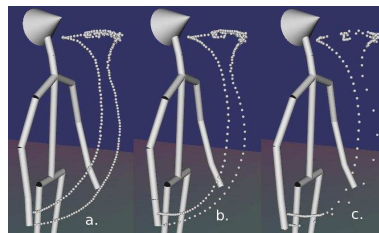


Figure 3.5. Trajectories of the right hand during knocking motions used in Publication I. Original motion (b) is modified to nearly constant speed (a) or with exaggerated acceleration (c).

Analysis of the answers showed that motion styles could be perceived from both acted and modified motions. This result agrees with results presented in related publications where styles have been perceived from even more simplified point-light representations [41, 92]. Though, it has been suggested that styles are seen as less intense from point-light representa-

tions than from representations showing the form of the characters [53]. The analysis also revealed that the intended styles were not always perceived as the most visible ones. This was true for both the acted and the modified motions. Also, the neutrally acted motions were not perceived as completely neutral, but contained for example relaxedness, confidence and hints about gender. These aspects seemed to greatly differ between the actors. In light of these results, the viability of relying only on expert actors or animators for providing a ground truth in the context of motion styles can be questioned. At least, it seems plausible that perception of motion styles can be more subjective than of actions such as walking or jumping.

From the three implemented modifications to motions, changing postures and scaling frequency bands were found to be useful tools for adjusting styles. One-to-one relationships between the modifications and the perceived styles were not found. Instead, combinations of modifications were observed to produce styles that they could not produce individually. The modification of timings was not found to have much impact on the perceived styles. A similar problem with retiming has been reported previously when trying to change the emotional content of captured motions [33]. It is also possible that the modification could be more effective when applied to other types of motions as it has been speculated that arm movements might not be an optimal way to express all the affects [73].

Based on the answers of the questionnaire, evaluation of motion styles benefits greatly from allowing several styles to be simultaneously rated instead of forcing the participants to select only one style or emotion. This is evident as the modifications to motion could affect the perception of several styles. Further analysis shows that styles tired and sad were used very similarly in the questionnaire, and the same applies to styles angry and masculine as shown in Figure 3.6. This implies that the said styles could be joined to a single dimension. However, if a different set of motions were given in the questionnaire, there could be situations where the style descriptions would be used separately. For example, another publication evaluating human motions has found a stronger connection between perceived sadness and gender [41].

In the end, giving general suggestions on what are the best dimensions to be included in future questionnaires would require a much wider range of actions and acted combinations of styles. With the lessons learned from Publication I, the questionnaire in Publication IV allowed free descrip-

tion of motions instead of trying to create a long list of all possible style dimensions.

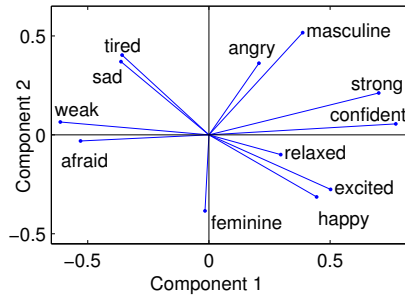


Figure 3.6. Maximum likelihood estimate for common factors for perceived styles in the questionnaire of Publication I. Loadings of the original dimensions are plotted into a two-dimensional model.

3.3 Semantics of human motion

Publication IV presents an analysis of how motion styles are described when people are allowed to use their own words in the descriptions. The idea was to explore how unanimous people are when they describe human motion, and to find good ways to model motion-related vocabularies based on numerical motion data.

Overall experiment settings of Publication IV and Publication I are similar. In both studies motions were recorded and modified to create a set of stimuli which was evaluated in a questionnaire. However, in Publication IV the creation of the stimuli is based on interpolations between acted examples producing a denser and more even set of motions. Also, in the questionnaire people were asked to write in natural language what the character is doing and how it is doing it, thus guiding the answers much less than the explicit scales of Publication I.

The acted examples were performed by two actors who were asked to run, walk and limp with styles sad, slow, regular, fast, and angry. The final stimuli were produced by interpolating pairs and triplets of motions. The motions were then animated as stick figures. The animations were shown one by one in the questionnaire and the participants were asked to describe the animation in writing with a verb or phrase (such as ‘swimming’ or ‘mountain climbing’) and from zero to three modifiers (such as ‘colorfully’ or ‘very colorfully’). The questionnaire was answered by 22 participants in Finnish.

The answers revealed that while the actor instructions had 3 verbs and 5 styles, the participants described the motions more varyingly with 88 verbs and 233 modifiers. Further analysis showed that the most common words explained a large portion of the word usage, but there was also a long tail of rarely used words. In practice the results imply that, if a single annotator describes motions, the most common descriptions will be covered, but many seldom used ways to describe motions will be missed.

To analyze the verbal descriptions against numerical motion data, features based on coordinates, velocities, accelerations, rotations as quaternions, and distances between body parts were calculated. The distributions of the verbs and modifiers are plotted on the PCA dimensions of the numerical features in Figure 3.7 and in Figure 3.8. The figures show that verbs appeared in continuous areas while modifiers could be scattered to separate clusters. The three synonyms for limping in the Finnish language ‘ontuu’, ‘nilkuttaa’, and ‘linkuttaa’ also appear consistently in the same area of Figure 3.7.

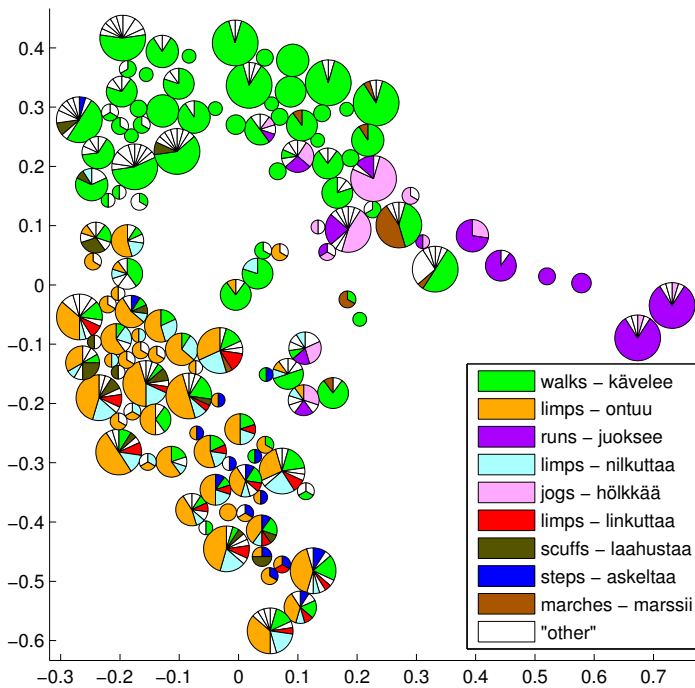


Figure 3.7. Distributions of most common verbs for each motion from Publication IV mapped on the first and second normalized PCA components. Each pie represents descriptions of one motion, the surface area of the pies is proportional to the number of given descriptions, and the position of the pies reflect the style of the motions.

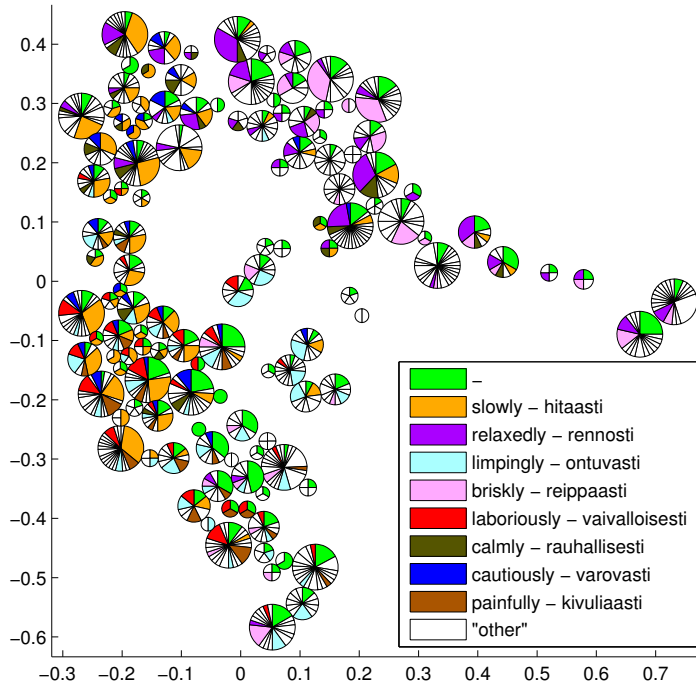


Figure 3.8. Distributions of most common modifiers for each motion from Publication IV mapped on the first and second normalized PCA components in similar manner as in Figure 3.7.

The data from Publication IV allows making suggestions related to symbol grounding, and more specifically on how words can be tied to measurable phenomena [30]. The results imply that numerical definitions for verbs can be based on linking a verb to a section of a feature space. This correspond to the idea that verbs can be modeled as convex regions in a conceptual space [19]. Furthermore, a hierarchy could be formed to model relationships between generic and more specific verbs. For example, ‘limping’ could be a special case of ‘walking’ as in Figure 3.7 the motions described as ‘limping’ were very often described also as ‘walking’, while the opposite was not true.

In turn, modifiers that are most often adjectives or adverbs, did not form convex regions as has been suggested [19]. For example, in Figure 3.8 ‘slowly’ can be seen in two disconnected areas, one where motions can be described as ‘slow walking’ and another where the term ‘slow running’ is appropriate. Another example is ‘briskly’, that appeared separately as ‘brisk walking’ and ‘brisk limping’. This suggests that modifiers could be best modeled as transitions in a numerical feature space which basically means comparisons between motions. This idea was taken to practice in Publication V.

Following the suggested ways to ground words related to motions requires that the concepts can be classified as verbs or modifiers. The classification can be obvious for example with concepts ‘to walk’ and ‘slowly’. However, a case such as ‘limping’ can be more ambiguous as the data shows that in the Finnish language the concept was used both as a verb ‘to limp’ - ‘ontua’ and as a modifier ‘walk limpingly’ - ‘kävellä ontuen’.

The ambiguity between actions and styles is typical in natural language. Expressions such as ‘limping’ or ‘walking’ are not purely descriptions of functions, but also implicitly set restrictions on the style. Furthermore, there might not be a single correct way to label actions, rather the labels may vary depending on what group of motions is considered. In synthesis of style variations, important considerations are if two motions can be interpolated without artifacts, and do the interpolations belong to the same action category as the original motions. While these properties may often be satisfied by motions described with the same verb, the natural language labels are neither absolute guarantees nor restrictions due to their ambiguous nature.

A possible idea for future work would be to more accurately measure the shift in the feature space caused by adding a modifier such as ‘slow’ to a verb. Information about this issue could be beneficial for automatic generation of labels for recorded motions. Measuring the shifts would require a more dense sampling of the motion space than in the existing dataset to allow reliable estimates.

3.4 Controlling styles

Publication V presents a method for controlling motion style with relative commands such as ‘do the same, but more sadly’ as shown in Figure 3.9. This work builds on top of the conclusion of Publication IV that motion styles could be best modeled comparatively as relations between motions. The central idea in Publication V is to show that embracing the relative nature of motion styles enables making accurate adjustments to synthesized motion styles.

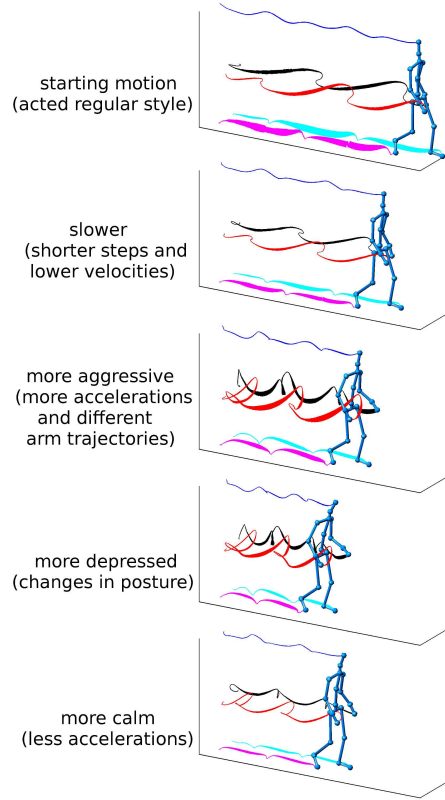


Figure 3.9. Control of walking style with relative style commands used in Publication V. Starting from an acted motion on top, each picture shows incremental changes towards the bottom.

3.4.1 Implementation of relative style control

Publication V makes a distinction between absolute styles that can be perceived from a single motion and relative styles that are seen from differences between motions. The goal was to find correspondences between perceived relative styles described with phrases of natural language and numerical models of styles. Practical steps included acting example motions, annotating pairwise differences between motion pairs, calculating numerical features from the example motions, and creating numerical definitions for the used natural language labels. The numerical definitions are called style vectors [104] as they tell the direction in a numerical feature space where the related style grows more intense. The style vectors in turn can be used to control parametric motion synthesis methods. An example of iterative control of interpolation-based synthesis is given in Figure 3.9.

Modeling human motion with numerical features enables automatic evaluation of motion style. In Publication V, the numerical modeling starts from calculation of per-frame values including positions in a local coordinate system of a character, velocities, accelerations, distances between body parts, and joint rotations as quaternions. These values are then turned into features representing short motion sequences by decomposition to frequency bands, averaging over time, and normalization to make actions performed with left and right sides of the body equal. Total number of individual features was 4816.

Creating numerical definitions of relative styles based on pairwise comparisons requires taking into account that differences on several styles can often be perceived simultaneously from one pair of motions. In an ideal case, the styles that are different from semantical and perceptual points of view could also be adjusted separately. Having to adjust styles in groups based on similarities that occurred by chance in a set of example motions would be less than ideal.

On a practical level, this calls for dividing numerical motion features into ones that are essential for a style, meaning the features that behave systematically in all examples of the style, and to incidental ones that only correlate with the style in most cases. There is no universal set of essential features, but different sets of features are essential for different styles.

In the implementation, simultaneously appearing styles were taken into account in three parts of the process. In acting example motions, the actor was asked to perform pairs of styles in addition to single styles to provide data that would contain also non-stereotypical combinations of styles. In annotation of style differences, the annotator was allowed to write zero to three styles per motion pair. In the creation of style vectors, an elimination step was included that removes the features that do not systematically appear in all examples of an annotated relative style. The harsh elimination of features was possible as the large number of individual features allowed at least part of them to be preserved. These considerations allowed producing style vectors that contain the essential aspects of the perceived styles, while ignoring incidental correlations between styles that may appear randomly or be caused by preferences of individual actors.

To test the process in practice, an actor was asked to perform walking motions with pairwise combinations of styles fast, slow, relaxed, tense,

angry, sad, limping, and excited. In the annotation, the following styles were perceived at least in five motion pairs: fast, slow, aggressive, lazy, excited, energetic, calm, limping, healthy, depressed, busy, relaxed and tense. Style vectors were then created for these perceived styles.

A set of style vectors can be used for controlling styles of animated motions produced with parametric synthesis. The process (Fig. 3.10) starts from initial parameters which produce a desired action. Next, if the user is not satisfied with the style, a style adjustment such as faster, slower or more aggressive can be selected. The system must then find which adjustment to the synthesis parameters produces the desired change of style. This is done automatically by synthesizing new motions with offsets to each parameter separately, calculating numerical features of the new motions, and finding a combination of the parameter offsets that matches the style vector of the desired style. Based on this data, the new synthesis parameters can be calculated with an off-the-shelf solver as the feature values of the motions representing parameter offsets form a Jacobian matrix. The style control process works in principle with any parametric motion synthesis method. To test the control method, interpolation-based parametric synthesis producing varied walking styles was used as an example case.

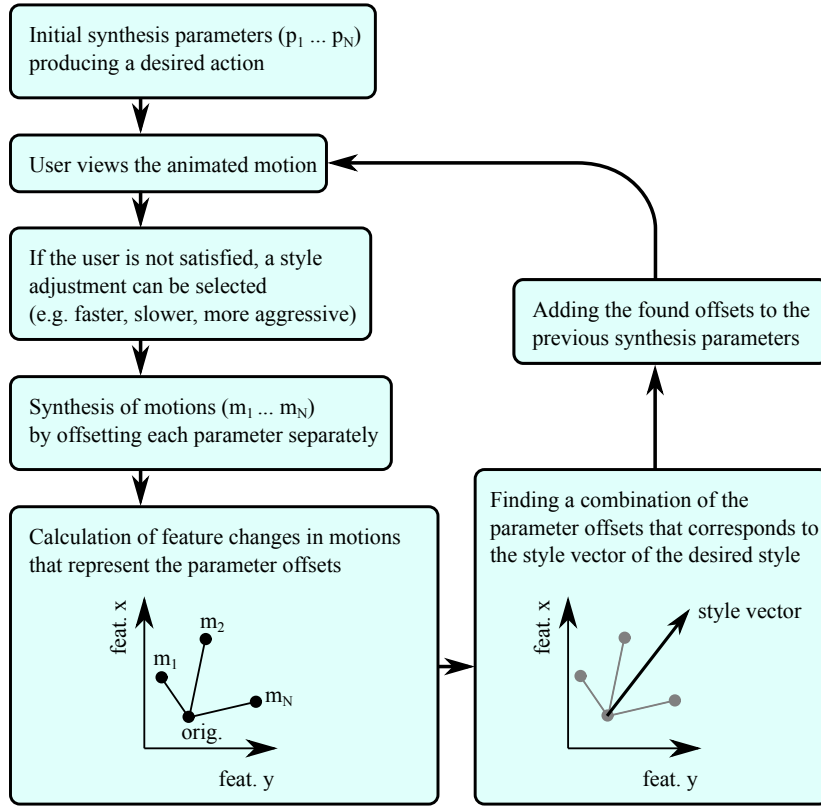


Figure 3.10. Process of controlling a style of single character motion

3.4.2 Evaluation of relative style control

For every method that claims to produce motions with specific styles, an evaluation is needed to validate the claims. Furthermore, it is important to have people who have not been involved in creation of the motions performing the perceptual validations. As explicated in Section 2.7, this has not been done for most of the published methods that synthesize motions with varying styles. The work presented in Publication V seeks to turn the trend around.

To validate the method for controlling style, two experiments with human participants and one numerical assessment were performed. The two experiments were based on crowdsourced questionnaires as that allowed having hundreds of participants. This was deemed necessary as the definitions of the styles were based on opinions of only one annotator.

The first experiment aimed at testing the accuracy of the style vectors when they are used in detecting style differences from previously unseen motions performed by different actors. New sets of motions, similar to

those used in the creation of the style definitions, were performed by four actors. The motions were then presented to participants as pairs in a questionnaire to create a ground truth on what style differences humans can see between the motions. Styles of the same motion pairs were evaluated using the style vectors, and the results of the evaluation were compared to the ground truth. This comparison showed that differences in styles fast, slow, aggressive, lazy, excited, energetic, calm, limping, healthy, depressed, and busy were successfully detected with accuracies over 90% while the chance level was 50%. This result shows that the used approach generalizes from one actor to others. The styles relaxed and tense had accuracies 77.1% and 59.5%, respectively. These accuracy levels were deemed too low to be useful in practice, and the two styles were pruned from further validations.

The next validation was a numerical test aimed at exposing the effects of the elimination of incidentally correlating features. In this test, correlations between the style vectors were calculated in two cases, once without the elimination procedure (Fig. 3.11) and once with it (Fig. 3.12). Comparing these two cases shows that without the elimination the style vectors are heavily correlated, and can be divided into two groups where one group points towards faster style and another towards slower style. Using the elimination procedure creates much lower correlations between the style vectors, thus making more refined control of styles possible. Also, correlations that remain after the elimination, such as a positive correlation between styles slow and lazy, and a negative correlation between styles limping and healthy, are reasonable as the meanings of the words are tightly linked.

The third validation explored style control in a practical case where interpolation synthesis was used. The test setting contained 35 starting motions that were an even sample from the parameter space. From each of these 35 motions, the style control system was used to create new motions towards the following eight styles: limping, healthy, depressed, slow, calm, aggressive, busy, and fast. Next, the initial walking motions and the adjusted versions were compared pairwise in a crowdsourced questionnaire to find what style differences people actually see in the pairs. The novelty of this test setting is that it does not force the participants to choose between the styles, and the choice is not forced even implicitly by giving a long list of alternatives, but in each comparison only one style was given.

	limping	healthy	depressed	slow	lazy	calm	aggressive	energetic	busy	excited	fast
limping	1.0	-0.9	0.7	0.8	0.9	0.5	-0.3	-0.6	-0.8	-0.6	-0.8
healthy	-0.9	1.0	-0.4	-0.7	-0.6	-0.5	0.4	0.6	0.6	0.5	0.7
depressed	0.7	-0.4	1.0	0.8	0.8	0.2	-0.2	-0.5	-0.6	-0.2	-0.7
slow	0.8	-0.7	0.8	1.0	1.0	0.7	-0.6	-0.8	-0.9	-0.6	-1.0
lazy	0.9	-0.6	0.8	1.0	1.0	0.6	-0.4	-0.7	-0.9	-0.6	-0.9
calm	0.5	-0.5	0.2	0.7	0.6	1.0	-0.9	-0.9	-0.7	-0.8	-0.8
aggressive	-0.3	0.4	-0.2	-0.6	-0.4	-0.9	1.0	0.9	0.5	0.6	0.6
energetic	-0.6	0.6	-0.5	-0.8	-0.7	-0.9	0.9	1.0	0.8	0.6	0.8
busy	-0.8	0.6	-0.6	-0.9	-0.9	-0.7	0.5	0.8	1.0	0.7	0.9
excited	-0.6	0.5	-0.2	-0.6	-0.6	-0.8	0.6	0.6	0.7	1.0	0.7
fast	-0.8	0.7	-0.7	-1.0	-0.9	-0.8	0.6	0.8	0.9	0.7	1.0

Figure 3.11. Correlations between style vectors from Publication V without elimination of incidental features. Correlations stronger than ± 0.15 are in green and red backgrounds, and correlations stronger than ± 0.5 are in bright versions of the colors.

	limping	healthy	depressed	slow	lazy	calm	aggressive	energetic	busy	excited	fast
limping	1.0	-0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1
healthy	-0.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
depressed	0.0	0.0	1.0	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.1
slow	0.0	0.0	0.1	1.0	0.7	0.0	0.1	-0.1	-0.1	-0.3	-0.6
lazy	0.0	0.0	0.2	0.7	1.0	0.4	-0.3	-0.6	-0.4	-0.6	-0.7
calm	0.0	0.0	0.0	0.0	0.4	1.0	-0.7	-0.7	-0.3	-0.3	-0.2
aggressive	0.0	0.0	0.0	0.1	-0.3	-0.7	1.0	0.7	0.4	-0.1	-0.1
energetic	0.0	0.0	0.0	-0.1	-0.6	-0.7	0.7	1.0	0.6	0.2	0.2
busy	0.0	0	0.0	-0.1	-0.4	-0.3	0.4	0.6	1.0	0.3	0.3
excited	-0.1	0.1	0.1	-0.3	-0.6	-0.3	-0.1	0.2	0.3	1.0	0.8
fast	0.0	0.0	0.0	-0.6	-0.7	-0.2	-0.1	0.2	0.3	0.8	1.0

Figure 3.12. Correlations between style vectors from Publication V after the elimination of incidental features.

The results of the third validation shown in Figure 3.13 tell that all the intended styles were perceived as more intense. However, simultaneous changes in several other styles were also seen in many cases. Pairs of styles that can be considered opposites, such as fast and slow, explain part of the simultaneous changes. The remaining simultaneous changes highlight the importance of a versatile synthesis method, as even a system that could perfectly detect the differences between styles does not help if the controlled synthesis method cannot produce all the styles independently from each other. Thus, interpolation synthesis can limit the

achieved separation of styles as it does not enable free transfer of styles from one motion to another, but all the produced motions are between existing examples.

		Perceived Styles							
		limping	healthy	depressed	slow	calm	aggressive	busy	fast
Intended Styles	limping	0.3	-0.6	-0.1	0.2	0.0	-0.2	-0.2	-0.3
	healthy	-0.2	0.1	-0.1	0.0	0.1	0.0	0.0	0.0
	depressed	0.6	-0.4	0.4	0.3	-0.1	-0.3	0.1	-0.3
	slow	0.2	-0.7	0.6	0.6	0.5	-0.5	-0.3	-0.6
	calm	-0.1	-0.2	0.4	0.6	0.5	-0.5	-0.5	-0.4
	aggressive	0.1	0.4	-0.3	-0.2	-0.7	0.8	0.3	0.2
	busy	0.1	0.1	-0.3	-0.6	-0.5	0.4	0.5	0.5
	fast	-0.2	0.6	-0.3	-0.7	-0.5	0.5	0.3	0.6

Figure 3.13. Changes in mean ratings from an evaluation of style adjustments from Publication V where original range was from -2 to 2. Numbers on white do not statistically differ from zero ($p=0.05$), significant positive differences are green and significant negative differences red.

Based on the evaluations it could be argued that reducing correlations between style vectors using the feature elimination would not be an optimal approach. A possible alternative could be to treat the style vectors solely as mathematical entities and to make them orthogonal with small offsets pushing them gradually apart. However, this approach could destroy also meaningful relationship such as the negative correlation between styles limping and healthy (Fig. 3.12). Without relying on the data, it would be impossible to know if preserving the correlation between limping and healthy is more important than preserving a correlation of equal strength between styles limping and lazy (Fig. 3.11). A similar argument speaks against using PCA components as vectors that represent style words. Since PCA components are always orthogonal to each other, they cannot be used to represent styles such as slow and lazy which from a semantical point of view can be expected to be partially correlated with each other.

3.5 Discussion

Human motion has been viewed in this chapter as a phenomenon containing continuously varying styles with possibility of several styles occurring simultaneously. Also, a method for controlling motion style of a single

character has been presented that is compatible with this view about human motion. The method is based on relative definitions of styles built from comparisons between motions. A similar comparative approach has been previously used for example in measuring values such as honesty and responsibility as they can be represented as choices between given alternatives [2]. Motion style and values can be seen as similar phenomena as it may be difficult to rate them with absolute scales. At least, ratings of motion styles are likely to be more fuzzy and subjective than classifying actions visible in human motion.

The relative definitions allow styles to be well defined regardless of the amount of style seen from an individual motion. This is different from the views presented in the context of felt affects [72] and words describing styles [19] where affects and style-related adjectives have been considered to be well defined only near the extremes. The compact nature of the relative definitions implies that they could be the actual way people model styles mentally. For example, a relation such as ‘slower’ or ‘more aggressive’ needs to be learned only once and can be used in the context of many actions, while thinking of styles as regions in a conceptual spaces [19] would require learning the combinations of each style and action separately.

While the speculation about mental models is not in the core of this dissertation, the same reasoning applies to using action recognition methods for recognition of styles. The action recognition methods model actions as regions of a numerical feature space. Applying this conceptualization to styles would divide them to several clusters as was noted in Section 3.3. This would again require having separate examples for each combination of style and action to learn all occurrences of styles.

While adjusting style with relative steps from an initial motion was shown to be possible in Publication V, it can be asked if the approach is the most intuitive way to adjust styles. An alternative way is to define styles through examples and model them as dimensions that could be adjusted for example with sliders [79]. This approach also enables extrapolation of new styles that combine predefined styles [10] as is illustrated on the left side of Figure 3.14. A potential problem emerges when a combination of styles turns out to look unnatural as illustrated on the right side of Figure 3.14. In that case, it is not possible to determine a globally applicable minimum and maximum for a style, but rather the allowed ranges would depend on other styles undermining the intuitiveness of the

control mechanism.

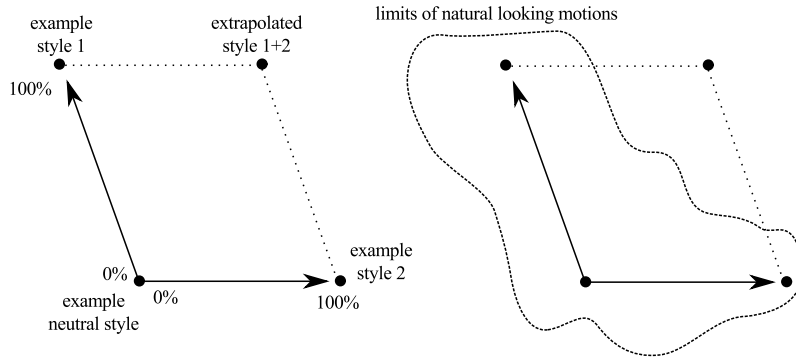


Figure 3.14. On the left is an example-based extrapolation scheme, on the right is an example where motions fall outside the limits of natural-looking motion.

A question related to synthesis of styles is why the produced styles have not been commonly validated with independent observers. Based on the questionnaires in Publications I and V, it is apparent that perceptual evaluations can give important information as all the tested styles were not successfully perceived. A possible reason is that a commonly used and easily reproducible methodology for evaluating style content does not exist. This is possible as for example the evaluations in Publication V would have been much more difficult to run without the crowdsourcing platforms created in recent years. A more pessimistic possibility is that styles are viewed as hazy and subjective concepts, and that it is easier to leave the issue to an artist than trying to give styles more concrete definitions.

While the presented research shows that embracing continuous and overlapping nature of styles is possible in practice, it is also apparent that the properties are not completely universal truths, but have their exceptions. The style fast is a good example of a continuous style dimension as in an animation it is almost always possible to increase velocity. A more complicated case would be the style natural. Starting from an unnatural motion and going towards a more natural style would likely reach a peak of naturalness, and after the peak further adjustment could turn the motion back to unnatural. A similar complication has been suggested to exist when describing motions as more or less symmetric [92]. Exceptions to the possibility of simultaneous appearance of styles can be found from cases that require the same body part to be in different poses or move in opposite manners.

The novel sets of features presented in Publications IV and V can be useful in recognition of motion styles. While the features presented in this

dissertation have been used in previous publications, they have not been used together. Also, the specific focus on motion style instead of actions and gestures separates the presented feature sets from other published feature sets [75]. However, it is not certain if the features cover all sides of motion styles, and it might be possible to get the same performance with a smaller set of features. Also, the features may not work outside the context of motion style. For example, the normalization used in Publication V that made the left and the right sides of body numerically equal can harm detection of symbolic gestures as the left hand is considered tainted in some cultures.

In the end, adjusting motion style of a single character can be useful in animations, but it does not solve all cases as interpretations of styles may vary depending on behavior of other characters. This side of motion style is elaborated in the next chapter.

4. Styles in interaction between characters

4.1 Background

This chapter presents research about motion style that emerges from interactions between characters as presented in Publications II and III. The common themes in the publications are interpretations of motion styles when modulated by the context of the motion, and treating motion style as an aspect that can vary continuously in time. The goal is to create a framework that allows maximal expression through motion style without limitations set by other modalities such as speech, facial expressions or symbolic gestures. More specifically, instead of a turn-based approach where behavior is planned over long time spans [95], a model that allows the behavior to be in a continuous flux is used. First, a general framework is designed and shown to be feasible with a proof-of-concept system. Then an extension of the system is presented with more attention given to the process of authoring the interaction.

4.2 Experiment on continuous bodily interaction

Publication II presents and evaluates a proof-of-concept system that enables bodily interaction between a human and a virtual character. The idea is to take a basic interaction loop (Fig. 4.1) and to view it as an enactive system. In practice, three aspects of the enactive paradigm [17] are emphasized. The first aspect is the aim to create interaction containing a continuously flowing stream of actions instead of a series of discrete actions. The second aspect is to avoid turn-based action and reaction, but to have the roles blur together. The third aspect is to enable sense-making with the interaction, in other words, the interaction should result in in-

creased understanding of the other party. In Publication II, the technical implementation of such a system in the context of human motion is explored, and the nature of the resulting sense-making is evaluated with an interview of participants trying out the system.

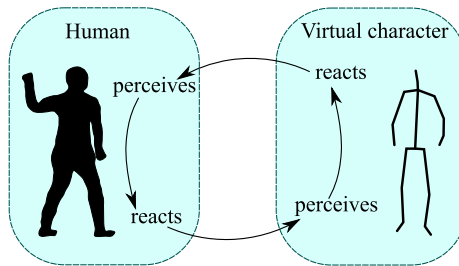


Figure 4.1. A basic loop of continuous interaction between a human and a virtual character



Figure 4.2. Bodily interaction between a human and a virtual character from Publication II

The actual setup had a virtual character projected on a large screen and a human participant wearing a motion capture suit as is shown in Figure 4.2. The motions of both parties were represented numerically with two motion descriptors that aim to be abstractions understandable to humans. The first descriptor was Quantity of Motion (QoM) [13] that is an estimate of the total amount of movement. The second descriptor was distance from the imaginary glass wall separating the parties.

Behaviors for the virtual character were created by mapping observed descriptor values of a human to desired descriptor values of the virtual character. An example of a single mapping is given in Figure 4.3. In practice, a user interface was created that has visual representations of the descriptor spaces for both the human and the virtual character as is shown in Figure 4.3. Creating a single mapping was done by clicking related points with a mouse. A fully working behavior rule required that the descriptor space of the human was evenly covered which means in

case of two dimensions a minimum of four mappings from the corners of the input space. Mappings for the intermediate input values were created by interpolation.

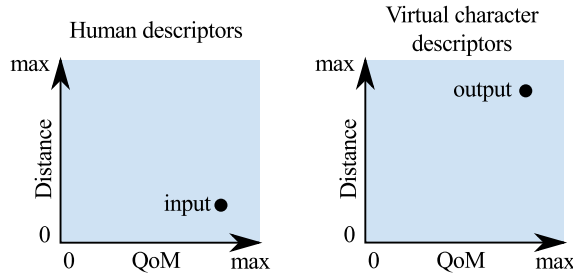


Figure 4.3. Example of an interaction mapping where fast movement of a human at a close distance (input) makes a virtual character want to perform fast movements at a long distance (output).

To realize the desired descriptor values with an animation, motion synthesis based on a graph [46] was used. The graph contained variedly performed motions including walking, jumping, and running that were segmented to approximately one second long clips. Synthesis of the motions was done automatically by evaluating alternative sequences from the graph and playing the one that matched best the desired descriptors.

To test the approach, seven volunteers were asked to engage in free interaction with the virtual character. Six different interaction rules were presented to the participants. Three of the rules were versions of imitation where the QoM of the virtual character was either unrestricted, or limited to only low values or to only high values. The fourth rule mapped the QoM of the human to the Distance of the virtual character making the character back off when the human did motions with high QoM. The fifth rule inverted the QoM of the human for the virtual character, and made the character respond by high QoM, for example by jumping and waving hands, when the human is standing still. The sixth case had the virtual character perform completely random motions regardless of what the human did.

Interview of the participants revealed that the interaction rules were perceived as different attitudes. In particular, taking a step backwards after fast movement from the human was described as scared behavior. Also, the participants said that they occasionally imitated the actions of the virtual character. These observations suggest that the implemented system did enable enaction between the participants and the virtual characters. However, drawing more detailed conclusions related to the inter-

action is not possible as the unguided interaction resulted in great variation between the actions of the participants. This highlights the need for further development of experiment methodology related to enaction.

From a technical point of view, the proposed framework was satisfactory as it separates the implementation of the motion synthesis and the mechanism for making behavior rules for the virtual character. This allowed both parts to be developed separately, and to be reused even when one of parts would be completely replaced. The graph-based motion synthesis introduced at times too much lag which caused the virtual character to be reacting to old events thus making its behavior incomprehensible. This highlighted the need to optimize the motion graph [78, 103], and to include interpolations between parallel motions [83, 32] to enable more responsive behaviors.

From a point of view of an animator, the method for creating behavior rules was usable, but the range of possible behaviors was limited by having only two motion descriptors as that is a too simplistic model for human motion. Also, extending the method to cases with a higher number of descriptors was seen as potentially problematic as visualizing high dimensional mappings may not be possible in a user friendly way. These sides of the framework were developed further in Publication III.

4.3 Authoring expressive interaction

Publication III builds on the framework that enables bodily interaction between a human and a virtual character, as presented in Publication II. The range of possible behaviors is expanded by allowing the virtual character to perform more varied motions, and by adding new numerical motion descriptors that enable more precise control of the behaviors. The publication also presents a method for authoring interaction rules by defining them through recorded actions and reactions, thus solving several problems related to modeling interaction in a high dimensional feature space.

In Publication II, the virtual character was restricted to a small volume visible through a projected display. This restriction was removed in Publication III and the character was allowed to stand, walk, turn, jump and generally move around on a flat floor. The motions were synthesized with the same motion graph based approach as in Publication II. To enable control of the new behaviors, motion descriptors for turning left/right and

moving forward/backward were added. Also, a new version of Quantity of Motion (QoM) called Non-transitional QoM (NtQoM) was introduced that estimates the energy used for body language or other expressive motions, disregarding locomotion. Examples of the high and low values of the descriptors are shown in Figure 4.4.

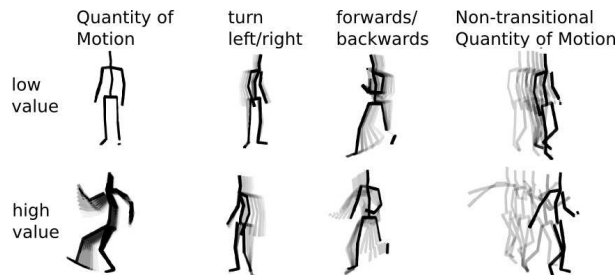


Figure 4.4. Examples of motion descriptors calculated from a single character

The less restricted interaction allowed new types of relationships between a human and the virtual character. These were represented with the descriptors for distance between the characters, facing angle, and approach/retreat as is visualized in Figure 4.5.

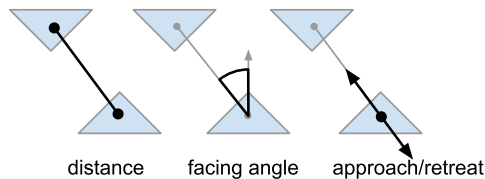


Figure 4.5. Examples of motion descriptors that represent relations between characters

While the new descriptors allowed more precise control of the motion, they also made creating behavior rules with the mouse driven interface difficult. The three most pressing problems were the difficulty of conceptualizing the numerical values as concrete motions, the combinatorially increased amount of required mappings per behavior rule, and the increased chance to set descriptor values that are not realizable as a physically plausible motion. These problems may explain why purely machine learning based continuous interaction authoring has been proposed before [39], but systems allowing animators to edit the interaction rules have not been published. The problems were solved by introducing a new way to author behavior rules that is based on recorded actions and reactions between humans such as shown in Figure 4.6.

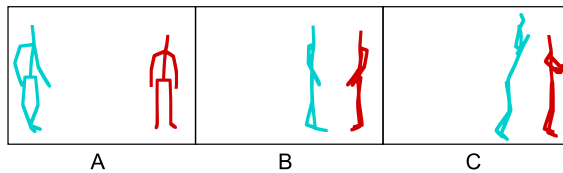


Figure 4.6. An action and reaction sequence with the red character representing a desired reaction of a virtual character to the action of the cyan character, where the reaction is to do nothing when the distance is high (A), to turn towards the other character when distance is low (B), and to turn away when the other character is aggressive (C).

Creating the behavior rules was done by an animator, who first needed to identify moments that are good examples of the desired behavior. Then the animator had to decide which descriptors are relevant in the selected examples and which may vary freely. After this the system can pick the values for the descriptors from the motion data, and the rule is ready. Technically, the rules were implemented with Radial Basis Functions (RBFs) [12] as they enable interpolation of the scattered data points. A scaling factor was added to the RBFs to enable adjusting the amount of effect each dimension has on the results.

The method was shown to work by creating behaviors such as taking interest in another character, but turning away when the other performs aggressive motions as in Figure 4.6. Also, a behavior rule that made the virtual character actively start the interaction was demonstrated. This shows a definite increase to the range of behaviors compared to the previous version of the system introduced in Publication II.

The problem of an animator having to conceptualize numerical descriptor values as concrete motions is removed by the new method for authoring behaviors as the example motions provide the exact values. The example motions also prevent the user from giving physically impossible combinations of descriptor values. Furthermore, the ability to allow part of the descriptors vary freely in the behavior rules breaks the curse of dimensionality. In other words, the amount of required mappings does not grow combinatorially in relation to the total number of motion descriptors. It can be concluded that the presented approach, which is between machine learning and manual authoring, allows creating many expressive behaviors while still requiring only a small number of example motions.

4.4 Discussion

A framework for creating a continuous loop of expressive bodily interaction between a virtual character and a human has been presented in this chapter. Furthermore, the framework has been extended with a method for authoring interaction rules based on acted example motions. While the method has been demonstrated with the more unpredictable case of interaction between a human and a virtual character, the system can also be used for creating interaction between two virtual characters.

The interviews with participants interacting with the system in Publication II and the examples of interaction in Publication III show that the proposed approach can create interactive behaviors interpreted for example as emotional expressions. However, performing more user experiments could be useful in finding out the full range of possible styles and determining the usability of the authoring method in practical cases. Also, an obvious improvement for the framework would be to include the numerical definitions for natural language based styles of a single character from Publication V. This would allow working with styles such as ‘aggressive’ and ‘depressed’ instead of more abstract descriptors such as Quantity of Motion.

The presented evaluations of expressive interaction are less rigorous than the ones used in case of style perceived from single characters. The main additional challenge is in repeatability of the interaction as the human participants are unlikely to act twice in exactly the same way. This prevents performing similar side-by-side comparisons of virtual characters as were used in Publication V. It is possible to record the interactions, and view them later on from a third-person point of view, but this may not give the same impressions.

Authoring interaction with examples of recorded actions and reactions is a novel approach in the presented research. While actions and reactions have been used before [39], the presented approach can be used with less examples as part of the learned behavior relies on expertise of a human observer. The approach also enables tweaking the learned behavior if the example data does not match exactly with the desired behavior. The method for authoring interaction requires that the used modality allows continuous interaction, but is not limited to bodily motion. Therefore, the method could be reused for example in the context of facial expression or tone of speech.

Generally speaking, the presented interaction framework emphasizes tight coupling between a virtual character and a human participant. In other studies, coupling through bodily motions has been used with the goal of creating more co-presence and rapport, and thus a more pleasant interaction [8, 77, 37]. A difference here is that, in the presented work, negative emotions such as being scared of someone are also considered. The idea behind this is that coupling and synchronization of movements may create a sense of co-presence, but whether the felt presence is positive or negative may also depend on other aspects such as perceived motion style.

The bodily interaction allowed by the framework is low-level and can even be described as almost animalistic. While the immediate bodily reactions can be seen as a foundation of human behavior, more high-level behaviors are also needed for building a virtual character that resembles a real human. A technical step towards this direction would be to allow varying emotional states which all would have their own interaction rules, and allowing different behaviors towards different people. Also, an ability to perform and to recognize culture dependent symbolic gestures would be expected from an imitation of a real human.

As interaction with a virtual character is often considered to be multi-modal, merging the proposed bodily interaction framework with existing multimodal frameworks [95] is needed for building practical applications. A major difference between the approaches seems to be in planning behaviors over time. While the proposed framework takes full benefit of expressive motion styles by allowing them to vary continuously over time, other modalities such as spoken utterances or symbolic gestures need to be given longer blocks of time. Furthermore, a suggested way to enable continuous interaction with existing multimodal frameworks is to add support for interrupting a chosen behavior [105]. The interruptions do not make sense if style is viewed as a continuously changing aspect of motion. A possible solution could be to have a tight interaction loop in parallel to a loop that assigns behaviors to modalities requiring longer blocks of time. This could reduce the frequency of required interruptions as part of them could be replaced with modulation of the style.

5. Conclusions

This dissertation presents two novel methods that start from the capture of acted motions and produce expressive performances as the final result. The methods work in the contexts of motion style of a single character and styles emerging from interaction between characters. By applying these methods it is possible to produce virtual characters that can interact fluidly while still allowing the expressiveness of their motions to be controlled. This enables virtual characters to simulate human interaction more naturally which can be useful in games that contain character developments. More serious applications include practicing real-life situations containing emotional interactions between humans.

The results imply that gaining full expressive power of motion style requires treating it as equally important as other modalities such as facial expressions and speech. Also, artificial limitations such as only considering emotional styles or controlling styles with mechanisms designed for less continuous and fluid modalities should be avoided. For animations of virtual characters, this means that the impact of motion style should not be limited to modulation of actions only, but the ranges of styles an action allows should be considered already when selecting the action to be performed.

The work draws attention to two aspects of motion style that should be taken into account in design of systems including human motion. The first is the relative nature of verbal descriptions of styles that can be numerically modeled through comparisons between motions. This allows the styles to be well defined even when a style of a motion is not intense enough to be easily described with absolute terms. The second important aspect of motion style is its continuous and fluid nature over time. This has impact on the design of multimodal interaction as fluidly changing style is possible only if a tight interaction loop is built between the parties

of the interaction.

There are two limitations on the generalizability of the results. The first is the limited range of considered actions as the main focus has been on locomotion. The second is that the final version of the bodily interaction framework was not fully tested with users. Also, it would be always possible to get more reliable results by having more actors, annotators and evaluators in the experiments.

In the future, tools for animation of single character motion can be refined by combining the control mechanism presented in this work with new synthesis methods. The methods could be based on, for example, extrapolation of style differences, algorithmic modifications to motions, and physics-based simulation of motion. Tools for authoring motion based behaviors for interactive characters can also be improved. The presented interaction framework could be extended to allow several emotional states that appear as different behaviors. Furthermore, allowing different behaviors towards different people, and blending the behaviors when facing a group of people could also make the interaction more realistic. A final challenge would be to integrate the interaction framework based on bodily motion seamlessly with other expressive modalities such as speech and facial expressions.

Bibliography

- [1] Agrawal, S., Shen, S., van de Panne, M.: Diverse motion variations for physics-based character animation. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 37–44. SCA '13, ACM, New York, USA (2013)
- [2] Alwin, D.F., Krosnick, J.A.: The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly* 49(4), 535–552 (1985)
- [3] Amaya, K., Bruderlin, A., Calvert, T.: Emotion from motion. In: Graphics Interface. pp. 222–229 (1996)
- [4] Arikan, O.: Compression of motion capture databases. In: ACM SIGGRAPH 2006 Papers. pp. 890–897. SIGGRAPH '06, ACM, New York, USA (2006)
- [5] Arikan, O., Forsyth, D.A., O'Brien, J.F.: Motion synthesis from annotations. In: ACM SIGGRAPH 2003 Papers. pp. 402–408. SIGGRAPH '03, ACM, New York, USA (2003)
- [6] Aviezer, H., Hassin, R.R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., Bentin, S.: Angry, disgusted, or afraid?: Studies on the malleability of emotion perception. *Psychological Science* 19(7), 724–732 (2008)
- [7] Bernhardt, D., Robinson, P.: Detecting affect from non-stylised body motions. In: Paiva, A., Prada, R., Picard, R. (eds.) *Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, vol. 4738, pp. 59–70. Springer Berlin Heidelberg (2007)
- [8] Bevacqua, E., Stanković, I., Maatallaoui, A., Nédélec, A., De Loor, P.: Effects of coupling in human-virtual agent body interaction. In: Bickmore, T., Marsella, S., Sidner, C. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 8637, pp. 54–63. Springer International Publishing (2014)
- [9] Blumberg, B.M., Galyean, T.A.: Multi-level direction of autonomous creatures for real-time virtual environments. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques. pp. 47–54. SIGGRAPH '95, ACM, New York, USA (1995)
- [10] Brand, M., Hertzmann, A.: Style machines. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques.

- pp. 183–192. SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., New York, USA (2000)
- [11] Bruderlin, A., Williams, L.: Motion signal processing. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques. pp. 97–104. SIGGRAPH '95, ACM, New York, USA (1995)
 - [12] Buhmann, M.D.: Radial basis functions: theory and implementations, vol. 12. Cambridge University Press (2003)
 - [13] Camurri, A., Lagerlöf, I., Volpe, G.: Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies* 59(1-2), 213–225 (2003), applications of Affective Computing in Human-Computer Interaction
 - [14] Castellano, G., Mancini, M., Peters, C., McOwan, P.: Expressive copying behavior for social agents: A perceptual analysis. *Systems, Man and Cybernetics, Part A: Systems and Humans*, IEEE Transactions on 42(3), 776–783 (May 2012)
 - [15] Chi, D., Costa, M., Zhao, L., Badler, N.: The emote model for effort and shape. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. pp. 173–182. SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., New York, USA (2000)
 - [16] Clavel, C., Plessier, J., Martin, J.C., Ach, L., Morel, B.: Combining facial and postural expressions of emotions in a virtual character. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 5773, pp. 287–300. Springer Berlin Heidelberg (2009)
 - [17] De Jaegher, H., Di Paolo, E.: Participatory sense-making. *Phenomenology and the Cognitive Sciences* 6(4), 485–507 (2007)
 - [18] Esteves, C., Arechavaleta, G., Pettré, J., Laumond, J.P.: Animation planning for virtual characters cooperation. In: *ACM SIGGRAPH 2008 Classes*. pp. 53:1–53:22. SIGGRAPH '08, ACM, New York, USA (2008)
 - [19] Gärdenfors, P.: A semantic theory of word classes. *Croatian Journal of Philosophy* (41), 179–194 (2014)
 - [20] Geijtenbeek, T., Pronost, N.: Interactive character animation using simulated physics: A state-of-the-art review. *Computer Graphics Forum* 31(8), 2492–2515 (2012)
 - [21] Glardon, P., Boulic, R., Thalmann, D.: Pca-based walking engine using motion capture data. In: *Computer Graphics International, 2004. Proceedings*. pp. 292–298. IEEE (2004)
 - [22] Gleicher, M.: Retargetting motion to new characters. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques. pp. 33–42. SIGGRAPH '98, ACM, New York, USA (1998)
 - [23] Gleicher, M.: Motion path editing. In: Proceedings of the 2001 Symposium on Interactive 3D Graphics. pp. 195–202. I3D '01, ACM, New York, USA (2001)

- [24] Grassia, F.S.: Practical parameterization of rotations using the exponential map. *Journal of Graphics Tools* 3(3), 29–48 (1998)
- [25] Grassia, F.S.: Motion editing: Mathematical foundations. In *course: Motion Editing: Principles, Practice, and Promise*, SIGGRAPH (2000)
- [26] Gratch, J., Marsella, S.: Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In: *Proceedings of the 5th International Conference on Autonomous Agents*. pp. 278–285. AGENTS '01, ACM, New York, USA (2001)
- [27] Grochow, K., Martin, S.L., Hertzmann, A., Popović, Z.: Style-based inverse kinematics. In: *ACM SIGGRAPH 2004 Papers*. pp. 522–531. SIGGRAPH '04, ACM, New York, USA (2004)
- [28] Hachimura, K., Takashina, K., Yoshimura, M.: Analysis and evaluation of dancing movement based on LMA. In: *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*. pp. 294–299 (Aug 2005)
- [29] Hämäläinen, P., Eriksson, S., Tanskanen, E., Kyrki, V., Lehtinen, J.: On-line motion synthesis using sequential monte carlo. *ACM Trans. Graph.* 33(4), 51:1–51:12 (Jul 2014)
- [30] Harnad, S.: The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42(1), 335–346 (1990)
- [31] He, Z., Liang, X., Wang, J., Zhao, Q., Guo, C.: Flexible editing of human motion by three-way decomposition. *Computer Animation and Virtual Worlds* 25(1), 57–68 (2014)
- [32] Heck, R., Gleicher, M.: Parametric motion graphs. In: *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games*. pp. 129–136. I3D '07, ACM, New York, USA (2007)
- [33] Heloir, A., Kipp, M., Gibet, S., Courty, N.: Evaluating data-driven style transformation for gesturing embodied agents. In: Prendinger, H., Lester, J., Ishizuka, M. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 5208, pp. 215–222. Springer Berlin Heidelberg (2008)
- [34] Hodgins, J.K., Wooten, W.L., Brogan, D.C., O'Brien, J.F.: Animating human athletics. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*. pp. 71–78. SIGGRAPH '95, ACM, New York, USA (1995)
- [35] Hsu, E., Pulli, K., Popović, J.: Style translation for human motion. *ACM Trans. Graph.* 24(3), 1082–1089 (Jul 2005)
- [36] Huang, H.H., Seki, Y., Uejo, M., Lee, J.H., Kawagoe, K.: Modeling the multi-modal behaviors of a virtual instructor in tutoring ballroom dance. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 7502, pp. 489–491. Springer Berlin Heidelberg (2012)
- [37] Huang, L., Morency, L.P., Gratch, J.: Virtual rapport 2.0. In: Vilhjálmsson, H., Kopp, S., Marsella, S., Thórisson, K. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 6895, pp. 68–79. Springer Berlin Heidelberg (2011)

- [38] Ikemoto, L., Forsyth, D.A.: Enriching a motion collection by transplanting limbs. In: Proc. of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 99–108. Eurographics Association (2004)
- [39] Jebara, T., Pentland, A.: Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In: Computer Vision Systems, Lecture Notes in Computer Science, vol. 1542, pp. 273–292. Springer Berlin Heidelberg (1999)
- [40] Jia, L., Yang, Y., Tang, S., Hao, A.: Style-based motion editing. In: Digital Media and its Application in Museum Heritages, Second Workshop on. pp. 129–134 (2007)
- [41] Johnson, K.L., McKay, L.S., Pollick, F.E.: He throws like a girl (but only when he’s sad): Emotion affects sex-decoding of biological motion displays. *Cognition* 119(2), 265–280 (2011)
- [42] Kaipainen, M., Ravaja, N., Tikka, P., Vuori, R., Pugliese, R., Rapino, M., Takala, T.: Enactive systems and enactive media: embodied human-machine coupling beyond interfaces. *Leonardo* 44(5), 433–438 (2011)
- [43] Kim, Y., Neff, M.: Component-based locomotion composition. In: Proc. of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 165–173. Eurographics Association (2012)
- [44] Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on* 4(1), 15–33 (2013)
- [45] Kovar, L., Gleicher, M.: Flexible automatic motion blending with registration curves. In: Proc. of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 214–224. Eurographics Association (2003)
- [46] Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques. pp. 473–482. SIGGRAPH ’02, ACM, New York, USA (2002)
- [47] Lasseter, J.: Principles of traditional animation applied to 3d computer animation. *SIGGRAPH Comput. Graph.* 21(4), 35–44 (Aug 1987)
- [48] Lau, M., Bar-Joseph, Z., Kuffner, J.: Modeling spatial and temporal variation in motion data. *ACM Trans. Graph.* 28(5), 171:1–171:10 (Dec 2009)
- [49] Lin, Y.H., Liu, C.Y., Lee, H.W., Huang, S.L., Li, T.Y.: Evaluating emotive character animations created with procedural animation. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 5773, pp. 308–315. Springer Berlin Heidelberg (2009)
- [50] Lindén, K., Carlson, L.: Finn wordnet-wordnet på finska via översättning. *LexicoNordica* 17(17) (2010)
- [51] Liu, C.K., Hertzmann, A., Popović, Z.: Learning physics-based motion style with nonlinear inverse optimization. *ACM Trans. Graph.* 24(3), 1071–1081 (Jul 2005)

- [52] Liu, G., Pan, Z., Li, L.: Motion synthesis using style-editable inverse kinematics. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 5773, pp. 118–124. Springer Berlin Heidelberg (2009)
- [53] McDonnell, R., Jörg, S., McHugh, J., Newell, F.N., O’Sullivan, C.: Investigating the role of body shape on the perception of emotion. *ACM Transactions on Applied Perception (TAP)* 6(3), 14 (2009)
- [54] Menache, A.: *Understanding motion capture for computer animation*. Morgan Kaufmann (2000)
- [55] Min, J., Chai, J.: Motion graphs++: A compact generative model for semantic motion analysis and synthesis. *ACM Trans. Graph.* 31(6), 153:1–153:12 (2012)
- [56] Min, J., Liu, H., Chai, J.: Synthesis and editing of personalized stylistic human motion. In: *Proc. of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games*. pp. 39–46. ACM (2010)
- [57] Mizuguchi, M., Buchanan, J., Calvert, T.: Data driven motion transitions for interactive games. In: *Eurographics 2001 Short Presentations*. vol. 2, p. 6 (2001)
- [58] Mori, H., Hoshino, J.: ICA-based interpolation of human motion. In: *Computational Intelligence in Robotics and Automation, 2003. Proc. 2003 IEEE International Symposium on*. vol. 1, pp. 453–458. IEEE (2003)
- [59] Müller, M., Röder, T., Clausen, M.: Efficient content-based retrieval of motion capture data. *ACM Trans. Graph.* 24(3), 677–685 (2005)
- [60] Mäkäpäinen, M., Kätsyri, J., Takala, T.: Exaggerating facial expressions: A way to intensify emotion or a way to the uncanny valley? *Cognitive Computation* 6(4), 708–721 (2014)
- [61] Neff, M., Fiume, E.: Modeling tension and relaxation for computer animation. In: *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. pp. 81–88. SCA ’02, ACM, New York, USA (2002)
- [62] Neff, M., Fiume, E.: From performance theory to character animation tools. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.) *Human Motion, Computational Imaging and Vision*, vol. 36, pp. 597–629. Springer Netherlands (2008)
- [63] Neff, M., Kim, Y.: Interactive editing of motion style using drives and correlations. In: *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. pp. 103–112. SCA ’09, ACM, New York, USA (2009)
- [64] Ortony, A., Turner, T.J.: What’s basic about basic emotions? *Psychological review* 97(3), 315 (1990)
- [65] Oshita, M.: *Smart motion synthesis*. In: *Computer Graphics Forum*. vol. 27, pp. 1909–1918. Blackwell Publishing Ltd (2008)

- [66] Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective multimodal human-computer interaction. In: Proceedings of the 13th Annual ACM International Conference on Multimedia. pp. 669–676. MULTIMEDIA '05, ACM, New York, USA (2005)
- [67] Parent, R.: Computer animation: algorithms and techniques. Morgan Kaufmann (2012)
- [68] Pasch, M., Poppe, R.: Person or puppet? the role of stimulus realism in attributing emotion to static body postures. In: Paiva, A.C., Prada, R., Picard, R.W. (eds.) Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science, vol. 4738, pp. 83–94. Springer Berlin Heidelberg (2007)
- [69] Pejsa, T., Pandzic, I.S.: State of the art in example-based motion synthesis for virtual characters in interactive applications. In: Computer Graphics Forum. vol. 29, pp. 202–226. Blackwell Publishing Ltd (2010)
- [70] Pelechano, N., Allbeck, J.M., Badler, N.I.: Controlling individual agents in high-density crowd simulation. In: Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 99–108. SCA '07, Eurographics Association, Aire-la-Ville, Switzerland (2007)
- [71] Perlin, K., Goldberg, A.: Improv: A system for scripting interactive actors in virtual worlds. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. pp. 205–216. SIGGRAPH '96, ACM, New York, USA (1996)
- [72] Picard, R.W.: Affective computing: challenges. International Journal of Human-Computer Studies: Applications of Affective Computing in Human-Computer Interaction 59(1–2), 55 – 64 (2003)
- [73] Pollick, F.E., Paterson, H.M., Bruderlin, A., Sanford, A.J.: Perceiving affect from arm movement. Cognition 82(2), B51–B61 (2001)
- [74] Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
- [75] Poppe, R., Van Der Zee, S., Heylen, D.K., Taylor, P.: Amab: Automated measurement and analysis of body motion. Behavior Research Methods 46(3), 625–633 (2014)
- [76] Pullen, K., Bregler, C.: Motion capture assisted animation: Texturing and synthesis. ACM Transactions on Graphics (TOG) 21(3), 501–508 (2002)
- [77] Reidsma, D., van Welbergen, H., Poppe, R., Bos, P., Nijholt, A.: Towards bi-directional dancing interaction. In: Harper, R., Rauterberg, M., Combetto, M. (eds.) Entertainment Computing - ICEC 2006, Lecture Notes in Computer Science, vol. 4161, pp. 1–12. Springer Berlin Heidelberg (2006)
- [78] Ren, C., Zhao, L., Safonova, A.: Human motion synthesis with optimization-based graphs. In: Computer Graphics Forum. vol. 29, pp. 545–554 (2010)
- [79] Rose, C., Cohen, M., Bodenheimer, B.: Verbs and adverbs: Multidimensional motion interpolation. Computer Graphics and Applications, IEEE 18(5), 32–40 (1998)

- [80] Rosen, D.: Animation bootcamp: An indie approach to procedural animation. (a keynote talk). In: Game Developers Conference Europe 2014. San Francisco, USA (17-21 March 2014), <http://www.gdcvault.com/play/1020583/Animation-Bootcamp-An-Indie-Approach>, accessed: 7th November 2014
- [81] Russell, J.A., Weiss, A., Mendelsohn, G.A.: Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology* 57(3), 493–502 (1989)
- [82] Shapiro, A., Cao, Y., Faloutsos, P.: Style components. In: Proceedings of Graphics Interface 2006. pp. 33–39. Canadian Information Processing Society (2006)
- [83] Shin, H.J., Oh, H.S.: Fat graphs: Constructing an interactive character with continuous controls. In: Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 291–298. SCA '06, Eurographics Association, Aire-la-Ville, Switzerland (2006)
- [84] Shoemake, K.: Animating rotation with quaternion curves. *SIGGRAPH Comput. Graph.* 19(3), 245–254 (Jul 1985)
- [85] Talbot, C., Youngblood, G.: Spatial cues in hamlet. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 7502, pp. 252–259. Springer Berlin Heidelberg (2012)
- [86] Taylor, R., Torres, D., Boulanger, P.: Using music to interact with a virtual character. In: Proceedings of the 2005 Conference on New Interfaces for Musical Expression. pp. 220–223. NIME '05, National University of Singapore, Singapore, Singapore (2005)
- [87] Tikka, P., Vuori, R., Kaipainen, M.: Narrative logic of enactive cinema: Obsession. *Digital Creativity* 17(4), 205–212 (2006)
- [88] Tilmanne, J., Dutoit, T.: Continuous control of style and style transitions through linear interpolation in hidden markov model based walk synthesis. In: Gavrilova, M., Tan, C. (eds.) *Transactions on Computational Science XVI, Lecture Notes in Computer Science*, vol. 7380, pp. 34–54. Springer Berlin Heidelberg (2012)
- [89] Tilmanne, J., Moinet, A., Dutoit, T.: Stylistic gait synthesis based on hidden markov models. *EURASIP Journal on Advances in Signal Processing* 2012(1), 72 (2012)
- [90] Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., Swartout, W.: Ada and Grace: Direct interaction with museum visitors. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 7502, pp. 245–251. Springer Berlin Heidelberg (2012)
- [91] Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision* 2(5), 371–387 (2002)
- [92] Troje, N.F.: Retrieving information from human movement patterns. Understanding events: How humans see, represent, and act on events pp. 308–334 (2008)

- [93] Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*. pp. 91–96. SIGGRAPH '95, ACM, New York, USA (1995)
- [94] Urtasun, R., Glardon, P., Boulic, R., Thalmann, D., Fua, P.: Style-based motion synthesis. *Computer Graphics Forum* 23(4), 799–812 (2004)
- [95] Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A., Pelachaud, C., Ruttkay, Z., Thórisson, K., van Welbergen, H., van der Werf, R.: The behavior markup language: Recent developments and challenges. In: Pelachaud, C., Martin, J.C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 4722, pp. 99–111. Springer Berlin Heidelberg (2007)
- [96] Wallbott, H.G.: Bodily expression of emotion. *European journal of social psychology* 28(6), 879–896 (1998)
- [97] Wang, J.M., Fleet, D.J., Hertzmann, A.: Multifactor gaussian process models for style-content separation. In: *Proceedings of the 24th International Conference on Machine Learning*. pp. 975–982. ICML '07, ACM, New York, USA (2007)
- [98] Welman, C.: Inverse kinematics and geometric constraints for articulated figure manipulation. Ph.D. thesis, Simon Fraser University (1993)
- [99] Witkin, A., Popovic, Z.: Motion warping. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*. pp. 105–108. SIGGRAPH '95, ACM, New York, USA (1995)
- [100] Wu, X., Ma, L., Zheng, C., Chen, Y., Huang, K.S.: On-line motion style transfer. In: Harper, R., Rauterberg, M., Combetto, M. (eds.) *Entertainment Computing - ICEC 2006, Lecture Notes in Computer Science*, vol. 4161, pp. 268–279. Springer Berlin Heidelberg (2006)
- [101] Xu, J., Takagi, K., Sakazawa, S.: Motion synthesis for synchronizing with streaming music by segment-based search on metadata motion graphs. In: *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. pp. 1–6 (2011)
- [102] Zhang, Z.: Microsoft kinect sensor and its effect. *MultiMedia, IEEE* 19(2), 4–10 (2012)
- [103] Zhao, L., Normoyle, A., Khanna, S., Safonova, A.: Automatic construction of a minimum size motion graph. In: *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. pp. 27–35. SCA '09, ACM, New York, USA (2009)
- [104] Zhuang, Y., Pan, Y., Xiao, J.: A Modern Approach to Intelligent Animation: Theory and Practice, chap. Automatic Synthesis and Editing of Motion Styles, pp. 255–265. Springer (2008)
- [105] Zwiers, J., van Welbergen, H., Reidsma, D.: Continuous interaction within the saiba framework. In: Vilhjálmsson, H., Kopp, S., Marsella, S., Thórisson, K. (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 6895, pp. 324–330. Springer Berlin Heidelberg (2011)

Errata

Publication I

- At the start of the section 2.2 the text “Neff and Kim state... ” should be “Neff and Fiume state...”.
- In acknowledgments, the character ä had been dropped out from the name of Timo Idänheimo.
- In the sentence of the fifth section "A similar ineffectiveness of retiming motions has been noticed earlier when trying to change the emotional content of captured motions [12].", the citation should be to Heloir et al. [6] instead of the [12].

Publication I

Klaus Lehtonen and Tapio Takala. Evaluating Emotional Content of Acted and Algorithmically Modified Motions. In *24th International Conference on Computer Animation and Social Agents (CASA 2011)*, Chengdu, China, Transactions on Edutainment VI, Lecture Notes in Computer Science, Volume 6758, pages 144-153, May 2011.

© 2011 Springer Science and Business Media.

Reprinted with permission.

Note: Due to copyright restrictions, this electronic version of the dissertation does not contain the final versions of the publications, but instead the versions that have not been edited by the publishers.

The final publications can be found from the web sites of the publishers.

Evaluating Emotional Content of Acted and Algorithmically Modified Motions

Klaus Lehtonen and Tapio Takala

Aalto University,
School of Science,
Department of Media Technology,
Espoo, Finland
klaus.lehtonen@tkk.fi, tapio.takala@tkk.fi

Abstract. Motion capture is a common method to create expressive motions for animated characters. In order to add flexibility when reusing motion data, many ways to modify its style have been developed. However, thorough evaluation of the resulting motions is often omitted. In this paper we present a questionnaire based method for evaluating visible emotions and styles in animated motion, and a set of algorithmic modifications to add emotional content to captured motion. Modifications were done by adjusting posture, motion path lengths and timings. The evaluation method was then applied to a set of acted and modified motions. The results show that a simple questionnaire is a useful tool to compare motions, and that expressivity of some emotions can be controlled by the proposed algorithms. However, we also found that motions should be evaluated using several describing dimensions simultaneously, as a single modification may have complex visible effects on the motion.

Keywords: computer animation, motion capture, emotional motion, evaluation of motion style, motion editing

1 Introduction

We can see emotions in facial expressions, poses and motions of animated characters. Motion capture is a convenient way to record the visible emotions. However, doing new motion capture for each different case requires time and money and is not practical in all situations [1]. New ways to reuse motion data can lower the costs. Reusing and modifying motions are especially beneficial in games, because game characters need to move and react in many different ways in real time.

Much research has been done to change the style of captured motions, but there has been less effort in validating the effects of the modifications. Many published modifications create changes in motions, but how significant the changes are to people viewing the motions has not been investigated [2–5]. In a case where motion style of modified motions was evaluated with many viewers, it was found out that the modification was not always noticed by the viewers [6].

We present a questionnaire based method for evaluating the emotional and stylistic content of motions with a large audience. This evaluation is used to compare the effects of three algorithmic modifications to motion data. These modifications change the posture, motion path length, and timing of the motions. The results show that a questionnaire is an effective method for measuring emotional content of motions and to compare the modification methods. The questionnaire can also reveal if a modification produces unwanted side effects to the style of the motions.

2 Related Works

2.1 Evaluating Motions

We found an earlier evaluation of modified motions that used valence, arousal and dominance as the dimensions in their questionnaire [6]. However, some stylistic properties like masculinity and femininity are not easily mapped to the three dimensions. In another study, dynamic and geometrical features of dancing motions were assessed using Laban motion analysis [7]. This approach requires expert knowledge from the evaluators. Enabling untrained people to evaluate everyday actions requires a simpler approach.

Evaluating emotions seen in facial expressions is a common practice [8]. Facial expressions are a simpler case than bodily motions, because the face is mainly used for communication and much less for other actions, whereas in bodily motion both appear. Questionnaires used to evaluate facial expressions often force labeling one face image with one emotion. This is a justified approach as prototypical faces are associated with basic emotions in commonly used models. Defining similar prototypical bodily motions is hard and, since motions can have many stylistic properties in addition to visible emotions, the task is even harder.

Based on these observations we decided to limit our study to assessing emotional content of bodily motions only. We animated motions with a simplified human model without face and fingers. In the questionnaire we asked evaluation of multiple simultaneous dimensions instead of a single choice.

2.2 Modifying Motions

Neff and Kim state that important components affecting motions are posture, transitions between poses, simplification and exaggeration [9]. We wanted to use modifications that would directly affect these components. Some of these aspects of motion can be changed in motions that are defined with a low number of keyframes per second (around 1 Hz) [10]. For example transitions between poses in keyframes can be easily edited without changing the keyframes themselves. Some methods cannot be used straight away with motion defined with a high number keyframes per second (over 10 Hz), such that motion capture produces, as transitions between the keyframes are too fast.

Many modification methods of motion styles are based on the difference between two recorded motion signals [2–4]. The way these methods treat the motion data differ greatly, but the results are all dependent on the styles of the input motions. In our study, we used a method based on differences between still poses for changing postures of characters. For other changes, however, we used procedurally generated modifications.

Bruderlin and Williams presented multiresolution filtering that we used for exaggerating and diminishing the length of motion paths [5]. The transitions between poses can be changed by the speed transform described by Amaya et. al. [2]. However, the speed transform is not a direct method as it requires two motions for defining the changes made to a third motion. To make the method more direct, we developed a new heuristic approach that defines the changes in timings based on the velocity of the input motion.

3 Implementation

3.1 Capturing Motions

To create suitable motions, we asked one female and one male actor to perform a short walk and a knocking motion with ten different styles that were used earlier by Pollick et. al.: *afraid, angry, excited, happy, neutral, relaxed, sad, strong, tired* and *weak* [11]. Both of the actors had acted in theater performances, but they did not have previous experience with motion capture. We also captured several standing poses. For recording the motions, we used Optitrack Full Body Motion Capture system, consisting of twelve cameras capturing motions with 100 frames per second. The system gives coordinates and orientation of the hips, and rotations for 18 joints.

3.2 Modifying Motion Paths

To change the length of the motion paths, multiresolution filtering described by Bruderlin and Williams was used [5]. The basic idea is to divide an original motion signal into frequency bands that are modulated separately in a way that adds no phase shift. Then we can return to the original signal format by summing up the modified frequency bands. We observed that multiplying the middle bands (between 1.0 and 12.5 Hz) makes the motion paths exaggerated or reduced.

With this method we can make a short and slow walk longer and faster or vice versa. During the process, the modification exaggerates or diminishes the poses as seen in Figure 1. Our hypothesis was that this modification would make motions look more *excited, tired* and *weak*. The hypothesis was based on the findings that high movement activity has been connected with *elated joy, hot anger* and *terror*, while low movement activity has been connected to *boredom* and *contempt* [12].

Multiresolution filtering suffers from stretching bones if applied to joint positions as input signals. Using joint rotations can also be problematic as unnatural rotations

may happen when the angles are near a gimbal lock. To avoid these problems, we represented the joint rotations with two orthogonal vectors pointing the direction of the bones. This way the bone lengths are not affected by the modification and gimbal lock is not an issue.

Multiplying frequency bands can make the feet of the character slide. This was fixed in post-processing by recalculating the coordinates of the hips. The fix was done by measuring the original supporting periods of the feet and forcing the feet to stay still during these periods in the new motions.

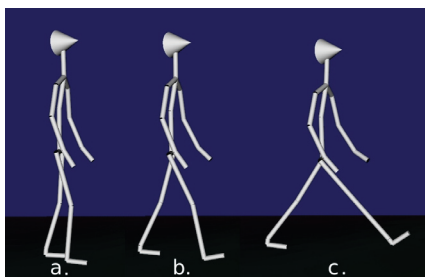


Fig. 1. Changes in pose when a walking motion (b) is modified by making motion paths shorter (a) and longer (c).

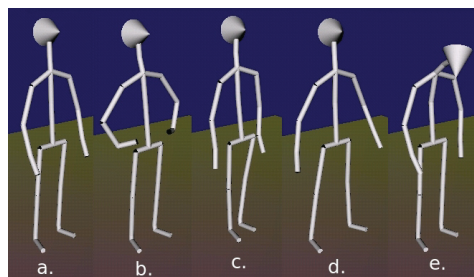


Fig. 2. Still poses used in changing posture of animated characters. Pose (a) is the neutral pose of the act

3.3 Modifying Posture

For creating motions with varying posture we need a motion with neutral posture, a neutral still pose and varying expressive poses (Fig. 2). Next we calculate the differences between joint rotations in the *neutral* pose and the other poses. Then we can add the desired difference to each frame of the motion to change the posture of the animated character. If we need to reduce or exaggerate the change of the posture, we can multiply all the changes with a constant. We made a hypothesis that modification to posture could make motions look more *sad* or more *confident*, as earlier studies show that those emotion types have stereotypical positions of upper body, shoulders and head [12].

With a motion like a regular walk the change of posture works quite well as it is. If the motion has parts where pose of the character is very different compared to the neutral pose, the technique can have unwanted side effects. For example, if the character is reaching forward with a straight arm, the arm can become twisted by the change of the posture. This can be fixed by fading the changes gradually out when the end of a limb goes too far from its position in the *neutral* pose. The fade-out must be done to all the joints that affect the position of the limb.

3.4 Modifying Timings

Amaya et al. adjusted the speed of motions by time warping the motions according to differences between two reference motions [2]. Their approach depends on the quality of the original motions and requires motion segmenting to find links between the reference motions. To bypass these requirements, we developed two heuristics that define the time warps without any reference motions. The heuristics aim to produce motions with either constant speed or added acceleration while keeping the total duration of the motion unchanged (Fig. 3). Our hypothesis was that these changes would make motions look more *relaxed* and *angry*, respectively. The hypothesis was based on findings that high movement dynamics are connected with *hot anger* [12].

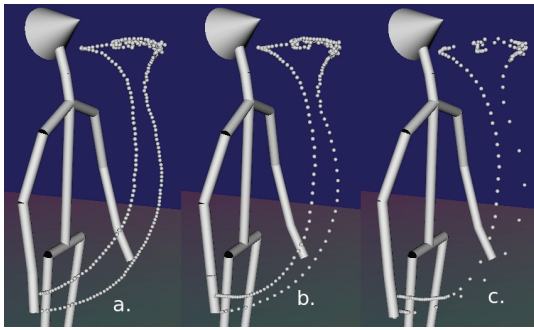


Fig. 3. Trajectories of the right hand during a knocking motion. Original motion (b) is modified to nearly constant speed (a) or with exaggerated acceleration (c).

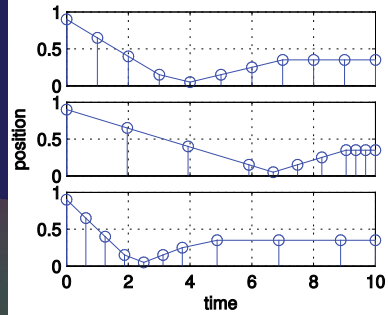


Fig. 4. Timewarping motion: evenly timed original frames (top) are repositioned in time to form constant speed (middle) or added acceleration (bottom).

Transforming a motion to constant speed can be done by taking evenly timed motion samples, spreading out frames with fast movement and compressing frames with slow movement, as depicted in figure 4. New motion is then formed by resampling between these as keyframes. The frames where the position does not change at all, would be compressed completely together to get constant speed, but this would cause all pauses of the motion to be skipped. More natural motion results if the pauses are preserved by enforcing a minimum duration between frames (as seen in the last four frames in the middle plot of figure 4).

In practice time warping is done separately for each limb of the character. When constant speed for each frame is desired, we can use the velocity values of the limb limited by a minimum value as the durations between frames and then scale all frames to preserve original total duration of the motion. When adding acceleration to a motion, we use original durations of the frames and increase the durations next to frames that are local minima of velocity and then scale as in the previous case. This makes the pauses of the motion last longer and other parts faster as seen in the lowest plot of figure 4.

3.5 Combining Modifications

To use the modifications interactively, the path lengths of an original motion are modified to produce a motion with short motion paths and another with long motion paths. Next, the sliding feet are fixed. Then, the timings of the three motions are modified to produce versions with constant speed and added acceleration. Once these steps are done we can create new motions by interpolating between the produced motions.

Modification of posture is done last to the interpolated motion. Modifying path lengths, fixing the sliding feet and modifying timings requires off-line processing. When they are done an animator controlling the system has real-time feedback as the interpolation of motions and changing the posture are both very fast operations.

4 Questionnaire

Evaluation of emotional and stylistic content of motions is necessary for comparing the motions and for testing methods that create modifications. We were also concerned with the validity of a questionnaire based evaluation. This led to three research questions: 1. Can acted styles and emotions be distinguished by viewing motions animated with a stick figure? 2. Do the three implemented modifications change emotions seen in the motions? 3. What are suitable dimensions to be rated when evaluating motions?

The captured acted motions were used as material for the first question. For the second question, we created pairs from the *neutral* motions to each of the acted motion styles. The pairs were created according to our hypotheses of the effects of the modifications and intended to have the same emotional and stylistic content as the acted motions. We also created motions with intended styles *masculine* and *feminine* that attempted to change the gender of the characters. Motion *energetic sadness* was created by combining modifications that we had hypothesized to increase *sadness* and *excitement*. The emotion *happy* was omitted because it was not known which modifications could affect *happiness*. The final combinations of the modifications are in figure 7.

40 videos were created from the acted and modified motions. The videos were shown in randomized order. The participants were able to play the videos many times, but it was instructed that viewing the videos once should be usually enough.

We included in the questionnaire all the emotional descriptions given to the actors. However, *neutral* was not included as it was assumed to fall in the middle of a scale. *Confident* was not part of the original set of emotions recorded with actors, but it was necessary to have a pair for *afraid*. To make the questionnaire simpler to answer, we combined opposite motion styles, making scales between pairs *sad-happy*, *tired-excited*, *angry-relaxed*, *weak-strong*, *afraid-confident* and *masculine-feminine*.

For each video, the participants were asked to evaluate how these adjectives describe the character in the video in a scale with five steps. The middle choice was the default and it was instructed to be used if neither of the alternatives feels good or if the participant is unsure which one is better.

The questionnaire was made with a server side PHP script and the videos were embedded in the web page as Flash objects. 28 non-paid participants were recruited through social media. 8 of them were female and 20 were male.

An answer to our first research question can be found by examining a confusion matrix between the intended styles and the perceived styles as shown in figure 5. Each acted style has the ratings of videos by both actors except the neutral male and neutral female columns, which show how the ratings for *neutral* motions differ between the actors.

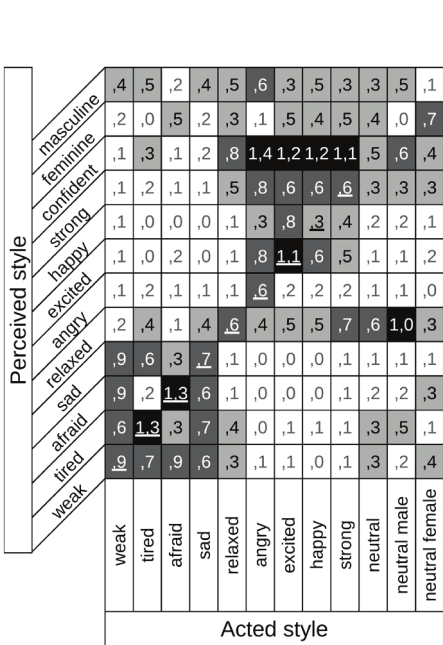


Fig. 5. Confusion matrix between acted and perceived styles, showing on average how much each style was seen in each video on a scale from 0 to 2. Underlined scores tell how well an intended style was recognized. Strongest perceptions are on dark background.

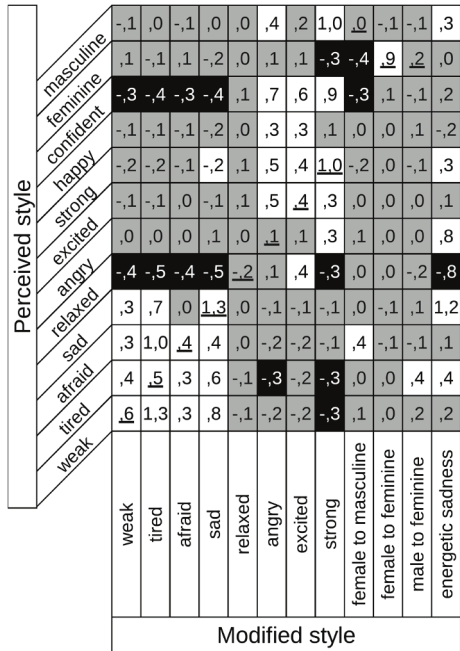


Fig. 6. Difference of scores between original and modified motions. Maximum possible change is 2. Positive changes higher than 0.25 are on white background and negative under -0.25 are on black background. Values for the intended styles are underlined.

The second research question calls for a comparison between ratings of the original *neutral* motions and the motions made by modifying them. The results are shown as differences of their scores in figure 6.

The third research question was about finding suitable characterizing dimensions to be rated when evaluating motions. We analysed how well the dimensions were chosen by estimating their common factors (using Matlab function *factoran*). Results in figure 8 show that the dimensions are not independent and some of them are effectively the same.

Modifications		Intended styles	
A	very short motion paths	afraid	B, F, G
B	short motion paths	angry	C, D
C	long motion paths	energetic	C, H
D	increased acceleration	excited	C, D
E	constant speed	feminine	F, G
F	hands close to body	masculine	I, J
G	feet close to each other	relaxed	E
H	head looking down	sad	B, H
I	feet more apart	strong	C, J
J	elbows broadly	tired	A, H
		weak	A, F, G

Fig. 7. Intended styles and the modifications done to achieve them.

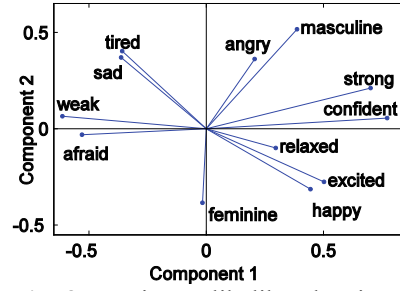


Fig. 8. Maximum likelihood estimate for common factors in the questionnaire. Loadings of the original dimensions are plotted into a two-dimensional model.

5 Discussion

The ratings of the acted videos in figure 5 tell that it is possible to see many styles from a motion that does not have hands or facial expressions. The styles are not seen one at a time, but rather a motion seems to fit a range of styles. This can be partly explained by close relations between the styles that are seen in figure 8, but it is also possible that motion alone does not separate all different styles and emotions as well as facial expressions do. Based on these findings, evaluation of motions requires multiple dimensions to be accurate and selecting more than one description should be allowed.

Figure 6 shows that the three modifications can change emotions and styles seen in motions. The effects of the modifications are not constrained to a single emotion or style. When we compare the intended effects and the perceived styles in figure 6, we can see that the *strong* and *sad* modifications worked well. A contributing factor could be that both modifications had postures that are easily identifiable. Modifications *weak*, *tired*, *afraid* and *excited*, created mainly with changes to motion paths, did add at least a little of the intended style, but the modifications *relaxed* and *angry* did not work as intended. *Anger* and *relaxedness* were hypothesized to be created with changes to the timings of the motions. It is possible that walking and knocking motions are not suitable material for the modification, but based on this data, the modification of timings has to be considered quite useless. A similar ineffectiveness of retiming motions has been noticed earlier when trying to change the emotional content of captured motions [12].

The modification *energetic sadness* in figure 6 created both *sadness* and *angriness*. The modification that was intended to create *sadness* did not create *angriness* and the only difference between the *sad* and the *energetic sad* modifications was in the length of motion paths. This suggests that the modifications do not have only one-to-one

relations with emotions, but combinations of modifications can be used to create emotions that the modifications cannot produce separately. This also suggests that the effects of the modifications depend greatly on the styles and emotions of the input motion.

The evaluation shows that modifications to posture and motion paths are good tools for artists, but it does not reveal if the modifications can be used without a human checking the results. A systematic evaluation of different kinds of motions with all combinations of the modifications would be necessary for assessing the reliability of the modifications.

The dimensions in the questionnaire are not perfect as pairing two descriptions to one axis prevents those from being selected at the same time. A common factor analysis of the styles shown in figure 8 tells that there is redundancy in the descriptions we used. In this data set the pairs *tired-sad*, *weak-afraid*, *excited-happy* were closely related. We could simplify the questionnaire by removing one description from each pair without losing much information.

During the analysis two weaknesses were found from the questionnaire. One is related to calculating statistically meaningful variances, which is limited because we had only five steps in each dimension. The second weakness is that all descriptions cannot be used at the same time as one dimension joins two descriptions. A better solution could be to have dimensions that have continuous range instead of steps and descriptions that are in form 'happy - not happy'. However, this would increase the number of decisions a test participant has to do.

6 Conclusions and Future Work

We captured acted motions with different styles and emotions and then tried to produce similar stylistic effects with three algorithmic modifications applied to neutral motions. The motions were evaluated with a questionnaire to determine what styles and emotions are visible in them. This kind of evaluation was not done in any of the papers related to modifications to motion that we have found.

We found evaluating motions with a questionnaire to be a good tool for comparing motions. It was observed that many styles can be seen from one motion and that a single modification can affect many styles. This confirmed that comparison of motions is more meaningful, when examining several describing dimensions simultaneously, instead of only concentrating to one dimension. In this sense evaluating motions is different from evaluating facial expressions as forced choice of one description is not enough for motions. What are the best dimensions for evaluating emotions and styles in motions remains a question yet to be answered.

The results show that modifying posture allowed creating the *strong* and *sad* motions. Changing the length of the motion paths helped in creating *weakness*, *tiredness*, *fear* and *excitement*. Changing timings of the motions was not found to affect the content of the motions significantly. Adjusting timings might be helpful if different types of motions were studied.

When combining modifications, the effects on the style of the motion were not always the same as the sum of effects of the modifications separately. This suggests

that modifying already emotional motions could reveal more about the modifications than modifying only *neutral* motions. Similar phenomena could also be present when interpolating emotional motions. Also, creating totally *neutral* motions by motion capture is hard as the physical appearance of actors always affects captured motions. In the end, bodily motions alone cannot express all emotions. Therefore, other methods such as facial expressions must also be used when making complete animations.

Acknowledgments. This work was partially funded by aivoAALTO project of Aalto University and by Academy of Finland, project Enactive Media (128132). Also, thanks to Timo Idänheimo for his help during the work.

References

1. Menache, A.: Understanding Motion Capture for Computer Animation and Video Games. Academic Press. 238 p. (2000)
2. Amaya, K., Bruderlin, A., Calvert, T.: Emotion from motion. Proc. Graphics interface GI'96, pp. 222-229. Canadian Information Processing Society. (1996)
3. Hsu, E., Pulli, K., Popovic, J.: Style translation for human motion. Proc. SIGGRAPH '05, ACM Transactions on Graphics 24 (3), pp. 1082-1089. (2005)
4. Shapiro, A., Cao, Y., Faloutsos, P.: Style components. Proc. Graphics Interface GI'06, pp. 33-39. Canadian Information Processing Society. (2006)
5. Bruderlin, A., Williams, L.: Motion signal processing. Proc. SIGGRAPH '95, ACM Transactions on Graphics 14 (3), pp. 97-104. (1995)
6. Heloir, A., Kipp, M., Gibet, S., Courty, N.: Evaluating Data-Driven Style Transformation for Gesturing Embodied Agents. Proc. Intelligent Virtual Agents (IVA '08). In: Prendinger, H., Lester, J., Ishizuka, M. (Eds.), LNCS, Vol. 5208, pp. 215-222. Springer-Verlag, Heidelberg. (2008)
7. Hachimura, K., Takashina, K., Yoshimura, M.: Analysis and evaluation of dancing movement based on LMA. IEEE International Workshop on Robot and Human Interactive Communication, pp. 294-299. (2005)
8. Ruttkay, Z.: Cultural dialects of real and synthetic emotional facial expressions. AI & Society 24 (3), pp. 307-315. (2009)
9. Neff, M., Fiume, E.: From Performance Theory to Character Animation Tools. In: Klette, R., Metaxas, D., Rosenhahn, B. (eds.) Human Motion: Understanding, Modelling, Capture, and Animation. Springer. (2008)
10. Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. Proc. SIGGRAPH '00, pp. 173-182. ACM Press/Addison-Wesley Publishing Co., New York, USA. (2000)
11. Pollick, F., Paterson, H., Bruderlin, A., Sanford, A.: Perceiving affect from arm movement. Cognition, Vol. 82, Issue 2. (2001)
12. Wallbott, H.: Bodily expression of emotion. European Journal of Social Psychology, Vol. 28, pp. 879-896. (1998)

Publication II

Roberto Pugliese and Klaus Lehtonen. A Framework for Motion Based Bodily Enaction with Virtual Characters. In *11th International Conference on Intelligent Virtual Agents (IVA 2011)*, Reykjavik, Iceland, Lecture Notes in Computer Science, Volume 6895, pages 162-168, September 2011.

© 2011 Springer Science and Business Media.

Reprinted with permission.

Note: Due to copyright restrictions, this electronic version of the dissertation does not contain the final versions of the publications, but instead the versions that have not been edited by the publishers.

The final publications can be found from the web sites of the publishers.

A Framework for Motion Based Bodily Enaction with Virtual Characters

Roberto Pugliese and Klaus Lehtonen

Department of Media Technology,
School of Science,
Aalto University,
Espoo, Finland
`{roberto.pugliese,klaus.lehtonen}@tkk.fi`

Abstract. We propose a novel methodology for authoring interactive behaviors of virtual characters. Our approach is based on enaction, which means a continuous two-directional loop of bodily interaction. We have implemented the case of two characters, one human and one virtual, who are separated by a glass wall and can interact only through bodily motions. Animations for the virtual character are based on captured motion segments and descriptors for the style of motions that are automatically calculated from the motion data. We also present a rule authoring system that is used for generating behaviors for the virtual character. Preliminary results of an enaction experiment with an interview show that the participants could experience the different interaction rules as different behaviors or attitudes of the virtual character.

Keywords: enaction, motion capture, bodily interaction, authoring behaviors.

1 Introduction

Authoring believable behaviors for virtual characters is a crucial step towards the creation of immersive gaming experiences. In social encounters behaviors emerge as humans react to actions of others in a continuous feedback loop. This process is sustained by bodily interaction among the different parties. Human-computer bodily interaction is possible even at consumer level with latest sensor technology.

We are interested in behaviors that can be observed in and activated by bodily motion and how to use this as a medium of interaction with a virtual character. Our interests are not in traditional goal-oriented interaction or in symbolic language. For these reasons an enactive loop, where both parties can continuously affect the other through actions and the style of motions, was chosen as the model of interaction instead of using discrete gestures.

We present a framework that allows bodily interaction between a human and a virtual character in an enactive loop. The implementation takes a long motion capture sequence as input and automatically segments it into a motion library,

indexed by motion styles, that is used to animate the virtual character. We also present a rule authoring system that is used for generating behaviors for the virtual character.

2 Related Works

In this section we explore earlier works related to enaction and to techniques that enable interaction through motion with animated characters.

2.1 Enaction

Enactive Media is an approach to design modalities of human-machine interaction. While traditionally interactivity has been approached with theories and tools for goal-oriented tasks, the enactive paradigm focuses on a tight coupling between machine and the user, here a participant or enactor. The process is a feedback loop: the actions performed by the enactor affect the medium that in turn affects the following actions of the enactor. The coupling is sustained by means of bodily and spatial involvement, or enactment [1]. An enactive system may involve even a community of agents in participatory sense-making [2].

We want to create a process where the participant will be able to notice different behaviors in the virtual character as a response to his or her own behavior. The rules governing the interaction do not need to be explicit but they can be learned by interacting, in accordance with the original definition of enaction from Bruner [3], that is to learn by doing. While an enactive account for human-computer interaction has been provided in other fields such as facial expressions of virtual characters and movie creation [4], based on psychophysiological input, an implementation of enaction with a virtual character based on bodily motion is yet absent.

In our enactive setting no assumption about the meaning of gesture is done a priori but meaning is actively constructed by the participant and emerges from the enactive loop. This calls for representing the quality of the interaction and the motion style in an objective and non-hierarchical way. We borrow the spatial ontology (ontospace) approach by Kaipainen et al. [5] as a solution. An ontospace is defined by ontological dimensions (ontodimensions) that correspond to descriptive properties of the content repertoire, which in our case are motion clips.

2.2 Interaction through Motion with Animated Characters

Animating characters is possible with motion graphs that contain captured motion segments and a list of allowed transitions between the segments [6]. A motion graph can be constructed automatically from a large corpus of motions and can be used to produce arbitrarily long continuous motions [7].

In a previous work, full-body interaction with a virtual creature meant mainly giving commands and instructions to virtual creatures and the set up did not

allow symmetrical interaction [8]. Similarly, Improv [9] allows interaction with virtual characters. It allows creating scripted sequences of animation and interaction by means of if-statements based on the properties of the characters. These earlier systems concentrated mainly on goal-oriented actions.

In human-computer bodily interaction one needs to extract motion cues able to describe the human motion in a machine friendly way. We follow a methodology based on previous work in the field of analysis of expressive gesture in music and dance performances [10]. The process has camera-based tracking and calculation of motion features that serve as descriptors for motions. Those descriptors include amount of movement and body contraction/expansion.

3 Framework

The system that we have built simulates a situation where two persons are separated by a glass wall and are able to interact only through bodily motions. This creates an enactive loop and allows replacing one or two of the persons with a virtual character (Fig. 1), in our system rendered as a stick figure. The enactive system reacts to human motion (input) by triggering a recorded motion clip (output).

3.1 Enactive Loop

Any motion clip (either recorded or realtime captured) can be associated to a point in an ontospace based on its values of the motion descriptors as coordinates. Before the enaction (Fig. 2) can start the output ontospace is filled with acted motions (Z). The loop starts by mapping the human motion into the input ontospace with the descriptors (A). Then a rule system determines the desired position of the virtual character in the output ontospace (B). Next the animation engine searches for the closest motion to the desired position from the acted motions (C). The virtual character then proceeds to play the motion (D). As the last step the human observes the motion of the virtual character (E), which affects the motion of the human, etc.



Fig. 1. Live enaction.

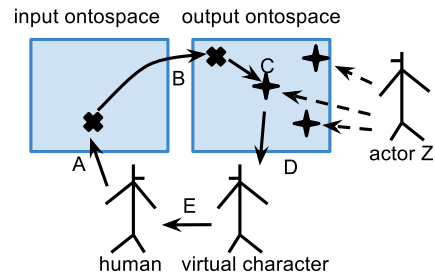


Fig. 2. Enactive loop sustained by the human and the virtual character.

3.2 Motion Analysis and Generation Using Descriptors

For the enactive loop we needed a virtual character that moves with varying motion styles and reacts to the motion style of the human. We use Quantity of Motion (QoM) and Distance descriptors for motions. The former is used as an estimation of the energy and the motion style while the latter characterizes the interaction between the social spaces of the human and the virtual character.

Our definitions are the following: Quantity of Motion (QoM), sum of the frame to frame displacements of all the joints in the body of the character divided by the number of frames in the motion, minus the minimum amount of motion required to move from the starting position to the end position; Distance, the distance of the center of the body from the wall separating the characters.

These descriptors allow placing every possible motion in a two-dimensional ontospace (Fig. 3), constituting a simple case that still allows authoring behaviors. To normalize the descriptors values Distance was scaled linearly, but for QoM we used a log-like function. This takes into account that humans perceive very small changes in the amount of motion if the overall speed is low, but for high speeds the change needs to be much larger to be noticed [11].

Our virtual character is a program that takes desired descriptor values as input and then generates an animated motion sequence that fits to the desired values. To be able to do this we created a motion library containing idle standing (concentration of dots in Fig. 3), walking and running (extremes of Distance in Fig. 3) and jumping actions (high QoM in Fig. 3). These actions were acted with varying styles to evenly populate the ontospace with motion segments. Total of six minutes of motion was automatically segmented to create a motion graph with approximately one second long clips that allow smooth transitions to many other clips. The segmentation was based on finding frames of motion that have a similar pose and speed. After playing a clip the number of alternative following clips ranged with our motion library from 2 to 240.

3.3 Authoring Rules

In our methodology, authoring the rules corresponds to finding a meaningful transformation of the input ontospace, the one of the human, into the output ontospace, the one of the virtual character. The transformation is a mapping defined by example point-pairs in the input and output ontospace. To make the mapping work for inputs in between the example points, we search for the k -nearest neighbors in the input space and determine the output with a weighted interpolation of the corresponding points in the output space. For this, we made a GUI for creating mappings between the ontospaces by specifying examples of corresponding point pairs (A, B, C and D in Fig. 4). In the case of a two-dimensional ontospace this means clicking a point in the input space and then clicking the desired output in the output space. An obvious mapping is the identity transformation which makes the virtual character imitate the motion style of the human. Once a rule is defined, it can be used with a larger motion library without any extra manual work.

In the case of a large number of descriptors, authoring the rules with a mouse can become a tedious and difficult task. A promising alternative approach could be to first record actions with motion capture and then to record the responses to those actions. With this motion data, it should be possible to populate the input and output spaces and obtain behaviour rule.

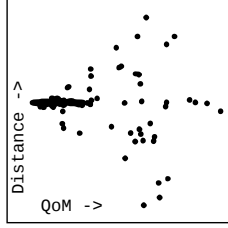


Fig. 3. Ontospace populated with motion segments (dots). The coordinates of the dots are the QoM and the Distance.

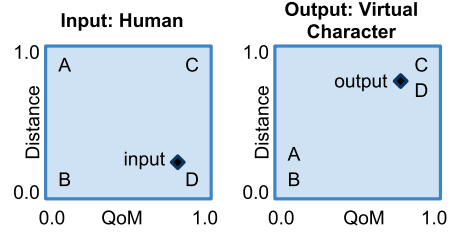


Fig. 4. An example of authoring the behaviour rules with point pairs A, B, C and D.

4 Enaction Tests and Interview with the Participants

In order to validate our methodology and evaluate the effectiveness of our implementation, we conducted enaction tests where a participant had to bodily interact with a virtual character projected in front of them. The participants were 7 unpaid volunteers, 5 male and 2 females of age from 25 to 55.

At the beginning of the experiments the participant was inside a motion capture room and informed about the area where he or she could move and the fact that the virtual character was a stick figure which is able to see the human as a stick figure. No explicit goal of the experiment was stated besides the suggestion to freely explore the bodily interaction with the virtual character.

Six conditions were presented. The first one was always a plain imitation rule used to familiarize the participant with the setting and the interaction paradigm. The next 6 conditions were other rules in randomized order. Those rules were created from a mathematical point of view to try different mappings of descriptors and restricting the virtual character to a limited area of the ontospace.

Condition A was a plain imitation rule. Conditions B and C were imitations with QoM of the virtual character limited to low values in condition B and to high values in condition C. Condition D mapped the QoM of the human to the Distance of the virtual character, as in Fig. 4. This causes the virtual character to back off when the human does motions with high QoM. Condition E inverted the QoM of the human for the virtual character. This makes the virtual character have high QoM, for example, by jumping and waving hands when the human is standing still. In condition F the virtual character played random motion clips without being affected by the human.

In each condition the participant was free to experiment with that rule for 2 minutes. After the experiment we interviewed all the participants to get detailed information about their experiences.

5 Interview and Discussion

Evaluating bodily motions during enaction in an objective manner is a more difficult task than evaluating pre-recorded videos of motions. The main reason is that the conditions are not fully controllable and repeatable because by definition the outcome strongly depends on what the participant does.

On questions related to quality of the interaction with the virtual character we found out that all the participants felt there was interaction in some conditions and also that they could identify different behaviors. There was a general agreement about a character that showed a recognizable scared behavior. This character belonged to the condition D which causes the virtual character to back off when the human does motions with high QoM. Another often mentioned behavior was aggressiveness. This was probably caused by the motions with high QoM such as jumping and waving hands.

The participants said that in some conditions it was hard to understand what made the character react. Besides the condition F with a randomly acting character, this could be explained by that the reaction time of the character could become too slow if the character was playing a long motion. We are aware that two descriptors are not enough to properly describe human actions. We realized that a too simple system makes the participant focus mainly on discovering the rules and the descriptors rather than being in the flow of enaction.

The participants said that their own behavior was affected by the behavior of the virtual character and many of the participants said that they started to mimic the gestures seen in the character. Most of the time, participants moved more when the virtual character was active and less when the character was passive. These facts indicate that the interaction we designed is effectively a case of enactive loop, where both parties affect each other.

6 Conclusions and Future Work

We have presented a framework to design bodily interaction with virtual characters based on the concept of enaction and an authoring tool to specify different behaviors for them that gradually emerge during and due to the interaction. Behaviors are created by mapping the input ontospace of the human, described by motion descriptors, into the output ontospace of the virtual character, populated with automatically evaluated motions. Preliminary tests with participants showed that experiencing different interaction rules as different behaviors or attitudes of the virtual character is possible even in the simplest case of a two-dimensional motion descriptor space.

Defining motion styles with motion descriptors allows using large amount of captured motion without adding more work as no manual annotation is required.

In the future, we plan to add new motion descriptors and differentiate different parts of the body. The manual process of authoring behaviors could be replaced by acting them out in the case of a large number of motion descriptors. Furthermore, we intend to use interpolation among different rules to create virtual characters changing their behaviors during the enaction.

Acknowledgments. This work is part of the Enactive Media project, funded by the Academy of Finland, decision number 128132. We want to thank the Enactive Media team Mauri Kaipainen, Niklas Ravaja, Tapio Takala, Pia Tikka and Rasmus Vuori for the fruitful discussions and intellectual contribution to this paper.

References

1. Varela, F., Thompson, E. and Rosch, E. *Embodied Mind: Cognitive Science 2. and Human Experience*. MIT Press, Cambridge, MA, (1991)
2. De Jaegher, H., & Di Paolo, E.A. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4), 485-507, (2007).
3. Bruner, J. *Toward a theory of instruction*. Belknap Press of Harvard University Press, Cambridge, MA, (1966)
4. Kaipainen, M.; Ravaja, N.; Tikka, P.; Vuori, R.; Pugliese, R.; Rapino, M.; Takala, T.: *Enactive Systems and Enactive Media. Embodied human machine coupling beyond interfaces* . Leonardo, (2011) (in press)
5. Kaipainen, M., Normak, P., Niglas, K., Kippar, J. and Laanpere, M. Soft ontologies, spatial representations and multi-perspective explorability, *Expert Systems* 25, 5, 474-483 (2008)
6. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In *Proc. of SIGGRAPH '02*. ACM, New York, NY, USA, 473-482. (2002)
7. Zhao, L., Normoyle, A., Khanna, S., Safonova, A.: Automatic construction of a minimum size motion graph. In *Proc. of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '09)*, Dieter Fellner and Stephen Spencer (Eds.). ACM, New York, NY, USA, 27-35. (2009)
8. Blumberg, B., Galyean, T.: Multi-level direction of autonomous creatures for real-time virtual environments. In *Proc. of SIGGRAPH '95*, Susan G. Mair and Robert Cook (Eds.). ACM, New York, NY, USA, 47-54. (1995)
9. Perlin, K., Goldberg, A.: Improv: a system for scripting interactive actors in virtual worlds. In *Proc. of SIGGRAPH '96*. ACM, New York, NY, USA, pp. 205-216. (1996)
10. Camurri, A., Mazzarino, B., Ricchetti M., Timmers, R., Volpe, G.: *Multimodal Analysis of Expressive Gesture in Music and Dance Performances, Gesture-Based Communication in Human-Computer Interaction, Lecture Notes in Computer Science*, vol. 2915/2004, pp. 357-358, (2004)
11. Levine, S., Krhenbhl, P., Thrun, S. and Koltun, V.: Gesture controllers. In *Proc. of SIGGRAPH 2010*, Hugues Hoppe (Ed.). ACM, New York, NY, USA, Article 124, 11 pages. (2010)

Publication III

Klaus Förger, Tapio Takala and Roberto Pugliese. Authoring Rules for Bodily Interaction: From Example Clips to Continuous Motions. In *12th International Conference on Intelligent Virtual Agents (IVA 2012)*, Santa Cruz, USA, Lecture Notes in Computer Science, Volume 7502, pages 341-354, September 2012.

© 2012 Springer Science and Business Media.

Reprinted with permission.

Note: Due to copyright restrictions, this electronic version of the dissertation does not contain the final versions of the publications, but instead the versions that have not been edited by the publishers.

The final publications can be found from the web sites of the publishers.

Authoring rules for bodily interaction: From example clips to continuous motions

Klaus Förger, Tapio Takala, and Roberto Pugliese

Department of Media Technology,
School of Science,
Aalto University,
Espoo, Finland

`{klaus.forger,tapio.takala,roberto.pugliese}@aalto.fi`

Abstract. We explore motion capture as a means for generating expressive bodily interaction between humans and virtual characters. Recorded interactions between humans are used as examples from which rules are formed that control reactions of a virtual character to human actions. The author of the rules selects segments considered important and features that best describe the desired interaction. These features are motion descriptors that can be calculated in real-time such as quantity of motion or distance between the interacting characters. The rules are authored as mappings from observed descriptors of a human to the desired descriptors of the responding virtual character. Our method enables a straightforward process of authoring continuous and natural interaction. It can be used in games and interactive animations to produce dramatic and emotional effects. Our approach requires less example motions than previous machine learning methods and enables manual editing of the produced interaction rules.

Keywords: animation, motion capture, bodily interaction, continuous interaction, authoring behavior

1 Introduction

Virtual characters are common in modern games and their bodily motions can reflect the emotions and attitudes between characters. Real-time motion synthesis and synchronization with external stimuli enables making the characters interactive. The possibilities for using bodily interaction have increased as even consumer level sensor technology allows capturing bodily motions. Sometimes a motion does not mean much out of context, but can have a lot of meaning if displayed as a synchronized reaction to an action [1]. A good example is a virtual character taking a step backwards in isolation versus taking a step backwards as a reaction to aggressive behavior of another character.

Expressive motion based interaction between two virtual characters is possible with a library of recorded and annotated motions. For example, if a character moves in a way that was annotated as angry, then another character could react

by selecting a motion that was annotated as scared. Similar interaction between a human and a virtual character requires ability to evaluate the style of previously unseen motions in real-time. Only part of the emotional content in a motion library can be annotated in advance as it can vary depending on the context of the motion.

In this paper we explore how motion captured examples of human interaction can be used in authoring interaction rules for virtual characters. These rules allow real-time generation of expressive behaviours in a continuous interaction loop. The idea is that there should be no frozen pauses during an interaction sequence as can happen in task-based approaches, but instead all idle moments could be used to reflect the attitudes of the participants. Continuous interaction scheme could be also used for subtle control over the style of motion. For example a walking character could immediately react to an observed aggressive action by changing the style of the walk from neutral to careful. The amount of visible aggression could be continuously mapped to the amount of carefulness.

We concentrate on the case where humans and virtual characters have equal amount of information from each other. The virtual characters observe humans through features that characterize different qualities of human motion. The features, from now on referred as motion descriptors, can be calculated in real-time. We show how using example motion pairs makes authoring of interaction rules a straightforward process. The example motions also help avoiding impossible combinations of motions descriptors. Moreover, we show that selective use of motion descriptors allows solving the curse of dimensionality, that arises from modeling human motion simultaneously with several descriptors.

We next present related work, and then describe our implementation in three parts. First is the calculation of motion descriptors. The second is the interaction rule authoring where observed descriptors are mapped to descriptors of desired reactive behaviours. The last part is motion synthesis that turns the descriptors to actual motions of a virtual character. In the fourth section, we present a use case of the process of authoring behaviours. The last sections contain discussion, conclusions and future work.

2 Related work

Real-time motion synthesis can be done by creating a motion graph from captured motions and playing one motion segment after another according to the graph [2]. Furthermore, if the captured motions are annotated, it is possible to control the motion synthesis [3]. This can happen by selecting motion segments from the graph that correspond to attributes used in the annotation. We use motion synthesis that is based on a motion graph. Our graph includes information about the motion style that is used in controlling the motion synthesis. Therefore, it is similar to the metadata motion graphs, which have been used for synchronizing human motion with beats in streaming music [4].

Manual annotation of motion can be very time consuming. Therefore automatically calculated motion descriptors for human motion have been developed

[5, 6]. The descriptors can measure for example the amount of motion, acceleration or qualities of the pose of a character. Similar values have been also calculated from the relational motion of hands representing two entities such as small animals [7]. We use motion descriptors for annotation of recorded motion in a motion graph and real-time motions. We also extend the relative descriptors from relations between hands to the case of relations between two human characters.

Earlier systems that allowed interacting with virtual creatures were often targeted at goal oriented interaction [8] or interaction scripted with if-else clauses [9]. A more fluid model of interaction was allowed by a probabilistic method that uses pairs of recorded actions and reactions to learn how to react to human movements [10]. A similar system based on example motions has been used for teaching cleaning robots socially acceptable motion styles [11]. Our method for defining interaction fits in between these older methods as it takes advantage of example action-reaction motions and manual definitions.

The importance of usability has been noted in earlier works that present tools for authoring behaviours of virtual creatures [12, 13]. These tools assume that a range of low level behaviours such as wandering, following and actions that display emotional states are available. The tools allow applying the low level behaviours to crowds and joining them together to form more complex patterns. Our method could be used for authoring the said low level behaviours. One requirement for the earlier tools has been that using them should not require coding experience or understanding complex models that govern the behaviours. This requirement is taken into account in the design of our method.

Publications related to virtual characters include works on embodied conversational agents (ECAs) [1]. Considerable effort has been taken towards a unified Behaviour Markup Language (BML) that can be used when creating ECAs [14]. These works mainly view bodily motion as a way to make verbal conversations more believable. In this paper, we consider varied situations beyond conversations, and study non-verbal interaction where bodily motion is the only channel of communication. We also extend the scope of behaviours from friendly and believable characters to ones that could be considered anti-social and annoying as those can be required in for example games containing dramatic sequences.

A proposal has been made to extend the BML from describing the behavior of a single virtual human to the case of continuous interaction between two characters [15]. Continuous interaction is a core aspect of our paper, but we have a different point of view on what part of motion we want to control. The proposal suggests developing an XML based approach that could be useful in defining and controlling discrete reactions and gestures. However, we concentrate on motion style that we abstract with motion descriptors which have a continuous range of values that vary from frame to frame. For this reason we use a more continuous control mechanism.

Our work builds on earlier work about defining interaction rules using motion descriptors [16]. The earlier system showed that Radial Basis Functions (RBF) [17] can be used to map the input motion descriptors to output descriptors. The

output descriptors where then used to control a motion synthesis engine. These parts created a virtual character that reacts to observed human motions. The system used only two input and output descriptors and therefore had limited capability to create interesting interaction. In this paper, we show that using more than two descriptors allows much more varied interaction. However, it also introduces the curse of dimensionality as the number of the combinations of the descriptors grows exponentially compared to the number used descriptors. That in turn forced us to find new ways to create the interaction rules.

3 Implementation

Our system is based on an interaction loop where two characters can observe each other and react to the actions they witness. The steps from the observed motions to the synthesized motion are shown in Figure 1. First, the values of the input descriptors are calculated from the observed motion of another character (or human) and from the character’s own motion. Secondly, the input descriptor values are mapped according to the interaction rules into the desired output descriptor values. Then the motion synthesis engine creates a motion that fits the output descriptor values as well as possible. Next, the newly synthesized motion can be observed by another virtual character or by a human.

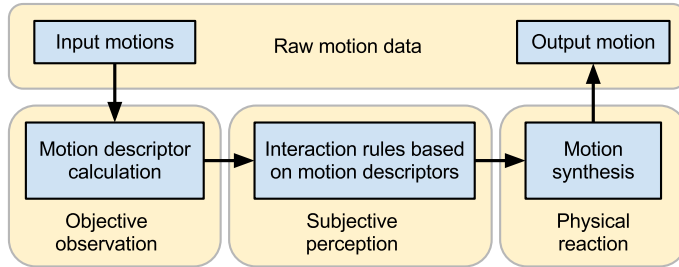


Fig. 1. Model of tasks performed by an interactive virtual character system and the flow of information in the system.

3.1 Motion descriptors

In our implementation, a motion descriptor is a function that takes as input the observed 3D motion data and calculates a value between zero and one for each frame of the motion. Ideally, the descriptors would tell all about the motion style of an action and ignore unimportant aspects. In practice, the descriptors are limited to what can be easily defined mathematically and calculated in real-time. The descriptors act as objective measures that are not affected by any internal states of the characters.

In this work we concentrate on behaviours that include standing, walking, jumping and generally moving around on a flat floor. For these types of motions, relevant motion descriptors for an isolated character include Quantity of Motion (QoM) [6], turning left/right and moving backwards/forwards. Examples of motions corresponding to high and low values of the descriptors are shown in Figure 2. The QoM estimates the total amount of energy in the motion and is calculated as the sum of instantaneous velocities of all body parts. We also use a variant of QoM called non-transitional QoM (NtQoM) that estimates the energy used only for body language or other expressive motions, disregarding locomotion.

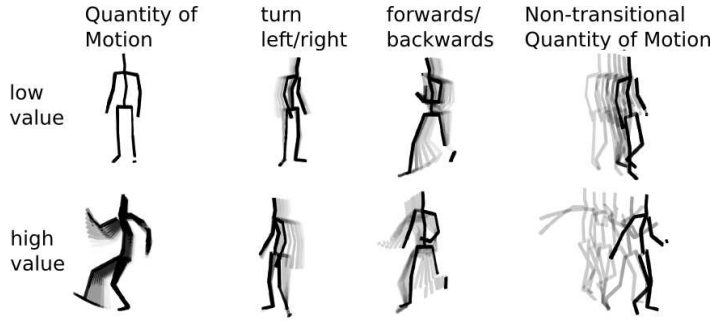


Fig. 2. Examples of motion descriptors calculated from an isolated character.

Another class of descriptors are relational descriptors, i.e. those that compare motion of two characters. Of these we use the distance between the characters, their facing angle, and approach/retreat as illustrated in Figure 3. The facing angle is normalized to be zero for face-to-face characters and one when a character's back is turned towards the other. The values of approach/retreat are linearly relational to the velocity along the direction to the other character. The extreme values of approach/retreat are normalized in order to get zero when the character is moving towards the other character at 7 m/s and one when moving away at the same speed. The value 7 m/s is reasonable limit for bodily interaction as a forceful jump without a run can have approximately that velocity. All the other descriptors are also normalized in a similar manner. For the distance there is only one common value for both characters, but the characters have their own values for the facing angle and approach/retreat.

We used a commercial motion tracking system (NaturalPoint OptiTrack) with 24 cameras working at 100 frames per second to capture the motions from which we calculated the descriptors. We were also able to calculate all the used descriptors from the data Microsoft Kinect provides using OpenNI library. However, the lower accuracy and higher amount of errors in the Kinect data made

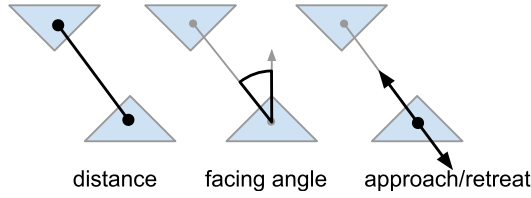


Fig. 3. Relational motion descriptors between two characters.

using it impractical. Especially calculating QoM from the noisy data is unreliable.

While the motion descriptors estimate different aspects of the motion, they are not independent from each other. Some dependencies are direct. For example, if QoM is zero, it limits all other descriptor values to those corresponding to no movement. Another type of dependency is dynamic. For example, between QoM and facing angle all value combinations are physically possible, but changing facing angle becomes impossible if the QoM remains zero. Because of these physical constraints, it is useful to pick descriptor values from recorded motions. A randomly selected set of descriptor values might be impossible to synthesize into motion of a virtual character.

3.2 Interaction rule authoring

The interaction rules define the reactions of a virtual character to observed motions. The reactions can vary depending on the own position and motion of the virtual character. A rule can reflect physical properties such as being strong or weak. Also, the mental state of the character, such as sadness or aggressiveness, can be built in the behavior produced with a rule.

In practice, the rules are used for projecting a frame of an observed motion to a desired reaction. Here we consider the frame to include positional information and instantaneous velocities of body parts. Our method uses motion descriptors to abstract the frames of input and output motions. In an earlier publication the mappings were created with manually defined example point pairs [16]. A mapping consisted of one point in the input corresponding to one point in the output space. After the mappings for the rule were defined, projecting a point from the input to the output was done with Radial Basis Functions (RBF), that is a sparse data interpolation method [17]. This process of creating a rule required filling the input space evenly with points along every dimension.

The standard RBFs approximate a function from a high dimensional space to a single dimension using example points where the output value is predefined [17]. In practice, if the point that is being projected is close to only one of the example points, the output will have the value that is linked to that example point. Should the point be in the middle of two example points, the output would be an average of the linked output values weighted by the inverse of distances to those example points. The case of projecting from a high dimensional space to another high

dimensional space with RBFs requires only repeating the standard case for each of the output dimensions [17]. The RBFs were chosen as the interpolation method as it is a simple approach to code and cheap to calculate.

The projection using point pairs and RBFs works well up to two descriptor dimensions, but faces the curse of dimensionality in the combinatorial sense with a higher number of dimensions. This means that the amount of point pairs required to cover the input space grows exponentially with the number of dimensions. High dimensionality also hinders visualizing and editing points as three spatial dimensions is the limit on human vision. Growing dimensionality also makes motion synthesis harder as each descriptor dimension sets new requirements for the produced motions.

When experimenting with high dimensional interaction rules, we observed that creating the mappings rarely requires using more than three dimensions simultaneously. In fact, many interesting and useful mappings require only using one or two dimensions, but the set of required dimensions varies between mappings. This observation lead us to the conclusion that the problems caused by a high dimensionality could be solved by allowing the author of the rules to select which dimensions are relevant to each of the mappings individually.

Mathematically, our solution requires pairing each set of descriptor values with a scaling vector and a modification to the projection done with RBFs. The scaling vector indicates how important the related descriptor dimensions are. The input data we need to consider includes the point we want to project p and the example point pairs indexed as $[1...k...K]$. A point pair consists of an input point i_k and the related input scaling vector s_k , an output point g_k and the related output scaling vector u_k . The data we want to calculate is the output descriptor values o and the output scaling vector h . During real-time interaction p is the observed motion, o is the desired reaction and h tells how the output descriptors should be prioritized in the motion synthesis.

Next, we go through the changes needed in the standard case of RBF. Let us consider an input space with dimensions indexed as $[1...n...N]$ and an output space with dimensions indexed as $[1...m...M]$. The scaling vector s_k of an example input point i_k affects the calculation of the distance d_k between the point p , that is being projected, and the point i_k as follows:

$$d_k = \left| (i_k - p) \begin{bmatrix} s_{k_1} & 0 & \dots & 0 \\ 0 & s_{k_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_{k_N} \end{bmatrix} \right|. \quad (1)$$

If the scaling vector s_k is all ones, then the distance calculation returns to the unmodified case. If one of the values in the scaling vector is zero, then that input descriptor dimension is effectively ignored by the mapping. The output scaling vectors $u_{1...K}$, which are paired with example output descriptor values $g_{1...K}$, enable the creation of mappings that can be independent also on the output side of the projection. In practice, this means that the influence of a mapping (a point pair) can be limited to only part of the output descriptors $o_{1...M}$. The

values for the example output scaling vectors $u_{1...K}$ and the example output points $g_{1...K}$ of the mappings are used in the RBF interpolation resulting in the output descriptor values $o_{1...M}$ as follows:

$$o_m = \frac{\sum_k (g_{k_m} \cdot u_{k_m} \cdot (1 - d_k))}{\sum_k (u_{k_m} \cdot (1 - d_k))}. \quad (2)$$

Finally, the values $h_{1...M}$ of output scaling vector that forms a pair with output descriptor values $o_{1...M}$ is:

$$h_m = \sum_k (u_{k_m} \cdot (1 - d_k)) \quad (3)$$

As the values for descriptors and the scaling vectors are limited to the range from zero to one we also limit all the values in the calculations to the same range. The o and the h are given as input to the motion synthesis engine. The descriptor values o control type and style of synthesized motion and scaling vector h tells how the output descriptors should be prioritized and which can be ignored.

The practical work that the author of the rules must do when creating a mapping for a rule includes defining the values for input and output descriptors and scaling vectors. This can be done manually with the sliders of a graphical user interface shown in Figure 4 (E, F). However, we have found that it is not always easy to see which movement would correspond to given descriptor values or vice versa. This problem can be overcome by capturing an example action and reaction and selecting the descriptor values from them with the user interface. The time line/frame counter (fig. 4 D, upper slider) can be used to simultaneously browse through the values (fig. 4 B, E, F) and view an animation (fig. 4 A) of the example motions.

In example motions the action and the reaction do not always happen at exactly the same moment. Therefore, there can be a need to scroll the reaction forward in order to find the right descriptor values for it. This can be done with the lower slider in the user interface in Figure 4 (D). This offsetting is necessary especially for the relational descriptors that have only one value such as the distance between the characters. Without offsetting, those descriptors would by definition have the same value for both input and output.

Picking descriptor values from example motions helps the process of authoring rules, but picking the values for scaling vectors cannot be done in the same way especially in a high dimensional case. If more than one example pair of motions displaying the same interaction would be available, then some of the scaling values could be estimated based on the correlations in the examples. However, this would add much work in capturing the examples. For this reason the author of the rules needs to have a vision of the intended interaction that guides the selection of the scaling values.

Compared with the old method of creating rules [16] the new method allows rules to be made with much less mappings as the scaling can be used to ignore those descriptor dimensions which are not relevant. When using the old method,

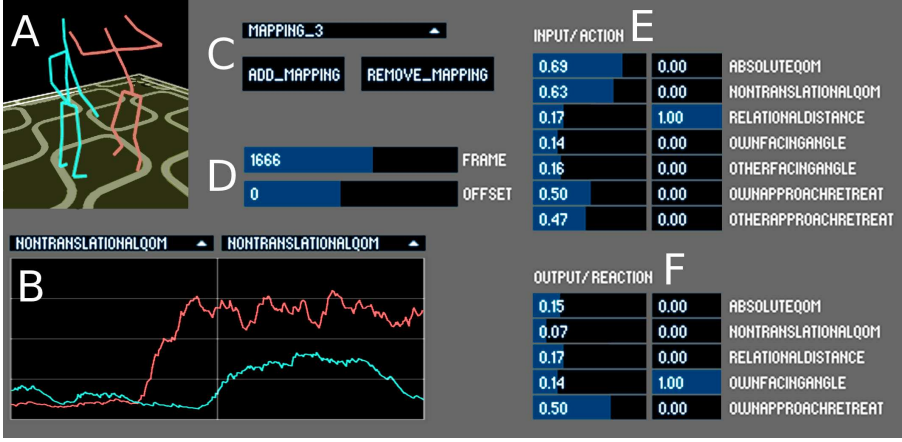


Fig. 4. A graphical user interface for authoring interaction rules that includes animation of the example motions (A), view of descriptor values over time (B), selection of mappings (C), sliders for the animation time and offset between motions (D), input/output descriptors (usually picked from animation) (left side E, F) and input/output scaling (manually defined) (right side E, F).

it would have been necessary to define mappings for all the combinations of descriptor dimensions, even those that should not affect the end result.

3.3 Motion synthesis

During real-time interaction, the motion synthesis engine takes the desired output descriptor values and creates a continuous motion following the values as closely possible. We use a motion graph based on recorded motions for the synthesis. All the motions in the motion graph are annotated using the motion descriptors. After this we can synthesize new motions by concatenating motion clips that fit well to the desired descriptor values.

The motion graph we used contains a motion library divided into motion clips and the possible transitions between the motion clips. The motion clips are half to two seconds long samples from a nine minutes recording. The recordings contained motions that are required for moving on a flat surface (standing, turning, walking, running) and a few expressive motions (waving hands, jumping). The motions were recorded many times with differing styles to get versions with both high and low QoM similarly as shown in Figure 2.

We create the motion graph by finding all transitions from a frame to another where the pose and velocities do not differ too much with the restriction that the transitions are at least half a second apart. We do not prune any transitions from the motion graph as it would increase the reaction time of the virtual character and reduce the amount of possible reactions. During real-time interaction, we evaluate as many motion clip sequences as is possible during half a second,

usually a few tens of thousands, and then select the sequence of clips that matches the desired descriptors best. The high number of possible clip sequences makes motions synthesis the most computing intensive part of the whole system.

One challenge here is that the desired values might require actions that cannot be performed simultaneously. For that problem, the scaling vector of the output values (Equation 3) helps as it tells which descriptors need to be prioritized and which can be ignored. The constraints set by the human body and physics offer another challenge as they should not be broken when natural looking motions are desired. The concatenative synthesis we use always produces physically plausible motions, but allows only approximate following of the desired descriptors.

During real-time interaction, selecting the next motion clips requires searching for the sequence of clips where the deviation from the desired motion descriptors is minimal. For the descriptors that can be calculated from an isolated character, this can be done by using the descriptor values that were calculated when preparing the motion graph. The relative descriptors cannot be calculated in advance as they vary depending on the other character. Also, the relative descriptors cannot be used directly in the search as the virtual character can only decide its own motion. Therefore, the relational values for distance, facing angle and approach/retreat are transformed into position, direction of movement and facing in the absolute coordinate system. These values can then be used as parameters to be optimized in the search for the next motion clip.

4 Example cases of authoring behaviours

The simplest type of examples of authored behavior contains only one mapping between the input and the output descriptors. Let us consider a character that turns its face to another character. For this behaviour, the input should have all the scaling values of the descriptors set to zero and the output facing angle with value zero and weight one. During real-time interaction, having only this mapping would set the desired facing angle to zero. All the other descriptors would be ignored by the motion synthesis engine as they would have zero scaling values. A character following this rule would create an impression that it is aware of the position of the other character, but it would ignore all other aspects of the motion.

A more interesting character could be one that acknowledges the other character by turning to face it, but would get offended and turn away if the other character misbehaves. To create this behavior, an example pair of an action and a reaction shown in Figure 5 can be used. The behaviour rule can be created by scrolling through the action and reaction and by creating mappings from all the significant parts of the motion pair.

The motions start by having the characters far away from each other (fig. 5 A). A desired reaction in this case could be to ignore the far away character. This can be turned into a mapping where the input has the distance between the characters with weight one and the output has all the descriptors weighted zero.

In the next part of the motions, the characters have come near each other and the reaction character has turned to face the action character (fig. 5 B). This can be turned into a mapping that has a low distance and low non-transitional QoM on the input side and low facing angle on the output side. The last part of motion has the action character waving hands forcefully and the reaction character responds by turning away (fig. 5 C). This part corresponds to a mapping where the input side has a low distance and high non-transitional QoM and the output has a high facing angle.

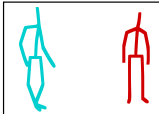

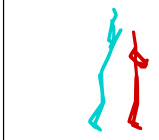
		INPUT		OUTPUT			
A			Scaling	Value		Scaling	Value
		Distance			Distance		
		NtQoM			NtQoM		
		Facing			Facing		
B			Scaling	Value		Scaling	Value
		Distance			Distance		
		NtQoM			NtQoM		
		Facing			Facing		
C			Scaling	Value		Scaling	Value
		Distance			Distance		
		NtQoM			NtQoM		
		Facing			Facing		

Fig. 5. Significant frames (A-C) from an example action (on the left side) and reaction (on the right side) motions, descriptor values picked from those parts of the motions and the scaling values decided by the author of the rule.

After the mappings are defined, the rule is ready to be tested. The testing can happen by seeing if changing the values of input descriptors produce sensible output values. This can show if any errors were made while defining the rules. However, testing the rules with real-time interaction is also important as it can show if there are problems in the synthesis of the output descriptor values. The synthesis can fail if the virtual character is not able to find any possible motions that would fit the output descriptors quickly enough. This can be a problem especially if the motion graph has long motions that cannot be interrupted. Another possible problem is that an action can be so short that the input descriptors show the action for too short time to cause a reaction. This calls for more careful descriptor design and possibly descriptors that are calculated as an average over a period of time instead of just individual frames of motion.

Variations to the presented example behaviour could be that instead of turning away when provoked the character would start to be aggressive. The roles could be also swapped and then the virtual character would be the one starting the provocation.

5 Discussion

The example case shows that using recorded action and reaction motions guides the work flow of authoring interaction rules. The rules can be authored and tested with graphical and bodily user interfaces. Therefore, the requirements for the author creating the interaction rules do not include coding experience or learning an XML dialect. To further develop the usability of the system, user tests should be done with the users authoring new rules and interacting with virtual characters following those rules.

The used reactive interaction rules work in real-time and they are good for interactive background characters. Characters that need more intelligence could be built by adding reasoning capabilities and internal state into the virtual character and selecting the interaction rules based on the internal state. However, the interaction rules alone are not enough as the believability of the authored behaviours is heavily dependent on the capabilities of the motion synthesis engine.

It is challenging to synthesize motions that are realistic and balance the sometimes conflicting demands of expressiveness in real-time. The used motion graph approach is not ideal as it only allows playing motions a clip at a time, while perfect following of the desired output descriptors would require a more continuous synthesis method with a shorter reaction time. The situation could be helped by real-time filtering of the produced motion or using physics based motion synthesis when a sudden reaction is needed.

In other than research applications, using only bodily motions is not enough for creating a complete virtual character. We feel that modalities such as facial expressions and tone of voice could be added to the current authoring system. Also, defining musical interaction could be possible. The main requirement is that it must be possible to create continuous signals to describe the medium and to drive a synthesis engine with those signals. Combining the continuous interaction with discrete gestures could be more challenging. That would require deciding whether the continuous changes in the motion style only modulate the gestures or could they also interrupt or prevent the gestures.

One shortcoming in the presented approach is that the author of the behaviours has to assume the connections between the descriptor values that represent motion style and actual emotions visible in human motion. For example angry motions are likely to have high QoM, but there are many motions with high QoM that might not look angry. A possible solution could be to develop new descriptors that are learned from annotated data with machine learning techniques. The new descriptors could be estimates for visibility of emotions like anger and sadness. Simultaneous use the planned emotional descriptors and the ones that we have presented could allow more precise authoring of emotional reactions.

6 Conclusions and future work

In this paper we introduced a new way to author behavior rules for interactive virtual characters using bodily motions as the medium. We showed that the simultaneous use of several motion descriptors enables creation of expressive interaction rules. Since the descriptors are continuous in both the time and motion style domains, the produced interaction has a chance to be fluid and natural. The problems that emerge from the use of several descriptors include the curse of dimensionality and increased risk of physically impossible descriptor combinations. We solved the curse of dimensionality by adding scaling vectors for each set of descriptor values in each mapping. This reduced the amount of required mappings per interaction rule dramatically.

We also introduced a way to create mappings based on an example with action and reaction motions. This approach can reduce the risk of defining impossible descriptor combinations. The example also guides the creation of the interaction rules and makes the process a straight forward one. Even when using example motions, our method allows manual fine tuning of the rules.

In the future we intend to develop motion descriptors that measure visibility of emotions like sadness and anger from human motions. A promising approach is to create descriptors from annotated motion data by machine learning techniques. We see the development of these descriptors as a required step in order to go from expressive bodily interaction to emotional bodily interaction.

Acknowledgments. This work has received funding from the projects Enactive Media (128132) and Multimodally grounded language technology (254104) which are both funded by the Academy of Finland.

References

1. Huang, L., Morency, L., Gratch, J.: Virtual Rapport 2.0. In: Vilhjálmsson, H., Kopp, S., Marsella, S., Thórisson, K. (eds.) IVA 2011. LNCS vol. 6895, pp. 68-79 (2011)
2. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. ACM Transactions on Graphics (SIGGRAPH '02), vol. 21, is. 3, pp. 473-482. ACM, New York (2002)
3. Arikan, O., Forsyth, D. A., O'Brien, J. F.: Motion synthesis from annotations. ACM Transactions on Graphics (SIGGRAPH '03), vol. 22, issue 3, pp. 402-408 (2003)
4. Xu, J., Takagi, K., Sakazawa, S.: Motion synthesis for synchronizing with streaming music by segment-based search on metadata motion graphs. Conf. on Multimedia and Expo (ICME), 2011 IEEE International, pp. 1-6 (2011)
5. Hachimura, K., Takashina, K., Yoshimura, M.: Analysis and evaluation of dancing movement based on LMA. In: IEEE International Workshop on Robot and Human Interactive Communication 2005 (ROMAN 2005), pp. 294-299. IEEE (2005)
6. Camurri, A., Mazarino, B., Ricchetti, M., Timmers, R., Volpe, G.: Multimodal Analysis of Expressive Gesture in Music and Dance Performances. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 20-39. Springer, Heidelberg (2004)

7. Young, J., Ishii, K., Igarashi, T., Sharlin, E.: Puppet Master: designing reactive character behavior by demonstration. In Proc. of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '08), pp. 183-191. Eurographics Association, Aire-la-Ville, Switzerland (2008)
8. Blumberg, B., Galyean, T.: Multi-level direction of autonomous creatures for real-time virtual environments. In: Mair, S.G., Cook, R. (eds.) Proc. of SIGGRAPH 1995, pp. 4754. ACM, New York. (1995)
9. Perlin, K., Goldberg, A.: Improv: a system for scripting interactive actors in virtual worlds. In: Proc. of SIGGRAPH 1996, pp. 205216. ACM, New York (1996)
10. Jebara, T., Pentland, A.: Action Reaction Learning: Automatic Visual Analysis and Synthesis of Interactive Behaviour. In: ICVS 1999. LNCS, vol. 1542, pp. 273-292. Springer, Heidelberg (1999)
11. Young, J., Ishii, K., Igarashi, T., Sharlin, E.: Style-by-demonstration: Teaching Interactive Movement Style to Robots. In ACM Conf. on Intelligent User Interfaces (IUI '12), pp. 41-50. ACM, New York, USA (2012)
12. Metaxas, D., Chen, B.: Toward gesture-based behavior authoring. In: Proc. of the Computer Graphics International 2005 (CGI '05), pp. 59-65. IEEE Computer Society, Washington, DC, USA (2005)
13. Ulicny, B., Ciechomski, P., Thalmann, D.: . Crowdbush: interactive authoring of real-time crowd scenes. In: Proc. of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation (SCA '04), pp. 243-252. Eurographics Association, Aire-la-Ville, Switzerland (2004)
14. Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thórisson, K. R., Welbergen, H., Werf, R. J.: The Behavior Markup Language: Recent Developments and Challenges. In: Pelachaud, C., Martin, J., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS vol. 4722, pp. 99-111. Springer, Heidelberg (2007)
15. Zwiers, J., Welbergen, H. V.: Continuous interaction within the SAIBA framework. In: Vilhjálmsson, H., Kopp, S., Marsella, S., Thórisson, K. (eds.) IVA 2011. LNCS vol. 6895, pp. 324-330. Springer Berlin / Heidelberg (2011)
16. Pugliese, R., Lehtonen, K.: A framework for motion based bodily enaction with virtual characters. In: Vilhjálmsson, H., Kopp, S., Marsella, S., Thórisson, K. (eds.) IVA 2011. LNCS vol. 6895, pp. 162-168. Springer, Heidelberg (2011)
17. Buhmann, M. D.: Radial Basis Functions : Theory and Implementations. Cambridge University Press, Cambridge, United Kingdom (2003)

Publication IV

Klaus Förger, Timo Honkela and Tapio Takala. Impact of Varying Vocabularies on Controlling Motion of a Virtual Actor. In *13th International Conference on Intelligent Virtual Agents (IVA 2013)*, Edinburgh, UK, Lecture Notes in Computer Science, Volume 8108, pages 239-248, August 2013.

© 2013 Springer Science and Business Media.

Reprinted with permission.

Note: Due to copyright restrictions, this electronic version of the dissertation does not contain the final versions of the publications, but instead the versions that have not been edited by the publishers.

The final publications can be found from the web sites of the publishers.

Impact of Varying Vocabularies on Controlling Motion of a Virtual Actor

Klaus Förger¹, Timo Honkela², and Tapio Takala¹

¹ Department of Media Technology,

² Department of Information and Computer Science,
School of Science,
Aalto University,
Espoo, Finland

{klaus.forger,timo.honkela,tapio.takala}@aalto.fi

Abstract. An ideal verbally controlled virtual actor would allow the same interaction as instructing a real actor with a few words. Our goal is to create virtual actors that can be controlled with natural language instead of a predefined set of commands. In this paper, we present results related to a questionnaire where people described videos of human locomotion using verbs and modifiers. The verbs were used almost unanimously for many motions, while modifiers had more variation. The descriptions from only one person were found to cover less than half of the vocabulary of other participants. Further analysis of the vocabularies against the numerical descriptors calculated from the captured motions shows that verbs appeared in closed areas while modifiers could be scattered to disconnected clusters. Based on these findings, we propose modeling verbs with a hierarchical vocabulary and modifiers as transitions in the space defined by the numerical qualities of motions.

Keywords: motion capture, natural language, virtual actors

1 Introduction

Animations and computer games have characters that act out scenes which an animator has designed. When creating these scenes, animators need believable human motion and ways to control the motion. To satisfy this need for human motion, many collections of captured motion have been made available [1]. Word based searching can be used to find suitable motions without a need to browse through the whole database. This way of searching corresponds to an ideal situation in which an actor would be always ready to act out motions based on short descriptions. In this paper, we concentrate on the effects of varying vocabularies on the motion searches.

In addition to words, human motion databases could also be searched by giving example motions or giving numerical requirements as search expressions. However, we limit the scope of the paper to collections of human motion where every motion clip is annotated with at least one written search term. The annotations can be the instructions given to an actor or opinions of persons viewing

the motions. A potential problem is that a third person might not use or even understand the same vocabulary which was used in the annotations.

To find out how much variation there is in the vocabularies of people describing human motion, we constructed a questionnaire containing several different kinds of human locomotion. We asked people to describe the animated motion with one verb and from zero up to three modifiers which were adjectives or adverbs. Data from the questionnaire shows that variation between vocabularies of different people is large enough to cause potential misunderstandings.

We also present further analysis of the vocabularies against the numerical descriptors calculated from the captured motions. This analysis shows that verbs appear in closed areas whereas modifiers can be scattered to disconnected clusters. Based on these findings, we discuss what are the best ways to model the vocabularies.

2 Related Work

Controlling virtual actors with natural and unrestricted language requires creating links between the describing words and physical motions. A simple approach for creating the links is manual annotation which means writing labels for every motion. The task can be made easier by calculating descriptor values which reflect the quality of the motion [2]. The motion descriptors allow generalizing annotations as we can assume that two motions that are numerically close to each other are likely to be annotated in the same way. In this paper, we use motion descriptors when comparing motions.

Many methods and systems designed for controlling virtual characters assume that there is a small selection of allowed commands [3–5]. More fine grained control of both style and length of motions performed by a virtual character could be desired. This can be achieved with real-time interaction rules between two virtual characters, as the rules are based on continuous parameters [2]. However, the set of parameters can feel artificial to the end user, especially if the parameters are derived from the numerical qualities of the motions. Motion analysis frameworks such as Laban Motion Analysis (LMA) assume that the user knows a set of expert terms for describing human motion such as the Laban notation [6]. It has been found that systems allowing the use of natural language can reduce the amount of expertise and time needed in controlling virtual actors [7]. A challenge in natural language processing is that people can have subjective views on the meaning of words [8]. Our interests are in finding out how much manual annotation and analysis of motion is needed to enable controlling a virtual actor with natural language.

The assumption that, a small amount of motion classes is enough, does not appear only in systems that control virtual characters. Commonly used motion databases are often based on a selection of words given to the actors who perform the motions [9]. This can result in databases with plenty of motions, but where all the motions belong to stereotypical categories. A reason for taking shortcuts in annotation is that manual annotation can take a lot of time and effort [10].

As a motion database with annotations by several persons was not available, we decided to create one.

There are methods for creating new motions with different styles by using a selection of parameters which may be stylistic and emotional [11] or related to the trajectories of the motions [12]. These methods enrich a motion database as they create new motions by blending existing ones. We decided to use motion blending as it allows producing motions between stereotypical classes.

Three questions of interest were left open by the related works. How sufficient annotations from a single person are when building natural language descriptions? Do people describe the same motions with several synonyms? Do people have different opinions about the meaning of the used words? To answer these questions, we created a motion collection and a questionnaire which are presented in the next sections.

3 Motion Data Generation

To study natural language descriptions of human motion, we first needed a collection of motions to be described. We chose locomotion as it appears commonly in animations and it also allows displaying many motion styles. In order to create a set of motions that would have variation in both verbs and modifiers, we decided to use a mix of acted motions and interpolations between those motions. We recorded short locomotion sequences with two actors using Optitrack motion tracking system. The actors were asked to perform walking and limping with styles 'sad', 'slow', 'regular', 'fast' and 'angry'. Running was recorded with only the styles 'slow' and 'fast' as the limited capture area made recording running challenging. To make the motions easy to interpolate, the actors were instructed to always start from the same position with their right leg and to perform the motions towards the same direction.

The blended motions were produced with three steps which were initial alignment, time warping and interpolation. In the first step, the supported and lifted phases of the feet were detected and aligned among the motions. The second step was time warping the motions to make them synchronized. The aligned frames between the supported and lifted phases were matched and the rest of the frames were re-sampled to get a smooth frame rate. As the last step, the coordinates of the root joints were interpolated linearly and the joint rotations were interpolated as quaternions with the slerp algorithm [13].

We used two-way and three-way interpolation to create the blended motions. In the two-way case, three new motions are created with steps of 25%. In the three way case, we created all the two-way combinations, three motions with the percentages 70%-15%-15% and one motion with an even split of 33%-33%-33%. Ideally, we would have created blends from all possible combinations of the original motions, but that would have resulted in too many to be viewed reasonably. Also, some motions like fast running and slow walking were too different to be interpolated. We ended up creating blends between the motions that had a similar style and also between motions that had the same intended verb. The combinations used in the blends are shown in Figure 1.

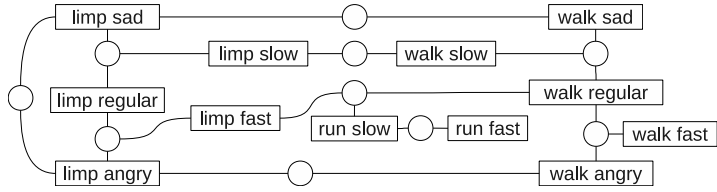


Fig. 1. The boxes show the original captured motions with the instructions given to the actors, the circles represent the combinations used in motion blending.

4 Questionnaire and Methods for Analysis

The idea of the questionnaire was to collect verbs and modifiers that describe the motions. The questionnaire was web based and all the motions were shown as videos with a stick figure character as shown in Figure 2. The duration of the videos ranged from 3 seconds (fast running) to 12 seconds (slow limping). Finnish language was used in the questions and the answers. The participants were gathered through work contacts and social media.



Fig. 2. An example of the stick figure representation portraying an angry walk.

The task given to the participants was to describe the seen motion with one verb or phrase (such as ‘swimming’ or ‘mountain climbing’) and from zero up to three modifiers (such as ‘colorfully’ or ‘very colorfully’). To make the answering easier we divided the videos into three sets and the participants could answer as many sets as they liked. Set A included all unmodified motions and had 24 videos, set B had 40 motions which were 50%-50% interpolations and the set C had the rest 60 motions. The total amount of videos was 124.

Our first research question is: Is the collective vocabulary used by a group of annotators larger than the target vocabulary given to actors of the motions and larger than the vocabulary of a single annotator? An answer to this question helps in deciding how much effort should be put into developing better search terms for motion databases. A way to find an answer to this question is to calculate how much of the collective vocabulary would be covered by the terms given to the actors and the words used by a single participant.

The second research question is: Can the variation in the collective vocabulary be decreased by finding synonyms? An answer to this question is important as joining search terms requires only a small change in a motion database. This question requires qualitative grouping and analysis of the vocabulary. Comparison of the distributions of the words over the motion samples can also help recognizing synonyms. We use FinnWordNet [14] as the source of the definitions

of the words. The translations of the Finnish words into English are also based on the FinnWordNet as it contains professionally made translations.

The third research question is: Do people have different opinions about the meaning of the used words? If there are large variations in how people use the same words, it would make building an optimal motion search much harder as the subjectivity would have to be taken in to account. Answering this question calls for plotting the distributions of the describing words on a space that is defined by numerical qualities of the motions.

To form a space which is based on the qualities of the motion, we calculate describing values called motion descriptors which include coordinates, velocities, accelerations and rotations as quaternions of each joint. From the velocities we used both absolute values and the velocities separately along the x, y and z axes. Also, we included the distances between pairs of body parts in a set that includes hips, neck, head, elbows, hands, knees and feet. To remove the variation caused by physical differences between the actors, we removed the personal means of descriptors as that has been found to help classification of motions [15].

5 Results

The participants of the questionnaire consisted of 9 females and 13 males with ages between 21 to 70 years. For the participants, the previous experiences with human motion were mainly linked to sports related hobbies. All 22 participants completed the set A, 10 also completed the set B and 2 participants did all the three sets of videos. Varying inflections which do not affect the meaning in this context such as 'walk' and 'walking' were cleaned from the data.

In the analysis, we have two points of view to the vocabularies. The first is the plain vocabulary where all the used words are considered equally important. The second is the shared vocabulary in which a word used by N persons is N times more important than a word used by one person. The distribution of the shared vocabulary is shown in Figure 3. From the figure we can see that 88 unique verbs and 233 unique modifiers were used by the participants. It also shows that the most common words explain a large part of the word usage, but there is also a long tail of rarely used words. For example nine most used verbs explain 50% of the shared vocabulary, but in order to reach to 90% one must consider 65 verbs.

Coverage of the words which were given to the actors and the words used by an average annotator are shown in Figure 4. Analysis of the vocabularies in Figure 4 is limited to the 24 videos in the set A as we needed to have annotations from all the participants to make a fair comparison. For the other analyses all the motion sets were used. Acted verbs plotted in Figure 4 have only coverage of 3% in the plain vocabulary as the three verbs given to actors are only a small part of the total 88 used unique verbs. However, when considering the shared vocabulary the three words have coverage of 29%. This comes from the fact that walking (kävelee) was used by all the 22 participants, running (juoksee) by 19 and limping (ontuu) by 14, while the total sum of usage counts was 190.

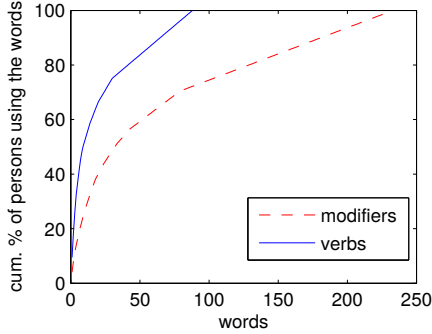


Fig. 3. Cumulative percentage of coverage of the shared vocabulary. The words are sorted from most used to the least used.

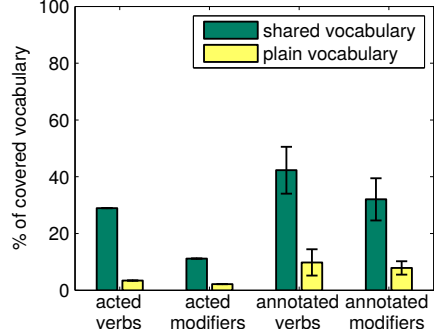


Fig. 4. Coverage of the plain vocabulary and shared vocabulary for verbs and modifiers given to the actors and average coverage of annotations from a single person. Standard deviation is shown for the averages.

Our first research question was related to how much the words given to the actors or the words of a single annotator cover of the overall vocabulary. The answer based on Figure 4 is that in the best case the words given to the actors cover a third of the vocabulary. Therefore, we can say that the set of words given to the actors of the motions would not enable making motion searches with natural language. The vocabulary of an average annotator does work better as it covers nearly 50% of the verbs in the shared vocabulary. Still, there is room for improvement. The coverage of the plain vocabulary is less than 10% which shows that having only one annotator will cause missing many rarely used words.

For finding synonyms, we used dictionary definitions of words and their translations to English as provided by FinnWordNet [14]. The words 'ontuu', 'nilkuttaa', and 'linkuttaa' are synonyms based on dictionary definitions and they all translate to 'limps' in English. This also shows that they could be considered to be alternative labels for the exactly same motions. From the modifiers we could not find synonyms as easily as from the verbs. Modifiers such as 'nopeasti' – 'fast' and 'kiirehtien' – 'hurriedly' can be considered to be similar, but whether they are synonyms is uncertain based on the data from the questionnaire.

For seeing the relationship between the numerical qualities of the recorded motions and the words used in the descriptions, we plotted the nine most frequent verbs (Fig. 5) and nine most frequent modifiers (Fig. 6) onto the PCA (principal component analysis) space based of the motion descriptors. To make the figures more readable we added small offsets to the overlapping pies to separate them. Web based versions of the two figures that also show the related animated motions are available at: <http://research.ics.aalto.fi/cog/mgl/>

Figure 5 shows that for many motions vast majority of the annotators are unanimous about the verbs. The three alternative words for 'limping' appear in the same area of the map and cause division between the annotators, but joining those words as synonyms would clean up the division. Two subjective divisions which cannot be accounted to synonyms are visible in the verbs. The first is

between 'jogging' and 'running'. It seems that the participants could not agree where to draw a line between the two actions. The second subjective division is between 'walking' and 'limping'. While 'walking' has an area that is almost unanimously 'walking', almost all of the 'limping' motions have also a small share of 'walking' in them.

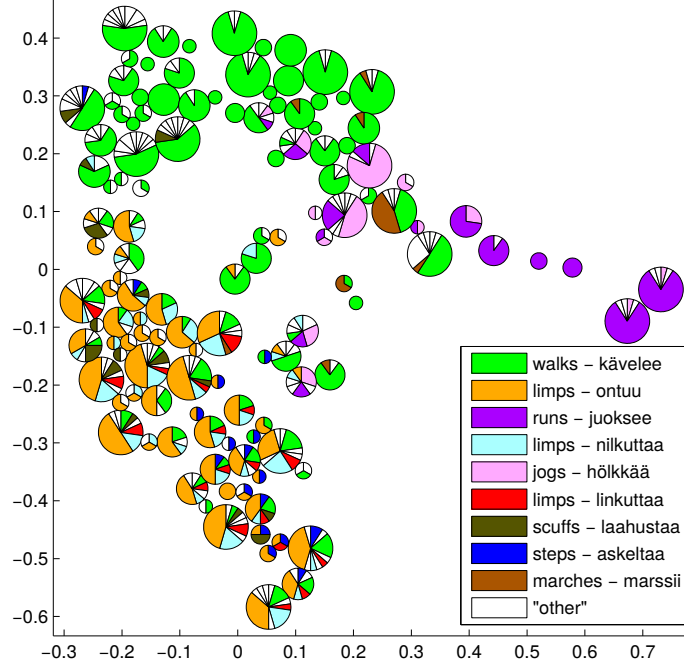


Fig. 5. Distributions of most common verbs for each motion mapped on the first and second normalized PCA components. The surface area of the pies is proportional to the number of answers and the position of the pies reflect the style of the motions.

Modifiers plotted in Figure 6 show that the participants were less unanimous in their answers than with verbs. There are even cases where almost all the participants gave different modifiers. Part of the variation can be explained by the fact the participants could give up to three modifiers. Still, even limiting the analysis to the first given modifiers, there would be no videos where one word would cover more than 50% of the answers if the video got more than two answers.

Many of the words are limited to a part of the PCA space. Verbs in Figure 5 form connected areas while modifiers can have disconnected distributions. For example the modifier 'slowly' appears mostly in the left side of Figure 6 where are the verbs 'walking' and 'limping', but also a few times near the center where the motions are described as 'jogging' or 'running'. The greater variation of modifiers is visible as the greater amount of the class 'other' than in the verbs.

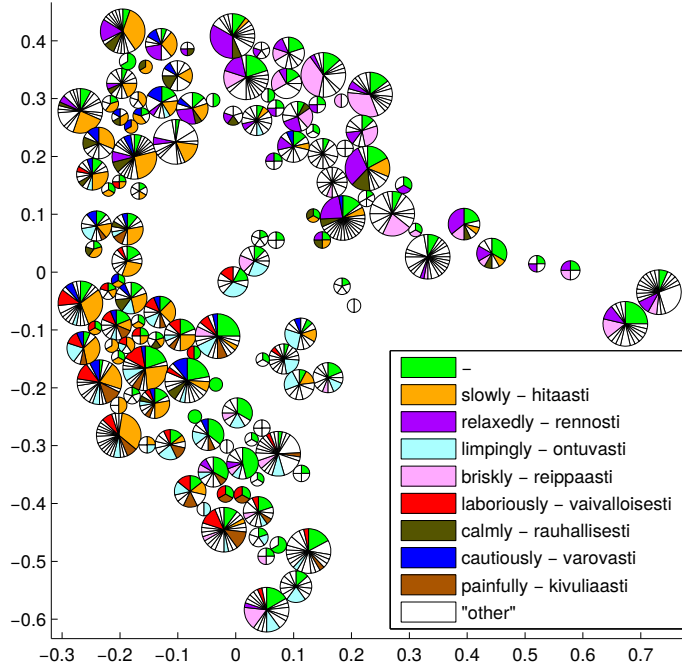


Fig. 6. Distributions of most common modifiers for each motion mapped on the first and second normalized PCA components. The surface area of the pies is proportional to the number of answers and the position of the pies reflect the style of the motions.

6 Discussion

How do the results of the questionnaire guide building a virtual actor that could be controlled with natural language? The first lesson is that relying only on the words given to the actors is not likely to cover the required vocabulary. Having one person annotate all the motions works better. However, the annotations of a single person are not enough in the cases where the borders between different verbs are subjective or when several synonyms exist. Modifiers are more challenging than verbs as the participants were far from unanimous and the modifiers did not always form continuous areas in the descriptor space.

For the verbs, hierarchical style of description could be beneficial as that would allow using words in a general sense and in a more specific sense. For example a parent category 'walking' could be divided into subcategories 'limping' and 'walking'. This way part of the subjectivity could be taken into account without needing more than one annotator. In practice, this could be achieved by giving the annotators two motions and a task to describe the motions with one verb.

Verb-modifier combinations could act as the most specific level of the description hierarchy. However, this would mean annotating a large amount of verb-modifier combinations. A more practical approach to handle modifiers could be to treat them as transitions in the descriptor space instead of areas of the space.

For a user instructing a virtual actor, this would mean first saying a verb and then saying a modifier to adjust the style of motion towards a desired direction. This approach could fix the problems caused by discontinuities in the distributions of modifiers. For example starting from walking and moving repeatedly towards a faster motion style would end up in a running motion. To find out what transitions correspond to which modifiers, a comparative task such as 'motion A is more X than motion B' should be given to the annotators.

While the questionnaire could always be made better, the main factor that speaks for the questionnaire is that the participants were able to freely select the words they used. If a selection of possible words had been given, it would have distorted the vocabularies of the participants. The decision to analyze the vocabularies as words-per-person instead of words-per-video makes our results more general. The counts for words-per-video are closely tied to the selection of videos, but the counts for words-per-person should not change dramatically even if part of the videos would be shown more times than others. One shortcoming in the questionnaire is the lack of repetitions. From data with repetitions, we could analyze how much of the variation in the descriptions is caused by difficulty of deciding between possible alternatives.

7 Conclusions and Future Work

In this paper, we presented results from a questionnaire in which participants were asked to describe videos of human locomotion with one verb and from zero up to three modifiers which were adjectives or adverbs. We analyzed the vocabulary as such and also in connection with numerical motion descriptors calculated from the motions. The results show that the original words given to the actors of the motions did not cover the used vocabulary of the participants viewing the motions. The vocabulary of a single annotator had better coverage, but the data would not help in cases where several synonyms exist for a verb or when the exact definition of a verb is not shared between the participants. The results also show that the modifiers used in describing the motions contain more variation than the verbs.

The main use case we considered was a virtual actor that can be controlled with natural language. Based on our results, we conclude that just linking each motion with the describing words would not allow controlling a virtual actor accurately. The linking would not take into account that meaning of verbs can be subjective and that modifiers are used variedly. The improvements we are planning include building a hierarchical vocabulary for verbs and modeling modifiers as transitions in the space defined by the numerical qualities of the motions. Realizing these improvements requires changing the annotation method from annotation of one motion at a time to annotation where similarities and differences are described between two motions.

Acknowledgments. This work has received funding from the Hecse doctoral program and the project Multimodally grounded language technology (254104) which is funded by the Academy of Finland.

References

1. Ahad, M., Tan, J., Kim, H., Ishikawa, S.: Action dataset - A survey. Proc. of SICE Annual Conference 2011 (SICE 2011), pp. 1650-1655. (2011)
2. Förger, K., Takala, T., Pugliese, R.: Authoring Rules for Bodily Interaction: From Example Clips to Continuous Motions. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 341-354. Springer, Heidelberg (2012)
3. Blumberg, B., Galyean, T.: Multi-level direction of autonomous creatures for real-time virtual environments. In: Mair, S.G., Cook, R. (eds.) Proc. of SIGGRAPH 1995, pp. 4754. ACM, New York (1995)
4. Perlin, K., Goldberg, A.: Improv: a system for scripting interactive actors in virtual worlds. In: Proc. of SIGGRAPH 1996, pp. 205-216. ACM, New York (1996)
5. Vilhjálmsón, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thórisson, K. R., Welbergen, H., Werf, R. J.: The Behavior Markup Language: Recent Developments and Challenges. In: Pelachaud, C., Martin, J., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS vol. 4722, pp. 99-111. Springer, Heidelberg (2007)
6. Hachimura, K., Takashina, K., Yoshimura, M.: Analysis and evaluation of dancing movement based on LMA. IEEE International Workshop on Robot and Human Interactive Communication 2005 (ROMAN 2005), pp. 294-299. IEEE (2005)
7. Talbot, C., Youngblood, G.: Spatial Cues in Hamlet. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. Conf. (eds.) IVA 2012. LNCS, vol. 7502, pp. 252-259. Springer, Heidelberg. (2012)
8. Honkela, T., Raitio, J., Lagus, K., Nieminen, I., Honkela, N., Pantzar, M.: Subjects on objects in contexts: Using GICA method to quantify epistemological subjectivity. In Proc. of International Joint Conference on Neural Networks (IJCNN 2012), pp. 2875-2883. (2012)
9. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing, Vol. 28, Issue 6, pp. 976-990. (2010)
10. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently Scaling up Crowdsourced Video Annotation. International Journal of Computer Vision, Vol. 101, Issue 1, pp. 184-204. Springer US. (2012)
11. Rose, C., Bodenheimer, B., Cohen, M. F.: Verbs and Adverbs: Multidimensional Motion Interpolation Using Radial Basis Functions. Computer Graphics and Applications, IEEE, vol.18, no. 5, pp. 32-40. (1998)
12. Kovar, L., Gleicher, M.: Automated Extraction and Parameterization of Motions in Large Data Sets. In: Marks, J. (ed.) Proc. of SIGGRAPH 2004, pp. 559-568. ACM, New York. (2004)
13. Shoemake, K.: Animating rotation with quaternion curves. ACM SIGGRAPH computer graphics, Vol. 19(3), pp. 245-254. (1985)
14. Lindén, K., Carlson, L. FinnWordNet - WordNet på finska via översättning. (In English: FinnWordNet - Finnish WordNet by Translation). LexicoNordica - Nordic Journal of Lexicography, vol. 17, pp. 119-140. (2010)
15. Bernhardt, D., Robinson, P.: Detecting affect from non-stylised body motions. In: Paiva, A., Prada, R. (eds.) Affective Computing and Intelligent Interaction, pp. 59-70. Springer, Heidelberg. (2007)

Publication V

Klaus Förger and Tapio Takala. Animating with Style: Defining Expressive Semantics of Motion. *The Visual Computer*, Online First Articles, February 2015.

© 2015 Springer Science and Business Media.

Reprinted with permission.

Note: Due to copyright restrictions, this electronic version of the dissertation does not contain the final versions of the publications, but instead the versions that have not been edited by the publishers.

The final publications can be found from the web sites of the publishers.

Animating with Style: Defining Expressive Semantics of Motion

Klaus Förger · Tapio Takala

pre-print version

Abstract Actions performed by a virtual character can be controlled with verbal commands such as ‘walk five steps forward’. Similar control of the motion style, meaning how the actions are performed, is complicated by the ambiguity of describing individual motions with phrases such as ‘aggressive walking’. In this paper, we present a method for controlling motion style with relative commands such as ‘do the same, but more sadly’. Based on acted example motions, comparative annotations, and a set of calculated motion features, relative styles can be defined as vectors in the feature space. We present a new method for creating these style vectors by finding out which features are essential for a style to be perceived and eliminating those that show only incidental correlations with the style. We show with a user study that our feature selection procedure is more accurate than earlier methods for creating style vectors, and that the style definitions generalize across different actors and annotators. We also present a tool enabling interactive control of parametric motion synthesis by verbal commands. As the control method is independent from the generation of motion, it can be applied to virtually any parametric synthesis method.

Keywords Computer animation · Human motion · Motion style · Motion synthesis · Style vector · Feature extraction · Feature selection · Verbal description of motion style

1 Introduction

An animator would often like to control virtual characters the way a theater director does, giving verbal commands rather than manipulating individual limbs like a puppeteer. Goal-oriented actions can be created with existing motion synthesis methods [7, 12, 15], even by scripting the requirements in natural language [14]. Different styles, meaning how actions are performed, can be produced with parametric and example-based methods [8, 17, 18, 20, 23]. However, controlling style with verbal attributes has received less attention. Many motion synthesis methods do not have a direct relationship between input parameters and the resulting styles. To fill this gap, we present a method that allows accurate control of motion style with high-level natural language commands. A similar approach has been applied in controlling color themes to create affective changes in images [24].

We define motion style to be a visually recognizable aspect of captured or synthesized motion. Furthermore, we define absolute style as one that can be perceived from individual motions and relative style as that perceived from differences between motions. Motion styles can be modeled numerically or described with natural language. In this work we seek correspondences between these two, in order to computationally define, identify and control styles in animation.

Judgements about styles are more vague and subjective than about goal-oriented actions. For example when a character tries to reach an object, we can measure if the hand touches the object, but it is less clear if the hand motion is seen as aggressive, gentle or nervous. Several styles may be perceived in one action. Often there is a gradual change from one style to another, such as from a lazy to an energetic walk. Styles can be

characterized by physical adjectives (e.g. fast or slow) and emotional expressions (sadly, aggressively, etc.). In natural language we may describe absolute styles with phrases such as 'slow movement' or 'walking like Mick Jagger', and relative styles by comparative forms such as 'more aggressive'.

Automated identification of styles is possible by associating verbal descriptions with recorded example motions, which in turn are represented by numerical features. An absolute style can be represented as a collection of individual example motions and modeled as a statistical distribution. Analogously, a relative style can be represented as a collection of motion pairs showing differences in that style, and the distribution of differences can be modeled as a vector in feature space [27].

The main contribution of this paper is a new and more accurate method for constructing vector based definitions of relative styles. The basic idea is for each style to find the essential features that in all examples unanimously change when the amount of perceived style changes, and to ignore other features. To accomplish this we need systematical acting of example motions, perceptual annotation of the styles, and individual feature selection for each style.

We also present an implemented system for controlling parametric motion synthesis with the style definitions. The control is indirect as we automatically generate variations of a motion and evaluate which variation shows the desired style best, and then change the synthesis parameters accordingly. Therefore, the style control is independent of the synthesis method. Furthermore, we show with user tests that the produced style vectors accurately predict perceptual evaluations of styles and that the style definitions generalize from one actor to others. Promising results have been achieved with relative styles fast, slow, aggressive, lazy, excited, energetic, calm, limping, healthy, depressed and busy.

We limit our practical experiments to human locomotion, such as walking or running, characterized by physical adjectives and emotional expressions. However, the method for evaluating style is not limited to locomotion and may be extended to non-cyclic motions. We leave out symbolic aspects of conversational gestures that require knowledge of a specific culture to be correctly understood. However, the manner how gestures, such as waving a fist, are performed could still be controlled with our method.

In the following sections, we first review previous work on motion style. Then we present our method, detailed by calculation of low level motion features, creation of the style definitions, and the style-based control of motion synthesis. Finally, a study is described on how well the style definitions and motions produced by style-

controlled interpolation synthesis match human perceptions. We conclude with limitations and potential extensions of the method.

2 Related Work

In this section, we review techniques for editing style in captured motion, and studies on the perception and verbal description of styles. Based on these, we discuss how style semantics and low-level motion synthesis methods have been matched.

2.1 Motion Style in Computer Animation

Traditional motion capture does not separate style from action but the motion is replayed as it is. Only space-time constraints necessary for retargeting the motion to a different character are imposed [7]. All stylistic variations are performed by the real actor. If needed for later use, they are stored in a database and then selected by indexing with a style attribute [12].

One way to approach style explicitly is to model it as the difference between a specific and a regular action. As two captured motions seldom are in the exactly same phase, warping in space and time is usually needed to make them comparable. Hsu et al [8] used machine learning to construct a linear time-invariant model with example motion pairs to model the stylistic difference between the motions. The model enables transforming new neutrally acted motions to the learned style in real-time.

When changing motion styles, we do not always need to have a specific motion sample as a target. Instead, we can try out how editing low level motion data affects the perceived styles. Bruderlin and Williams [2] proposed equalization in frequency space as a tool, demonstrating for example calm and nervous movement resulting from low and high pass filtering, respectively. Min and Chai [14] developed a generative graph model for motion synthesis, separating finite structural variations for content actions (such as walking) and continuous style related variations (such as walking speed and step size). In these works expressive style is not modeled explicitly, and thus cannot be controlled directly.

Yet another approach is to model a motion signal as a sum of editable components. This allows both analysis of various important features, and synthesis as recombination of components. Fourier spectrum edited by filtering [2] is one example of this type of modeling. Alternatively, action sequences can be statistically modeled as combinations of base functions produced with Principal Component Analysis (PCA) [20, 23] or Independent

Component Analysis (ICA) [18]. With PCA and additional statistical analysis, emotional and gender-related styles such as nervous, sad, relaxed, male and female have been successfully identified [20,21].

Our method was inspired by these works. Particularly, we adopted from Bruderlin and Williams [2] the frequency components as motion features.

2.2 Perception and Verbal Description of Style

One of the earliest systems enabling semantic control of motion style identified verbs as distinct actions and adverbs as versions of the actions in different styles [17]. We basically follow this, although the distinction is not strict. Some verbs include a stylistic aspect, against which adverbs tend to be relative modifications. For example, scuffing may imply dampened motion, and slow running may be almost the same as fast walking, and still all these are variations of the same action of locomotion. There are also complex interactions between different styles conceptualized as adverbs, as one tends to imply another. For example, the perceived gender of a moving character can be affected by the perceived amount of anger and sadness [10].

Many methods exist for recognition of actions based on groups of individual examples [16]. As absolute style can be represented by individual examples, the same methods could be applied. However, we concentrate on relative style as that allows precise iterative fine tuning of styles.

In a recent study about natural language in describing human motion, verbal annotations of motion samples were related to their calculated low level features such as distances between body parts, velocities, accelerations and absolute positions [6]. Plotting the results against PCA components of the features (Fig. 1) indicates that verbs tend to be localized in partly overlapping clusters, whereas adverbs are less unanimously annotated (Fig. 2). This coincides with the intuitive understanding that unlike verbs that can be used alone, adverbs are linguistic modifiers that tend to reflect as directions rather than locations in the feature space. Fine control of style with absolute definitions would require dividing all verb clusters to smaller pieces such as slow walking, aggressive walking and sad walking. This would require a lot more samples and annotations than defining only generic actions.

Motion style in dancing can be described with Laban notation, based on expert terms related to effort and shape of motions [3]. The definitions of expert terms need to be learned explicitly, while natural language does not require additional training. Also, the terms

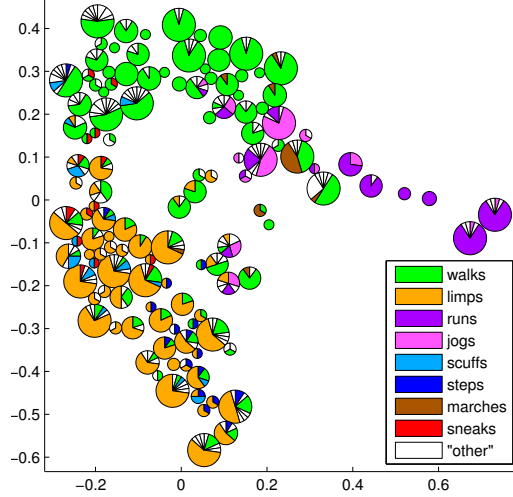


Fig. 1 Motions, annotated with verbs, mapped on the first and second normalized PCA components of numerical motion features [6]. The surface area of the pies is proportional to the number of annotations and the distances between the pies reflect the similarity of the motions

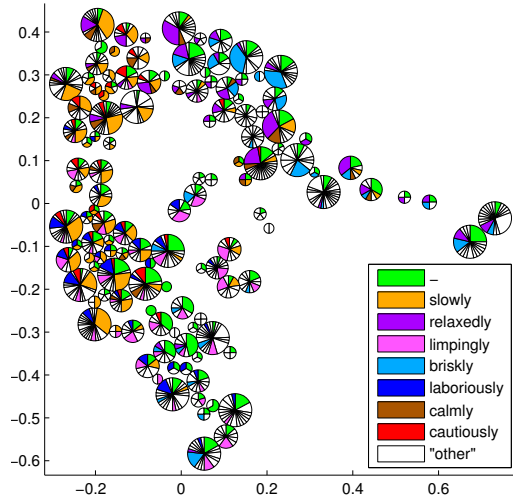


Fig. 2 Same motions as in Fig. 1, annotated with adverbs, mapped on the first and second normalized PCA components of numerical motion features [6]. ("-" colored with green means that no adverb was given.)

that have precise definitions inside dancing may not be sensible in other motion categories. This is why we use laymen terms instead of expert definitions.

Psychological studies on perception and recognition of affects in bodily motion mostly assume discrete non-overlapping classes of emotions or use abstract affective dimensions instead of natural language [11]. For this reason they are not directly applicable to controlling continuous motion in animation. Important considerations for psychologists are whether acted or authentic emotions should be studied and if the ground truth comes from actors or observers [11]. These questions are more straightforward in animation as synthetic virtual characters do not have real intentions or feelings. What counts is observers' perception alone.

For animators, the stimuli used in studies about emotions in human motion may look very simplified, often consisting only of point-lights on a black background [10,21]. One reason for simplified appearance is that giving too many details, such as facial expressions, may divert attention away from the motion or modulate the perception of emotions [1,5]. In our work, we compromise and use a stick figure. It lacks details but helps in perceiving postural differences between motions.

Our general goal is to allow animators to adjust style of synthesized motion by words in natural language. In earlier research, this approach has been taken with action commands (verbs) such as 'walk five steps and pick up the object' [14]. An alternative non-verbal approach has been to sketch key-poses of a motion sequence, allowing more precise positioning of the actions but still lacking control over other style related attributes [26]. In this paper, we focus on refining the actions by relative commands such as 'do the same, but more slowly and sadly'.

2.3 Matching Style Semantics and Synthesis

Given a verbal description, a corresponding motion can be produced in different ways. A rich database of motion samples acted in all possible styles would be easy to use but impractical to generate. More viable is a parametric model, mapping verbal instructions to navigation in the parameter space of a synthesis engine.

Motion interpolation is a parametric method that can produce a continuous range of styles between compatible original samples [17]. However, the results of interpolation cannot be accurately predicted from the parameters and verbal descriptions of the original samples when styles are mixed. For example, interpolation between sad and aggressive motions could end up looking neutral or showing sadness in the pose and aggression in accelerations.

Although motion inaccuracy, such as foot sliding, is a problem in goal-oriented actions, it can be alleviated by sophisticated interpolation methods [15]. The same has not been possible with styles. In lack of automatic evaluation, manual annotation of several interpolated samples is necessary to make reliable predictions, and the number of possible interpolations grows combinatorially with the number of new original motions.

Modeling motion styles with a functional decomposition (PCA or ICA) allows direct synthesis by recombination of the desired components [20,18,23]. These methods offer orthogonal parameters which can be tuned independently to reach a desired style. However, the parameters do not automatically match with natural language descriptions of styles that may be partially synonyms or opposites. Every parameter can affect several perceived styles depending on how the styles were correlated in the original motions used in calculating the components. For example, adjusting emotional styles described with phrases such as sadness or relaxedness can also affect styles related to the body shape of the character [20]. Another problem is that although these methods enable extrapolation of motion from one sample to new situations, such as a different speed, extrapolation carries a risk of producing motions that are not physically realistic if not used carefully. For reasons stated above, we think component based methods are not suitable for describing relative styles with natural language.

Treating motion signals as frequency bands is another candidate for style synthesis [2,22]. For example, Bruderlin and Williams [2] report that amplifying high frequencies can add "a nervous twitch" to a walking motion. However, this may not be the case for all input motions. Assigning meaning to the parameters may be even more difficult than with interpolation or component based methods, as the frequencies may have different effects depending on the input motion.

2.4 Vector Based Style Definitions

Relative differences in style between motions can be modeled in a numerical feature space as style vectors representing the direction of increasing perceived style. Zhuang et al. [27] defined style vectors statistically as differences between means of motion samples performed by an actor repetitively in different styles (Fig. 3a). Styles of new motions can then be compared by calculating their difference in projection onto the style vector (Fig. 3b).

The idea of style vectors is to provide a numerical measure for relative style differences which in turn can

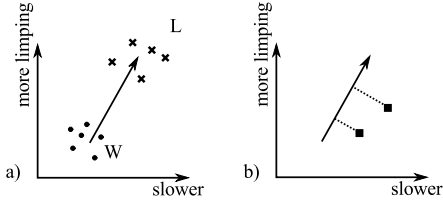


Fig. 3 a) a training set of walking (W) and limping (L) motions, with the style vector that points toward the learned limping style. b) styles of two motions compared by projecting them onto the style vector

be used for iteratively adjusting motion synthesis parameters towards a desired style. Previously a natural language description for the style vectors was more of an afterthought, and the descriptions were not validated in practice [27]. In our work, we follow the same principle, but consider an accurate match between numerical and linguistic descriptions to be vital for a usable style definition. This pushed us to develop the method further.

3 Catching the Essence of a Style

Our aim is to let animators control motion styles by computational features. But how do we know which of them are relevant for a style? As some styles are related to postures and others to limb velocities, the same set of features is not relevant for all.

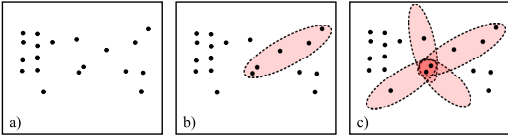


Fig. 4 Features used to model styles. a) the set of all computed features b) the features that correlate in stereotypically acted examples, c) subsets formed by acting the same style in different ways - the essential features are in the intersection

In the set of all potential features (Fig. 4a), we want to identify those relevant for each particular style. To find them, we may ask an actor to perform motions in varying intensities and calculate which features consistently change when the style gets stronger (Fig. 4b).

However, if some features correlate with multiple styles, they cannot make a distinction between those. For example, the style vector in Figure 3 would judge a slower but otherwise normal walk as limping, because limping typically is a slow motion.

We propose a solution where an actor performs variations of one style combined with other simultaneous

styles instead of just repetitions of one style (for example, 'sad+fast' and 'sad+aggressive' in addition to plain 'sad'). This way we can identify the essential features common to all cases where a style difference appears (intersection of ellipses in Fig. 4c).

As a lot of irrelevant features may get dropped out with this refinement, our approach requires the number of original features to be high in order to ensure that at least some essential features can be found.

4 Motion Synthesis with Refined Style Vectors

Below we present a method for calculating style vectors that more accurately identify different styles. We first describe the feature set used for evaluating styles in motion, and then give details on how style vectors are constructed from acting a set of sample motions through perceptual annotation to calculation of the vectors (Fig. 5).

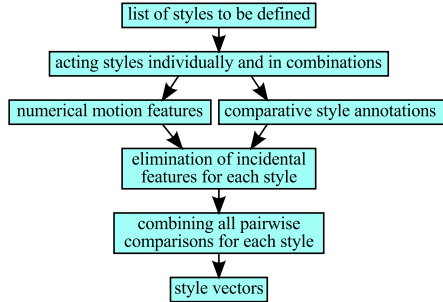


Fig. 5 Overall process of creating style vectors

We also present a system for controlling interpolation based motion synthesis by style vectors. Motion control starts from one sample that presents the desired action in any style. An animator can then use natural language based descriptors to adjust the motion towards the desired style while keeping action the same.

Our method does not rely on mapping one style to one synthesis parameter. Instead, we build style control as gradual navigation in the parameter space by solving a parameter combination that best produces a desired change in style. Virtually any synthesis method can be used, as we treat motion synthesis as a black box, containing possible post-processing steps such as inverse kinematics.

4.1 Motion Features

Computational comparison of motions requires numerical motion features. Our aim is to find generic features that can be used for detecting style in the data captured from any type of human movement. Raw motion capture data consists of time varying signals with values for each frame. From that, we calculate a set of features where each value represents a short motion segment (approximately 2-10 seconds long) as styles are partly dynamic properties which cannot be seen in single frames. In order to accurately identify many styles, we need a lot of potential features, out of which a suitable subset is defined for each style.

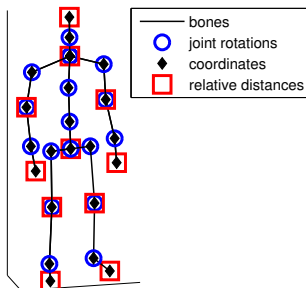


Fig. 6 Skeleton structure of the motion capture data showing bones of constant length, the 18 rotating joints, and the 22 points of body, the coordinates of which are used in calculation of our features, either as such or as 55 relative distances (see Table 1)

We model the human body as a hierarchical skeleton structure (Fig. 6) with constant bone lengths and joint rotations that vary in each frame. The lowest level of per frame data includes coordinates, velocities, accelerations and rotations at joints (expressed as quaternions). From the velocities we take both absolute values and the components along axes of the character’s local coordinate system. Also, we include all pairwise distances between pelvis, neck, head, elbows, hands, knees and feet. This set of motion signals has been useful in recognition of action verbs [6].

To expand the set to be more suitable for motion style, we also calculate how the signals vary in frequency domain [21]. Following the filtering method by Bruderlin and Williams [2], we divide the initial signals into seven frequency bands. From the original signal captured with 100 Hz sample rate we extract approximately the ranges 0.1–0.5–1.1–2.2–4.5–9–18–50 Hz. Thus, we have 301 motion signals (Table 1) in eight versions (original and the seven frequency bands) making 2408 signals altogether. To summarize the signals as num-

bers that describe whole motion segments, we take their means and standard deviations over all frames in the segment. With this the number of dimensions in our feature space becomes 4816.

As similar movements can be performed using the left or the right side of the body, we consider these to be of identical style. To make the 4816 features the same in both cases, we first checked which of them already are mirror invariant. For the rest, taking absolute values makes equal the features directly related to sideways motion. Instead of velocities for the left and the right hand, we sort them pairwise to get velocities of the slower hand and the faster hand. For features related to sideways motion of paired limbs, we multiply the value of one side with -1 and then apply the sorting.

To make our features invariant of body size, we divide all coordinate values by the height of the actor, which also scales velocities and accelerations to comparative ranges. Other normalizations between actors are not applied as they could harm the identification of styles related to bodily structures.

4.2 Creating Vector Based Style Definitions

For defining styles, we asked an actor to perform a regular walk and eight style variations relative to that: fast, slow, relaxed, tense, angry, sad, limping, and excited. We also asked the actor to perform combinations of two styles (all except fast+slow and relaxed+tense as those styles can be considered mutually exclusive), making altogether 35 motions. Our amateur actor was able to perform the style combinations with noticeable variation, although it required him to consciously analyze different aspects of individual styles and devise a way to combine them.

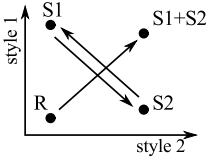
To obtain perceived differences in styles, motions were annotated in pairs using a questionnaire, where each page displayed two video samples for visual comparison. The annotator provided up to three words to describe their differences and quantified them on a scale ‘a little/somewhat/a lot more’. To avoid biased answers, we did not offer any predefined choices for the words.

For avoiding excessive manual work, we limited annotation to the motion pairs that most likely show differences between styles. As depicted in Figure 7, we included those where a double style (i.e. actor instructed to present two styles simultaneously) is compared with a regular motion (26 pairs), and all possible combinations of single styles (56 pairs). The latter comparison was done in both ways (shown separately as swapped pairs) to encourage the annotator to name opposite differences. Altogether this reduced our questionnaire

Table 1 The motion signals from which the features representing motion segments are derived

Signals	Consisting of	Number of dimensions
Positions in local coordinate system of the character	22 body parts with 3 values each (The center of pelvis is the root joint for which only elevation coordinate is taken into account.)	64
Absolute velocities	22 body parts with 1 value each	22
Velocities along local coordinate axes	22 body parts with 3 values each	66
Absolute accelerations	22 body parts with 1 value each	22
Distances between pelvis, neck, head, elbows, hands, knees and feet	pairwise combinations of the 11 body parts with 1 value each	55
Joint rotations as quaternions	18 joints with 4 channels each (The base of neck and shoulders has three overlapping joints. The three bones starting from the central hip represent the pelvis and share the same rotation.)	72
Total		301

from 1190 possible comparisons to only 82. The annotation was done by one of the authors. To ensure that the results are not biased, we later made a validation by crowdsourcing.


Fig. 7 Motion pairs used in the comparative annotation: double styles (S1+S2) against regular motion (R), and single styles both ways against each other (S1 and S2)

From the annotation data, we selected 13 most common verbal descriptions that appeared in at least five example pairs: fast, slow, aggressive, lazy, excited, energetic, calm, limping, healthy, depressed, busy, relaxed and tense. For these styles we proceeded to calculate style vectors. Eighteen other verbal descriptions appeared in the annotation data less than five times.

For each style we collected the results of annotation in form of Table 2, with one row for each pairwise comparison where the style was seen (N varying from 5 to 25 depending on how many motion pairs got labeled with the style). The vector \mathbf{c}_x consists of the differences of numerical features between the compared motions. The perceived style difference a_x is a value scaled from 'a little/somewhat/a lot more' to 1, 2 or 3 respectively. In the last column A_x is the sum of all difference values given in the comparison for any styles. In our case, as the motion pairs were shown only once during the questionnaire and the annotator may give at most three styles per motion pair, the maximum value for A was 9.

From this table we identify those features that agree in all comparisons, i.e. we select those y for which $c_{x,y}$ has the same sign in all rows $x=1\dots N$. These are the

Table 2 Summary of collected data for one annotated style, with a row for each pairwise comparison in the questionnaire.

Vector of feature differences in the displayed motions	Perceived style difference	Sum of all perceived differences
$\mathbf{c}_1 = \langle c_{1,1}, c_{1,2}, \dots, c_{1,4816} \rangle$	a_1	A_1
$\mathbf{c}_2 = \langle c_{2,1}, c_{2,2}, \dots, c_{2,4816} \rangle$	a_2	A_2
\dots	\dots	\dots
$\mathbf{c}_N = \langle c_{N,1}, c_{N,2}, \dots, c_{N,4816} \rangle$	a_N	A_N

essential features for recognizing the particular style. The other features, which are incidental, we eliminate from the style vector, thus effectively reducing dimensionality of the feature space. However, as the essential features are not the same for all styles, we retain all original features, only weighting them for this style by multipliers defined as:

$$\begin{aligned} m_y &= 1 \text{ if } \forall x : c_{x,y} \geq 0 \vee \forall x : c_{x,y} \leq 0 \\ m_y &= 0 \text{ if } \exists x : c_{x,y} > 0 \wedge \exists x : c_{x,y} < 0 \end{aligned} \quad (1)$$

The multipliers are then used to create eliminated versions \mathbf{s}_x of the difference vectors \mathbf{c}_x :

$$\mathbf{s}_x = \begin{bmatrix} m_1 & 0 & \dots & 0 \\ 0 & m_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_{4816} \end{bmatrix} \mathbf{c}_x \quad (2)$$

The final style vector \mathbf{u} is formed as a weighted mean of the reduced difference vectors

$$\mathbf{u} = \frac{1}{N} \cdot \sum_{x=1}^N \left(\left(\frac{a_x}{A_x} \right)^2 \cdot a_x \cdot \mathbf{s}_x \right) \quad (3)$$

where we use the style difference a_x as weight, normalized by its proportion of all style differences given in that comparison (A_x). The previous method [27] treated all motion examples equally and used an unweighted mean as a style vector. We explicitly try to utilize the variance in motions and emphasize the features that contribute to a style. Therefore, we give more

weight to the comparisons where the amount of the annotated style is large (multiplication with a_x) and makes up a large percentage of all styles (squared term).

4.3 Vector Based Control of Motion Synthesis

Out of many possible synthesis methods, we selected motion interpolation as it is widely used in animation software and games for producing blends between motions. Also, interpolation is less prone to unnatural results than those methods that extrapolate outside the range of recorded examples.

The parameters of interpolation tell how much the end result should resemble each input motion. To avoid extrapolation, the parameters must be non-negative and sum to 100%. As the input motions for the interpolation, we took the same 35 locomotions with varying styles that were used in creating the style vectors. We matched the times when the feet get on and off the ground. After time warping, root positions of the character were interpolated linearly. For joint rotations, normalized linear interpolation (*nlerp*) of quaternions [19] was applied. Acceleration spikes that may appear as side effects of time warping were smoothed in a post-processing step. Note that time warping was needed for the interpolation synthesis only. For evaluation of styles – the essential part of our method – it is sufficient that the motions contain the same actions; even the number of cyclic repetitions could vary.

What is an optimal control method for motion synthesis depends on the predictability and cost of synthesizing individual motions. A brute force approach would be to produce style variations randomly and pick one closest to the desired style. Instead, we evaluate the effect of offsetting each synthesis parameter individually and then solve the best combination of changes to the parameters, effectively performing a gradient search.

Motion interpolation is a locally stable synthesis method, meaning that adding a small offset to a parameter has a small predictable effect on the produced motion. Then we can model the effects of parametric changes linearly with a Jacobian matrix:

$$\mathbf{J}\mathbf{x} = \mathbf{u} \quad (4)$$

where each column of \mathbf{J}_k is the vector of partial changes in feature values \mathbf{u} caused by changing the corresponding parameter x_k alone.

Looking for a desired style change \mathbf{u} , the required parameter change can be found by solving this equation for \mathbf{x} . An exact solution is unlikely as the number of parameters is much lower than the number of motion features. Pseudoinverse is a suggested solution

in inverse kinematics, but we used an off-the-shelf least squares solver (*lsqnonneg* in Matlab) as it easily finds an approximate solution with minimal error while keeping the synthesis parameters inside the interpolation range. The steps for finding new synthesis parameters with a desired style are listed in Algorithm 1.

Algorithm 1 Finding new synthesis parameters

```

1: Start with arbitrary parameters (param) that produce the
   desired action
2: while user not satisfied do
3:   User selects a desired style change (style vector u)
4:    $\mathbf{J} = \text{CONSTRUCTJACOBIANAT}(\textit{param})$ 
5:    $\mathbf{x} = \text{SOLVELINEARSYSTEM}(\mathbf{J}, \mathbf{u})$ 
6:    $\mathbf{x} = \text{SCALETOLIMITMAXIMUMPARAMETERCHANGE}(\mathbf{x})$ 
7:    $\textit{param} = \textit{param} + \mathbf{x}$ 
8:    $\textit{param} = \text{SCALETO100PERCENT}(\textit{param})$ 
9:    $\text{SYNTHESIZEANIMATIONWITH}(\textit{param})$ 
10: end while
```

We built a user interface (shown in Online Resource 1) for trying out the style control in practice. It shows an animation of the current motion and allows relative adjustment towards a desired style. The user may control either the desired change in each style and let the algorithm tune the parameters, or adjust the 35 synthesis parameters directly. In our experience, the latter was more tedious especially when trying to simultaneously get more than one style visible in the motion.

Examples of motions produced with our system are shown in Figure 8 and as animations in Online Resource 1. They show how aspects such as step size, velocities, posture, and limb trajectories behave when the style is changed. The trajectories show that excess feet sliding did not appear even though inverse kinematics was not used.

The visualization method in Figure 8 appears to be a novel technique for presenting motion style in still images. In our opinion it shows the dynamics of motion better than a series of stick figures.

Our implementation in a multicore computer is fast enough for interactive applications. Main part of the computation is spent on synthesizing motion trials for calculating the Jacobian. As the motions are synthesized independently, the task can be distributed to the available cores in parallel.

5 Experimental Validation

In this section, we test how well the style vectors work in practice. In our first experiment the styles seen by human observers in a set of walking motions are compared with automatic evaluations done with style vec-

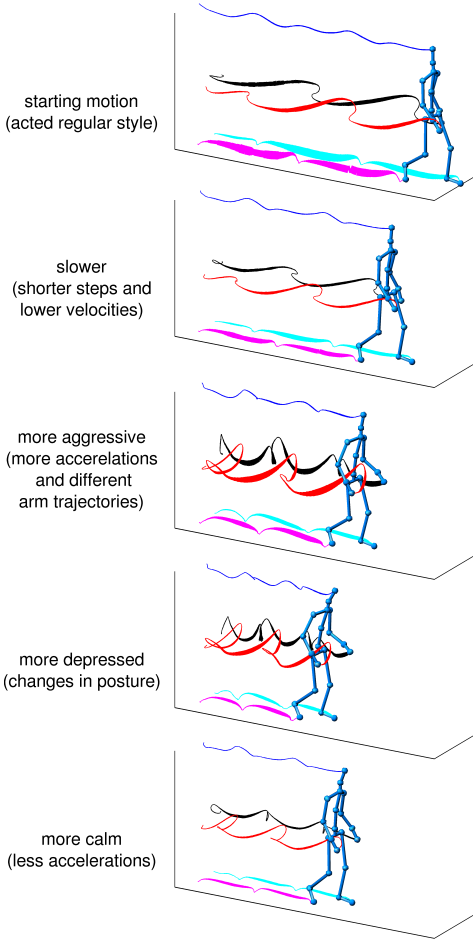


Fig. 8 Control of walking style by relative style commands. Starting from an acted motion on top, each picture shows incremental changes towards the bottom. Trajectories leading to the final pose are shown for head, hands and feet. Line thickness indicates velocity

tors. Next, we assess the impact of our feature elimination process to the quality of style vectors. In a second experiment, we test if human observers recognize style adjustments produced by motion synthesis with our method.

5.1 Validation of Style Definitions

Accurate style vectors should enable automatic evaluation of styles acted by new actors and the result should be agreed on by new human observers. To test this aspect, we produced a new set of locomotions, an-

notated the perceived style differences with a crowd-sourced questionnaire, and compared the annotations to the style evaluations produced with style vectors.

The new set of locomotions was performed by four actors and acted in similar styles as before when creating the style vectors. As some actors were not able to perform all style combinations properly, we enriched the motion set by also creating 50/50% interpolations between the actor’s motions, disregarding those where interpolation caused visible artifacts. From this set of 168 unique motions we produced 347 pairs that in our opinion differed in at least one of the annotated styles.

The numbers of motion pairs and unique motions for each style are shown in Table 3. Some pairs were used as examples of several style differences. Also, one motion sample may appear in several pairs.

The motion pairs were shown as stick figure animations on a web page. The observer was given one of the 13 style descriptions and asked to evaluate on a five-point scale (*much more* / *slightly more* / *equal amount* / *slightly less* / *much less*) how much one sample shows the given style compared with the other. The presentation order was balanced so that each pair of videos was shown twice, with the order of comparison reversed. We ran the questionnaire using the web-based crowdsourcing platform CrowdFlower and received a total of 10569 ratings randomly distributed among 456 participants. For quality control, we included a test in the start that required separating pairs of 100% identical videos from pairs that showed extremes of opposite styles. This way we could be sure that all the participants were at least able to view the videos.

The crowdsourcing service provided us with six or seven ratings for every combination of a style word and a video pair. At this point, we pruned the data by only keeping those combinations in which majority (at least four) of the participants had agreed on which of the videos had more of the mentioned style. This reduced the number of style word and video pair combinations from the original 1041 to 952 which we took as ground truth for our tests.

To measure the accuracy of style definitions we tested how well automatic evaluation of style differences agrees with the ratings of the majority of the questionnaire participants. Style difference of a motion pair was automatically evaluated by calculating the dot product between the style vector and the vector of feature differences between the two motions. The sign of the dot product was taken as indication of which motion shows more style. In this setup, the chance level for accuracy is 50%.

The results, shown in Table 3, tell that most of the style definitions reached at least 90% accuracy and sev-

Table 3 Accuracies of automatic style evaluation

Style word	Accuracy %	Number of motion pairs	Number of unique motions
fast	100	170	93
slow	100	166	89
aggressive	100	70	57
lazy	100	62	46
excited	100	28	30
energetic	98.8	80	49
calm	98.5	65	48
limping	97.1	68	57
healthy	96.8	63	47
depressed	92.0	88	46
busy	90.0	20	33
relaxed	77.1	35	39
tense	59.5	37	43

eral even got 100% of the test pairs correct. This is a good result as the style vectors were produced from motions of one actor and annotations of one person, while the evaluation set had four actors and hundreds of observers.

We also observe that the styles relaxed and tense were less accurately defined than the other styles. The use of style vectors for controlling motion synthesis sets an acceptability level for the accuracy. For example, if an animator asks for a more relaxed motion, with 77% accuracy the system would give a more relaxed motion only three times out of four. Therefore we dropped the relaxed and tense style definitions from the rest of the experiments.

5.2 Assessment of the Impact of Incidental Features

The main difference between our method and the previously published one [27] is the elimination of incidental features. If the elimination step works, it should remove false correlations between styles and preserve correlations only when the styles defined are semantically overlapping. In order to evaluate the impact of feature elimination, we calculated pairwise correlations between style vectors produced without elimination (Fig. 9) and compared them with those produced by our elimination process (Fig. 10).

The correlations in Figures 9 and 10 reveal that the elimination step does make the style vectors more independent from each other. For example, before elimination the styles limping and slow have a correlation of 0.8. This means that increasing the amount of visible limping would also increase the amount of slowness (cf. Fig. 3). However, after the elimination step, the styles limping and slow have a correlation that rounds to 0.0 meaning that with these style vectors, adjusting the

	limping	healthy	depressed	slow	lazy	calm	aggressive	energetic	busy	excited	fast
limping	1.0	-0.9	0.7	0.8	0.9	0.5	-0.3	-0.6	-0.8	-0.6	-0.8
healthy	-0.9	1.0	-0.4	-0.7	-0.6	-0.5	0.4	0.6	0.6	0.5	0.7
depressed	0.7	-0.4	1.0	0.8	0.8	0.2	-0.2	-0.5	-0.6	-0.2	-0.7
slow	0.8	-0.7	0.8	1.0	1.0	0.7	-0.6	-0.8	-0.9	-0.6	-1.0
lazy	0.9	-0.6	0.8	1.0	1.0	0.6	-0.4	-0.7	-0.9	-0.6	-0.9
calm	0.5	-0.5	0.2	0.7	0.6	1.0	-0.9	-0.9	-0.7	-0.8	-0.8
aggressive	-0.3	0.4	-0.2	-0.6	-0.4	-0.9	1.0	0.9	0.5	0.6	0.6
energetic	-0.6	0.6	-0.5	-0.8	-0.7	-0.9	0.9	1.0	0.8	0.6	0.8
busy	-0.8	0.6	-0.6	-0.9	-0.9	-0.7	0.5	0.8	1.0	0.7	0.9
excited	-0.6	0.5	-0.2	-0.6	-0.6	-0.8	0.6	0.6	0.7	1.0	0.7
fast	-0.8	0.7	-0.7	-1.0	-0.9	-0.8	0.6	0.8	0.9	0.7	1.0

Fig. 9 Correlations between style vectors without elimination of incidental features with values greater than 0.15 in green and less than -0.15 in red backgrounds respectively

	limping	healthy	depressed	slow	lazy	calm	aggressive	energetic	busy	excited	fast
limping	1.0	-0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1
healthy	-0.9	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
depressed	0.0	0.0	1.0	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.1
slow	0.0	0.0	0.1	1.0	0.7	0.0	0.1	-0.1	-0.1	-0.3	-0.6
lazy	0.0	0.0	0.2	0.7	1.0	0.4	-0.3	-0.6	-0.4	-0.6	-0.7
calm	0.0	0.0	0.0	0.0	0.4	1.0	-0.7	-0.7	-0.3	-0.3	-0.2
aggressive	0.0	0.0	0.0	0.1	-0.3	-0.7	1.0	0.7	0.4	-0.1	-0.1
energetic	0.0	0.0	0.0	-0.1	-0.6	-0.7	0.7	1.0	0.6	0.2	0.2
busy	0.0	0	0.0	-0.1	-0.4	-0.3	0.4	0.6	1.0	0.3	0.3
excited	-0.1	0.1	0.1	-0.3	-0.6	-0.3	-0.1	0.2	0.3	1.0	0.8
fast	0.0	0.0	0.0	-0.6	-0.7	-0.2	-0.1	0.2	0.3	0.8	1.0

Fig. 10 Correlations between style vectors after the elimination of incidental features

level of limping can be done without affecting the perceived slowness. The correlations that remain non-zero after the elimination are reasonable as those style pairs can be semantically considered close to synonyms or opposites (such as slow and lazy, or calm vs. energetic).

5.3 Validation of Synthesized Styles

The first validation experiment indicated that the style vectors correspond well to human perception when testing with acted motions. This implies that the vectors should allow accurate control of motion synthesis. To directly test it, we ran the first validation experiment again with motions produced by interpolation synthesis.

As starting motions for the test, we took equally spaced samples from the parameter space of the interpolation. Each sample was produced with 50% of one parameter and the remaining 50% equally divided for all others. This gave us 35 initial motions. From every initial motion we created eight modifications, each adjusted to display a fixed amount more of a particular style. This was done by offsetting the parameters by a vector calculated from Eq. 4. To reduce the human evaluation task, we used only the eight style vectors which did not show large positive correlations in Figure 10, *i.e.* limping, healthy, depressed, slow, calm, aggressive, busy and fast. Thus, we ended up with 280 pairs showing an initial motion and its adjusted version.

Perceptual evaluation of the 280 motion pairs was done with a similar questionnaire as in section 5.1. We received a total of 4485 individual ratings randomly distributed among 314 participants.

Answers of the questionnaire were scaled so that ± 2 means *much more/less style*, ± 1 *slightly more/less style*, and 0 stands for *no change in style*. From this data, the mean scores for every combination of intended and perceived styles were calculated, and statistically significant differences from zero with p-value 0.05 were identified. The means are based on 70 or 71 evaluations. Figure 11 shows a confusion matrix of the results.

		Perceived Styles							
		limping	healthy	depressed	slow	calm	aggressive	busy	fast
Intended Styles	limping	0.3	-0.6	-0.1	0.2	0.0	-0.2	-0.2	-0.3
	healthy	-0.2	0.1	-0.1	0.0	0.1	0.0	0.0	0.0
	depressed	0.6	-0.4	0.4	0.3	-0.1	-0.3	0.1	-0.3
	slow	0.2	-0.7	0.6	0.6	0.5	-0.5	-0.3	-0.6
	calm	-0.1	-0.2	0.4	0.6	0.5	-0.5	-0.5	-0.4
	aggressive	0.1	0.4	-0.3	-0.2	-0.7	0.8	0.3	0.2
	busy	0.1	0.1	-0.3	-0.6	-0.5	0.4	0.5	0.5
	fast	-0.2	0.6	-0.3	-0.7	-0.5	0.5	0.3	0.6

Fig. 11 Mean scores from evaluation of style adjustments. Scores on white do not statistically differ from zero ($p=0.05$), significant positive differences are green and significant negative differences red

If controlling the synthesis is successful, the intended style should get a significant increase due to the adjustment and even larger change than any other style. The diagonal of Figure 11 shows that all intended styles were actually perceived to increase. However, the change was not always the largest. This is understandable in cases where the initial motion already has plenty of the in-

tended style visible; then the increase cannot be very large as discussed in the next section.

6 Discussion

Our method for creating style vectors does require some talent and concentration from actors, people instructing the actors and the person annotating the motions. Therefore, the method does not replace the work of animation professionals. However, since the style vectors can be used for controlling styles of new motion sets, the fruits of the labor can be enjoyed by people who are not experts in motion capture techniques.

Our animated demo (Online Resource 1) and the related experiments show that style vectors enable control of several styles simultaneously. How intense the styles eventually get, is up to the acted motions and the synthesis method used. Interpolation limits expressivity to that of the input motions while extrapolation may produce more intense but sometimes unnatural style.

We model a relative style with one style vector, but acknowledge that a global vector is not sufficient in all cases. For example 'natural' is a property that has a maximum from which there are many, even opposite ways to get away, and its negation (unnatural) is ambiguous. Local style vectors that always point to the maximum (or away from it for respective negations) could work better than a global vector. We were able to define 'healthy' with a global style vector as our set of examples had limping as the only unhealthy movement. However, asking an actor to perform in an unhealthy way could provoke a demand for more specific instructions. This may be the reason why the style vector for healthy did not score so well in our experiment (Fig. 11). As most starting motions of the test already looked quite healthy, it could not be improved much.

We acknowledge that low correlations between style vectors (Fig. 10) create expectations of better separation between styles than the results of the crowdsourced experiment imply (Fig. 11). Varying proficiency of the English language among the globally distributed participants may explain part of the overlapping use of style words. To our knowledge, all previous publications presenting style oriented motion synthesis have completely omitted a similar validation. Therefore, our work can be considered state-of-the-art in this respect.

A risk in our method is that a style vector can degenerate to zero if no essential features are left after elimination. This can happen if the style is ill-defined, poorly acted, or annotated inconsistently due to human errors. We do not consider the last reason to be a serious one as style definitions can be created by relatively low amount

of annotations by just one person. Therefore, correcting annotation mistakes does not mean much work.

The style vectors could be produced by different means than our process. We considered using Support Vector Machine (SVM) to find a hyperplane separating two style classes and applying its normal as the style vector. SVMs work well in classification of absolute concepts represented with individual examples such as verbs [4, 25]. For relative concepts a better option is Ranking SVM [9], but we did not adopt that either as the method by Zhuang *et al.* [27] or our refinement of it are simpler to implement and computationally less intensive.

Our method could be developed further by experimenting with new actions, styles, actors, low-level features and synthesis methods. Preliminary experiments on reusing style definitions with other actions have been promising. For example, definitions for styles slow and aggressive based on locomotion seemed to apply to hand waving or turning. However, trying to make a hand wave more limping created random looking results.

A practical use case for our method is communication with virtual characters. Bodily motions could drastically improve expressivity compared to facial expressions or symbolic messages alone. Another use case is browsing in a motion library. Starting from one motion with the desired action, its variations in style could be found with relative steps instead of having to watch all possible alternatives.

7 Conclusions

In this paper, the semantic meaning of verbally described styles has been grounded in numerical motion data more precisely than before. Our main contribution is the method producing more accurate style vectors by eliminating other features than those essential for recognizing a style.

We have presented a method for indirectly controlling motion synthesis by style words. We let an arbitrary synthesizer generate candidate motions, evaluate them with style vectors, and select the best. For a stable synthesis method, such as interpolation, the desired changes in style can be mapped to offsets in synthesis parameters. Controlling a large number of parameters this way is more user friendly than adjusting them directly.

Our evaluation of the method shows that style definitions created from motions of one actor and annotated by one observer, accurately predict styles observed by other people in motions performed by other actors.

In a practical application, a virtual actor could be first commanded to perform an action such as 'walk'

or 'run' and then the performance could be fine-tuned by relative commands such as 'more limping' or 'more aggressively'.

Preliminary results suggest that the method generalizes many styles over motion categories, such as from locomotion to turning in place, but further research is needed to find the precise requirements for successful transfer of style.

Acknowledgements This work has been supported by the HeCSE graduate school and the project Multimodally grounded language technology (254104) funded by the Academy of Finland. The Mocap toolbox by Neil Lawrence [13] was used in this research.

References

1. Aviezer, H., Hassin, R.R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., Bentin, S.: Angry, disgusted, or afraid?: Studies on the malleability of emotion perception. *Psychological Science* **19**(7), 724–732 (2008)
2. Bruderlin, A., Williams, L.: Motion signal processing. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '95*, pp. 97–104. ACM, New York, NY, USA (1995)
3. Chi, D., Costa, M., Zhao, L., Badler, N.: The emote model for effort and shape. In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '00*, pp. 173–182. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (2000)
4. Cho, K., Chen, X.: Classifying and visualizing motion capture sequences using deep neural networks. In: *Proceedings of the 9th International Conference on Computer Vision Theory and Applications, VISAPP2014* (2014)
5. Clavel, C., Plessier, J., Martin, J.C., Ach, L., Morel, B.: Combining facial and postural expressions of emotions in a virtual character. In: Z. Ruttkay, M. Kipp, A. Nijholt, H. Vilhjálmsson (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 5773, pp. 287–300. Springer Berlin Heidelberg (2009)
6. Förger, K., Honkela, T., Takala, T.: Impact of varying vocabularies on controlling motion of a virtual actor. In: R. Aylett, B. Krenn, C. Pelachaud, H. Shimodaira (eds.) *Intelligent Virtual Agents, Lecture Notes in Computer Science*, vol. 8108, pp. 239–248. Springer Berlin Heidelberg (2013)
7. Gleicher, M.: Retargetting motion to new characters. In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '98*, pp. 33–42. ACM, New York, NY, USA (1998)
8. Hsu, E., Pulli, K., Popović, J.: Style translation for human motion. *ACM Trans. Graph.* **24**(3), 1082–1089 (2005)
9. Joachims, T.: Optimizing search engines using click-through data. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02*, pp. 133–142. ACM, New York, NY, USA (2002)
10. Johnson, K.L., McKay, L.S., Pollick, F.E.: He throws like a girl (but only when hes sad): Emotion affects sex-

- decoding of biological motion displays. *Cognition* **119**(2), 265–280 (2011)
11. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on* **4**(1), 15–33 (2013)
12. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. In: *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '02*, pp. 473–482. ACM, New York, NY, USA (2002)
13. Lawrence, N.: Mocap toolbox for matlab. Available on-line at <http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/mocap/> (2011)
14. Min, J., Chai, J.: Motion graphs++: A compact generative model for semantic motion analysis and synthesis. *ACM Trans. Graph.* **31**(6), 153:1–153:12 (2012)
15. Mukai, T., Kuriyama, S.: Geostatistical motion interpolation. In: *ACM SIGGRAPH 2005 Papers, SIGGRAPH '05*, pp. 1062–1070. ACM, New York, NY, USA (2005)
16. Poppe, R.: A survey on vision-based human action recognition. *Image and Vision Computing* **28**(6), 976–990 (2010)
17. Rose, C., Cohen, M., Bodenheimer, B.: Verbs and adverbs: Multidimensional motion interpolation. *Computer Graphics and Applications, IEEE* **18**(5), 32–40 (1998)
18. Shapiro, A., Cao, Y., Faloutsos, P.: Style components. In: *Proceedings of Graphics Interface 2006*, pp. 33–39. Canadian Information Processing Society (2006)
19. Shoemake, K.: Animating rotation with quaternion curves. *SIGGRAPH Comput. Graph.* **19**(3), 245–254 (1985)
20. Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision* **2**(5), 371–387 (2002)
21. Troje, N.F.: Retrieving information from human movement patterns. *Understanding events: How humans see, represent, and act on events* pp. 308–334 (2008)
22. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '95*, pp. 91–96. ACM, New York, NY, USA (1995)
23. Urtasun, R., Glardon, P., Boulic, R., Thalmann, D., Fua, P.: Style-based motion synthesis. *Computer Graphics Forum* **23**(4), 799–812 (2004)
24. Wang, X., Jia, J., Cai, L.: Affective image adjustment with a single word. *The Visual Computer* **29**(11), 1121–1133 (2013)
25. Wu, J., Hu, D., Chen, F.: Action recognition by hidden temporal models. *The Visual Computer* **30**(12), 1395–1404 (2014)
26. Yoo, I., Vanek, J., Nizovtseva, M., Adamo-Villani, N., Benes, B.: Sketching human character animations by composing sequences from large motion database. *The Visual Computer* **30**(2), 213–227 (2014)
27. Zhuang, Y., Pan, Y., Xiao, J.: *A Modern Approach to Intelligent Animation: Theory and Practice*, chap. Automatic Synthesis and Editing of Motion Styles, pp. 255–265. Springer (2008)



ISBN 978-952-60-6350-8 (printed)
ISBN 978-952-60-6351-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS