Perceptually Guided Processing of Style and Affect in Human Motion for Multimedia Applications

by

S. Ali Etemad

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Ph.D.)

in

Electrical and Computer Engineering

Carleton University
Ottawa, Ontario

Abstract

Computer graphics and animation, as a direct result of advancements in hardware and software, have become broad and demanding areas of research. Animation of human motion is a major component in the field that has attracted many due to its significance in movies, games, and virtual environments. We propose that processing features for style and affect, which are fundamental determinants of personality and naturally appearing motion, should be carried out through perceptually guided processing techniques. In this dissertation, we employ this approach and develop a set of tools for extraction, synthesis, and analysis of affective and stylistic motion features.

Temporal alignment is one of the most common issues in processing motion data. Accordingly, we first propose a new time warping technique for motion. The proposed method outperforms several existing techniques and has advantages such as precise alignment, low distortion, smooth warped motion trajectories, and high customizability.

Many motion processing techniques utilize incremental (joint-to-joint) processing of motion sequences. In addition, some systems process only selected joints or regions of the body. Hence, it is imperative to verify whether partial or subsets of computational solutions can lead to perceptually accurate results. Accordingly, we investigate and validate the notion of additivity in perception of affect from motion.

A system capable of extracting style/affect features from motion data using spline optimization is then introduced. Our method has several advantages over existing techniques, namely extracting the features as three separate movement, posture, and time components, which are the perceptual and functional sources for stylistic/affective motion. Our method also performs in Cartesian or joint-angle spaces rather than

Eigen/latent sub-spaces which is the common trend in existing techniques.

Towards synthesis of style/affect features, a perception-based expert-driven approach is used. Gaussian radial basis functions (RBFs) are first introduced as mathematical constructs for stylistic/affective features. A user interface is then developed using which animators can utilize these basis functions to synthesize the desired stylistic/affective features. Through analysis and in depth study of data collected from several animators, expert-driven perceptual shortcuts for generation of different stylistic/affective themes are derived. The features also shed light on various aspects of execution and perception of style/affect.

A unified system capable of both classification and translation of stylistic/affective features in motion is subsequently developed using ensembles of Gaussian RBF neural networks. The recognition module of the system outperforms several other classifiers and the style translation module produces results that are validated as perceptually accurate by viewers.

Finally, to provide a set of guidelines for animators, based on which stylistic/affective features can be modified or added to motion data, an empirical paradigm is proposed. The paradigm discusses the characteristics required to make motion scenes perceptually valid with respect to motion features and context. The existing body of literature, sensible examples, user-based case studies, and opinions of experienced animators support this paradigm. A provided study on relative significance of the different components of the paradigm enables accurate usage of the model.

To my wife Nasim,

my parents,

and my brothers

“Somewhere, something incredible is waiting to be known.”
– Carl Sagan

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisor Dr. Ali Arya without whom this study would have been impossible. His guidance, encouragements, and support made all the difference. I would also like to thank Dr. Avi Parush for his insightful discussions regarding perceptual processes through the course of this research.

I would like to thank the committee members of my thesis defense as well as the anonymous reviewers at journals and conferences who reviewed our work and provided insightful feedback and discussions. Their time and attention has definitely enhanced the quality of this presented work.

I wish to thank all of the funding agencies and corporations that financially supported my research. In particular, I would like to thank the Natural Sciences and Engineering Research Council (NSERC) of Canada and Ontario Centres of Excellence (OCE).

Last but not least, I wish to deeply thank my wife Nasim for all her unconditional love and support. Her endurance of my long hours at work, her support and encouragements for when I stumbled, and her love at every moment was what kept me going. I would also like to thank my parents and brothers for everything they have done for me throughout the years.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A	arms/hands
AL	arms/hands + legs/feet
ANN	artificial neural network
ANN	artificial neural network
ANOVA	analysis of variances
AU	action unit
BM	body motion
BVH	Biovision hierarchy
CCA	canonical correlation analysis
CMU	Carnegie Mellon University
CoTW	correlation-optimized time warping
CTW	canonical time warping
DDTW	derivative dynamic time warping
DOF	degree of freedom
DTW	dynamic time warping
FACS	facial action coding system
FM	face motion
GTW	generalized time warping
GUI	graphical user interface
H	head/neck
HA	head + arms/hands
HAL	head/neck + arms/hands + legs/feet

HL	head/neck + legs/feet
HMM	hidden Markov model
HT	head/neck + torso
HTA	head/neck + torso + arms/hands
HTL	head/neck + torso + legs/feet
ICA	independent component analysis
IK	inverse kinematics
IMW	iterative motion warping
KNN	K-nearest neighbor
l.s.	left side
L	legs/feet
LCTW	local canonical time warping
LDA	linear discriminant analysis
LMA	Laban movement analysis
LTI	linear time-invariant
MLP	multilayer perceptron
n.s.	not significant
PC	principal component
PCA	principal component analysis
PCC	Pearson's correlation coefficient
PF	primary feature
PID	proportional-integral-derivative
PL	point-light

PT	primary themes
PV	perceptual validity
r.s.	right side
RBF	radial basis function
RBFNN	radial basis function neural network
SD	standard error
SE	standard error
SF	secondary feature
SIFT	scale-invariant feature transform
SNR	signal to noise ratio
ST	secondary theme
SVM	support vector machine
T	torso
TA	torso + arms/hands
TAL	torso + arms/hands + legs/feet
TL	torso + legs/feet
UTW	uniform time warping

Chapter 1.

Introduction

1.1 Background and Motivation

Human motion studies are currently drawing much deserved attention. This is due to advancements in computing capabilities, which have lead to a significant growth in applications such as animated movies [1], interactive games [2], interactive gesture-based everyday applications [3], and virtual worlds [4, 5] that use character motion and animation as a key component.

Historically, character animation is created by the traditional technique of key-framing [6], or more recently motion capture [7] for more complicated movements. Key-framing refers to the process of defining the key moments of a movement by animator while the

computer tool performs interpolation to create in-between frames, while motion capture is the process of tracking and recording real motion performed by an actor using different sensors or cameras which will then be applied to computer models. For surveys on human motion analysis, we refer the interested reader to [8, 9, 10, 11]. Nevertheless, regardless of the technique used to create animation, the majority of animated behaviors in interactive or non-interactive multimedia applications are simply playback of pre-animated or pre-recorded sequences. This inability of animation systems in procedurally generating realistic behaviors limits the applications in terms of response to user interactions in real-time by creating appropriate content. Even in non-interactive applications like animated films, there is little support from intelligent animation software, despite the fact that automating the process of character animation would reduce the production cost and effort, and provide animators with a more effective and systematic way to generate content. Such procedural generation of character animation requires variety of computational models, algorithms and other technical considerations.

On the other hand, multimedia content (animated videos, games, virtual worlds, etc.) are mostly, if not all, directed towards a human audience. As a result, it is imperative to carefully consider the human mind and psychology, which is in the receiving and interpreting end of generated or processed material [12]. Animated human motion is no exception in this regard as its validity and quality needs to be tested against its reception by the audience; and an effective way to significantly enhance or ensure its perceptual quality, is to apply perception-based techniques in the systems that process them. In fact, it has been previously demonstrated that in different digital environments, the type of character, clothing, and complexity of the scene play a secondary role with respect to the

perceptual quality of motion [13]. The main goal of this research is to process human motion such that the outcome causes the intended perceptions. Whether motion features are intended for a particular action (for example running) or attribute (for example fatigue), they need to be consistent with or derived from the audience's perception. As a result, we study and utilize this notion, referring to it as *perceptually guided* or *perceptually valid*. This argument, which forms the motivating grounds for this research, is presented in Figure 1.1. The research proposed here aims at procedural analysis and generation of character animation, taking into account both technical and perceptual requirements.

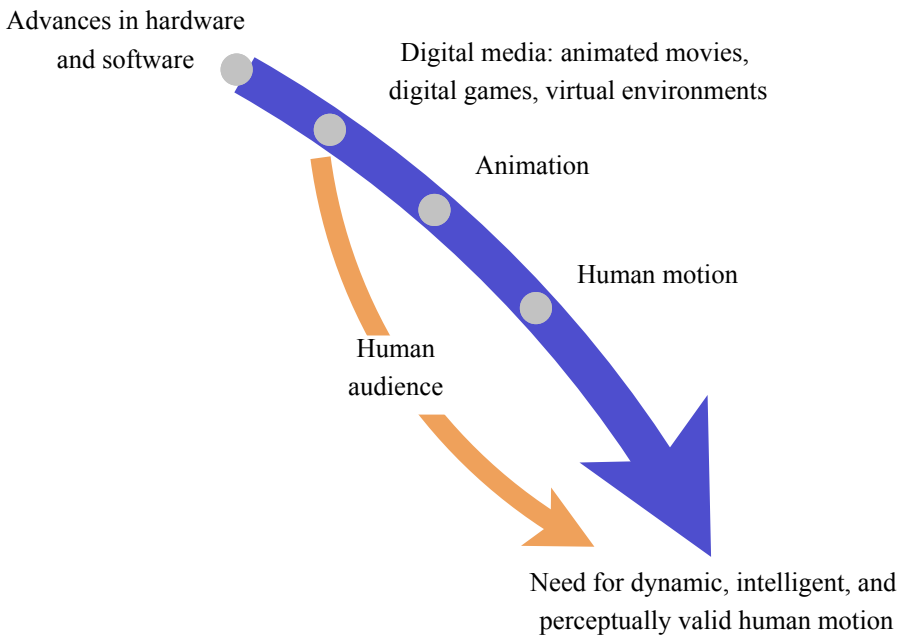


Figure 1.1. The motivating grounds for this dissertation.

Human motion can be considered as a combination of two sets of themes, represented by corresponding features [14]. While the primary themes (PTs) specify actions like walking or running, the secondary themes (STs) relate to affect, style, or individual characteristics

in which those actions are performed. These stylistic variations, or secondary features (SFs), are constantly present in motion data and have been noticed by researchers as early as Darwin [15]. For a review on perception of biological motion and associated themes, we refer the reader to [16, 17].

In multimedia content, different characters, based on their roles and attributes display different types of STs in motion. For example, in an animated film or a virtual world, a female character should walk differently compared to a male character. As another example, in a computer game, based on the energy level and health of a character, different styles of actions need to be displayed (tired, energetic, healthy, injured, etc). In all these cases, it is important that the viewer perceive the action the way it was intended both in terms of PTs (what's being done) and the secondary ones (who is doing it and how). This means that the motion not only needs to be physically or functionally correct, but also perceptually valid.

The complex nature of human motion, as well as the wide range of STs [18], contributes to the difficulty of the analysis, modeling, and synthesis of perceptually valid affective/stylistic human motion. Furthermore, the features that convey these themes are extremely difficult for machine learning approaches to extract and analyze due to their personalized nature, small spatiotemporal significance, and often, lack of sufficient and consistent training data.

Generally, motion studies focusing on STs can take three main routes. Analysis of the features in terms of both execution as well as perception is the first route. For these studies, there have been significant amounts of research carried out by the physiology and

psychology [16] communities as well as the multimedia community [19]. These studies can be used indirectly towards computational methods for extraction and modeling means as well as interpreting the outcomes of the other two components of motion studies. The second route is synthesis or modeling of features. This component of motion studies often draws interest from the multimedia community and animators in particular [20]. Synthesis, which is the second category, can be used to generate motion sequences, or alter existing ones, to achieve desired functional and perceptual properties. Finally, extraction of SFs is the third possible route [21]. Extraction of features is often of little practical significance alone. In fact, in most cases, it is a necessary step towards analysis, classification, and modeling of motion and its related features. In conjunction with analysis, extracted features can be recognized and interpreted for classification purposes, or studied for psychology purposes. Towards synthesis, extracted features can be used for transfer onto existing motion. Figure 1.2 illustrates the three routes and how they are linked in common studies of STs in motion.

Based on the presented arguments, we can conclude that developing perception-based platforms and tools for the study of STs and associated features in human motion is a critical and vital step towards achieving perceptually valid content. The general goal of this dissertation is to tackle the two major computational components of Figure 1.2, namely *extraction* and *synthesis* of SFs in human motion in order to create perceptually valid animation. However, while working along this path, *analysis* studies are sometimes required to design accurate systems as well as to validate and interpret the results.

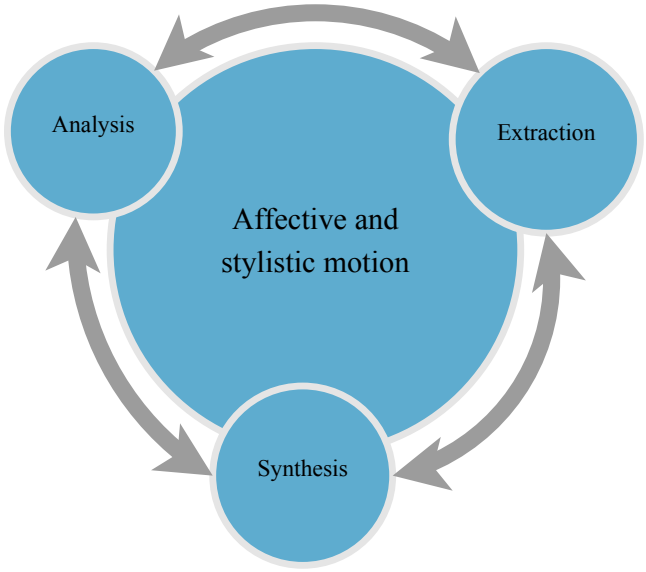


Figure 1.2. Extraction, analysis, and synthesis of SFs are the three key components of motion studies focusing on STs.

1.2 Problem Definition and Challenges

A vast corpus of computational methods has been proposed for processing human motion [8, 10, 11, 22]. While most proposed techniques are effective and result in computationally sound outcomes, as discussed earlier, methodologies that incorporate human perception as well as expert-knowledge are often neglected. Accordingly, incorporating expert-knowledge and perceptually guided procedures in existing techniques for extraction, analysis, and synthesis of stylistic motion defines the main problem in this research. Developing the tools necessary for processing stylistic motion, while maintaining perceptual quality and validity, is the challenging task that we have tackled in this dissertation.

Specifically, following are some of the main challenges associated with the problem at

hand, which we address in this dissertation:

i. Most existing techniques that deal with human motion, especially those that use relative editing or modeling require temporal alignment. Time warping is often used for this purpose [23]. Various time warping techniques have so far been developed and explored. Ideally, however, time warping should benefit from low distortion, smoothly warped sequences, spatial and temporal customizability, and accurate alignment, which most exiting methods do not provide.

ii. Generally, any attempt at extracting or synthesizing SFs will most likely take place on a spatially incremental basis, in other words one body joint at a time [24]. This type of processing is naturally viable from a computational perspective. However, perceptual implications are unknown should imperfect solutions or subsets of the complete solution sets be utilized. Such partial solutions may result in unexpected perceptual outcomes. As a result an in depth investigation of perception of affect from single-joint-affective motion vs. multi-joint-affective motion is required. In other words, it is not known whether the sum of impacts of different joints (or limbs) on perception of particular themes is proportional to the impact of the sum of the joints (or limbs).

iii. Extraction of SFs in spatiotemporal domain is an extremely difficult task as motion sequences can be carried out in a variety of different styles and be very personalized. Moreover, existence of three different classes of features, namely posture, movement, and time [19], further complicate the task.

iv. Most existing methods either focus on analyzing the psycho-physiology of motion [16] or generation of features purely from a computational perspective [25]. As a

result, expert-driven and perception-based knowledge that can take a step towards perceptually accurate results, is often the missing link in the proposed methodologies. Nevertheless, such approaches are recently becoming more popular [13].

v. Taking an expert-driven approach towards the problem of producing synthetic features for style and affect, faces the significant challenge of developing a platform and basis using which accurate and analyzable data can be collected. Engaging adequate number of experts in the field for such an approach is another challenge.

vi. Our brains and nervous systems perform both control/execution and perception of affective/stylistic motion as a single unit. Inspired by this concept, developing a single system capable of performing both classification and synthesis of SFs can be a challenging task. In fact, classifiers are not ideal tools for synthesis of features as they will often require significant reconfiguration for this purpose.

vii. A variety of different factors ranging from context and story to consistency of features need to be considered when animators generate or edit motion sequences. These factors can all have significant influences on perceptual quality and validity of depicted scenes. Yet, a unified paradigm, based on which animators can systematically determine the critical factors that need consideration in motion, is yet to be put forth.

1.3 Contributions

In order to achieve motion sequences with STs that retain high perceptual quality and

validity, the three components of motion studies, namely extraction, analysis, and synthesis need to be carried out through perceptually guided computational frameworks. Such approaches are often overlooked as they are difficult to materialize and be put into practical contexts, and yet can lead to effective and efficient solutions. Through the following, we mention the different contributions and solutions to problems mentioned in the previous section, taking our general approach into account.

- As described earlier, most motion processing systems utilize some sort of time warping to temporally align motion data that is being used. In this research, a time warping technique is proposed that outperforms existing popular methods from several perspectives. The proposed method benefits from high perceptual quality due to minimized distortion, better alignment, readily smoothed motion curves, and spatial and temporal customizability, which users can utilize to align motion sequences based on content and requirements of the application. This contribution corresponds to problem *i* from the previous section.

- Towards studying the problem *ii* depicted in the previous section, a novel experiment is designed which studies the spatially incremental processing of motion and explores the notion of additivity in the influence of individual limbs in perception of affect. The results of this experiment show that the amount of affect perceived from multiple limbs is highly correlated with the sum of perceived affects from individual limbs. This notion entails interesting conclusions concerning perception of affect and is also utilized in subsequent synthesis procedures as they occur on a spatially incremental (joint-to-joint) basis.

- In order to extract SFs, a model is first formalized for defining the relationship between actions and SFs. Using this model, and inspired by the way in which affective and stylistic motion is physiologically performed and perceived, a novel optimization-based method is proposed and implemented that separately extracts movement, posture, and temporal features from human motion. The extracted features can subsequently be added to other sequences, resulting in SF translation. This contribution solves problem *iii* mentioned in the previous section.

- To address problem *iv* and *v*, and generate perceptually guided and valid SFs, a user interface is developed, with which experienced animators tune Gaussian radial basis functions (RBFs) to generate different SFs in human motion. These generated features are recorded and summarized to create an expert-driven set of perceptual shortcuts that can be used to synthesize a dynamic range of affective/stylistic behaviors with scalable intensities. Moreover, through this process, it is demonstrated that specific mathematical functions, namely Gaussian RBFs, are extremely effective and efficient in modeling SFs in human motion.

- Motivated by problem *vi* presented in the previous section, a system is developed in which ensembles of Gaussian RBF neural networks carry out both classification and synthesis (translation) of SFs. To the best of our knowledge, an alternative unified system capable of performing both tasks does not exist.

- Finally, to address problem *vii*, an empirical paradigm called Perceptual Validity is proposed which provides a set of principles for animators to take into account when synthesizing, editing, or recording motion sequences. The paradigm takes into

account different components of motion and the scene and provides guidelines on the purposes for which motion features need to be tuned. The proposed paradigm is supported by arguments from related literature, case studies, and opinions of experienced animators.

1.4 Dissertation Outline

Chapter 1 presents the general background regarding this dissertation along with the motivating grounds for the conducted research. The problems addressed in this study and related challenges are described in detail. Our contributions to the field along the path are mentioned. Finally, an overview of the organization of this dissertation is put forth.

Chapter 2 surveys the related work on human motion processing. Different aspects of this research area are taken into account. Studies on perception of human motion are first reviewed, followed by motion computing techniques such as alignment, interpretation, and synthesis.

Chapter 3 presents the general proposed approach. Discussions on the problems addressed in this dissertation is provided, following the overall methodological approach. An overview on research methods and tools utilized to design and implement each component is also presented.

Chapter 4 deals with the notion of temporal alignment in human motion. As most motion processing methods require temporal alignment of actions, a novel time warping technique is developed. The proposed method outperforms most existing methods, both

perceptually and computationally. Low distortion, high temporal and spatial customizability, and better alignment are some of the benefits of our proposed technique. This warping method is later used in several other chapters.

Chapter 5 studies and validates the notion of spatially incremental processing of motion. In other words, it is illustrated that the perception of affect from multi-joint-affective motion is highly correlated with that of the sum of single-joint-affective motions. This study is necessary, given that the proposed methods in this dissertation deal with individual body joints for extraction, translation, and generation of SFs.

Chapter 6 presents a system capable of extracting spatiotemporal SFs from motion data using spline optimization. A model is first introduced that describes the relationship between SFs and performed actions. The formulated model forms the basis for this chapter and Chapter 8. A method is then developed, using which the three major components of SFs, namely movement, posture, and temporal features are extracted separately. The system benefits from high generalization. The extracted features using this highly practical system can be analyzed for psychology studies or be used in SF translation.

Chapter 7 uses expert-driven perceptually guided models as synthetic SFs. Initially, Gaussian RBFs are proposed as pre-defined mathematical constructs for modeling the features. Through rigorous perception and computational experiments, the precise modeling performance of these functions is illustrated. A graphical user interface is developed using which experienced animators can utilize the basis functions introduced earlier to synthesize the desired SFs. Using analysis and in depth study of data collected

from several animators, perceptual shortcuts for generation of different STs in human motion are derived.

Chapter 8 utilizes the concept of the Gaussian RBFs described earlier for recognition and translation of SFs. A neural network setting is employed for this purpose. The developed neural network scheme is used as a unified system capable of both classification and translation of features. High classification rates and perceptual quality of the style translation outputs confirm the effectiveness of the proposed system.

Chapter 9 proposes an empirical conceptual paradigm for perceptually valid character motion animation. Some existing literature, sensible examples, user-based case studies, and opinions of experienced animators support the model. This proposed paradigm provides animators with a set of guidelines and components that need to be taken into account for synthesis and presentation of perceptually valid motion sequences.

Chapter 10 provides the summary of contributions and concluding remarks regarding the different sections of this dissertation. Possible directions towards which the future research may be planned are also suggested.

Appendix A describes a typical motion capture system and the process of recording motion capture data. The data structure is then defined followed by a description of the different datasets used in this dissertation.

Appendix B presents the questionnaire format including the participant consent form and sample questions used for user evaluations in this study.

Appendix C states the specifications of the hardware and main software with which

different systems of this dissertation are implemented. A brief overview of the different implemented routines that can be valuable resources for other researchers in the field is provided. Finally, a discussion on the main issues that were encountered during implementation and ways of overcoming the problems is provided.

Figure 1.3 presents the layout of this dissertation. Blue arrows represent logical procedural flow of chapters while gray arrows denote the direct utilization of tools and techniques developed in some chapters in others. For example, the time warping method developed in Chapter 4 is utilized in Chapters 5, 6, 7, and 8.

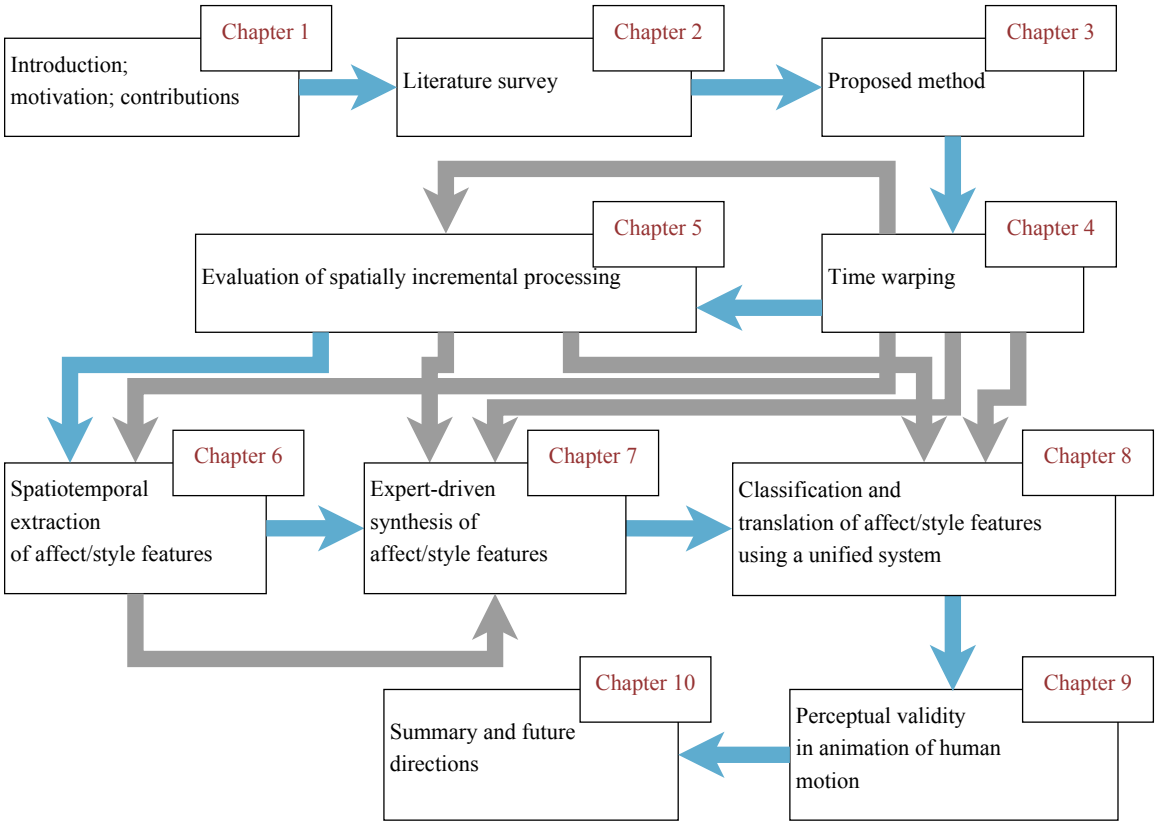


Figure 1.3. Thesis outline and organization of chapters. Blue arrows present procedural flow while gray arrows denote use of tools developed in each chapter.

1.5 Publications based on This Research

The presented materials in this dissertation have been published or submitted for possible publication to the following. Copyright permissions are acquired from the publishers where direct text, figures, or tables are used in this document.

Journals:

- S. A. Etemad and A. Arya, "Correlation optimized time warping for motion," *submitted*. [Chapter 4]

- S. A. Etemad, A. Arya, and A. Parush, "Additivity in perception of affect from limb motion," *Neuroscience Letters*, vol. 558, pp. 132-136, 2014. [Chapter 5]

- S. A. Etemad and A. Arya, "Extracting movement, posture, and temporal style features from human motion," *Biologically Inspired Cognitive Architectures*, vol. 7, pp. 15-25, 2014. [Chapter 6]

- S. A. Etemad and A. Arya, "Expert-driven features for style and affect in motion: Synthesis, analysis, and perception," *submitted*. [Chapter 7]

- S. A. Etemad and A. Arya, "Classification and translation of style and affect in human motion using RBF neural networks," *Neurocomputing*, vol. 129, pp. 585-595, 2014. [Chapter 8]

- S. A. Etemad, A. Arya, A. Parush, and S. DiPaola, "Perceptual validity in animation of human motion," *submitted*. [Chapter 9]

Conferences:

- S. A. Etemad and A. Arya, "A customizable time warping method for motion alignment," *Proceedings of the 7th IEEE International Conference on Semantic Computing*, pp. 387-388, 2013. [Chapter 4]

- S. A. Etemad, A. Arya, and A. Parush, "Spatial perceptual weights of energy-related features in animation of human motion", *Proceedings of Computer Graphics International*, S15, 2011. [Chapter 5]

- S. A. Etemad and A. Arya, "Separation and extraction of energy variants from human motion using temporal minimization," *Proceedings of the IEEE Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, pp. 73-77, 2011. [Chapter 6]

- S. A. Etemad and A. Arya, "Modeling and transformation of 3D human motion," *Proceedings of the 5th International Conference on Computer Graphics Theory and Applications*, pp. 307-315, 2010. [Chapter 6]

- S. A. Etemad and A. Arya, "Mining expert-driven models for affective motion", *ACM CHI Conference on Human Factors in Computing Systems (Workshop on Gesture-based Interaction Design: Communication and Cognition)*, 2014. [Chapter 7]

- S. A. Etemad and A. Arya, "Motion style translation with radial basis function networks," *Proceedings of the International Conference on Multimedia and Human Computer Interaction*, No. 36, 2013. [Chapter 8]

- S. A. Etemad and A. Arya, "Perceptually valid motion for avatars," *Proceedings of the International Conference on Multimedia and Human Computer Interaction*, No. 72, 2013. [Chapter 9]

Technical Reports:

- S. A. Etemad and A. Arya, and A. Parush, "Spatial perceptual weights of secondary variants in human motion for multimedia applications," *Technical Report, Carleton University, SCE-13-01*, 2013. [Chapter 5]

- S. A. Etemad and A. Arya, "Extraction of secondary features from gait trajectories using spatiotemporal spline optimization," *Technical Report, Carleton University, SCE-11-05*, 2011. [Chapter 6]

Chapter 2.

Related Work

2.1 Background

Chuck Jones, the famous animator at Warner Brothers (http://www.warnerbros.com/) animator, has said: “Believability. That is what we were striving for” [26]. Ollie Johnston and Frank Thomas, from Disney Studio’s so-called “Nine Old Men”, in their book *The Illusion of Life* state that: “Disney animation makes audience really believe in characters. There is a special ingredient in our type of animation that produces drawing that appear to think and make decisions and act on their own volition; it is what creates the illusion of life” [27]. Believable characters demonstrate believable behavior, which stems not only from physical realism, but also from the audiences’ perception of displayed content, and

form a major component of quality in animated content.

Other traits have been associated with quality and appeal in motion pictures, especially that of computer generated nature. From one standpoint, the notion of high quality content can be associated with the concept of aesthetics. The concept of aesthetics or beauty itself, however, is philosophically subjective and somewhat vague. As David Hume puts it in 1742, "beauty in things exists merely in the mind which contemplates them" [28]. Aesthetics experts, artists, and psychologists have offered many theories of what makes people consider an object beautiful, from evolutionary explanations to spiritual bases [29]. Some have associated this notion with the existence of expressive details and edits [30, 31, 32]. Furthermore, it has been illustrated that different factors such as generation of special effect scenes [33] and cinematic narrative discourse [34] play critical roles in synthesizing appealing multimedia content. Others have related beauty in art to the concept of creativity, and as a result, dynamic content generation [35, 36]. From another point of view, satisfaction and appeal may arise from being in one's comfort zone. Thus, some researchers have associated satisfaction (particularly in computer games) with exposure to specifically preferred content [37, 38]. Credibility and trust towards characters is another critical concept which has been directly or indirectly linked to believability and realism [39]. Naturalness of motion is another perspective from which appeal can be discussed. "Natural" phenomena are constantly occurring around us, and as a result, we become accustomed to them. Therefore, our perception of events and the concept naturalness are highly correlated [40, 41, 42].

In animation (or any other phenomenon for that matter), though it is incredibly difficult if not impossible, to quantify or conceptualize many of the concepts above, we believe the

human perception of the different aspects of displayed content can provide a decisive framework for assessment. Whether we ground these frameworks on aesthetics/beauty, naturalness, or believability, it is ultimately our perception of details that determines the contents' validity.

Generally, format-related quality and content-related quality are the two main components that define the quality of generated or recorded multimedia content. Format-related quality refers to elements such as resolution, frame rate, distortion, and others, for which there have been many computational frameworks proposed [43]. While these factors can in fact have indirect impacts on our perception of content, their main purpose is often the preservation of intended features and they are mostly not meant to determine or alter our understanding of the visualized material. Content quality, on the other hand, directly targets our understanding of, and appeal towards, the presented material. Different elements of the scene such as camera direction, lights, characters, and story can influence reception of the content. In this dissertation, we focus on content quality rather than format-related quality factors.

Human characters are one of the major elements in many animated and recorded scenes. Various parameters affect our perception of characters [44], among which, motion quality is vital and worth consideration. Animated movie characters, autonomous agents in real-time applications, and user avatars in virtual worlds are different versions of simulated humanoids that use human motion [45]. Animated characters can be used in non-interactive content such as movies as well as interactive applications such as games and virtual worlds where they can be employed as intelligent agents or user avatars [46]. Whether the interaction between audience/user and character is one-directional or bi-

directional, the generated scene must be believable to the audience, hence the need for perceptually valid content.

When dealing with biological motion, two categories are often addressed: face motion and body motion. This classification is carried out because a significant amount of information can be retrieved from facial expressions [47] despite their small spatial significance compared to the rest of the body. In this dissertation, we focus exclusively on body motion as it conveys a significant quantity of information, both in real life scenarios, as well as animated and digital characters. Figure 2.1 presents the overall schematic of parameters that influence multimedia content.

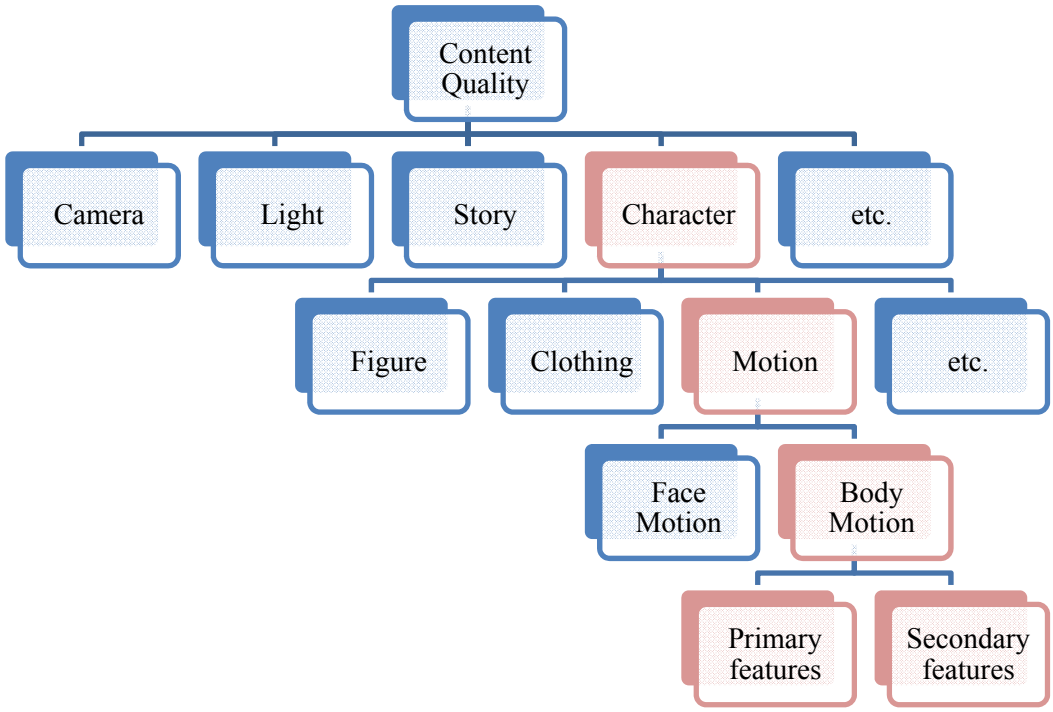


Figure 2.1. Different components of content quality.

A significant portion of communication between an agent and the audience is carried out

through non-verbal cues [48], e.g. tone, body language, etc. Such cues are often expressive [49] and convey personality, mood, emotion, energy, and others; and their critical role in "believable agents" has been well-emphasized [50]. Subsequently, affective computing has become one of the major areas of research for virtual agents [51, 52]. Different representations have been proposed for human expressive features, for example, Russell's model that describes basic emotions [53]. This model is used to describe some of the observations and findings in future sections of this dissertation. Figure 2.2 presents this model in the form of a circumplex.

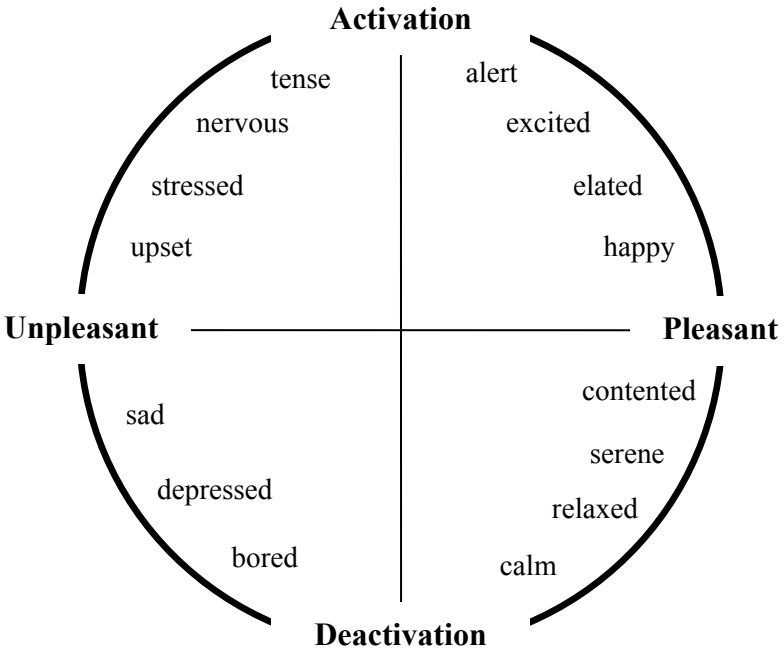


Figure 2.2. Model of emotions [53].

The following sections presents an inclusive review of works that have studies perception of secondary themes (STs) in biological motion, followed by computational techniques used to recognize/classify, control, and synthesize stylistic behavior in motion.

2.2 Human Motion: Perception

Humans perceive biological motion from an early age. It has been shown that infants, as young as 2-day old, look at biological human motion more than non-biological motion videos [54, 55, 56]. In fact, it is argued that humans have evolved internal neural/mental models that act as "life detectors" for recognizing biological motion [57].

While some studies were carried out on human motion between the years 1953 and 1970 [58, 59, 60, 61, 62], it was Johansson who in 1973 initiated the modern study of human motion perception [63]. By attaching small light sources to different body joints, and showing that naïve participants could rapidly recognize the motion of such orientation of lights (filmed in a dark room) as a human structure, the method of point light (PL) motion representation was introduced. Figure 2.3 presents a sample PL display of a human pose. The potential for the PL model in psychological and behavioral studies was quickly realized and its introduction helped with a variety of more detailed studies on biological motion. Since then, a significant amount of research has been carried out on perception of human motion (for a comprehensive review on the subject, the reader is referred to [16, 17]).

The course of human motion studies has since expanded through a wide range of hypotheses and findings. Regarding motion stimuli, the minimum time and discreetness of displayed joints required for gender classification from PL was investigated [64]. Different arrangements and setups of Johansson's PL were studies [65]. It has been illustrated that for detection of motion, body joints are most suitable for placing the point-lights, meanwhile the audience can distinguish motion even when the lights are on limbs

and in-between joints [66]. Perception of motion from different geometrical models such as a stick figure and polygonal figure was investigated [44]. In [67] the authors illustrated that geometric models affect perception of gender in neutral synthetic motion sequences, meaning including indicators of gender in the body figure would have an effect on gender perception. An interesting finding was that exaggerated female body features have a larger impact compared to the male features. Perception of emotions when specific body parts were hidden for the audience was later investigated and error-rates for different parts of the body and emotion classes were reported [19]. The authors showed that the upper body is most important when perceiving emotions.

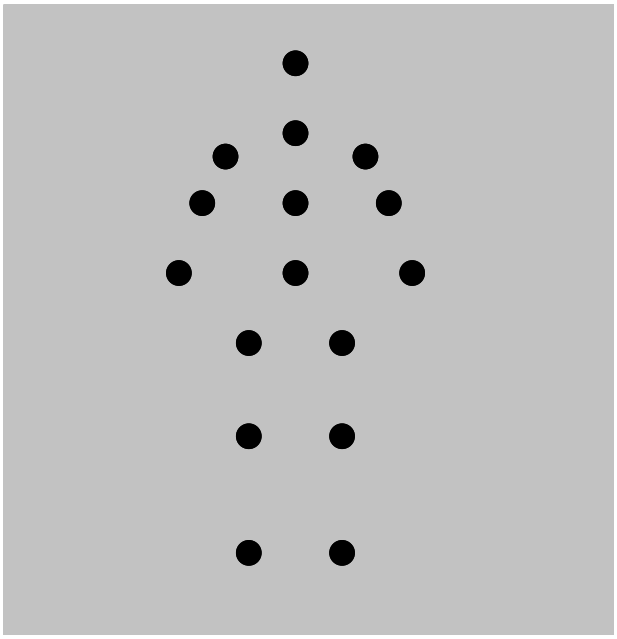


Figure 2.3. A sample PL display of a human pose originally proposed by Johansson [63].

A few years after the introduction of Johansson's method for motion representation [63], it was illustrated that subjects can successfully recognize their own walk as well as their friend's walks from PL [68]. Later studies confirmed the ability to recognize identity

from motion, showing a favor towards frontal views of the stimuli [69]. It was shown that self identification is almost view independent, while for identification of others, frontal and half profile views provide better stimuli compared to profile view [70].

Recognition of gender is among the very early studies and has since been widely explored. It was observed that the gender of a PL walker could successfully be classified by non-experts [71]. Barclay et al. studied temporal and spatial factors in perception of gender [64]. Saunders et al. studied the regions of the body that attract attention in gender recognition [72]. Troje studied perception of gender and revealed that dynamics plays a more critical role compared to posture in terms of gender-related cues [73]. Mather and Murdoch reported that while male walkers move their shoulders more than their hips and female walkers move their hips more than their shoulders, this extra "movement" is excess velocity and not displacement [74].

Affects are also known to be highly perceivable from motion and successful perception of emotions from motion has been widely illustrated [75, 76]. When perceiving emotions from arm movement, it was demonstrated that velocity increases gradually from weak to tired, sad, afraid, neutral, relaxed, happy, strong, angry, and excited [77]. The order does slightly vary for variations of the experiment, but the general trend is quite important and notable. Perception of emotions with static vs. dynamic stimuli was investigated and as expected, it was observed that dynamic stimuli are easier to perceive [78]. In [19] Normoyle et al. studied the effects of motion editing on perceived intensities of affects especially with respect to changes in dynamics and posture. Moreover, they concluded that the upper body contains more indicators for perception of affect and that while changes in posture can change the perceived affect, changes in dynamics impacts the

perceived intensities. Roether et al. investigated the perception of affect with respect to a variety of different kinematic and postural features [79]. The study in [80] reaffirmed the role of particular posture changes in perception of affect. Wallbott showed that movement and posture features in motion can sometimes be affect-specific [81]. Patterson et al. investigated the relationship among perceived affect and temporal changes in motion and confirmed the importance of temporal features [82]. Montepare et al. illustrated that in addition to speed, other movement features contribute to perception of emotions [83]. Barliya et al. investigated kinematics and perception of affective motion focusing on speed and dynamic features [84]. Crane and Gross studied and validated affect recognition from full-body motion reiterating the role of velocity [85]. Pollick et al. studied perception of both gender and affect from motion, showing that affect is easier to recognize [86].

From the abovementioned studies, and others, it is concluded that stylistic features in motion originate from, and can be classified into, three separate spatiotemporal sources: movement, posture, and temporal (speed) variations. Moreover, we can conclude that different body regions convey different amounts features towards perception of attributes from motion.

Table 1 summarizes important literature on perception and execution of motion, mostly in regards to STs. We have categorized the studies into the following general areas: internal models, influence of joint orientation/structure/body figure, perception of gender, perception of affect, and perception of identity.

Table 2.1. Summary of studies on perception and execution of motion. A: Internal models, B: Influence of joint orientation, structure, and body figure, C: Perception of gender, D: Perception of affect, E: Perception of identity.

Authors, year, reference	A	B	C	D	E
Johansson, 1973 [63]		✓			
Cutting & Kozlowski, 1977 [68]					✓
Kozlowski & Cutting, 1977 [71]			✓		
Barclay et al., 1978 [64]			✓		
Fox & McDaniel, 1982 [55]	✓				
Montepare et al., 1987 [83]				✓	
Bertenthal, 1993 [56]	✓				
Dittrich, 1993 [65]		✓			
Bertenthal & Pinto, 1994 [66]		✓			
Mather & Murdoch, 1994 [74]			✓		
Dittrich et al., 1996 [75]				✓	
Hodgins et al., 1998 [44]		✓			
Wallbott, 1998 [81]				✓	
Pollick et al., 2001 [77]				✓	
Patterson et al., 2001 [82]				✓	
Pollick et al., 2002 [86]			✓	✓	
Troje, 2002 [73]			✓		
Atkinson et al., 2004 [78]		✓		✓	
Troje et al., 2005 [69]		✓			✓
Clarke et al., 2005 [76]				✓	
Jokisch et al., 2006 [70]		✓			✓
Troje & Westhoff, 2006 [57]	✓				
Crane & Gross, 2007 [85]				✓	
Simion et al., 2008 [54]	✓				
McDonnell et al., 2009 [67]		✓	✓		
Roether et al., 2009 [79]		✓		✓	
Saunders et al., 2010 [72]		✓	✓		
Thrasher et al., 2011 [80]				✓	
Normoyle et al., 2013 [19]				✓	
Barliya et al., 2013 [84]				✓	

2.3 Human Motion: Computation

Computational processes on human motion fall into two main categories: (a) interpretation [87], and (b) synthesis [88, 89]. The former utilizes pre-existing recordings

such as videos or motion capture data to extract information regarding the classes of actions and actor attributes (e.g. gender, age, intention, etc.), retrieve and segment videos, and more, while the latter aims at generating or controlling motion data with desired characteristics. Figure 2.4 illustrates the overview of two categories of motion studies, where the flow of motion-to-information and information-to-motion is demonstrated for interpretation and synthesis respectively. Different techniques have been proposed and utilized for both categories of motion studies. In this chapter, we present a review of some of the important literature in both fields.

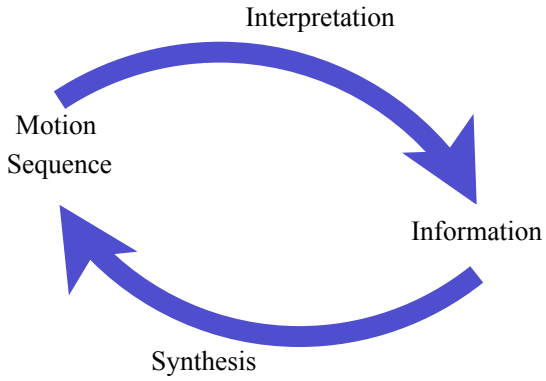


Figure 2.4. Two modalities for motion studies.

Generally, humans perform actions differently with respect to one another. It has been previously shown that factors such as gender, age, energy, health, ethnicity, and affect, influence the way we carry out actions [14]. As a result, even when motion data contain similar content, they vary in style, and these variations are manifested as spatiotemporal misalignments. On the other hand, in many cases, processing human motion requires motion data to be perfectly aligned. For example, it has been demonstrated that altering [90], blending [23], extraction [21], and translation [24] of motion require temporally

aligned data. Consequently, aligning motion is a critical step for both interpretation and synthesis procedures. Therefore, we first review the notion of motion alignment.

2.3.1 Motion Alignment

Uniform scaling [91], or uniform time warping (UTW), is naturally the most simple and possibly the most naïve method for aligning sequences. This technique is useful for length-matching of two or more time-series, and is not capable of aligning particular features in the process. However, this operation can be used as an important step in more advanced alignment techniques.

Dynamic time warping (DTW) [92] is one of the first non-linear warping techniques that employs the notion of similarity between two time series as a determinant for alignment. This is done through calculating the corresponding instances in the series and minimizing a distance objective function. Multiple variations and extensions of this method [93, 94, 95, 96] such as derivative dynamic time warping (DDTW) [93] have also been proposed. Generalized time warping (GTW) [97] has been proposed by Zhou and De la Torre as an effective extension to DTW. The extensions involve the capability of working with multiple modalities, for example motion capture data as well as recorded videos, more warping flexibility, and reduction in computational complexity.

The animation community has shown great interest in using or developing and enhancing time warping for aligning motion capture data. For example, Bruderlin and Williams used DTW for interpolating between sequences [98]. The motion warping method by Witkin and Popovic [99] is a variation of Bruderlin and Williams', and is intended to add small yet smooth changes to motion. The method proposed by Rose et al. manually selects key-

frames that need to be aligned, according to which in-between frames are aligned [23]. Gleicher used spatial constraints and inverse kinematics to preserve quality while retargeting motion [100]. The methods used by Kovar and Gleicher [101, 102] find correspondences between frames using distance minimization. Müller et al. used DTW towards their retrieval method [103]. This was done successive to segment-wise alignment using a proposed index. Müller and Röder employed DTW to derive motion templates used in classification and retrieval of motion capture data [104]. Zhou et al. used DTW in their proposed aligned cluster analysis that segments motion capture data [105]. Kim et al. employed Laplacian curve manipulation to perform warping based on user-defined constraints [106]. Raptis et al. utilized DTW in their gesture classification algorithm [107]. Cimen et al. used DTW prior to extracting affective descriptors from motion [108]. Heloir et al. proposed a DTW-based method along with weighted PCA for aligning motion of communicative gestures [109].

To address the problem of style translation, Hsu et al. proposed iterative motion warping (IMW) [24]. Style translation is the process of transferring the style of one particular motion sequence onto another. This process requires accurately aligned sequences. IMW is composed of space and time warping procedures (based on DTW) to address the problem. Taking into account the kinematics of motion, Hsu et al. utilized pose, velocity, and acceleration based feature vectors in their time warp objective function [110].

As one of the more recent techniques, Zhou and De la Torre proposed canonical time warping (CTW) and local canonical time warping (LCTW) based on canonical correlation analysis (CCA) and DTW [111]. Evaluated with synthetic facial expression videos and motion capture data, the method was shown to outperform other DTW-based

techniques including IMW, DTW, and DDTW.

Finally, different approaches have been developed for alignment in the computer vision community. Homography computations are popular means in this regard [112, 113, 114]. Similar to techniques developed for motion capture data, most computer vision methods of alignment rely on DTW. For example, Junejo et al. utilized a self-similarity matrix and proposed a video alignment technique based on DTW [115]. As another example, Lu and Mandal proposed a method based on computation of the trajectory of the object of interest and correspondence calculation using DTW [116].

2.3.2 Interpretation

Generally, three different communities have shown interest in interpreting motion from a computational standpoint: the vision community, the graphics community, and the behavioral/medical community. Members of the vision community often apply their findings to surveillance systems [117], sports (e.g. commentary applications, training, analysis) [118], and gesture-based interactive systems [119, 120], among others. The methodologies developed by the graphics community are most often used towards retrieval of motion capture data [102]. Finally, the behavioral sciences and medical community use motion computing for analyzing how humans see, perceive, and perform motion [79], as well as how particular disorders affect motion so that effective aid systems can be developed [121]. Vision-related research often uses video streams, graphics research frequently uses motion capture data, and the behavioral/medical related research uses both types of data along with medical imaging. Nevertheless, a new wave of studies is aiming to bring together these fields by developing systems that successfully fuse the two data types and utilize them in unified systems [122].

Motion interpretation consists of several popular sub-categories, the most important of which are modeling [123], classification [124], indexing and retrieval [103], segmentation [125], tracking [126], and feature extraction [127]. The three latter procedures are often used as components or pre-processing steps for other systems such as classification or retrieval [128].

Due to the increase in the availability of motion capture datasets, retrieval and indexing has recently become very popular. The goal of this process is to query and/or index different features in motion. Such motion features can range from actions themselves to emotional features, and can be characterized through geometrical, structural, and dynamic settings. A variety of techniques have been proposed in this regard which we describe below.

Many indexing techniques are grounded on the work of Faloutsos et al. [129]. Keogh et al. use time warping and bounding techniques for indexing of large datasets [130]. Liu et al. used k-nearest neighbor (KNN) for retrieval [131]. Their method utilized motion index trees of hierarchical joint features. Similarly, Deng et al. used body-part-based hierarchical trees along with string matching [132]. Krüger et al. utilized KNN classifiers to perform fast and efficient similarity search for indexing in very large datasets of motion [133].

Pre-defined features have often been used to eliminate the spatiotemporal variations in motion for content-aware retrieval. This approach enables for fast processing of motion capture files, which can be used for very large datasets. Müller et al. proposed spatiotemporal invariant geometric features for efficient and content-based retrieval of

motion [103]. Additionally, their model incorporates time segmentation. Müller and Röder utilized motion templates as a means for retrieving and classifying logically similar but spatiotemporally varied motions [104]. Both these mentioned techniques can be utilized for large motion datasets. Kapadia et al. developed a system using motion keys, which are defined as various structural and dynamic features. Their proposed method, which supports fuzzy operations, can be utilized for fast and efficient retrieval [134]. Finally, Müller et al. used motion templates as descriptors which capture consistent and variable features, along with genetic learning [135].

A great deal of overlap exists between classification/recognition processes and indexing/retrieval systems. While the former mostly emphasizes on motion semantics, the latter leans towards more advanced machine learning techniques. Moreover, the former has mostly attracted the animation community, while latter has attracted both computer vision and animation researchers alike. The three approaches of tracking, segmentation, and classification of vision-based motion are closely linked and quite popular.

Liu et al. used KNN to classify actions from a set of fused features [136]. Schindler and van Gool utilized motion snippets, anywhere from only 1 to 10 frames, to classify motion using support vector machines (SVMs) [137]. They illustrated that the snippets can achieve a recognition rate of 90% and 5-7 frames are sufficient for achieving a classification rate similar to when full sequences are used. Yu et al. employed artificial neural networks (ANN) for classification [138]. Motion detection was performed, followed by extraction of features that were used to train the ANN classifier. Probabilistic techniques such as hidden Markov models (HMM) have also been used for action recognition [139]. Ren and Xu used HMM for action recognition successive to feature

extraction [140]. Chan et al. proposed a fuzzy framework for action classification [141]. Similar methods have been implemented for classifying actor attributes such as gender and affect. For example, KNN, Naive Bayes, and SVM [142] was used for affect recognition and linear regression was utilized for gender and weight classification [122].

Feature extraction, in many cases, is used to analyze action or style features for behavioral studies. For example, principal component analysis (PCA) along with Fourier decomposition has been used for extracting gender-specific features [73] while PCA and regression analysis has been used in [79] for affective features. PCA and kinematic features were utilized for action recognition using KNN [143]. Features can also be used to enhance classification performance, for example, PCA features were used with linear discriminant analysis (LDA) [144], and scale-invariant feature transform (SIFT) was used with SVM [145]. When feature extraction is carried out, should the extracted features be related to style, they can also be transferred to other sequences. In such case, the process becomes a form of synthesis, namely style translation [21].

Laban movement analysis (LMA) [146, 147], is a notation language for defining motion features which categorizes human motion into Body (structure), Effort (energy), Shape, and Space. On a higher level, LMA explores Mobility/Stability, Inner/Outer, Function/Expression, and Exertion/Recuperation. In general, LMA defines and captures the entirety of motion, and is much more detailed than our PT/ST approach. However, for applications involving personalized variations, LMA provides many features and layers, which can be computationally expensive. We therefore believe that our two-layer approach is generally sufficient for such purposes. Nonetheless, LMA has been widely used in association with motion computing methods. Chi et al. has proposed an animation

system that utilizes LMA, more specifically Effort and Shape components, for adding expressive and natural features to gestures [148]. Motion retrieval [149, 150], segmentation [151], synthesis [152], and learning motion styles [153] have also been carried out using LMA.

2.3.3 Control and Synthesis

These methods aim at controlling [154], editing [155], modeling [156], or generating [157] actions or motion features (including SFs) for animation purposes [158]. Unlike classification methods, synthetic human motion has mostly interested the graphics community, and understandably so. Such methods are most likely utilized in creating animated movies, games, and other similar content.

Despite some early research on modeling human motion mostly through physics or control-based methods [159, 160, 161, 162], it was in 1978 that inspired by the new simplistic way of motion visualization, i.e. PL, the modern era of motion synthesis started. Cutting produced computer codes that animated a regular synthetic walk [163] along with feminine and masculine variants [164]. The graphics community has since taken this topic to new heights.

Different approaches are often considered for synthesis or control of motion features, namely rule-based techniques [90], control and physics-based methods [165], signal processing methods [98], and learning/optimization systems [25]. In the following, we review some of the important works based on these approaches.

Rule-based approaches were among the first attempts at synthesis of features. Amaya et al. adjusted amplitude and speed of motion to synthesize emotions in motion [90].

Bruderlin and Calvert developed a hybrid rule-based dynamic control system for gait generation [166]. They also proposed a knowledge-driven set of procedures for synthesizing different styles of human running [167]. Perlin utilized procedural texture synthesis for realistic real-time animation [168]. Perlin and Goldberg later proposed a system composed of separate animation and behavior engines for interactive generation of realistic animation based on author-defined rules [169]. Chi et al. developed a motion engine based on LMA that enables addition of Effort and Shape (different components of LMA) to motion for increased naturalness [148].

Signal processing methods are among the earliest routes towards synthesis and altering motion. Rose et al. used radial basis functions and interpolation/extrapolation methods for blending of styles [23]. Pullen and Bregler used motion capture data to add texture to keyframed animation [170]. Their proposed model adds mid and high-level frequency alterations to keyframed or synthesized signals through a process referred to as texturing. Bruderlin and Williams used multi-resolution filtering adapted from image processing to tune different frequency bands for altering motion features [98]. Component analysis methods have been widely used in extracting and adding style features to motion. For example, Urtasun et al. employed PCA decomposition to generate individual motion styles [171] while Shapiro et al. used independent component analysis (ICA) for decomposing motion to components and add stylistic ones to other sequences [172]. Liu et al employed ICA for decomposing motion into different subspaces, among which was those for style [21]. The features were then merged, altered, and transferred for generating new motion clips. Ahmed et al. generated realistic motion using wavelet analysis and multi-resolution blending [173].

As an early example of physics-based techniques, Boulic et al. generated a variety of walking motions by spatial, temporal, position, and configuration characteristics of walk models [174]. Tsai et al. later used inverted pendulum models for generating motion styles [165]. Grochow et al. proposed an inverse kinematics (IK) model that learned from probability distributions of motions in a dataset, and subsequently generated stylistic poses [175]. Coros et al. used IK and proportional-integral-derivative (PID) controllers to generate stylistic motion as well as motion for different character proportions [176]. Popovic [155], and Popovic and Witkin [177] edited dynamic properties of input motion to create a variety of realistic motions using space-time dynamics optimization. Fang and Pollard developed a system that generates physically valid motion [157]. Their method is based on the optimization of torque and force first order derivative objective functions. Wei et al. utilize a probabilistic framework along with a set of physical dynamics models and constraints to synthesize realistic motion [178].

Finally, in learning techniques, Hsu et al. proposed a system which learned style translation models using linear time-invariant (LTI) system identification [24]. Safonova et al. devised an optimization problem, using which motions from a dataset were depicted onto a lower dimension subspace that preserved the desired behavior features [179]. Brand and Hertzmann developed a system which learned style patterns from a dataset of dance motions using probabilistic models [25]. Their system is capable of synthesizing a variety of interpolated or extrapolated styles. Kovar et al. developed a framework in which a dataset of motion capture data is used to generate motion along user-defined graphs [180]. Their model uses an optimization problem to search for excerpts of available data and generates the required transitions. Liu and Popovic developed a system

in which realistic motion using linear and angular momentum constraints and learned KNN estimators [181]. Lee and Popovic utilized Markov models to learn behaviors from set of examples [182]. Ma et al. employed Bayesian networks to learn style parameters and latent variation [183]. Their statistical model is capable of interpolating motion styles and variations based on user defined parameters. Arikan and Forsyth proposed interactive cut-and-pasting of motion segments successive to graph search and followed by post-processing [184]. Arikan et al. later developed a system that enables synthesis of smooth and natural sequences using a pre-annotated dataset of motion [158]. Their algorithm is based on dynamic programming optimization.

Chapter 3.

Proposed Approach

3.1 Main Theory

John Lasseter of Pixar (http://www.pixar.com) has said: "When character animation is successful and the audience is thoroughly entertained, it is because the characters and the story have become more important and apparent than the technique that went into the animation. Whether drawn by hand or computer, the success of character animation lies in the personality of the characters" [185]. *Personality* or personal characteristics of digital characters are derived from a variety of factors that in Chapter 1 we categorized together as secondary themes (STs). STs can include emotions, gender, age, energy, and even attributes such as health, genetics, and social aspects, and others. Accordingly,

interpretation and synthesis of the features that display these themes, which we referred to in Chapter 1 as secondary features (SFs), is of critical importance. The general goal of this dissertation is processing of SFs in human motion. In particular, we aim at extraction, synthesis, and translation of SFs.

As we discussed in Chapters 1 and 2, to accurately process SFs, we believe perceptual quality is of critical importance and perceptually guided methods can provide efficient and effective solutions for many of the existing problems. We believe that for different components of SF-based motion processing, most notably time warping, SF extraction, SF synthesis, and SF translation, the mentioned approach is critical. Accordingly, we base this dissertation on perceptually guided/inspired and perceptually accurate methods.

In this chapter, we discuss our overall methodological approach for dealing with STs and associated features. We present discussions on the current status of motion processing systems based on the review of related work in Chapter 2, and argue how they can benefit from our approach. Overviews of methods and techniques used are briefly mentioned, and the tools required to execute and evaluate the developed systems are stated.

3.2 Methodology

3.2.1 Overall Methodological Approach

Animation of human motion is intended for a human audience, the approval of whom is essential for generated content. In Chapter 2, we reviewed some other important parameters such as naturalness, believability, realism, and aesthetic edits that can greatly influence our reception of motion animation. The common property amongst these

domains is their subjectivity and dependence on audience perception. Accordingly, despite the quantifiable nature of human motion, we believe the human perception, which is often qualitative, should play a leading role in development and testing of motion-related systems. Consequently, the study conducted through the course of this dissertation is composed of two main components, namely computational methods and perception studies, making our approach one of multidisciplinary nature. The computational methods include a variety of established processes from machine learning and optimization to statistical analysis. While computational components provide the main tools required to perform the designated tasks, perceptual cases lay the basis and grounds for the design of our proposed systems as well as user studies conducted to evaluate and validate their performance. In other words, we believe perceptually guided and perceptually compatible methods need to be considered for processing of motion and associated features.

3.2.2 Research Methods

- In Chapter 2, we observed that a significant portion of motion studies use time-warping as a critical tool for aligning motion features. Based on the review on time-warping methods presented in Chapter 2, it is evident that most existing warping methods used in the field of animation are developed based on dynamic time warping (DTW), and sometimes uniform time warping (UTW). These methods have considerable shortcoming [130, 186], ranging from the type of utilized objective functions to the type of stretching/compressing used in the warping procedure, for which we provide an in-depth discussion in Chapter 4. As we argue in Chapter 4, it is important to utilize warping techniques that are more grounded on our perception of

motion. Warping should ideally benefit from high customizability, which will enable fine-tuning of the warping process by the user or application. Low distortion is another desired feature that has dominantly been left unaddressed in most current warping techniques. An enhanced warping method that considers these factors, will most likely further decrease the artifacts that are known to be introduced due to warping [187]. Consequently, as the first step towards developing a set of computationally and perceptually accurate systems, a new time-warping method is introduced in Chapter 4 that addresses the mentioned issues.

- On the topic of perceiving motion and its embedded features, we observed in Chapter 2 that rigorous physiological and psychological research has been carried out to determine the way in which actions and stylistic/affective features are performed and perceived. While advanced physiological and psychological studies lie outside the realm of this dissertation, we find it necessary to investigate particular subjects that are required as pre-assumptions for our methods. The systems that we propose in future sections provide the capability of extracting and synthesizing style/affect features from individual joints. Accordingly, we believe it is essential to investigate how different parts of the body contribute to our perception of features and whether a limb-by-limb (or joint-by-joint) approach in extraction/synthesis of features is a valid one. Given the fact that most current systems use this type of spatially incremental processing, this study can have great implications in terms of efficiency. In other words, should the study prove that additivity holds for perception of STs from joints, existing methods can focus on only a subset of the human body and achieve perceptually acceptable results. This is investigated in Chapter 5.

- Despite the significant amount of work in interpreting motion and its related stylistic features, there remain particular areas that can benefit from perceptually guided frameworks. Existing techniques that aim at extracting or generating style and affect features are either based on observations [79], or directly computed from datasets of recorded motion data [172] using different learning or statistical tools. While the results acquired using such strategies have proven effective, they may not necessarily be efficient and perceptually optimal. Particularly, we believe a method that systematically extracts SFs based on the three categories in which they are spatiotemporally carried out (i.e. posture, movement, and time as described in Chapter 2), is a key missing approach. We address this problem in Chapter 6. Similarly for synthesis of features, we believe techniques that utilize expert-knowledge can provide efficient, perceptually effective, and insightful ways of generating style features in motion. In this area, however, a framework which enables the input of multiple experts (most likely animators) to be collected and merged to generate a feature synthesis system, is still missing. We address this in Chapter 7. Finally, from the variety of machine-learning or optimization-based methods for recognition or synthesis of motion features, to the best of our knowledge, a unified system capable of performing both has not yet emerged. Inspired by biological and cognitive architectures and in line with our perceptually guided approach, we believe such a system can be beneficial, both in terms of practicality and from a cognitive standpoint. This problem is addressed in Chapter 8.

- Motion processing techniques, including the ones developed in this dissertation, aim at altering or synthesizing features in motion to achieve particular goals. In Chapter

9, we summarize the common reasons for which motion features are often synthesized or modified. We present a comprehensive paradigm abiding by which results in perceptually valid animation of human motion. The paradigm consists of several components taking into account different aspects of the depicted scenes. Relative significance of the different components of the paradigm is also studied in detail, which can be of use to animators.

Figure 3.1 presents the methods proposed and developed in this dissertation. Descriptions on each of the systems and motivating factors with respect to computational and perceptual accuracy are provided in the figure.

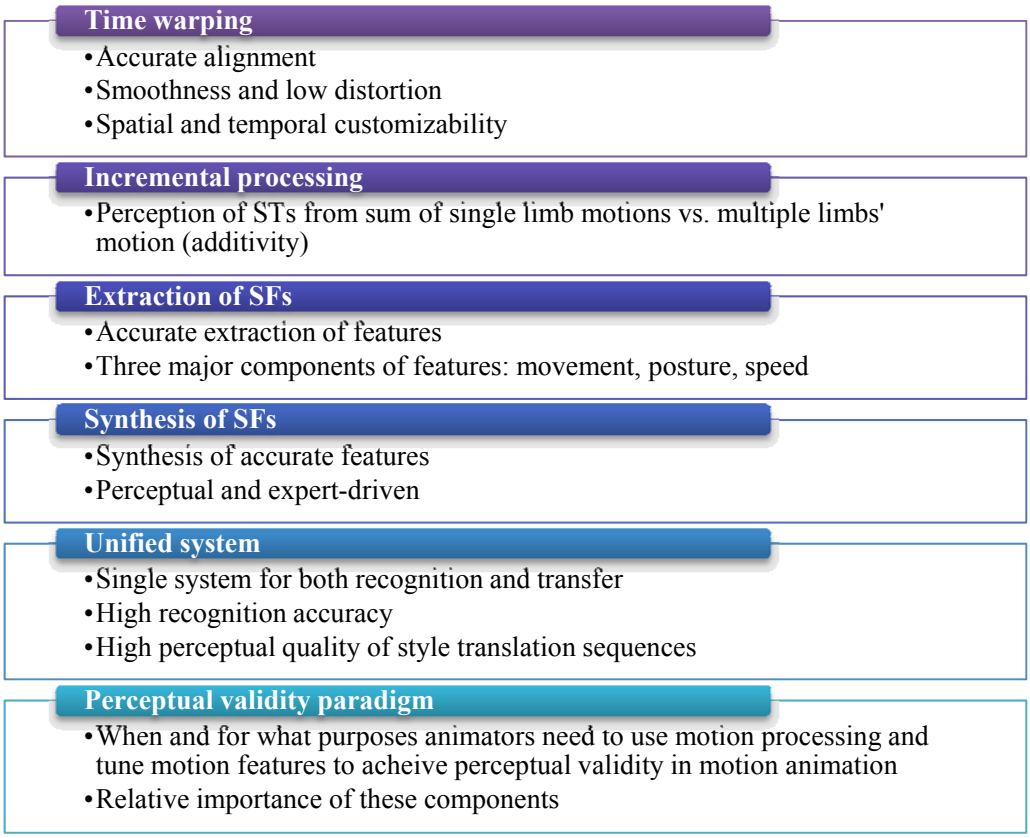


Figure 3.1. Summary of proposed and implemented methods.

3.2.3 Data Collection and Analysis

- Motion capture data comprise the main type of recorded motion used in this study. Pre-existing and publically available datasets are utilized. Additionally, a separate dataset is recorded which contains particular characteristics absent in pre-existing datasets and is used in parts of this study. For details regarding these datasets, we refer the reader to Appendix A.

- Generally, the methods developed in this research are illustrated through motion sequences. Video clips are rendered, posted online, and referred to in the text. Extracted frames are presented where necessary. The same clips or slightly modified versions of these videos are used in the perception studies. The characters in these motion sequences are represented by point-light (PL) and stick-figure. No particular mesh, skin, or clothes are applied in order to prevent any bias for the audience. Moreover, when displayed for the audience, no particular contexts are provided for the scenes to avert influencing the perception of the sequences. In other words, we assume that with lack of a given context, the audience will default to neutrality. Nonetheless, as we will argue in Chapter 9, context is a critical and determining factor in perception of motion scenes. Hence, additional research will be required on our proposed techniques and acquired perceptual feedback with different contexts taken into account.

- In the proposed methods, physical constraints have not been taken into account as we intended to evaluate and examine the performance of the proposed methods prior to significant post-processing. Nonetheless, due to the accurate performance of the proposed methods, physical constraints were not widely violated and perception was

not negatively affected. Thus, the need for related physical constraints is not felt.

- The systems and methods proposed in this dissertation are primarily evaluated using subject feedback through pre-defined questionnaires. Samples of these questionnaires are provided in Appendix B. Such evaluations, we believe, are key in ensuring perceptually accurate systems and are in line with utilizing the human perception as the grounds for computational motion processing techniques. Participants consist of experienced animators as well as those naïve towards motion studies. In Chapters 7 and 9, both experienced animators and naïve subjects participate while in Chapters 5, 6, and 8, only naïve participants provide the required feedback. Further details are provided in respective chapters.

- In some chapters of this dissertation such as Chapter 4 in which we propose an accurate and customizable time warping method, and Chapter 8 in which the proposed system performs classification, quantitative measurements for system evaluation are used. These measurements range from distance-based objective functions and correlation to number of true positive recognitions. In other chapters, user feedback is utilized to evaluate the performance of the systems.

- To analyze user feedbacks and ratings, quantitative statistical methods such as analysis of variances (ANOVA) are used. Correlation, regression, and other such statistical analysis techniques are also employed where needed.

3.2.4 Research Tools

Computational methods include optimization techniques such as dynamic programming and exhaustive search as well as machine learning systems such as K-nearest neighbor

(KNN) classifiers, support vector machines (SVMs), and artificial neural networks (ANNs). Statistical analyses include computation of mean, standard deviations, standard errors, histograms, regression, correlation, and ANOVA.

For user studies, paper-based questionnaires were used. Ethics approvals were secured. Samples of the used questionnaires are presented in Appendix B.

All computational algorithms including machine learning and optimization systems, statistical analysis, animation for user studies, and animation for the purpose of presentation in this dissertation are carried out in MATLAB. A description on implementation of the different systems that were developed throughout this dissertation is provided in Appendix C. There, we present the hardware and software specifications as well as the functions and programs that were implemented. In the future, some of the implemented routines will be released online for public use, as they can be useful resources. Finally, the implementation issues are briefly discussed in Appendix C.

Chapter 4.

Correlation Optimized Time Warping

4.1 Introduction

As mentioned in Chapter 2, motion data, a form of time-series, contain misalignments with respect to one another, even when they are similar in motion content and semantics. For example, Figure 4.1 illustrates three walks performed by the same person where the illustrated poses are clearly misaligned. Figure 4.2 (a) illustrates joint angle trajectories from two similar sequences where the misalignments are visible in the extrema. Similar to most of the reviewed works in Chapter 2, the methods presented in this dissertation (whether for interpretation, synthesis, or analysis) are no exception and need alignment as a critical pre-processing step.

To tackle the issue of misalignment in time-series, various techniques have been proposed based on application and context. For example dynamic time warping (DTW) was proposed to align speech signals [92], which became very popular in motion as well, while canonical time warping (CTW) [111] and iterative motion warping (IMW) [24] were proposed for motion data. A detailed review was presented in Chapter 2.

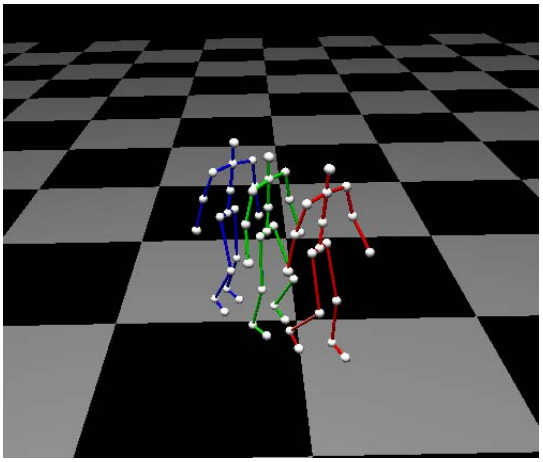
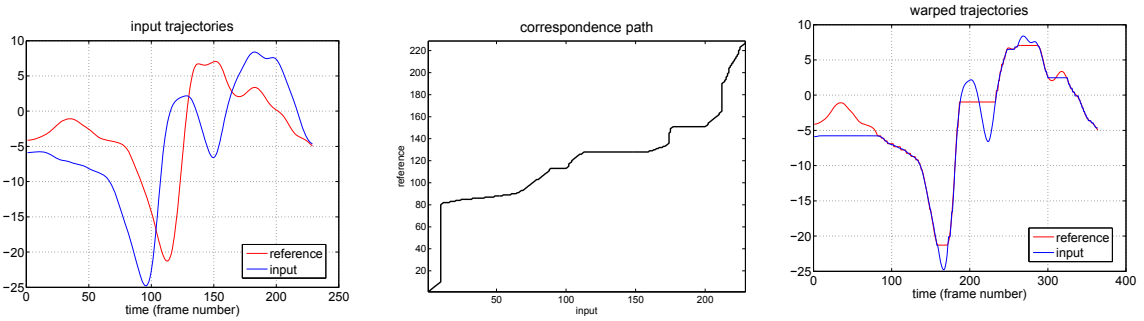


Figure 4.1. Misaligned poses are illustrated for three walk sequences performed by the same person.

First and foremost, the outputs produced by a warping system need to have both high perceptual and form-related quality. Most techniques such as DTW and CTW compute an alignment path, which determines the correspondence between instances of the input and reference trajectories. Figure 4.2 illustrates an example. Utilizing this correspondence function for warping does not produce smooth curves. Rather, identical unvarying consecutive frames (still-frames) are often used to compensate for timing where needed. Examples of this occur at the vertical and horizontal lines in the correspondence path presented in Figure 4.2. The output warped trajectories clearly illustrate the use of still-frame. It is however, desired that the warped sequences be smooth and distortion-free,

and the output sequences can be readily used for animation purposes. To remedy this, post-processing in the form of smoothing or interpolation is required. It should be noted however, that warping constraints, such as, slope constraints [94], weight factors [92], and path constraints [92, 95], have been widely proposed and used for existing techniques. While these modifications have often been imposed with the purpose of increasing efficiency and speeding up the procedures, they can serve towards prevention of excessive use of still-frames and over warping, hence distortion.

Another issue worth considering is that to achieve alignment, most methods such as DTW and CTW warp the reference and input together. An example of this is presented in Figure 4.2 where both input and reference are warped for aligning the trajectories. This can be considered a liability as it is often desired to warp an input motion independent of how the reference needs to compensate temporal misalignments.



(a) (b) (c)

Figure 4.2. Warping methods such as DTW use frame-wise correspondence and use still-frames to achieve alignment, warping both reference and input.

Most existing warping techniques such as DTW and CTW minimize the computed distance between corresponding motion trajectories through the process (utilize distance-

based objective functions). While distance is often used to quantify similarity between sequences, additional steps are sometimes taken to ensure such measures are correct representatives for comparison in motion data [104, 149]. In fact, it has been previously observed that distance-based measurements alone are not necessarily the best indicators of perceptually similar motion sequences [42, 130]. Distance-based objective functions analyze the proximity of entries of two trajectories at corresponding time instances (frames). Nevertheless, humans are known to look for shapes, forms, and patterns rather than investigate individual entries [188, 189]. As the audiences of motion processing systems are we humans, this property needs to be taken into account when aiming to achieve perceptually valid warped motion trajectories.

In addition to high perceptual and video quality in warped outputs, as well as a suitable characterizing factor for similarity/dissimilarity, an ideal warping technique must benefit from high customizability. Users or applications may need to tune the warping process. This tuning may be aimed at warping only particular regions of the body (spatial customization). Alternatively, the application may demand for different segments of the sequence to be warped differently (temporal customization), hence tweaking the temporal resolution of the warping process.

Finally, when warping two sequences, the selection of the reference depends on the application. On the other hand, in the event that multiple sequences need to be aligned, selection of the reference trajectory can be a difficult and influential issue. It is beneficial for the framework to allow for automatic selection of a reference that will demand the least warping from other sequences. This process can especially be useful when aligning a dataset for different purposes such as training a classifier. Moreover, this modality too

should be customizable to allow more emphasis on particular motion degrees of freedom (DOFs).

In this chapter, we propose correlation-optimized time warping (CoTW) for aligning motion sequences. The proposed method is inspired by and builds on the correlation optimized warping which was initially developed for aligning chemometric data [190]. This technique when previously utilized in other fields has illustrated good performance in different aspects such as peak shape and area preservation [191], and has been widely explored and developed in the fields of chemistry [187, 192, 193]. The method has also been used in image processing and biomedical image analysis [194]. To the best of our knowledge, this approach has not been the basis of any warping techniques for human motion data. In this chapter, we further develop and tailor this technique for human motion data and provide a new and robust method for aligning multiple motion sequences while addressing some of the shortcomings in currently available techniques.

In addition to robust alignment performance, CoTW has several advantages over most other time warping techniques: (a) it uses a more effective objective function based on correlation; (b) it allows for alignment to be customized both temporally and based on spatial regions of interest within the character body; (c) it reduces artifacts such as distortion and does not employ still-frames that appear in existing methods for timing adjustments; (d) optimal reference is automatically selected when multiple sequences are being warped. While some of the previous work partially address these issues, to the best of our knowledge, there is no warping technique that attends to them all. In the following sections, we study the parameters and the details of the method, and perform rigorous experiments showing the robustness of the method. Figure 4.3 presents a schematic of the

entire warping system and its different components. Other components of the system are described in the following sections.

The contents of this chapter have been published as [195, 196].

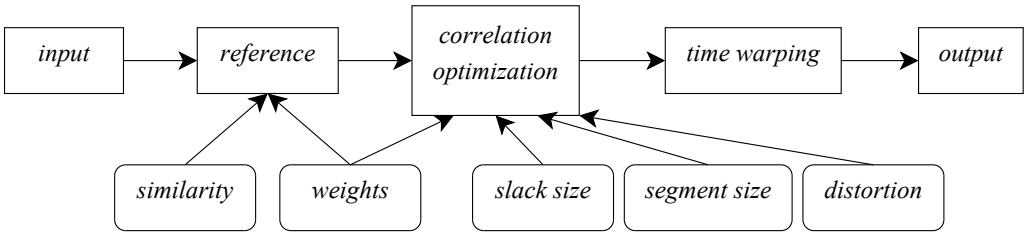


Figure 4.3. The overall schematic of the system is presented.

4.2 Proposed Method

4.2.1 Algorithm

Temporal modifications can be carried out linearly. For example, a motion trajectory with temporal length m (i.e. m frames) can be linearly compressed or stretched to new length m'. This correction may or may not align the critical features that are of importance. For instance, a fatigued-walk, which is slower than a normal walk, can be compressed to take the same temporal length as the regular walk. The strides however, will not necessarily be aligned (see Figure 4.1). This process, also referred to as uniform time warping (UTW), will be useful to simply length-match motion sequences. However, UTW can be used as a major building block in piece-wise or non-linear warping methods.

Based on the description of motion data provided in Appendix A, given a motion matrix $\boldsymbol{\mathcal{D}} = [\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \cdots \boldsymbol{\theta}_n]$ with n degrees of freedom (DOFs), the ith joint angle trajectory $\boldsymbol{\theta}_i$

with m frames is defined by $\boldsymbol{\theta}_i = \left\{\theta_i^{(t)} : t = 1, \cdots, m \in \mathbb{N}\right\}$. Accordingly for $\boldsymbol{\theta}_i$ the uniformly warped trajectory $\boldsymbol{\theta}_{UTW,i}$ is calculated using $\boldsymbol{\theta}_{UTW,i} = \boldsymbol{\theta}_i^T \mathbf{W}$ where $\boldsymbol{\theta}_{UTW,i} = \left\{\theta_{UTW,i}^{(t)} : t = 1, \cdots, m' \in \mathbb{N}\right\}$ and $\mathbf{W}$ is the $m \times m'$ warping matrix populated using linear interpolation factors required to warp $\boldsymbol{\theta}_i$ to achieve temporal length m'. If $m' > m$, the trajectory is stretched, and where $m' < m$ the trajectory is compressed. In addition to linear interpolation, non-linear methods can also be used for calculating $\mathbf{W}$. A sample compressing and stretching is illustrated in Figure 4.4 where linear interpolation is used.

Our proposed method uses UTW in two different stages. First, CoTW linearly warps the input trajectory using UTW to length-match the trajectory with respect to the reference. The input is then divided into a number of equal segments. Accordingly, $\boldsymbol{\theta}_i$ is rearranged as $\boldsymbol{\theta}_i = \left[\theta_i^{(1:\lambda)} \;\; \theta_i^{(\lambda:1+2\lambda)} \;\; \cdots \;\; \theta_i^{(c\lambda+1:m)}\right]^T$ where we have c segments each with a length of $\lambda \in \mathbb{N}$. Given $\boldsymbol{\theta}_i$ with a length of m, the number of segments is calculated using $c = \lfloor m/\lambda \rfloor$, $c \in \mathbb{N}$.

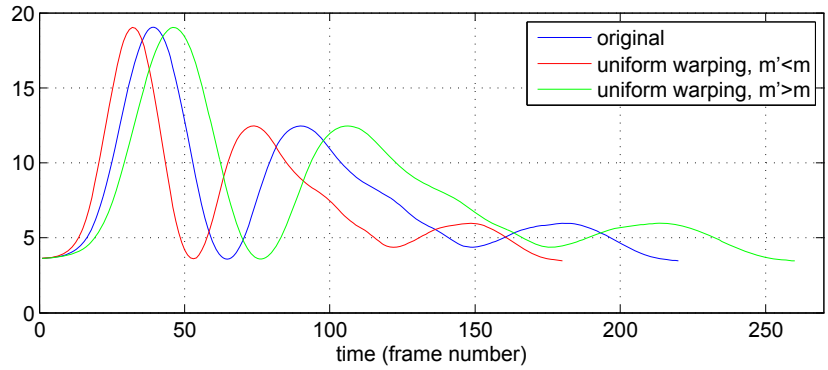


Figure 4.4. Linear stretching and compressing of a motion trajectory.

In addition to the segment size, a different parameter called slack size is introduced. We

denote slack size by $\delta \in \mathbb{N}$. This parameter determines how much each segment is permitted to warp in either direction. In other words, each segment of $\boldsymbol{\theta}_i$ will have a temporal length in the range of $[\lambda - \delta, \lambda + \delta]$ after CoTW warping. Specifically, segment i is warped by η_i where $\eta_i < \delta$. Accordingly, for assigned λ and $\eta_{1:c}$, input $\boldsymbol{\theta}_i$ is warped using:

$$\boldsymbol{\theta}_{CoTW,i} = \left[\theta_i^{(1:\lambda)}\mathbf{W}_1 \;\; \theta_i^{(\lambda:1+2\lambda)}\mathbf{W}_2 \;\cdots\; \theta_i^{(c\lambda+1:m)}\mathbf{W}_c\right]^T \tag{4.1}$$

where $\mathbf{W}_1$ to $\mathbf{W}_c$ are warping matrices with dimensions $\lambda \times (\lambda + \eta_i)$. The entries of $\mathbf{W}_i$ are populated with values required to linearly warp the designated segment to match the required length. We combine the segment-wise warping matrices to create the $m \times m$ global warping matrix:

$$\mathbf{W}_{\mathrm{g}} = \begin{bmatrix} \mathbf{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{W}_c \end{bmatrix} \tag{4.2}$$

where, ${\boldsymbol{\theta}_i}^T\mathbf{W}_{\mathrm{g}}$ warps the entire input $\boldsymbol{\theta}_i$ in segment-wise fashion. Since the input has already been length-matched with the reference, CoTW only allows combinations of warping degrees that result in the output having the length of m, in other words $\sum_{i=1}^{c} \eta_i = 0$.

Figure 4.5 illustrates the process where the original input is first uniformly warped (using UTW) and length-matched with the reference. The input is then divided into a number of segments. Each segment is warped by $\eta_i < \delta$. The objective function, which is described in the following sections, utilizes segments of the input and corresponding segments of the reference, and optimizes the set of η_i.

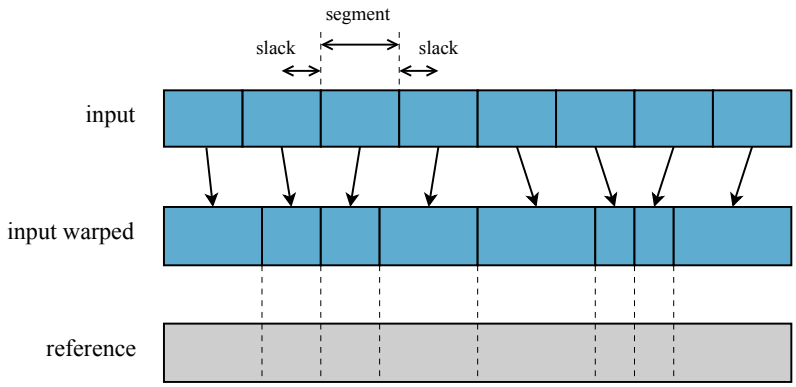
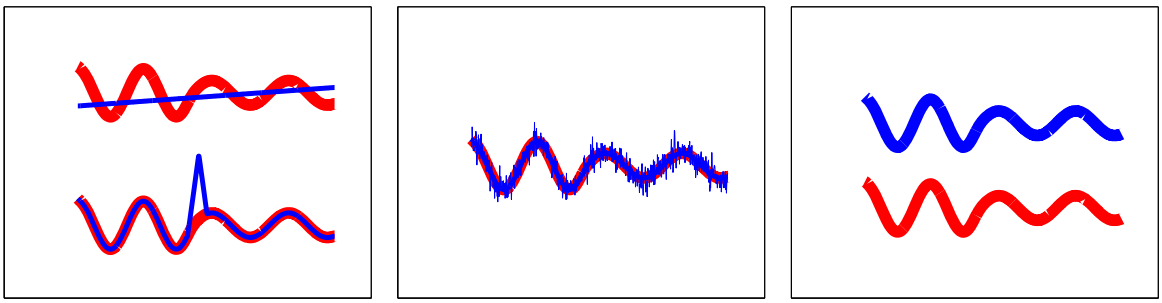


Figure 4.5. An input trajectory is divided into a number of segments. Each segment is allowed to warp, using UTW, by a bounding slack parameter. Warping is carried out with the aim of achieving maximized correlation with respect to the corresponding segments of reference.

4.2.2 Objective Function

As mentioned earlier, from a perceptual standpoint, distance functions may not perfectly capture and represent similarity between trajectories. From a computational perspective, distance-based objective functions (used in DTW for example), are known to only solve local problems and misalignments [130]. Let's assume two hypothetical joint angle trajectories $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ with lengths m. Using two typical distance-based objective functions such as $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ or $\sum_{i=1}^{m}\left|\theta_1^{(i)} - \theta_2^{(i)}\right|$, two identical trajectories with only one or few relatively distant entries, can result in an average distance identical to two trajectories where all of their entries are different (different shapes). However, semantically and perceptually, the first two trajectories are more similar than the latter. Figure 4.6 (a) shows an example where the pair shown in the bottom is more similar than the pair at the top, while the distances of the two pairs are similar. To illustrate another computational shortcoming, let us assume two identical trajectories, to one of which noise is introduced. If the magnitude of this noise is relatively small with respect to the

magnitude of the trajectory, the overall shape of the trajectory will not be affected, and the two trajectories remain relatively similar. Meanwhile, a distance-based analysis will point to the two being quite dissimilar. Figure 4.6 (b) illustrates the situation where a calculated distance is relatively large but the two trajectories are perceptually similar. This can be seen as an extension to the previous case. Finally, given two identical trajectories, one of which is spatially shifted, a distance-based measurement will indicate dissimilarity (based on the magnitude of the offset). Nevertheless, the two trajectories, especially from a motion perspective, as well as from a perceptual standpoint, are identical. Figure 4.6 (c) illustrates this situation where a relatively large distance is computed whereas the trajectories are identical in shape and form.



(a) (b) (c)

Figure 4.6. Three hypothetical situations where distance between the two trajectories fails to indicate similarity. In (a), from a shape and form standpoint, the pair at the bottom are more similar than the pair at the top. However, the norm-2 distances between trajectories in each pair are equal. In (b) relatively large distance is calculated for the pair while they are quite similar in shape. The blue trajectory is the noisy version of the red one. Finally, in (c) the distance between the pair is quite large given an introduced spatial offset. The two trajectories, however, are identical.

Similar to [190], we suggest and utilize Pearson's linear correlation coefficient (PCC) which is a numerical determinant of dependence of two variables or how similar the

shapes of two trajectories are. We derive an objective function based on PCC as the means of quantifying alignment. Specifically, for two motion trajectories $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ with m frames, the correlation coefficient is calculated as $\rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = cov(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/\sqrt{\left(var(\boldsymbol{\theta}_1)var(\boldsymbol{\theta}_2)\right)}$. The objective function is based on ρ as given by:

$$\rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{\sum_{t=1}^{m}\left(\boldsymbol{\theta}_1^{(t)} - \mu_{\boldsymbol{\theta}_1}\right)\left(\boldsymbol{\theta}_2^{(t)} - \mu_{\boldsymbol{\theta}_2}\right)}{\sqrt{\sum_{t=1}^{m}\left(\boldsymbol{\theta}_1^{(t)} - \mu_{\boldsymbol{\theta}_1}\right)^2 \sum_{t=1}^{m}\left(\boldsymbol{\theta}_2^{(t)} - \mu_{\boldsymbol{\theta}_2}\right)^2}} \tag{4.3}$$

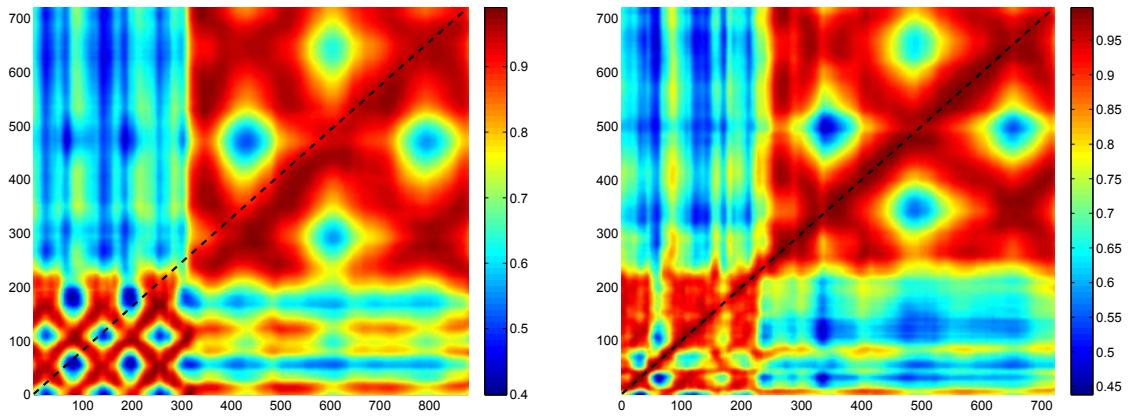
where μ represents the mean.

Revisiting the three cases where distance failed to represent similarity, we observe that in Figure 4.6 (a) PCC calculates $\rho = -0.0587$ for the top pair, which indicates dissimilar trajectories, while for the bottom pair $\rho = 0.8231$, indicating that the two are relatively similar. For Figure 4.6 (b), $\rho = 0.8231$, again pointing to the two trajectories being similar while distance measurements will calculate large values. Finally in Figure 4.6 (c), a calculated distance is very large while PCC calculates $\rho = 1.0$, pointing to identical shapes.

Another significant advantage of using a correlation-based function, is that PCC is a normalized value (maximum = 1). Therefore, calculated results for different motion representations such as Euclidean displacement vectors and joint angle trajectories will be comparable and can fit the same framework. This cannot be said about warping methods that utilize distance-based objective functions.

Figure 4.7 illustrates the correlation matrices between two sequences of motion before (a) and after (b) alignment. In these matrices, each entry is the correlation between postures

of the two sequences at given time instances. The diagonal line denotes a one-to-one correspondence between postures of the two sequences at corresponding frames. Evidently, the sum of all correlation values between corresponding postures has significantly increased after alignment. This is a testament that a correlation-based objective function can accurately represent motion alignment.



(a) (b)

Figure 4.7. Correlation matrices between two sequences of motion before (a) and after (b) warping. Each entry is the correlation value between postures of the two sequences at a given frame. Increased correlation entries on the diagonal line indicate that alignment increases correlation values between corresponding frames.

Our objective function calculates PCC for each segment of the length-matched input trajectory with respect to its corresponding section from the reference. The goal is to calculate the set of $\boldsymbol{\eta}$ that maximizes this function. For warping multi-dimensional data one approach is to warp each DOF separately. Using CoTW, despite being bound by δ, each segment will be warped by a different warping degree. As a result, synchronization between different DOFs of the motion sequence will be lost. To prevent this from happening, we calculate and combine the objective function for all n DOFs. Accordingly,

for length matched input and reference motion data, $\boldsymbol{\mathcal{D}}_{input}$ and $\boldsymbol{\mathcal{D}}_{reference}$, maximizing objective function:

$$J_{uniform} = \sum_{i=1}^{n} \rho\left(\boldsymbol{\mathcal{D}}_{input,i}, \boldsymbol{\mathcal{D}}_{reference,i} \mid s_k\right), \quad for\ k = 1{:}\,c \tag{4.4}$$

results in a uniform warping of all DOFs, where i represents the DOF, s denotes segment, and k represents the segment number.

Using Eq. (4.4), all DOFs maintain uniform and equal significance in the overall warping of the sequence. Incorporating a weight parameter in the objective function will result in alignment of motion sequences with more emphasis on particular DOFs. For example, warping with the aim of only aligning the arms/hands (and not the head or the feet), could be carried out using a weighted sum of ρ where the weights of joints other than arms/hands are set to zero. Accordingly, Eq. (4.4) will be updated as:

$$J = \sum_{i=1}^{n} u_i \,.\, \rho\left(\boldsymbol{\mathcal{D}}_{input,i}, \boldsymbol{\mathcal{D}}_{reference,i} \mid s_k\right), \quad for\ k = 1{:}\,c \tag{4.5}$$

where u_i is the weight associated with the ith DOF, and $\arg\max_{\boldsymbol{\eta}} J$ calculates the set of segment-wise warping degrees $\boldsymbol{\eta}$.

Ad-hoc means can be used to determine the weights, u_i. In gaits, for example, the importance of fingers and toes is significantly less than other limbs and joints. In sign language, on the other hand, the fingers are of critical importance. These considerations, among others, can be integrated into the process using the weight parameter. In the Results section, we suggest and test weights for aligning different regions of interest.

Previous studies such as [197] have investigated the relative importance of different joints in animation of human motion, which can provide suitable guidelines for tuning w_i.

4.2.3 Optimization

Given length matched input and reference motions $\boldsymbol{\mathcal{D}}_{input}$ and $\boldsymbol{\mathcal{D}}_{reference}$, and a set of assigned warping parameters (λ, δ), the optimal warped output sequence needs to be computed. Accordingly the optimal warping degree (η) for each segment needs to be calculated and utilized. In other words, all possibilities of η within $-\delta \leq \eta \leq \delta$ must be investigated for each segment. Dynamic programming is often used to solve problems that would require very large number of iterations if it were to be solved exhaustively. We use a 2-step backward-forward dynamic programming algorithm. The algorithm is derived from [190] and necessary modifications have been made based on descriptions provided earlier in this Chapter. The pseudo code of the algorithm is presented in Table 4.1. In this algorithm, c is the number of segments, m' is the new segment length successive to the initial UTW of the input which results in the length-matched version, F is a matrix which is populated using cost function values, v is the sum of objective function values, U is the lookup matrix containing the parameter values, and Y is the solution matrix. In this algorithm, all possible positions of a given segment are first inspected based on possible orientations of previous segments, and the optimum alignment is calculated. Iteratively, permutations that result in sub-optimal alignment are ruled out. As a result, the process always finds the set of warping degrees that best align the trajectory with respect to the reference.

Table 4.1. The dynamic programming optimization used in CoTW.

$F \leftarrow -\infty \times \mathbf{1}_{(c+1)\times(m'+1)}$
$F(c+1,1) \leftarrow 0$
for $i = c$ to 1
  $a \leftarrow max\big(i \times (\lambda - \delta), (c-i) \times (\lambda + \delta)\big)$
  $b \leftarrow min\big(i \times (\lambda + \delta), (c-i) \times (\lambda - \delta)\big)$
  for $j = a$ to b
    for $\eta = -\delta$ to $+\delta$
      $v \leftarrow F(i+1, j+\lambda+\eta) + J\big(\boldsymbol{\mathcal{D}}_{\text{input}}, \boldsymbol{\mathcal{D}}_{\text{reference}} | s_i\big)$
      if $v > F(i,j)$
        $F(i,j) \leftarrow v$
        $U(i,j) \leftarrow \eta$
      end
    end
  end
end
$Y(1) \leftarrow 1$
for $i = 1$ to c
  $Y(i+1) \leftarrow Y(i) + \lambda + U\big(i, Y(i)\big)$
end

4.3 Parameters and Distortion

In this section the impact of segment and slack sizes on the warped outputs are investigated. Given m' and λ, after dividing the trajectory into c segments, an extra segment may remain with the length of $m' - \lfloor m'/\lambda \rfloor \times \lambda$. This situation occurs when $\lfloor m/\lambda \rfloor \neq m/\lambda$. There are two possible approaches for dealing with this residual segment: (a) counting it as a separate segment, or (b) adding it to the last segment, making the cth segment a bit longer. Using the first approach, the length of the residual segment may become considerably smaller than other segments or the slack for that matter. In this case, warping it by larger values of η may cause significant distortion. We therefore use the second approach.

Figure 4.8 illustrates warping of a trajectory with $\lambda = 23$ and different δ sizes. The reference is a sinusoidal and the input is a sum of sinusoids. The boundaries of segments are displayed. Misalignment is seen as the peaks and valleys of reference and input occur at different time instances. It is evident that for $\delta = 1$, warping is minimal. As δ increases to $\delta = 7$, alignment improves. Beyond this value, however, there is no significant change in alignment. This is because, in this particular case, optimum warping occurs at $\eta = 7$, therefore, alignment remains unchanged for $\delta > 7$. Thus, we suggest that when manually tuning the warping parameters, it would always be safe to set δ to the maximum possible length. The method's boundary conditions do not allow $\delta > \lambda - 4$, and so, the maximum possible length of slack is $\delta = \lambda - 4$. Using the maximum δ, however, despite resulting in the optimum warping, may not always be suitable as it can cause distortion. The notion of distortion is described later in this section.

In regards to the segment length, let's initially assume that the entire trajectory is one segment, meaning $\lambda = m'$. This yields that the trajectory is not permitted to warp since the post-warp length must remain unchanged. Moreover, for $^{m'}/_{2} < \lambda \leq m'$, as discussed earlier, the residual segment will be added to its previous segment. Therefore, we conclude that only segment lengths of $\lambda \leq {}^{m'}/_{2}$ will result in practical warping. As λ decreases, the process will have more segments to warp in order to achieve greater correlation. Figure 4.9 illustrated the effect of segment size where for $\lambda = 45 > {}^{m}/_{2}$, no warping occurs. As the number of segments increases (λ decreases), alignment is improved. However, since decreasing the segment length results in constraining δ and

hence η, it does not always result in better alignment. For instance, in this example, the best alignment is achieved for $\lambda = 10$ where the fourth local maximum of the input is aligned with that of the reference.

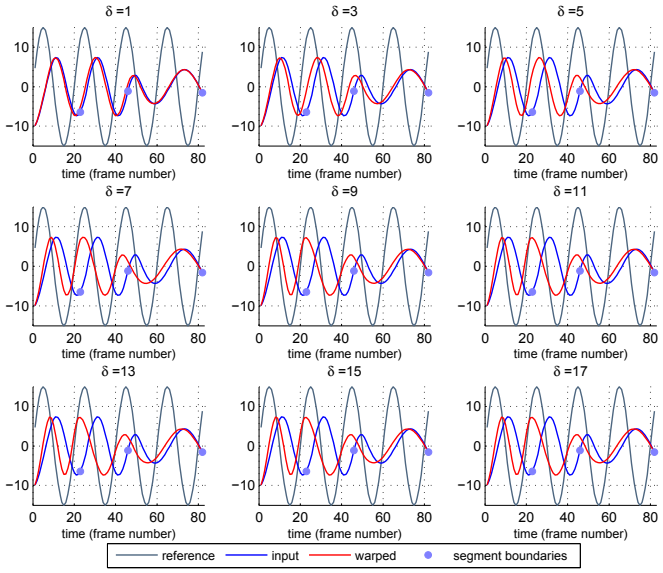


Figure 4.8. CoTW output with different values of δ.

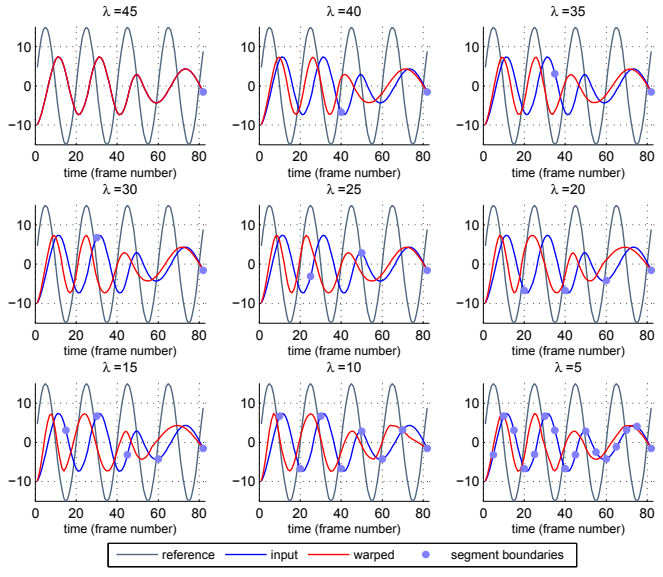


Figure 4.9. CoTW output with different values of λ.

In order to select λ and δ, we suggest an iterative exhaustive optimization problem, where for all combinations of λ and δ, the correlation of the output is calculated with respect to the reference. To further integrate the notion of joint customization, a weighted sum of DOF correlations, similar to Eq. (4.5) can be utilized. This step results in a correlation matrix $\boldsymbol{\rho}$ as illustrated in Figure 4.10. The figure depicts the correlations of an actual joint angle trajectory from a tired walk aligned with respect to an energetic walk. Here, the parameters resulting in the highest peak, $\lambda = 16$ and $\delta = 12$, achieve the best alignment.

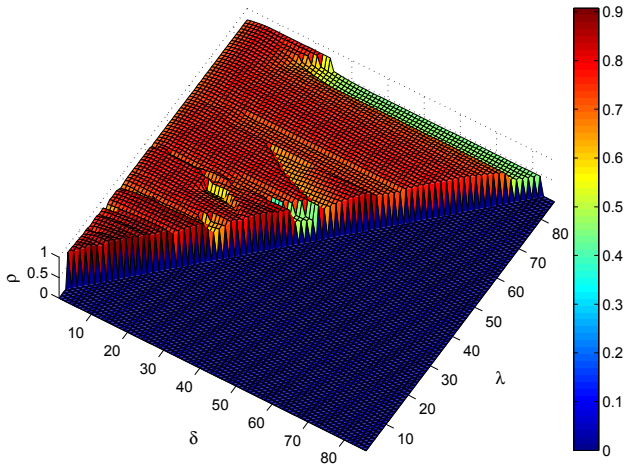


Figure 4.10. Correlation matrix for permutations of λ and δ.

Although maximum correlation can be achieved using the optimized parameters, the output motion could be distorted due to excessive warping. The algorithm by design employs two constraints with regards to the amount of warping applied to the segments: (a) the sum of warping degrees of all segments is equal to zero since the length of the motion will remained unchanged due to the initial length matching ($\sum_{i=1}^{c} \eta_i = 0$); (b) the slack parameter bounds the warping degree of each segment ($\eta \leq \delta$). Nevertheless,

should δ be relatively large, it is possible for a segment to be warped excessively. This can cause the motion to seem unnaturally slow or fast in the duration of that segment, as the slope of the trajectories would be affected significantly. An example of this is illustrated in Figure 4.11 (top) where the arrows point to two such instances. While such artifacts are not a point of concern for retrieval and classification purposes, for animation and editing applications, they need to be addressed.

To measure this distortion, the inverse of the signal to noise ratio (SNR) [198, 199] can be utilized as $\text{SNR}^{-1} = \frac{P_{noise}}{P_{signal}}$, where P represents the classical definition of power. In the context of this study, for input sequence $\boldsymbol{\mathcal{D}}_{input}$ which is length matched with the reference, this definition will translate to:

$$\text{SNR}^{-1} = \frac{\left\|\boldsymbol{\mathcal{D}}_{input,CoTW} - \boldsymbol{\mathcal{D}}_{input}\right\|}{\left\|\boldsymbol{\mathcal{D}}_{input}\right\|} \tag{4.6}$$

where $\boldsymbol{\mathcal{D}}_{input,CoTW}$ is the same sequence after CoTW warping. Note that we use the length matched version of the input sequence for computing the noise. This is to calculate the distortion caused by the non-uniform changes in the sequence rather than the changes that occur as a result of the initial length-matching. Eq. (4.6) indicates that to acquire minimum SNR^{-1}, $\boldsymbol{\mathcal{D}}_{input,CoTW}$ must approach $\boldsymbol{\mathcal{D}}_{input}$. As a result, minimizing this measurement during the warping process tends to cancel out the attempts made to carry out the non-uniform segment-wise warps. Moreover, our experiments showed that when SNR^{-1} with the described definition is minimized during the warping process, almost no warping occurs. As a result, we investigated with several variations of Eq. (4.6) to find a practical substitute for measuring distortion. For example, we used first and second

derivatives of the components in Eq. (4.6). We also experimented with using the difference of the norms instead of the norm of differences. Eventually, we concluded that Eq. (4.7) is the most suitable means of quantifying distortion and severe changes in the trajectories without taking into account the effects of uniform warping.

$$h = \left(\left\| \frac{\Delta \boldsymbol{\mathcal{D}}_{input,CoTW}}{\Delta t} \right\| - \left\| \frac{\Delta \boldsymbol{\mathcal{D}}_{input}}{\Delta t} \right\| \right) \cdot \left\| \frac{\Delta \boldsymbol{\mathcal{D}}_{input}}{\Delta t} \right\|^{-1} \tag{4.7}$$

This measurement entails that distortion is minimized when the norm of the slopes of $\boldsymbol{\mathcal{D}}_{input,CoTW}$ and $\boldsymbol{\mathcal{D}}_{in}$ converge. In other words, minimizing $\boldsymbol{h}$ prevents $\boldsymbol{\mathcal{D}}_{input,CoTW}$ from having drastic changes in slope, hence, less distortion. Similar to the $\boldsymbol{\rho}$ matrix, the distortion caused by different combinations of λ and δ can populate a matrix which we denote by $\boldsymbol{h}$. A weighted sum of the distortions of the DOFs can be used.

Equation 4.8 calculates the optimum λ and δ for warping the input where v is a weight factor for distortion:

$$\arg\max_{\lambda,\delta} \ \boldsymbol{\rho} - v \cdot |\boldsymbol{h}| \tag{4.8}$$

For applications where reducing the distortion is more central, larger values of v are used. As the tradeoff for minimization of distortion, suboptimal parameters are used, and best possible alignment may not be achieved. Nevertheless, the warped motion appears more natural. Figure 4.11 (bottom) presents CoTW warping, taking $\boldsymbol{h}$ into account. We observe that the two instances of distortion are prevented. Yet on the other hand, suboptimal alignment is achieved. For example, distortion minimization counters the alignment of the local minima at the 100th frame mark, where they were previously

aligned with $v = 0$. Figure 4.12 (a) shows the distortion matrix ($\boldsymbol{h}$) for the same previously used trajectories. Figure 4.12 (b) illustrate the correlation matrix with minimized distortion ($\boldsymbol{\rho - h}$).

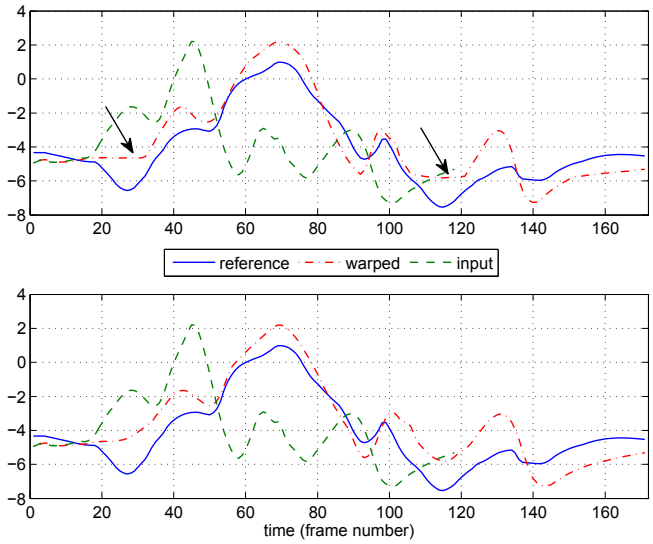
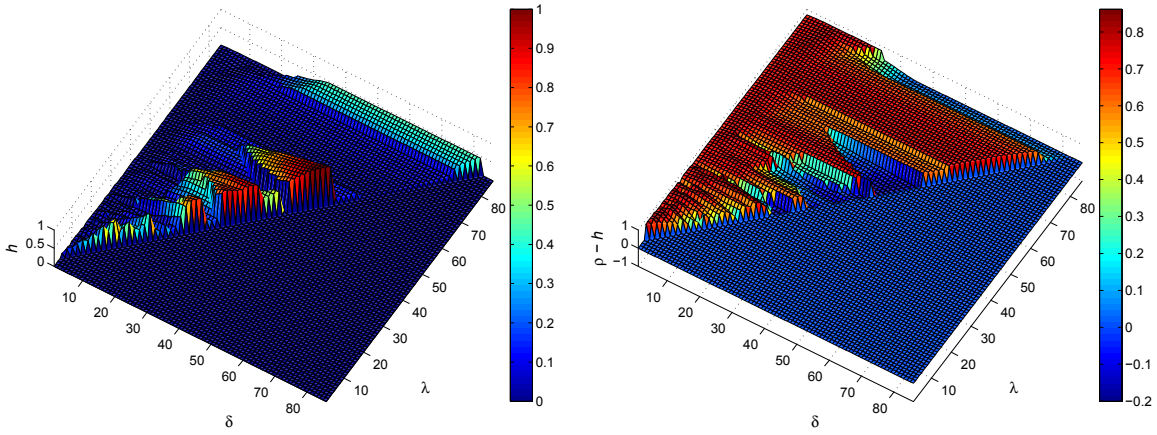


Figure 4.11. A motion trajectory warped without (top) and with (bottom) taking the distortion factor into account.



(a) (b)

Figure 4.12. Distortion (a) and subtraction of distortion from correlation (b) matrices for permutations of λ and δ.

In addition to calculation of the optimum parameters, λ and δ can be manually tuned. In many cases, the input sequence is composed of multiple actions and the goal of warping is to align each action with the corresponding action from the reference. In such cases, the number of actions in the input sequence can be a good determinant for the number of segments. For example, assuming the sequence is composed of two jumps and a kick, the number of segments can be set to 3. In other cases, the goal of warping may be to align sub-actions (sub-components of actions). In this case, the number of segments can be determined by the number of these components. For example, for a sequence consisting of a single jump, the number of segments can be set to 2. Here, segment 1 will correspond to the first half of the jump where the actor leaves the ground and reaches the maximum height of the jump. Segment 2 will then correspond to the second part of the jump during which the actor lands. This approach can significantly decrease the run-time due to leaving out the time-consuming exhaustive search. Moreover, the user can exercise this option to customize the warping process based on particular applications.

4.4 Automatic Reference Selection

When warping two sequences, the selection of the reference depends on the application. In the event that multiple sequences need to be aligned, for example when training a classifier, selection of the reference trajectory can be a difficult and influential issue. When warping multi-dimensional data, warping each DOF individually and with respect to a separate reference is one possible approach. Using DOFs of different sequences as references will cause the warped sequence to lose its synchronization among joints. In such cases, each DOF of the warped sequence will seem to be moving independent of

others. To avoid this artifact, we combine the information regarding the different DOFs of all sequences to select a single sequence as reference. In other words, we select a single sequence, and use all its DOFs as references for the respective DOFs of all other sequences.

In order to select a reference sequence, several approaches are possible. In addition to manual selection, which based on the application can be a valid approach, other techniques can be used. Skov et al. [193] mention several methods for this purpose. We utilize a similarity index. For a set of sequences, $\mathcal{F}$, where $\boldsymbol{\mathcal{D}}_{i,j}$ is the ith DOF of the jth length matched sequence, and u_i is the weight set described earlier, we calculate the similarity index through:

$$\mathcal{P} = \sum_{i=1}^{n} \left(u_i \prod_{j \in \{\mathcal{F}\}-r} \left| \rho\left(\boldsymbol{\mathcal{D}}_{i,r}, \boldsymbol{\mathcal{D}}_{i,j}\right) \right| \right). \tag{4.9}$$

Accordingly, $\arg\max_r \mathcal{P}$ determines the sequence which is most similar to all other sequences. Using this sequence will decrease the need for warping, potentially decreasing the imposed distortion. Note that we blended the similarity indexes of different DOFs using the same weight set described in previous sections to take into account the significance of each DOF.

We investigate the effectiveness of this index by comparing it with other techniques for selecting a reference. These methods include using the average trajectory, median, maximum-of-all, and minimum-of-all. Figure 4.13 presents motion trajectories from the xth axis of the right foot from 15 male walkers. To add differently shaped trajectories, we

included heavily smoothed versions of three of these trajectories as well. Smoothing was carried out using low-pass filtering. The trajectories computed as reference are shown in red for each method. It is observed that the mean, median, and similarity index select the smoothest references while minimum and maximum methods select trajectories with many local extrema. It is generally preferred that the reference be smooth, as well as being one of the *original* trajectories rather than a calculated trajectory based on others (mean or median). Figure 4.13 also illustrates all the trajectories warped with respect to the reference selected using the proposed similarity index.

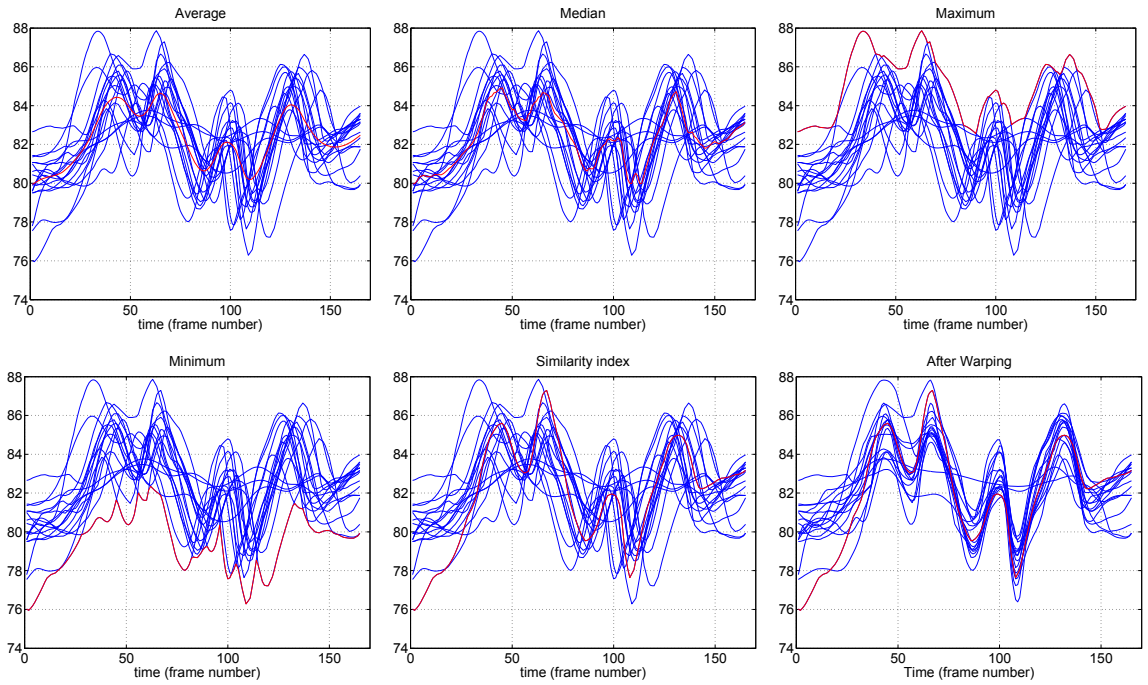


Figure 4.13. Different methods of determining a reference when multiple trajectories are available. The first five figures show the different possible methods, one of which is the proposed similarity index. In each case, the calculated reference is shown in red. The last figure presents the trajectories after warping with respect to the reference selected using the proposed similarity index.

4.5 Results and Discussions

4.5.1 Performance and Alignment

The proposed algorithm is implemented in MATLAB. To subjectively and quantitatively evaluate the results, other popular warping techniques namely UTW, DTW, and CTW are used. We use motions from the Carnegie Mellon University (CMU) dataset described in Appendix A. Different action classes are used. Complex actions such as climbing a ladder, dribbling a basketball, side twists, and kicking a ball as well as more simple actions such as jumping and walking are utilized. In most cases, the number of steps or actions in the input and reference vary, making it a relatively difficult alignment problem. Figure 4.14 illustrates how the absolute distances per frame per DOF compare for different warping methods and actions. Since for different sequences, the range of distances can significantly vary, they are normalized with respect to the maximum distance for each action class. In all cases, the maximum distance was observed for UTW, hence, all UTW normalized distances are equal to 1. Generally, DTW and CTW are not too different in terms of both distance and correlation, while CoTW shows less distance and increased correlation. The reduction in distance and increase in correlation show that overall, CoTW outperforms UTW, DTW, and CTW. To further investigate the significance of improved alignment by CoTW, we performed one-way ANOVA on both correlations and normalized distances. For distance, $F(3,20) = 7.07$, $p = 0.002$, which indicates significant effect for the warping method at $p < 0.005$. However, excluding UTW results as a naïve warping method, we obtain $F(2,15) = 0.99$, $p = 0.395$, which indicates DTW, CTW, and CoTW are not significantly different in terms of normalized distance. For correlation, $F(3,20) = 9.56$, $p = 0.0004$, which shows significant effect at

the $p < 0.0005$ level. Excluding UTW we obtain $F(2,15) = 6.46$, $p = 0.0095$, which means that CoTW has significantly improved correlation at the $p < 0.01$ level.

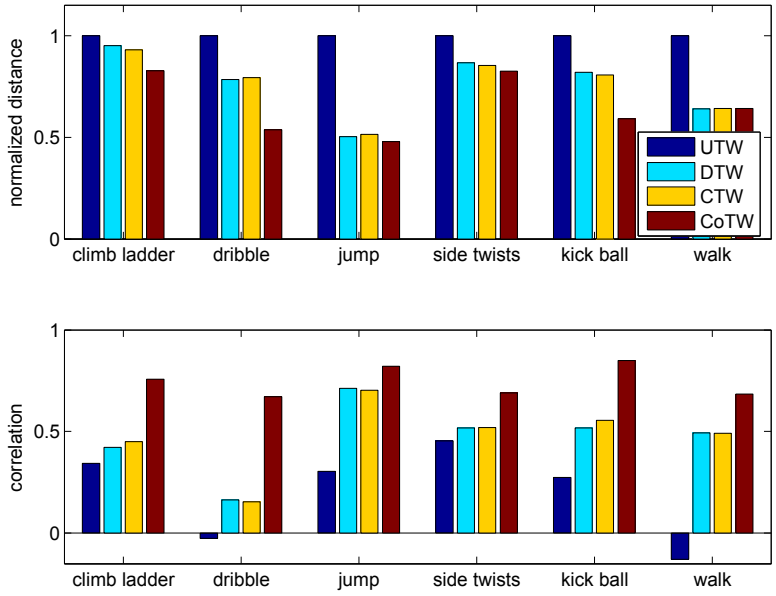


Figure 4.14. Normalized distance and correlations for different actions warped using UTW, DTW, CTW, and CoTW. In all cases, CoTW shows less distance and more correlation, indicating better alignment.

Video Clip A (https://www.youtube.com/watch?v=GamrGfQSWDM) presents the warping performance of CoTW as well as UTW, DTW, and CTW. The output videos of the warped sequences show that in general, DTW and CTW work well for aligning sequences. CoTW, however, outperforms the other methods. In Figure 4.15 (extracted from the Video Clip A), we illustrate the performance of CoTW where accurate alignment is achieved. In the climb ladder sequences, the input contains two extra steps. Interestingly, in the warped output, the character first climbs in an aligned fashion with respect to the reference, and then in order to make up for the extra steps, intensely slows down, almost to the point of stopping. The stop goes on for the duration of the excess

steps in the reference. The climb then resumes after this correction. While this may seem abnormal at first glance, it is in fact the best possible solution for such cases where the number of actions differs. The alternative solution is for the excess steps to be spread out through the entire sequence. This would result in misalignment in all sub-actions. The warped output behaves very naturally and accurate alignment with respect to the reference is achieved. The dribble sequences are composed of two different main actions, namely walking and dribbling. We observe that in the output, correction is mostly applied to the arms while the legs are only partially aligned. This is natural as uniform weights have been utilized for alignment and as a result, the maximized correlation enforces the warping process, regardless of the spatial regions. Had the weights been selected to highlight only specific regions of the body, warping would have been focused on specific joints. We test this in the following paragraphs of this section. In the side-ways twist motions, the last few frames of each twist appear to be misaligned in the output. This is not, however, actual misalignment. The two characters simply twist with different degrees. As a result, they appear to be misaligned whereas from a relative standpoint, alignment is achieved. Moreover, it should be noted that CoTW does not alter spatial variations, rather only the timing of actions. Therefore, the turning degree of the twist cannot be altered. In the walk sequences, the first and last frames of the output and reference differ, and have not been aligned. While in most examples, the input and reference motions start and end with similar poses, the walking input and reference start and end during the walk and with different poses. Referring back to the algorithm and Figure 4.8 as an example, we see that the first and last frames remain temporally unchanged after warping. Accordingly, these postures cannot be aligned through the

proposed method. The rest of the sequence, however, is aligned. Similar to the climb ladder, rapid changes in velocity, this time an increase, are observed. This is because the input contains two extra steps with respect to the references, and in order to compensate, faster strides are necessary.

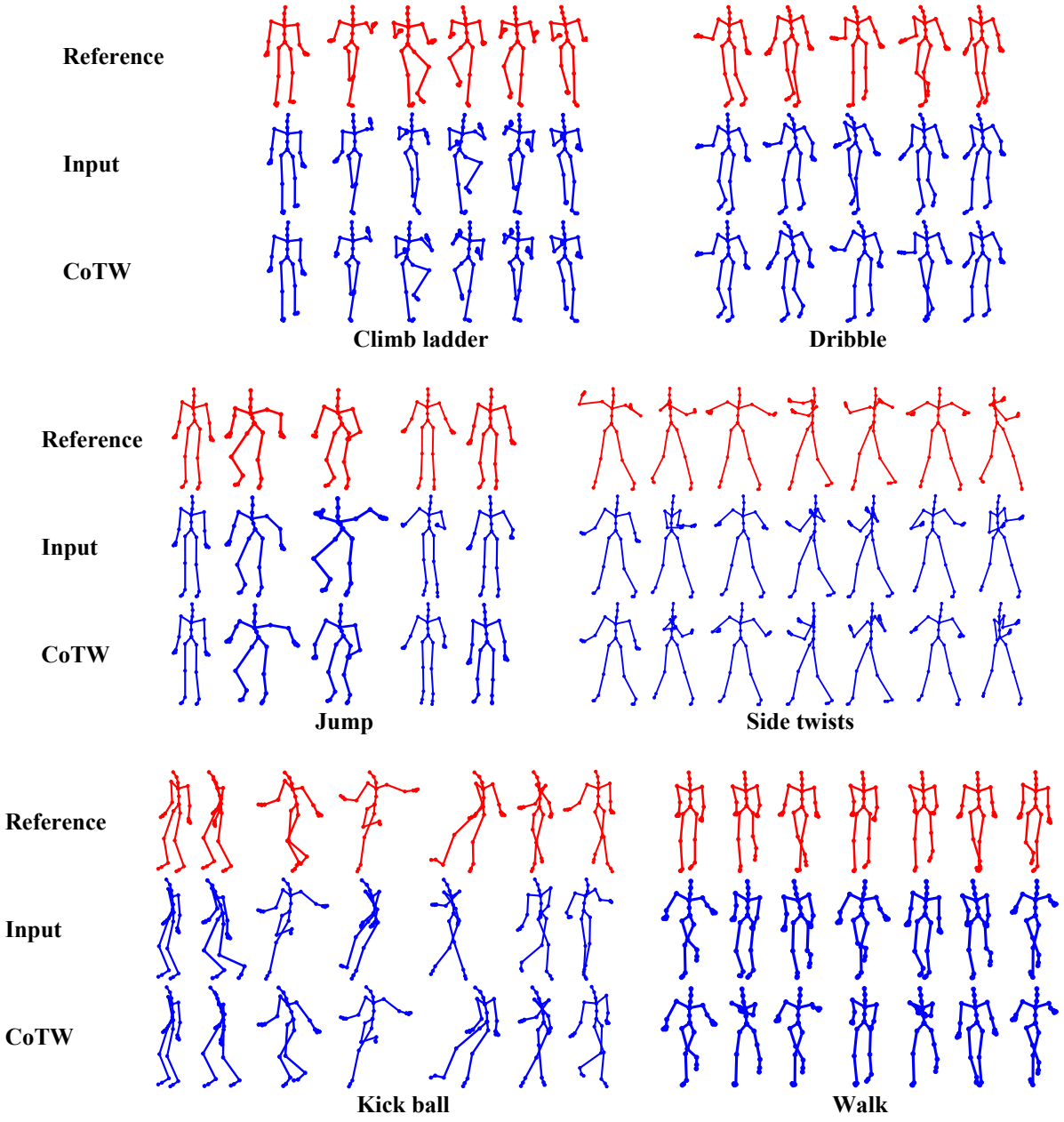


Figure 4.15. Different actions being aligned using CoTW (from Video Clip A).

While DTW and CTW are practical, fast, and generally accurate, for presented samples misalignments have occurred. Reasons for the misalignments can be the use of distance based objective functions as well as the complexity and non-equal number of actions in the sequences. These misaligned instances are clearly seen in Video Clip A. In Figure 4.16 (extracted from Video Clip A), we illustrate sample frames where DTW and CTW failed while CoTW has performed well. Moreover, it should be pointed out that in sequences such as climb ladder and walk where the number of actions significantly differs, instead of decelerating/accelerating the input to compensate for the timing differences, DTW and CTW use still-frames where particular frames are repeated. Moreover, DTW and CTW produce a warped version of the reference, with respect to which the input is aligned. Altering the reference to achieve alignment is an undesired property which CoTW does not have.

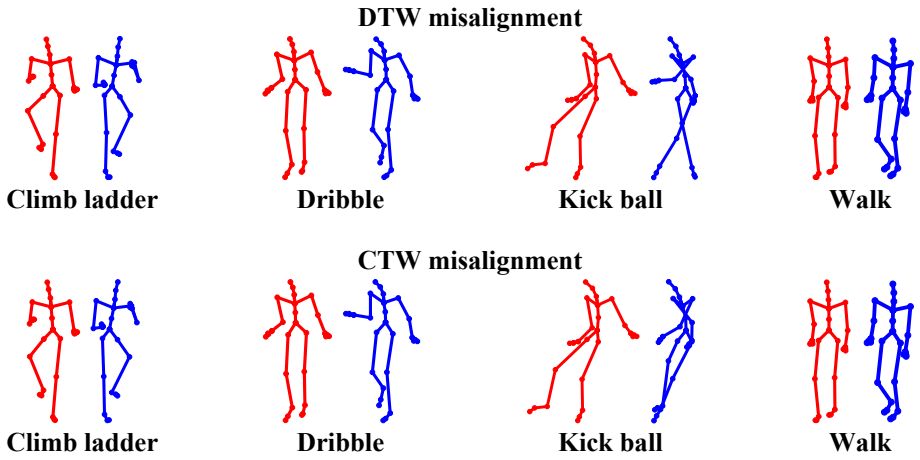


Figure 4.16. Imperfect alignment with DTW and CTW (from Video Clip A).

4.5.2 Customization

As described earlier, one of the main advantages of CoTW is its customizability. Here,

we illustrate how the weight parameter can be adjusted to efficiently and accurately warp different regions of interest in the input sequence. As an example, we use dribbling, an action that is composed of two separate regions performing different actions, namely the arms dribbling and the legs walking. Ad-hoc means are used to determine the weights presented in Table 4.2. Previous studies such as [197] have studied joints that are critical in motion animation and proposed weights for their significance, which can be utilized in our system. In Table 4.2, we used the semantic names of the joints rather than the exact names used in the CMU dataset files. For example, the name "Thigh" is used instead of "UpLeg".

Table 4.2. Suggested weights for warping focused on different regions of the body.

DOF	Full	legs/feet	arms/hands
global displacement	1.00	1.00	0.00
global orientation	1.00	1.00	1.00
hips	1.00	1.00	0.00
shoulders	1.00	0.00	1.00
thighs	1.00	1.00	0.00
spine	1.00	0.00	0.50
arms	1.00	0.00	1.00
legs	1.00	1.00	0.00
hands	1.00	0.00	1.00
feet	1.00	1.00	0.00
neck	1.00	0.00	0.00
head	1.00	0.00	0.00
fingers	0.00	0.00	0.00
toes	0.00	0.00	0.00

Video Clip A illustrates that when customizing the alignment process using CoTW, different joints can be accentuated with different significance levels. Few frames are presented in Figure 4.17. When the weights are uniformly distributed, a combination of a two-region alignment is achieved, partially aligning the walk and partially aligning the

dribble. When the regions of interest are the shoulder, arms, and hands, the process is performed with these regions being completely aligned while the legs are not. Similar observation is made when the regions of interest are the thighs, legs, and feet.

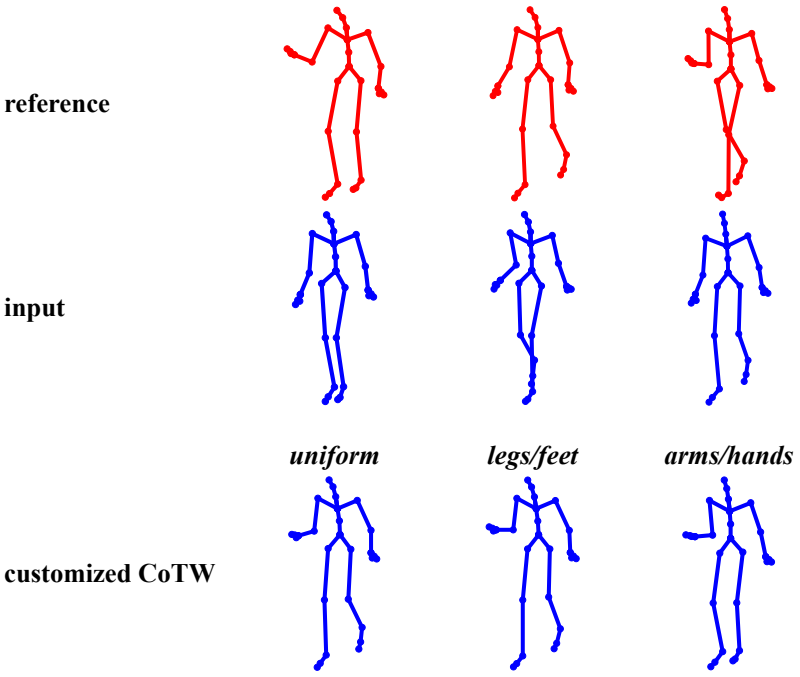


Figure 4.17. Customization of CoTW is presented. The input dribbling motion is warped with uniform weights, as well as the regions of interest being the walking or the dribbling (from Video Clip A).

Another customization capacity with CoTW is manual selection of segment and slack lengths. Previously, we illustrated that when a walking sequence is being warped, calculation and use of optimum λ and δ results in accurate alignment. Video Clip A and Figure 4.18 illustrate that when we manually assign the segment length to that of one step, the steps are aligned with acceptable precision.

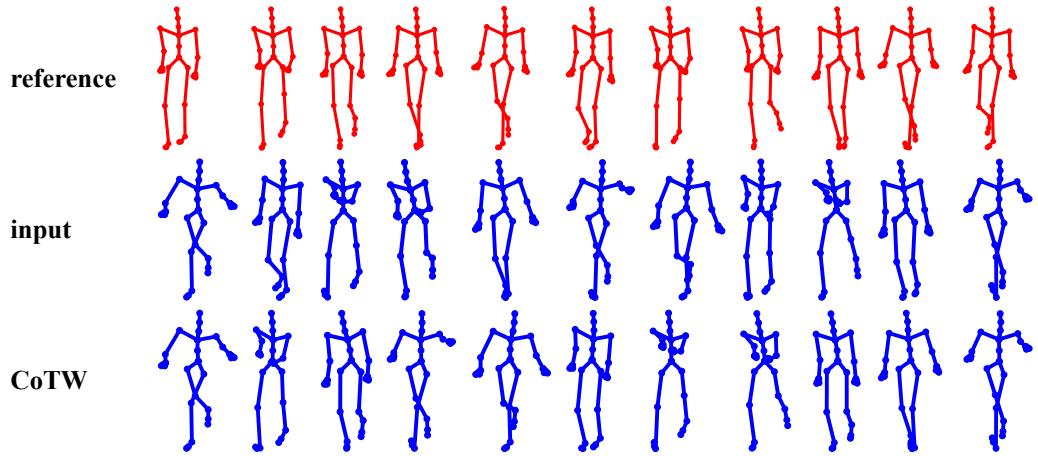


Figure 4.18. Manual tuning of λ and δ results in relatively acceptable alignment. Here λ is assigned as the approximate length of a single stride and $\delta = \lambda - 4$ (from Video Clip A).

4.5.3 Style Translation

Style translation is the process of transferring style features from one sequence of motion onto another [14]. This procedure can often be carried out using interpolation or extrapolation [23]. However, sequences contain relative misalignments which cause artifacts such as footskating [200] if style translation is carried out without time warping. Footskating is an artifact in which the legs and feet appear to be sliding and skating instead of taking solid and firm steps. As a result, style translation can be used as an evaluation factor for warping methods. We use neutral, macho, and marching style walks to perform this test. To also test the automatic reference selection configuration, two neutral walks are employed. The similarity index selects one of the two neutral walks as the reference. The other normal walk and the macho walk are warped with respect to the selected reference. By subtracting the stylistic walks from the reference and adding the resulting style features to the other neutral walk, style is transferred from one sequence onto another. Figure 4.19 illustrate frames of input (neutral) and output (macho and

marching) sequences, following alignment and style translation. The output video shows that the outputs demonstrate no or very little artifacts. A significant advantage of the CoTW method can be observed in this experiment where the output data do not require any post-processing, compared to other methods such as [14, 24] where the output needs to be cleaned and corrected.

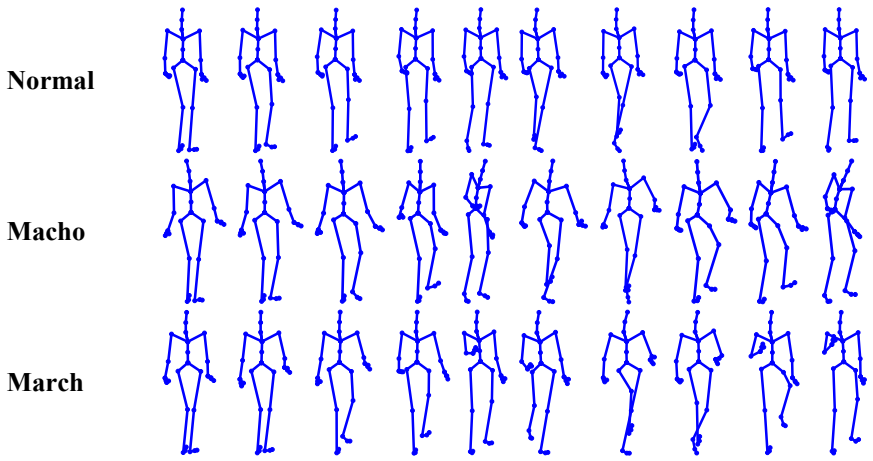


Figure 4.19. Style translation where a neutral input is converted to marching and macho walks.

4.5.4 Summary of Advantages

The proposed CoTW algorithm achieves robust alignment of sequences and outperforms UTW, DTW, and CTW especially when the number of actions in the input and reference are different. Additionally, CoTW demonstrated several advantages over existing warping techniques. While some previous works have addressed some of these issues, to the best of our knowledge, a single system that incorporates all of them does not exist. Following, the advantages of our method are summarized:

(1) CoTW uses correlation as a measure of similarity between motion trajectories, which

based on our earlier arguments, is significantly better compared to Euclidean, Manhattan, or similar distance functions which are often used.

(2) CoTW can be customized both spatially and temporally. Through the former, a weight set is applied to the character model, resulting in more alignment in specific joints or motion DOFs. The latter allows the user to select the length of segments of the sequence that will eventually be aligned. Nevertheless, optimum segment length and slack constraints can be automatically calculated and utilized.

(3) CoTW prevents over-warping of motion that causes distortion. As a result, CoTW outputs more naturally appearing warped sequences.

(4) Compared to other warping techniques which use still-frames to perform the warping, CoTW produces readily smoothed trajectories that can be used for animation purposes. Moreover, to achieve alignment, DTW and CTW warp the reference and input together whereas CoTW does not alter the reference.

(5) When multiple sequences are available, CoTW allows for automatic selection of a reference that will demand the least warping from other sequences. This process can especially be useful when aligning a dataset for different purposes such as training a classifier. Moreover, this modality too can be customized to allow more emphasis on particular motion DOFs.

4.5.5 Limitations

A computational limitation of CoTW occurs when one or more segments of a motion trajectory, whether input or reference, have zero variance. Based on Eq. (4.4) and

subsequently Eq. (4.3), the objective function cannot be calculated for zero-variance segments. This limitation does not exist in most distance-based objective functions. Another limitation of CoTW is that based on the proposed algorithm, the first and last frames remain unchanged. This entails that our method cannot align the first and last frames of motion sequences. To the best of our knowledge, this is the case with other existing methods such as DTW as well.

In terms of limitations and issues in the warped output sequences produced by CoTW, some misaligned gestures are observed in Video Clip A. These often occur when the input sequence is performed such that different DOFs of the body contain different types or temporally placements of misalignments with respect to the corresponding DOFs of the reference. As a result, all DOFs can't be perfectly aligned. While the introduced weight parameter (in Eq. 4.8) can align DOFs of interest, the tradeoff is that other DOFs may not be optimally aligned. A uniform weight vector, on the other hand, will aim at aligning all DOFs as much as computationally possible, which will result in some DOFs being sub-optimally aligned. Nevertheless, the overall instances of misaligned gestures with CoTW are rare compared to UTW, DTW, and CTW.

In addition to the above, it can be seen in Video Clip A that when using CoTW, the output motion accelerates/decelerates at certain points. This is especially visible in the climb ladder and walk sequences. While such instances seem unnatural, as the number of actions (for example, number of steps in the walk) are different between reference and input, parts of the input need to be accelerated/decelerated to compensate for the timing difference. Although the distortion minimization modality can reduce these changes in speed, alignment will not be as perfect. As discussed, other methods such as the DTW or

CTW use still-frames, whereas our method produces smooth but warped motion. Nevertheless, the solutions provided by both DTW/CTW and CoTW can still be seen as liabilities based on the application.

Runtime is another limitation of the proposed method. Since for a given segment and slack, different values of slack are evaluated, the runtime can grow rapidly, especially for large slack values. Even though dynamic programming prevents extremely long runtime, UTW, DTW, and CTW are all much faster than CoTW. Another reason for increased runtime of our method is that measuring correlation is computationally more demanding compared to measuring distance. Moreover, the exhaustive search for the optimum parameters will definitely further increase the runtime. For the six sequences presented earlier, Table 3 presents the average and standard deviations of runtimes for $\lambda = 40$ and $\delta = 20$. The results show that CoTW is the slowest among the tested approaches. The lengths of the 6 sequences were 310, 166, 155, 328, 228, and 102 frames.

Table 4.3. Average and standard deviations of runtimes for different warping techniques. The six sequences presented earlier were used.

warping	UTW	DTW	CTW	CoTW
time (s)	0.02±0.006	0.08±0.04	1.25±0.59	10.01±7.61

Faster computing algorithms and more efficient programming, as well as the parallel computing library in MATLAB can be utilized to reduce runtime. Lower level programming such as MEX files in association with MATLAB, or complete C/C++ implementations can significantly increase the speed. Finally, GPU implementations are known to speed up exhaustive search problems by more than 10 times [201]. This

approach can provide a practical solution for the run-time issue of CoTW, which can be explored in the near future. Nonetheless, while the other warping techniques (UTW, DTW, and CTW) will also perform faster on the GPU or using parallel computing, MATLAB in general performs slower for implementations with high computational complexity. As a result, we speculate that the impact of these measures on CoTW will be considerably more. Finally, in addition to runtime improvement, the notion of non-equal segment lengths, which can lead to even better alignment, can be investigated in the future.

Chapter 5.

Perception of Affect from the Motion of Single and Multiple Limbs

5.1 Introduction

In Chapter 2 we reviewed the history and state of the art in studies on human motion perception. In summary, it has been shown that a vast amount of detailed information can be perceived from human motion. For example, a variety of different styles and variations including identity [68], gender [71], emotions [75], and even weight of lifted objects [202], can be accurately recognized from simplistic representations of body such as point light display.

Most perception studies on affective and stylistic motion are based on full body motion. However, local limb motion is known to convey information regarding affect as well. For example, Pollick et al. [77] demonstrated the perception of different emotions from arm movement. In [203], we showed that different body parts convey sufficient perceivable features for recognition of walker attributes.

It is reported that mere kinematics of motion does not determine the perception of information from biological motion [204]. Instead, internal mental models are known to exist with which the viewer identifies human motion. These models might in fact be what Troje and Westhoff [57] refer to as evolved "life detector" processes that aim at detection of animals. Studies have shown that infants respond to biological motion more than non-biological [54, 55, 56]. These findings confirm the existence of internal models for recognizing and interpreting biological motion. Moreover, other properties such as ceiling and floor effects [205], as well as familiarity with particular motion cues can influence perception of motion.

In many motion processing techniques, especially procedural methods, such as [23, 90, 206], information is extracted from or modifications are made to the body, one limb or even one joint at a time. Moreover, in many instances, this approach does not extend to include the entire body, leaving the process spatially local. This is often done to increase efficiency, as in many cases, processing the entire body is not required to achieve the intended outcome from a perceptual perspective. In this dissertation, the methods developed in Chapters 6, 7, and 8, are all techniques that can be carried out on a limited number of joints. Especially, the method presented in Chapter 7 is concerned with perceptual models for generating stylistic features in motion. The generated features in

that chapter are applied to only a select number of joints to achieve the notion of perceptual models. In this chapter, we pose and investigate the following question: are the impacts of two limbs in perception of affect from motion equal to the sum of the impacts of the two limbs? In other words, are the influences of local limb motions linearly additive when perceiving affect or are there additional factors involved (see the model presented in Figure 5.1)? With respect to the research presented in this dissertation, the study in this chapter will answer whether or not incremental processing of motion is a perceptually accurate approach. Through collecting perception ratings for neutral sequences with only single affective limbs and comparing the sum of the ratings with sequences that contain multiple affective limbs, we investigate this notion. Our study shows that while additively does not hold in the classical sense, the results are highly correlated, and thus, incremental processing of motion leads to perceptually accurate outcome.

The contents of this chapter have been published as [203, 207, 208].

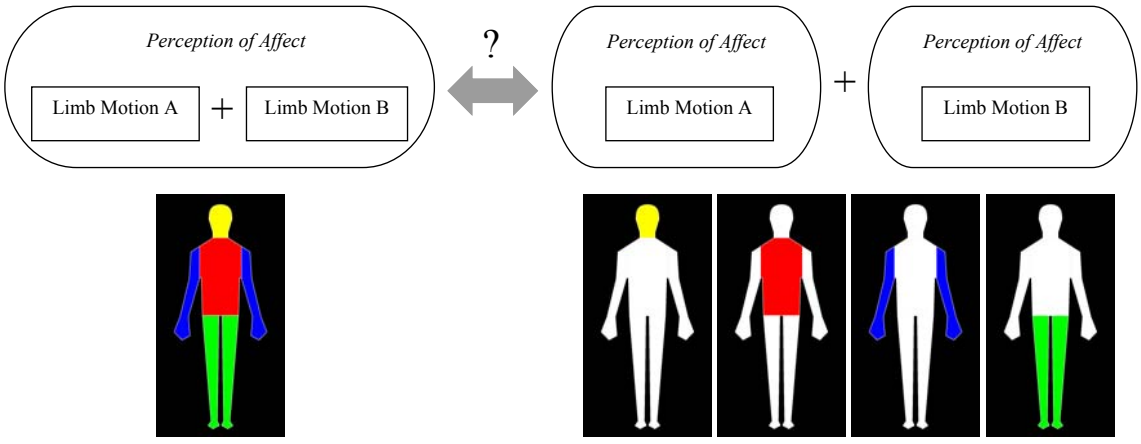


Figure 5.1. The question posed and addressed in this chapter: are the influences of limbs in perception of affect linearly additive?

5.2 Experiment Setup and Method

5.2.1 Process

To study the amount of conveyed affect from specific limbs, we define four general limbs for the human body: arms and hands, legs and feet, head and neck, and torso. These limbs are illustrated in Figure 5.1. To eliminate the impact of individual walker variations, we average multiple walk sequences. This process, which is described in the following sections, yields a single neutral walk, a single happy walk, and a single sad walk. Consecutively, the motion of each defined limb of the neutral walk is substituted with the corresponding limb motion of the happy walk, which results in 4 sequences that contain only single-limbs with affective features. Subjective ratings regarding the amount of affect are collected. The limb substitution process is repeated, this time for combinations of limbs (two-limb and three-limb permutations). Subjective ratings are collected again. By calculating the sum of ratings for single-limb substitutions and comparing them with multi-limb substitution ratings, additivity as defined earlier is investigated. The same process is repeated for the sad walk. Figure 5.2 presents the experiment process. In the following sections, details regarding the sequence averaging process, stimuli, participants, setup, and results are provided.

5.2.2 Average Walk Cycles

We investigate the posed question for happy and sad categories of emotion, which are at the two ends of the pleasantness vector of Russel's model of affect [53] (presented in Chapter 2). As a result, the results of this study are highly general. While we preferred to carry out the investigation on more classes of style and affect, such as feminine,

masculine, energetic, tired, and etc., to the best of our knowledge, the data required to expand the study was not readily available at the time that the study was being conducted. The data need to be in the form of stylistic/affective walk cycles, consistent with regards to the joint structures, carried out by several actors, and multiple times. These specifications are requires since the data, as mentioned, need to be temporally aligned and averaged. The HDM05 dataset described in Appendix A contains the required specifications. The dataset contains sequences of 5 actors repeatedly walking with neutral, happy, and sad styles. Nevertheless, we have no reason to believe that the conclusions drawn from this chapter are only limited to happy and sad affects and not the other mentioned motion styles.

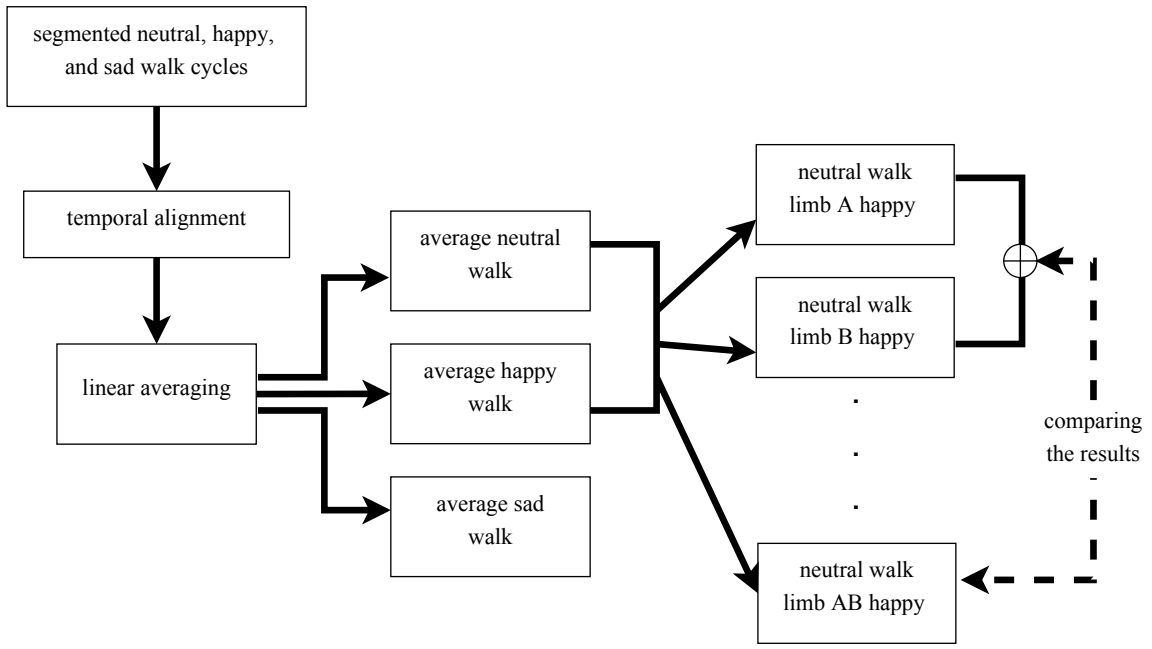


Figure 5.2. The process for the study conducted in this chapter.

The model used for the HDM05 data is composed of 96 degrees of freedom (DOFs). Some of the DOFs belong to minor joints such as fingers and toes, which are not used in

PL stimuli and most such experiments. We therefore modify the model and remove these extra joints from the model, reducing it to 54 DOFs. Moreover, by subtracting the global displacement vector from the sequence, an in-place treadmill-like sequence is achieved, similar to the sequences created and utilized by Troje [73].

By manually segmenting the sequences, 16 two-step walks in each affective category (neutral, happy, and sad) are achieved. When segmenting the sequences, we ensure that cycles start and end with similar postures and each cycle contains two complete strides. This is in accordance with the findings of [64] where it was shown that 1.6 to 2.7 seconds is required to achieve correct PL ratings. Moreover, since the beginning and concluding postures are almost identical, repeated display of the cycles are viewed as one continuous sequence. This property is displayed in Figure 5.3 where trajectories from neutral, happy, and sad walks construct a smooth closed loop.

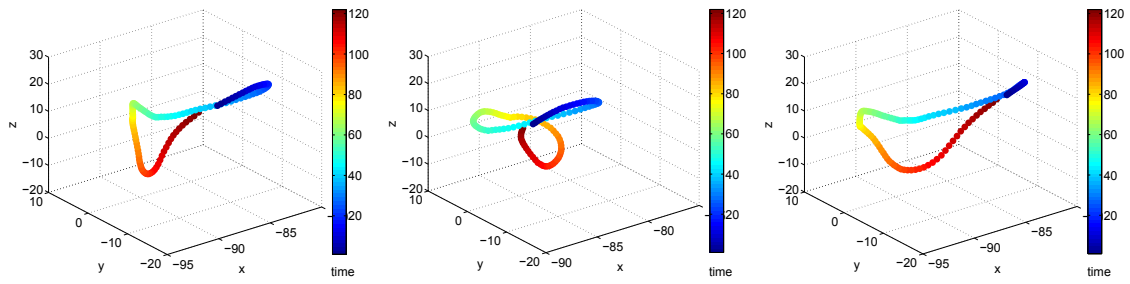


Figure 5.3. Trajectories from neutral, happy, and sad walks. Smoothly closed loops indicate proper segmentation of sequences for continuous replay of sequences.

The 16 walk cycles in each category are temporally aligned using correlation optimized time warping (CoTW). The details regarding this time-warping method were presented in Chapter 4. As illustrated, CoTW maintains low distortion, making it superior to alternative methods commonly used for motion data. Moreover, through producing

consistently smooth warped trajectories, perceptually sound output is achieved. Finally, CoTW enables for automatic selection of an optimum reference when multiple sequences are available, allowing for the least possible amount of warping to achieve temporally aligned sequences. By averaging the 16 cycles in each affective category, we achieve average neutral, happy, and sad walk cycles. Video Clip B (https://www.youtube.com/watch?v=Hxa-FQVy1k4) and Figure 5.4 illustrate each averaged sequence where high quality sequences, on par with the video of the average walker created in [73], are achieved.

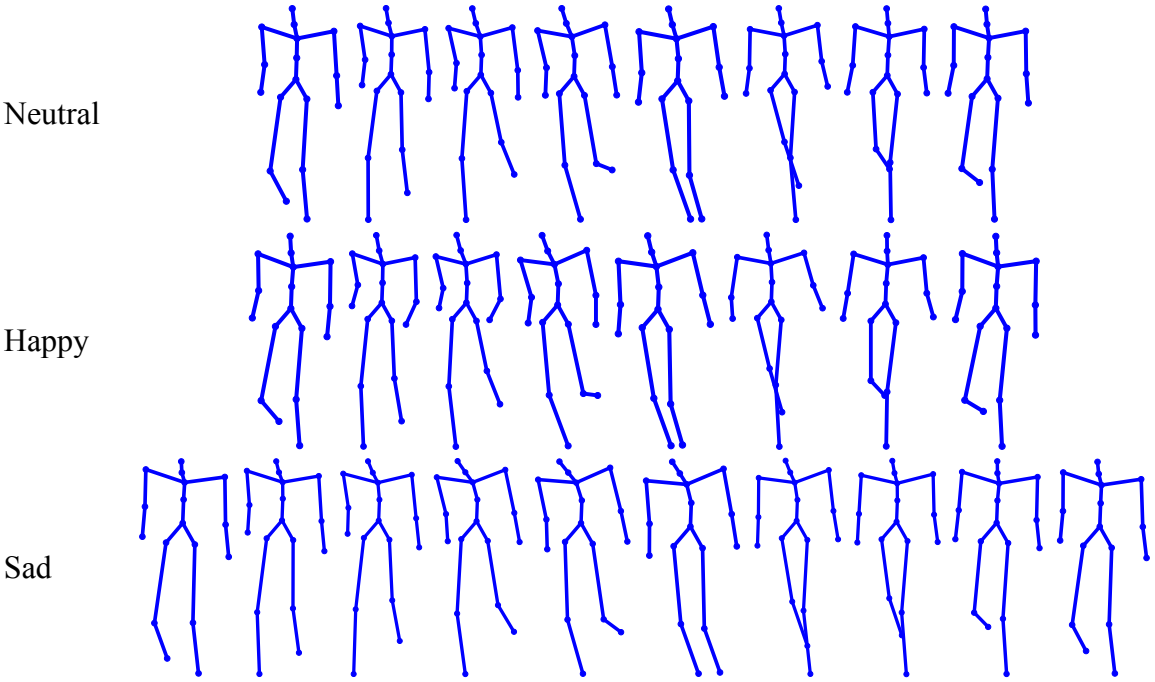


Figure 5.4. Frames from neutral, happy, and sad average walks (from Video Clip B).

The notion of average walk cycles has been previously used in the literature [73, 209]. While it can be argued that averaging the walk cycles may eliminate personal characteristics or expressiveness of the actions, the affective features that are being studied, exist in all of the sequences, and thus will not be removed. Furthermore, utilizing

non-averaged sequences might cause personalized traits in the sequences to point the audience towards particular classes of affect. In other words the happy, sad, and neutral sequences could contain additional features that can skew the perception results. Furthermore, the averaged cycles are later validated by the participants (described in Section 5.3). High ratings confirm the quality of the achieved sequences.

5.2.3 Stimuli

As mentioned in Section 5.2.1, motion of single limbs and combinations of limbs from the neutral average walk are substituted with corresponding limb motion of the affective average walks. These single and multiple limbs are presented in Table 5.1. Hence, a total of 16 sequences in each affective class (happy and sad) are created (14 presented in the table along with the initial neutral and fully affective sequence). These sequences form the stimuli for which participants provide ratings on the amount of perceived affect.

The obstacle that we stumbled upon when attempting to substitute limbs from one cycle with those of another, was the different temporal length of the sequences. To overcome this, the affective walks were speed-matched with the neutral walk using simple stretching/compressing of the sequences. This was carried out via standard interpolation techniques through the CoTW process (uniform time warping). While altering the speeds of affective sequences can have an impact on perception of affect, we argue that it does not influence the findings of this study. Generally, it has been demonstrated that human subjects successfully perceive emotions from affective sequences that have been speed-matched with neutral ones [79]. Moreover, as we describe in Section 5.3, the ratings are normalized such that the speed-matched fully affective sequences are associated with the maximum amount of emotion. This is similar to having an affective sequence with a

lower amount of emotion embedded within it to begin with. Moreover, the speed-matching process was carried out for the affective sequences as a whole, meaning the motion of all limbs are altered alike.

Table 5.1. Different limbs and combinations of limbs studied.

1 limb	*2-limb combinations*	*3-limb combinations*
head/neck (H) torso (T) arms/hands (A) legs/feet (L)	head/neck + torso (HT) head + arms/hands (HA) head/neck + legs/feet (HL) torso + arms/hands (TA) torso + legs/feet (TL) arms/hands + legs/feet (AL)	head/neck + torso + arms/hands (HTA) head/neck + torso + legs/feet (HTL) torso + arms/hands + legs/feet (TAL) head/neck + arms/hands + legs/feet (HAL)

5.2.4 Participants

25 individuals participated in this study. They were aged between 14 and 56 with a mean of 27.8 and standard deviation of 10, and were selected from both male and female groups. 10 were females and 15 were males. They were inexperienced towards human motion studies. No compensation was provided to the participants. Ethics approval was secured.

5.2.5 Setup and Tools

Like many other motion perception studies [208, 210], a stick-figure plus point-light model was used to illustrate the sequences for the participants. 6 points represented the arms and hands (3 for left and 3 for right), 6 points represented the legs and feet (3 for left and 3 for right), 3 represented the torso, and 2 represented the head and neck (1 each). The sequences were displayed on a 23.6 inch, 1080 HD, LED screen for all participants. Sequences were displayed from the frontal view with a 15º up-right incline. Viewing distance was assigned to the comfort of each participant where 40-80 centimeters from

the screen was mostly chosen.

Initially, participants were asked to determine the perceived affect for each of the three average walks. This was done in paper-based forced-choice format for validation of the three average walks. Forced-choice refers to the question format in which participants choose one or more of the multiple choices provided. Examples are given in Appendix B (for example in Sections B.2). The 32 sequences described in Section 5.2.3 were then displayed and participants were asked to rate the amount of happiness and sadness on a paper-based 7-point Likert questionnaire. Correcting the answers was permitted. Each walk cycle automatically repeated for as long as the participant needed to comfortably respond to the question. In most cases, 5-10 repeats was sufficient. To avoid any form of arrangement side effect, the order in which the sequences in each affective class were displayed was randomized. The order in which happy and sad categories were displayed was also randomized.

5.3 Results and Discussion

We first validated the three average and speed-matched walks which form the basis of this study. Table 5.2 presents the successful perception rates for the three sequences, indicating that the calculated average affective and neutral walk cycles are perceptually sound. The table illustrates that out of the 25 participants, only 2 perceived the neutral sequence as sad, and 1 perceived the happy sequence as neutral. Similar effects have been observed in other studies where sadness is found to be easier to distinguish compared to neutral and happy [142, 211]. We excluded the responses provided by these three

participants since they did not perceive the emotions as intended, indicating that any response regarding perception of affect from limb motion would most likely be inaccurate. In other words, existence of perceivable affective features in the entire body is a mandatory pre-requisite for this research, which is clearly not the case for the 3 participants in question. Thus, using the excluded responses would have been naturally incorrect.

Table 5.2. Validation of the three average walk cycles.

		perceived		
		Happy	Neutral	Sad
sequence	Happy	96.0%	4.0%	0.0%
	Neutral	0.0%	92.0%	8.0%
	Sad	0.0%	0.0%	100.0%

Participants rated each of the 14 generated sequences presented in Table 5.1 for happy and sad emotions, as well as the neutral and fully affective walks. For each participant, rating R is normalized to calculate $\bar{R}$ using:

$$\bar{R} = \frac{R - R_{min}}{R_{max} - R_{min}} \quad (5.1),$$

which maps the ratings between 0 and 1. For each participant, R_{max} and R_{min} are the maximum and minimum ratings provided by that participant in each affect class. In every instance, ratings for the neutral walks were mapped to 0 and ratings for the fully happy/sad walks were mapped to 1. Figure 5.5 presents the average normalized ratings and standard errors ($SE = SD/\sqrt{sample\ size}$). One-way analysis of variances, ANOVA, indicates a significant effect for limb motion in perception of happiness ($F(13,294)$ =

20.07, $p < 0.0001$) as well as sadness ($F(13,294) = 15.91$, $p < 0.0001$). Moreover, comparing the average of single limbs, 2-limb combinations, and 3-limb combinations indicates significant effect for the number of limbs for both happiness ($F(2,63) = 63.84$, $p < 0.0001$) and sadness ($F(2,63) = 23.37$, $p < 0.0001$). Finally, two-way ANOVA concludes that there is significant interaction between the class of emotion and $\bar{R}$ for different limbs and combinations of limbs ($F(13,588) = 11.96$, $p < 0.0001$).

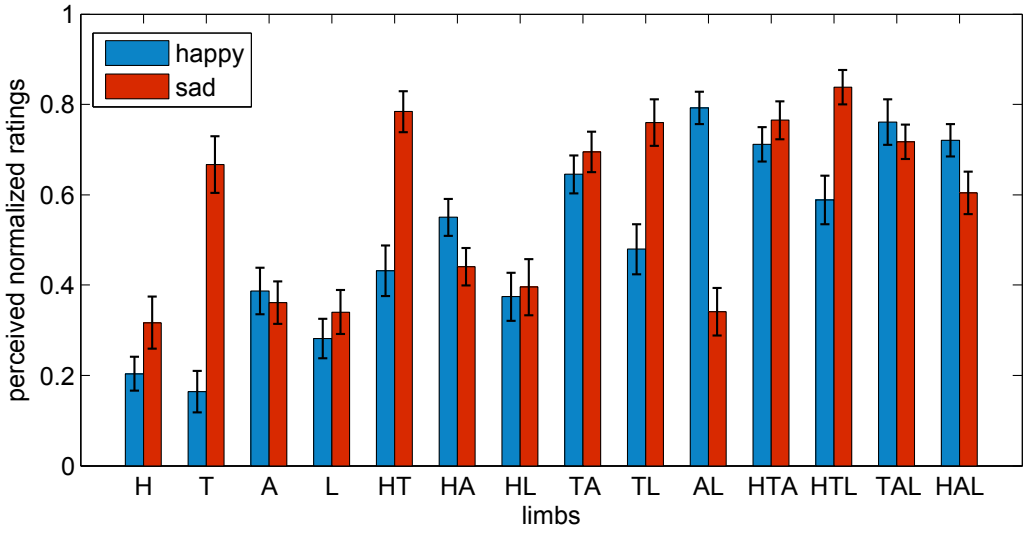


Figure 5.5. Average normalized ratings and standard errors for perception of happy and sad emotions from the created stimuli. Error-bars represent standard errors.

To observe the overall trend for ratings of sequences where single limbs, two-limb combinations, and three-limb combinations contain affective features, we average the ratings for these ratings. Figure 5.6 illustrates the outcome where the average ratings increase almost linearly with an increase in the number of affective limbs. It is observed, however, that a linear relationship approximates the average ratings for sad better than happy. However, two and three-region combinations are very similar for the two affect classes. This indicates that in general, it is more difficult for happiness to be perceived

from single limb motion while for sadness, the amount of perceived affect is proportional to the number of limbs that exhibit affective motion.

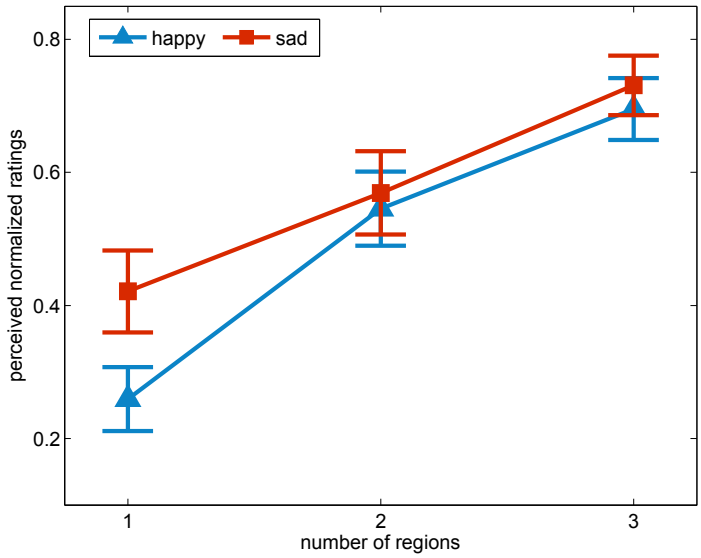
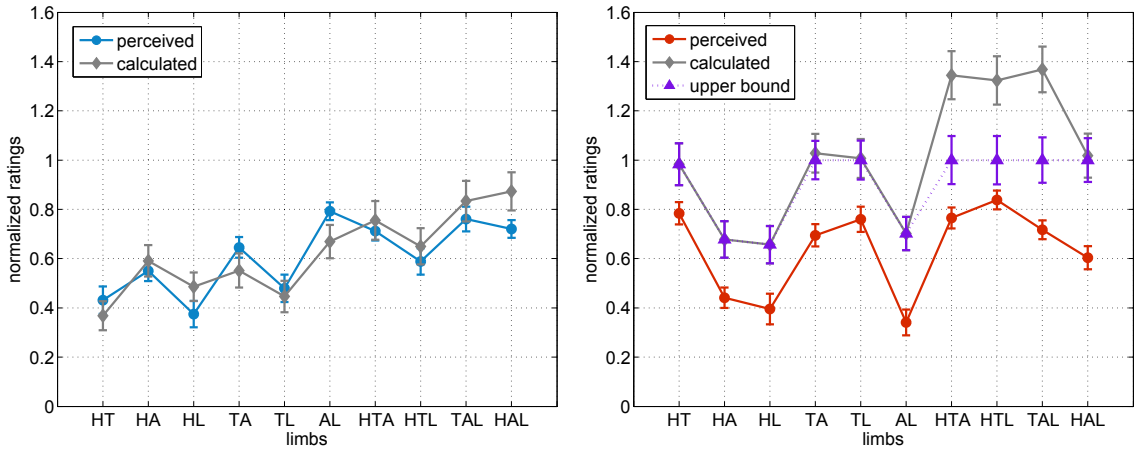


Figure 5.6. Average ratings when different number of regions contain affective features. Error-bars represent standard errors.

To investigate the concept of additivity, we use the single-limb $\bar{R}$ values from Figure 5.5 to calculate the combined limb ratings (HT, HA, HL, TA, TL, AL, HTA, HTL, TAL, HAL). Sums of average values are calculated using linear summation $\mathrm{M}_1 + \mathrm{M}_2$ while standard errors are added using Pythagorean summation $\mathrm{SE}_1 \oplus \mathrm{SE}_2 = \sqrt{{\mathrm{SE}_1}^2 + {\mathrm{SE}_2}^2}$. Similar operations are used for three-limb calculations. The results are presented in Figure 5.7 (a) and (b). For happiness, the calculated and perceived multi-limb ratings are very similar. This indicates that with an acceptable accuracy, additivity holds for the impact of limb motion on perception of happiness. In sadness, however, an offset appears and calculated ratings become greater than those perceived. Since the normalized ratings cannot be greater than 1, we manually assign an upper bound of 1 for the calculated

results, also shown in Figure 5.7 (b).



(a) (b)

Figure 5.7. Perceived and calculated normalized ratings for (a) happiness and (b) sadness. Error-bars represent standard errors.

We calculate Pearson's correlation coefficient (ρ) as a determinant of similarity between perceived and calculated results. For happiness, ρ = 0.84, for sadness ρ = 0.85, and for sadness with upper bound, ρ = 0.92, indicating that for both classes of emotion, the trends of calculated multi-limb ratings and perceived multi-limb ratings are very similar.

To better analyze the relationship between calculated multi-limb ratings and perceived multi-limb ratings, we plot the curves against each other. The results are illustrated in Figure 5.8. Linear regression is subsequently used to model the normalized ratings for further analysis. For happy, $0.96x + 0.04$ best models the relationship between the two measurements, indicating that the two sets of normalized ratings are highly related in a linear fashion. The same is observed for sadness where $1.29x + 0.19$ best models the relationship. The model is corrected as $0.79x + 0.40$ for sadness with an upper bound of 1. This indicates that while the linear model is generally accurate, the precision is lower

than that of happiness.

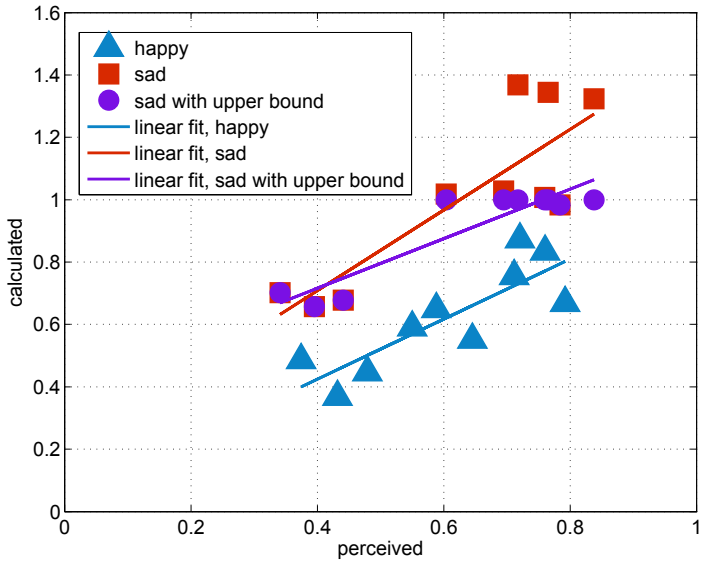


Figure 5.8. Perceived vs. calculated amount of affect from multi-limb motion. For sadness, an upper bound of 1 is taken into account for values exceeding the bound. Linear regression is used to model the observed relationships.

We calculate the offset for both affect classes by averaging the difference between the calculated and perceived values. Accordingly, we propose the following model for perception of affect from limb motion:

$$\bar{R}\left(\sum_i limb_i \,|affect\right) = max\left(\sum_i \bar{R}(limb_i|affect)\,, 1\right) - C \qquad (5.2),$$

where C denotes the offset. For happiness, C = 0.0168 with a standard error of 0.0288, for sadness C = 0.3769 with a standard error of 0.0486, and for sadness with upper bound, C = 0.2680 with a standard error of 0.0225. The results are presented in Figure 5.9. Insignificant error values indicate the accuracy of the proposed model and validate the use of an average offset.

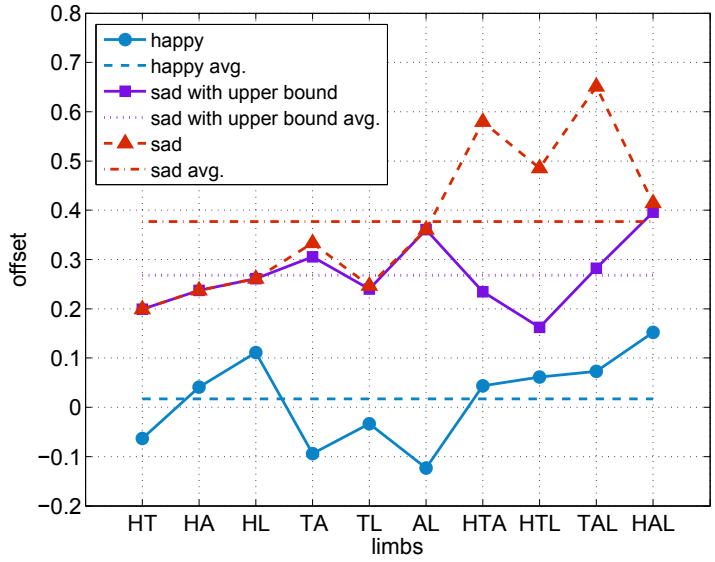


Figure 5.9. Difference values and average offsets for calculated and perceived ratings of happiness and sadness.

Motion dynamics have been shown to influence perception of emotions from motion. For example, it has been shown that velocity, phase, cadence, energy, and range of motion influence how affective motion is perceived [77, 82, 212]. However, if only kinematics were the determining factors in perception of affect, the offset in the proposed model would most likely not exist. While internal models that are particularly evolved for recognition of motion and social cues may be a cause, other reasons can induce the observed non-linear additive property. For example, it has been previously demonstrated that sad motion is easier to recognize compared to happy or neutral [142, 211]. Also, familiarity of subjects with single or combinations of specific limb movements in affective motion can cause super-additivity. Sub-additivity, on the other hand, may arise due to ceiling effects in motion perception. Nevertheless, the small deviation of each multi-limb rating's offset with respect to the average offset is an interesting observation. We believe further investigation is required to determine the exact reason for the offset

and its uniform nature.

In this chapter, we compared the sum of perceived affect from particular spatial regions with the amount of perceived affect from multiple regions of the body. It was observed that while the two were not equal, they were highly correlated, mostly with a linear relationship. As a result, additivity is a valid assumption for affective and stylistic motion features. Due to this property, procedures that extract information from, or synthesize/alter features in only a spatial subset of the body, can in fact lead to perceptually accurate results. Such approaches can be used to carry out the designated tasks in more computationally efficient manners. This property validates the procedure proposed in Chapter 7 and entails that the methods developed in Chapters 6 and 8 can be applied to only select body joints.

Chapter 6.

Extracting Movement, Posture, and Temporal Secondary Features

6.1 Introduction

In this chapter, we propose a framework for extracting spatiotemporal style features from motion. Earlier in Chapter 2, we reviewed the multitude of approaches for extracting and synthesizing motion features. Many of the proposed techniques such as learning-based methods highly depend on datasets of stylistic motion [24, 25]. Knowledge-based methods utilizing the observed differences in stylistic motion have also attempted to formulate the features responsible for conveying actor attributes and styles [90, 167].

Methods based on decomposition of motion trajectories have also been widely explored [21, 171, 172]. While the previously developed methods have proven practical for style translation and motion decomposition, they mostly explore latent subspaces, which make extracted features difficult to interpret and analyze. Extractions of features in spatiotemporal domains, on the other hand, has not been widely explored. Moreover, features responsible for conveying style and affect in motion are composed of three different components, namely: movement (also referred to as dynamics), posture (also referred to as structure), and timing (also referred to as speed or uniform temporal features) [19, 69, 73, 79, 80, 90]. Movement features are the changes to the motion trajectories and vary throughout the sequence. Posture features are those that often stay unchanged through the sequence. Rather, they are changes to the initial pose of the body with which the motion is carried out. The uniform temporal feature relates to the speed with which the sequence is performed. This feature is often correlated with the secondary theme in the sequence, meaning different styles are displayed with different speeds [74]. To the best of our knowledge, a systematic way of extracting all three components has not been developed. In this chapter, we propose a system that extracts all three components from input motion. These features can then be transferred onto other motion sequences for style translation, and can also be used by researchers in both psychology and multimedia fields for the study of stylistic human movement.

First, we describe the general relationship of actions and style features in motion. This model lays the foundation of the proposed feature extraction technique and subsequent style translations. The extraction system is then proposed. The system utilizes a dataset of neutral motion sequences and through maximizing a similarity index, selects a neutral

sequence that best corresponds to the inputs stylistic motion. The correlation-optimized time warping (CoTW) method proposed in Chapter 4 is then used in the process as the sequences need to be accurately aligned. A set of temporal cues are then distributed through the sequence while maximizing a correlation-based objective function. Spatiotemporal cubic splines are used to approximate the neutral component of the input. Movement and posture feature components are subsequently computed and extracted. The temporal feature is calculated as the third component of the feature set. The proposed algorithm is tested on various examples and the results are provided. As an application of the algorithm, style translation is carried out, through which, the feature set of stylistic inputs are transferred onto neutral ones. The overall schematic of the system is illustrated in Figure 6.1.

The contents of this chapter have been published as [14, 213, 214, 215].

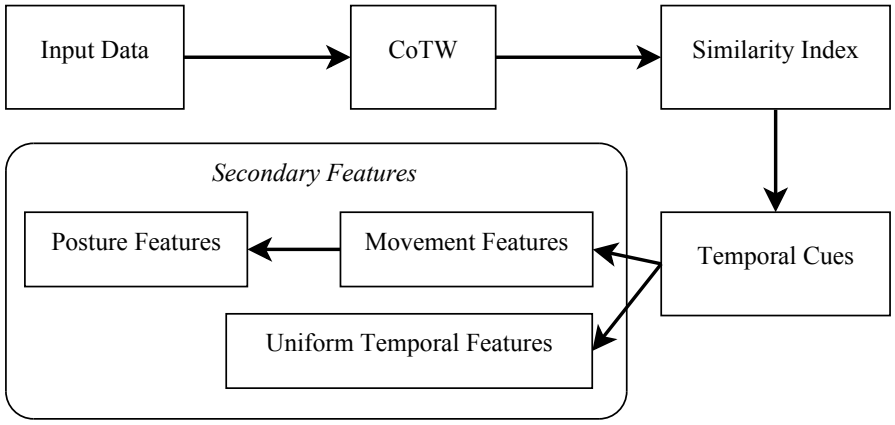


Figure 6.1. Overall process of the proposed system for extraction of style features. CoTW is carried out and similarity index is maximized followed by optimization of temporal cues and extraction of the three components of secondary features.

6.2 A Model for Action and Style

In order to develop a technique for extraction of style features, the relation between action and style is essential. Thus, a detailed look into this concept and a comprehensive model is required. As mentioned in Chapter 1, inspired by the works of Laban [216], we refer to the main action class in motion as primary theme (PT) and the affects, styles, or attributes associated with the actor performing the actions are referred to as secondary themes (ST). Accordingly, motion features that generate PT and ST are referred to as primary features (PF) and secondary features (SF) respectively.

Let us define:

$$\boldsymbol{\mathcal{D}} = M(\boldsymbol{P}, \boldsymbol{S}) \tag{6.1},$$

where the motion sequence $\boldsymbol{\mathcal{D}}$ is a function (M) of $\boldsymbol{P}$ and $\boldsymbol{S}$ which are the spatiotemporal feature sets corresponding to existing primary and secondary themes respectively. Earlier in Chapter 2 we reviewed some previous work in which interpolation/extrapolation methods [23, 217] were used for blending motion. Assuming two motion sequences $\boldsymbol{\mathcal{D}}_1$ and $\boldsymbol{\mathcal{D}}_2$ with equal temporal lengths, the interpolated motion is often in the form

$$\boldsymbol{\mathcal{D}}_{interpolated} = \alpha\boldsymbol{\mathcal{D}}_1 + (1-\alpha)\boldsymbol{\mathcal{D}}_2 \tag{6.2},$$

where α is the blending factor. Based on Eq. 6.1, we have

$$\boldsymbol{\mathcal{D}}_{interpolated} = \alpha M(\boldsymbol{P}_1, \boldsymbol{S}_1) + (1-\alpha) M(\boldsymbol{P}_2, \boldsymbol{S}_2) \tag{6.3}.$$

Generally, interpolated motion sequences are similar in PT and differ in ST, hence

$\boldsymbol{P}_1 = \boldsymbol{P}_2$ and $\boldsymbol{S}_1 \neq \boldsymbol{S}_2$. Accordingly:

$$\boldsymbol{\mathcal{D}}_{interpolated} = \alpha M(\boldsymbol{P}_1, \boldsymbol{S}_1) + (1-\alpha) M(\boldsymbol{P}_1, \boldsymbol{S}_2) \tag{6.4}.$$

Given that the blending techniques have been successful and output sequence motion itself is a perceptually valid sequence with blended STs, we can conclude:

$$M(\boldsymbol{P}_1, \alpha \boldsymbol{S}_1 + (1-\alpha)\boldsymbol{S}_2) = \alpha M(\boldsymbol{P}_1, \boldsymbol{S}_1) + (1-\alpha) M(\boldsymbol{P}_1, \boldsymbol{S}_2) \tag{6.5}.$$

The goal here is to define M such that this interpolation equation holds true. While various mathematical operators may result in this relationship, the simplest and yet logical approximation for M is the summation operator. Accordingly,

$$\boldsymbol{\mathcal{D}} = M(\boldsymbol{P}, \boldsymbol{S}) = \boldsymbol{P} + \boldsymbol{S} \tag{6.6}.$$

Substituting this definition in the interpolation relationship Eq. 6.5, for the left side (*l.s.*) and right side (*r.s.*) we get:

$$l.s.: \quad M(\boldsymbol{P}_1, \alpha \boldsymbol{S}_1 + (1-\alpha)\boldsymbol{S}_2) = \boldsymbol{P}_1 + \alpha \boldsymbol{S}_1 + (1-\alpha)\boldsymbol{S}_2 \tag{6.7}$$

$$r.s.: \quad \begin{aligned} \alpha M(\boldsymbol{P}_1, \boldsymbol{S}_1) + (1-\alpha) M(\boldsymbol{P}_1, \boldsymbol{S}_2) &= \alpha(\boldsymbol{P}_1 + \boldsymbol{S}_1) + (1-\alpha)(\boldsymbol{P}_1 + \boldsymbol{S}_2) \\ &= \boldsymbol{P}_1 + \alpha \boldsymbol{S}_1 + (1-\alpha)\boldsymbol{S}_2 \end{aligned} \tag{6.8},$$

which confirms our approximation for M. The additive model for primary and secondary themes is simple, intuitive, and reliable. Furthermore, in addition to Cartesian and joint angle spaces in which it has shown accurate performance through interpolation/extrapolation, it has also been applied in latent subspaces [21, 172, 218].

In this definition, Eq. 6.5 indicates that $\boldsymbol{S}$ is scalable. Furthermore, while on single-style

motions are studied in this dissertation, it is in fact possible to have multiple STs present in a sequence. Happy-feminine run, sad-masculine walk, and angry-old standing up are examples of such situations. Also, multiple PTs can be present in motion, for example, sad-masculine walk and wave. Therefore, based on these to observations, Eq. 6.6 can be expanded to:

$$\boldsymbol{\mathcal{D}} = M(\boldsymbol{P}, \boldsymbol{S}, \boldsymbol{w}) = \sum_{i=1}^{r_p} \boldsymbol{P}_i + \sum_{i=1}^{r_s} \boldsymbol{w}_i \, \boldsymbol{S}_i \tag{6.9},$$

where r_p and r_s represent the number of PTs and STs. $\boldsymbol{w}$ is the weight set associated with STs. As mentioned, in this dissertation, $\boldsymbol{r}_p, \boldsymbol{r}_s, \boldsymbol{w}$ are all set to 1 for simplification.

Based on Eq. 6.9, let's assume two sequences $\boldsymbol{\mathcal{D}}_1$ and $\boldsymbol{\mathcal{D}}_2$ with $r_p = r_s = w = 1$. Accordingly, we have

$$\Delta\boldsymbol{\mathcal{D}} = \boldsymbol{P}_2 - \boldsymbol{P}_1 + \boldsymbol{S}_2 - \boldsymbol{S}_1 \tag{6.10}.$$

By selecting sequences with similar primary themes $\boldsymbol{P}_1 = \boldsymbol{P}_2$ and one of the sequences to be stylistically neutral $(\boldsymbol{S}_1 = 0)$, we can extract the SF of the other:

$$\boldsymbol{S}_2 = \Delta\boldsymbol{\mathcal{D}} \tag{6.11}.$$

In other words, subtracting a neutral action from a sequence containing the stylistic version of the same action will result in the SFs.

One of the critical assumptions for this approach was $r_p = r_s = 1$. While it is difficult to perform a sequence that contains only one set of PT and ST, motion studies often make such an assumption for simplification [23, 24] and study actions and features one at a

time. The second assumption was $w = 1$. This indicates that ST must contain a normalized magnitude. While this notion is difficult to verify, should $w \neq 1$, for $\boldsymbol{\mathcal{D}} = M(\boldsymbol{P}, \boldsymbol{S}, w) = \boldsymbol{P} + w\boldsymbol{S}$ by allowing $\boldsymbol{S}' = w\boldsymbol{S}$ we can re-write the equation as $\boldsymbol{\mathcal{D}} = M(\boldsymbol{P}, \boldsymbol{S}') = \boldsymbol{P} + \boldsymbol{S}'$. This means that the ST can simply be assumed as a scaled version of the original. For example, for an energetic walk with $w = 0.5$, we can define the equivalent sequence which is half as energetic but with $w = 1.0$.

Another condition for Eq. 6.11 is $\boldsymbol{S}_1 = 0$. This means that one of the sequences needs to be stylistically neutral. While it is extremely difficult, if not impossible, to perform and record neutral motion, the notion has been widely used in the literature [90, 98, 109]. In most cases, actors are asked to display minimal stylistic behavior, often described through the setting of a regular daily scenario.

A motion sequence in general is composed of sub-sequences or sub-actions. Accordingly, motion sequence $\boldsymbol{\mathcal{D}}_i$ can be re-organized as $\boldsymbol{\mathcal{D}}_i = \left[\boldsymbol{\mathcal{D}}_i^{(1)}, \boldsymbol{\mathcal{D}}_i^{(2)}, \dots, \boldsymbol{\mathcal{D}}_i^{(k)}\right]$ where the sequence is composed of k sub-sequences. Substituting this formulation into Eq. 6.6, we get:

$$\boldsymbol{\mathcal{D}}_i = M\left(\left[\boldsymbol{P}_i^{(1)}, \boldsymbol{P}_i^{(2)}, \dots, \boldsymbol{P}_i^{(k)}\right], \boldsymbol{S}_i\right) \tag{6.12},$$

where i denotes the sequence number and $\boldsymbol{P}_i^{(j)}$ is the jth sub-sequence of the primary action in the sequence. For example, a two-stride walk ($k = 2$) is composed of two separate steps, where each can be considered as a sub-sequence. A critical condition that we posed earlier for calculating Eq. 6.11 was $\boldsymbol{P}_1 = \boldsymbol{P}_2$. As a result, as an example, for a two-stride walking sequence $\boldsymbol{\mathcal{D}}_1$ and $\boldsymbol{\mathcal{D}}_2$, the corresponding steps need to be similar

($\boldsymbol{P}_1^{(1)} = \boldsymbol{P}_2^{(1)}$ and $\boldsymbol{P}_1^{(2)} = \boldsymbol{P}_2^{(2)}$). This means, the corresponding halves of the two sequences must contain similar actions, i.e. single steps. The sub-sequences themselves can be broken down into temporally shorter sub-sequences. By induction, this process can be repeated such that $k = m$ where m is the number of frames or time instances in the sequences. Based on this definition and the fact that the two sequences are similar in length, corresponding frames of the two sequences must contain similar content (similar poses). This is the very definition of temporal alignment. Accordingly, for interpolation methods to produce correct results, the two sequences need to be aligned. This has been shown and widely used in the previous work [23, 98]. Moreover, this argument further validates our proposed model (Eq. 6.9). In this dissertation, we utilize the CoTW technique proposed in Chapter 4.

A final point to consider is that the defined SF is a set of spatiotemporal curves while the two sequences (stylistic input and the neutral sequences) were assumed to be equal in length. Naturally, however, this is hardly the case. In fact, the difference in length is caused by the behavioral STs in the input sequence. As a result, this type of feature cannot be extracted using the proposed model and an accompanying process is required to extract this feature.

When different segments of two motion sequences with different lengths are performed with different speeds, the temporal changes are manifested as spatiotemporal curves. However, there are instances where the length of the entire sequence is one of the influential components of the SF set [90]. In addition to movement and posture, we introduce a separate SF component which describes the relationship between the

temporal length of a stylistic sequence with respect to a normalized length. Based on this definition, Eq. 6.9 can be modified. If we represent all sequences by m' frames, the model can be updated as:

$$\mathcal{D} = M(\boldsymbol{P}, \boldsymbol{S}, \boldsymbol{w}, \boldsymbol{m}) = \left(\sum_{i=1}^{r_p} \boldsymbol{P}_i + \sum_{i=1}^{r_s} \boldsymbol{w}_i \, \boldsymbol{S}_i \right) \mathbf{W}_{m\prime \times m} \qquad (6.13),$$

where $\mathbf{W}_{m\prime \times m}$ is the uniform time warping (UTW) warping matrix that interpolates a sequence with initial length m' to achieve new length m as described in Chapter 4.

Following, we propose a method based on the proposed model for extracting the SF set of a given stylistic sequence as three separate components: posture, movement, temporal.

6.3 Proposed Method

6.3.1 Correspondence

The method proposed in this chapter only requires a single neutral action for extracting the three components of the SF. Many of the previous studies on feature extraction and editing, or synthesis of motion are also dependent on examples or datasets. For example, [24, 158, 170, 183, 184] are all as such.

Given a dataset of neutral sequences with PTs similar to that of the input stylistic sequence, selecting one to utilize in the system can be done arbitrarily. Nevertheless, the accuracy of the method and the quality of the extracted features can be increased by selecting a neutral sequence that is most similar with the input in terms PT. In other words, the goal is to approach $\boldsymbol{P}_1 = \boldsymbol{P}_2$ as much as possible. To select the optimum

neutral sequence from a dataset, we define the objective function:

$$G\left(\boldsymbol{\mathcal{D}}_{input}, \boldsymbol{\mathcal{D}}_{neutral}\right) = \sum_{i=1}^{n} \boldsymbol{u}_i \cdot \rho\left(\boldsymbol{\mathcal{D}}_{input,i}, \widetilde{\boldsymbol{\mathcal{D}}}_{neutral,i}\right) \quad (6.14),$$

where n is the number of DOFs. The ~ sign denotes the sequence after warping using CoTW as described in Chapter 4. $\boldsymbol{\mathcal{D}}_{neutral}$ is warped with respect to $\boldsymbol{\mathcal{D}}_{input}$ as the reference. $\boldsymbol{u}$ is a set of weights used for customization of the process similar to the one used in Eq. 4.5. Ad-hoc means as well as previous studies such as [197] can be used for determining this parameter. For the system proposed in this chapter, the weight of the shoulder and thigh joints are set twice the weight of hand and foot joints. The weights for fingers and toes can be set to zero due to their insignificance. As mentioned in Chapter 4, the CoTW process also utilizes a weight set based on which the significance of different DOFs in the warping process is decided. We assign the warping weight set equal to $\boldsymbol{u}$ since the aim of both parameters is to customize the processes based on the relative importance of different joints. Finally, the sequences $\widetilde{\boldsymbol{\mathcal{D}}}_{neutral}$ which maximizes $G\left(\boldsymbol{\mathcal{D}}_{input}, \boldsymbol{\mathcal{D}}_{neutral}\right)$ with respect to the other neutral sequences in the dataset is selected for the process.

6.3.2 Movement Features

We employ spatiotemporal cubic splines for modeling the neutral version of the stylistic input sequence. The non-linear nature and controllability of splines make them very suitable means for many applications where complex data are being modeled. These models have been widely used in geometric modeling, computer graphics, and motion processing [23] among other applications. Assuming a set of temporal control points (t_i)

with $r+1$ members, and their corresponding spatial values ($y_j = f(t_j), f: \mathbb{R} \to \mathbb{R}$), we have the following set of points $[t_j, y_j]$ for $j = 0,1,\dots,r$. Cubic splines are piecewise third-order polynomials in the form of:

$$F_j(i) = \sum_{k=0}^{3} c_{j,k}(i - t_j)^k \quad for \quad j = 0,1,\dots,r \tag{6.15}$$

where $i \in [t_j, t_{j+1}]$. Accordingly, for the function with r intervals, we have the spline approximation as:

$$F(i) = \begin{cases} F_0(i) & if\ t_0 \le i < t_1 \\ F_1(i) & if\ t_1 \le i < t_2 \\ \vdots & \\ F_{r-1}(i) & if\ t_{r-1} \le i < t_r \end{cases} \tag{6.16}$$

where $F: \mathbb{R} \to \mathbb{R}$ and $t_0, t_1, \cdots, t_r$ are the control points which the approximation must pass through [219]. This condition yields the two following set of constraints:

$$F_j(t_j) = f(t_j) \quad for \quad j = 0,1,\dots,r-1 \tag{6.17}$$

$$F_j(t_{j+1}) = f(t_{j+1}) \quad for \quad j = 0,1,\dots,r-1 \tag{6.18}$$

Additionally, to achieve a continuous and smooth approximation, the first and second derivatives must also be continuous. This gives the following set of equations:

$$F'_j(t_{j+1}) = F'_{j+1}(t_{j+1}) \quad for \quad j = 0,1,\dots,r-2 \tag{6.19}$$

$$F''_j(t_{j+1}) = F''_{j+1}(t_{j+1}) \quad for \quad j = 0,1,\dots,r-2 \tag{6.20}$$

Finally, the following boundary conditions must also hold true:

$$F'_0(t_0) = f'(t_0) \quad (6.21),$$

$$F'_{r-1}(t_r) = f'(t_r) \quad (6.22),$$

Eq. 6.17 to Eq. 6.22 yield $4r$ equations and $4r$ unknowns, solving which produces F. Based on the description of motion data provided in Appendix A, given a motion matrix $\boldsymbol{\mathcal{D}} = [\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \cdots \boldsymbol{\theta}_n]$ with n DOFs, the ith joint angle trajectory $\boldsymbol{\theta}_i$ with m frames is defined by $\boldsymbol{\theta}_i = \left\{\theta_i^{(k)} : k = 1, \cdots, m \in \mathbb{N}\right\}$. A set of spline control points ($\boldsymbol{t}$) when projected onto human motion trajectories ($\boldsymbol{\theta}_i^{(\boldsymbol{t})} = f(\boldsymbol{t})$) can be interpreted as temporal cues for segments of the motion sequence. Our goal is to optimize the set of cues $\boldsymbol{t}$ such that the spline approximation of $\boldsymbol{\mathcal{D}}_{input}$ resembles $\boldsymbol{\mathcal{D}}_{neutral}$. As the temporal cues are located on the joint angle trajectories of the stylistic input, a spline approximation using these cues, but with maximizing similarity with respect to the neutral sequence, will create a neutral version of the input sequence that is located on top of the input. In other words, this approximation excludes any spatial shifts. Based on the definition of movement and posture features mentioned earlier, this approximation can be used to extract movement features, leaving constant spatial shifts (posture features) intact.

Generally, when modeling a signal (or in the case of this study, joint angle trajectory) using cubic splines, using more control points leads to an approximation which follows the actual trajectory more precisely and accurately, reconstructing the higher frequency curves. Fewer and more widely distributed control points, on the other hand, result in a more loose and general approximation of the trajectory, leaving out higher frequency components. While non-uniform distribution of cues can be a more optimal solution, in order to measure a frequency value, uniform and evenly spaced cues are used in this

chapter. Henceforth we define a measure for the frequency rate of the temporal cues named ω (Eq. 6.23), where r is the number of temporal cues, m is the number of frames in the sequence, and f_s is the original sampling frequency of the signal during recording of the signal. In this research a constant sampling rate of $f_s = 60$ fps is used.

$$\omega = \frac{f_s}{m}\, r \tag{6.23}.$$

In order to approximate the neutral component of a stylistic input sequence, $\boldsymbol{\mathcal{D}}_{input}$, we calculate ω with the aim of maximizing similarity of the approximated sequence, $\boldsymbol{\mathcal{D}}_{approx}$, with respect to the warped version of the corresponding neutral sequence $\widetilde{\boldsymbol{\mathcal{D}}}_{neutral}$. Hence, using G as defined in Eq. 6.14,

$$\arg\max_{\omega}\ G\left(\boldsymbol{\mathcal{D}}_{approx}, \boldsymbol{\mathcal{D}}_{neutral}\right) \tag{6.24},$$

calculates the frequency of the temporal cues.

Figure 6.2 illustrates the correlation values for several joint angle trajectories of a stylistic sequence. It is shown that in all cases, for maximizing correlation, $\omega < 10$ Hz is required. The correlation values converge towards an asymptote which implies increasing the number of cues beyond a certain point does not result in significant changes in the model. This is because as the number of cues increases, the modeled approximation tends towards the actual stylistic signal rather than the neutral one. In other words, the goal is to stop increasing ω as soon as the modeled signal resembles the neutral signal.

Successive to calculating the optimal frequency for the temporal cues and approximating the neutral component ($\boldsymbol{\mathcal{D}}_{approx}$) of the sequence with splines, and based on Eq. 6.11, a

feature set is extracted using:

$$\boldsymbol{\Phi}_{movement} = \boldsymbol{\mathcal{D}}_{input} - \boldsymbol{\mathcal{D}}_{approx} \quad (6.25),$$

where $\boldsymbol{\Phi}_{movement}$ is the movement SF.

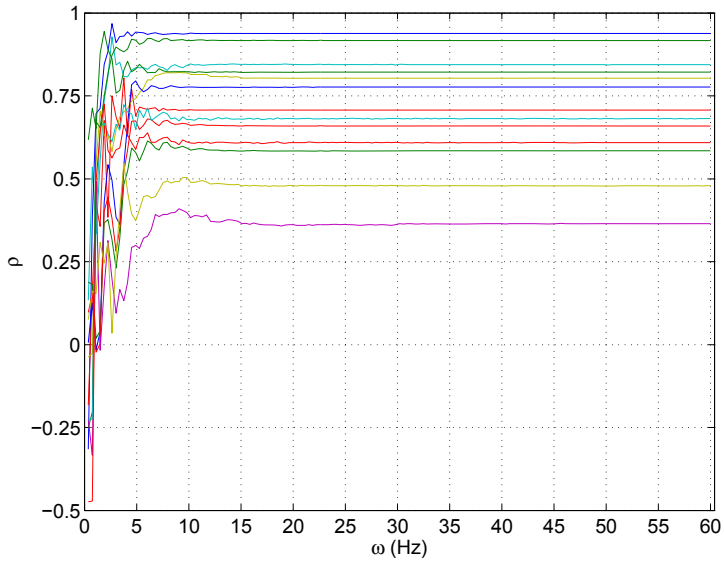


Figure 6.2. ρ vs. ω for several joint angle curves of a stylistic walk. The maximum of each curve (usually occurring between 3 to 8 Hz) is employed as the optimum frequency for cues.

The reason that this feature set is associated with *movement* is its varying nature throughout the sequence. This reasoning becomes more clear as we extract the posture feature in the following section. Figure 6.3 illustrates a motion trajectory being modeled with the optimal ω value (~ 3 Hz) and the movement SF set being extracted. It is observed that the approximated neutral trajectory using splines is very similar to the corresponding neutral trajectory. However, posture features, in the form of spatial offsets, remain to be extracted.

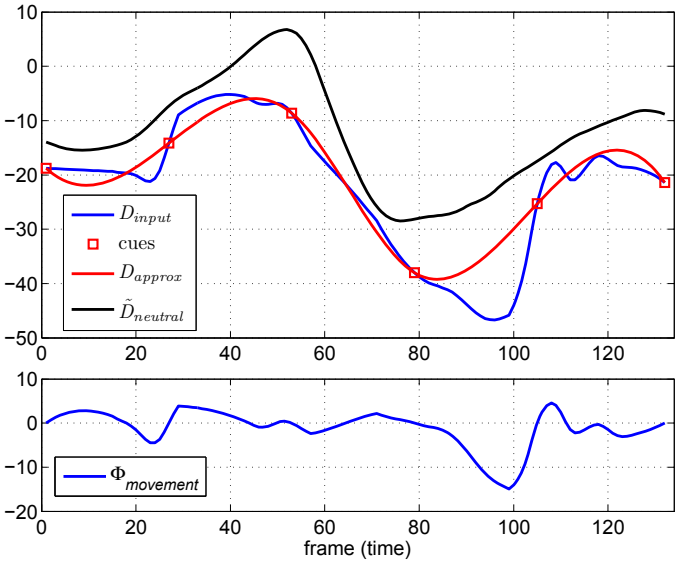


Figure 6.3. Extraction of movement SF from a motion signal with ω = ~ 3 Hz.

6.3.3 Posture Features

It was illustrated in Figure 6.3 that the temporal cues used to approximate the neutral component are spatiotemporally located on the stylistic input signal. Therefore, the extracted movement SF set is always positioned on the input and excludes any spatial offsets with respect to the corresponding neutral sequence. However, as illustrated in Figure 6.3, spatial offsets do exist as segments of the approximated signal need to be vertically shifted in order to fall on top of the neutral signal. Due to its unvarying nature, this spatial difference between the modeled sequence and the neutral sequence highly resembles the posture features described earlier. In other words, in the context of human motion, these features describe how the body structure of the neutral sequence should be modified through spatial shifts to become similar to the stylistic body structure. Subsequently, this feature set can be calculated by:

$$\boldsymbol{\Phi}_{posture,\,avg} = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{\mathcal{D}}_{approx}^{(i)} - \widetilde{\boldsymbol{\mathcal{D}}}_{neutral}^{(i)}\right) \qquad (6.26),$$

which results in a constant average posture feature. Here, m represents the length of the sequence. Figure 6.4 illustrates the average posture feature extracted for the joint angle trajectory used in Figure 6.3.

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

$$\boldsymbol{\Phi}_{posture,\,segment} = \frac{1}{m_{seg}}\sum_{i=1}^{m_{seg}}\left(\boldsymbol{\mathcal{D}}_{approx}^{(i)} - \widetilde{\boldsymbol{\mathcal{D}}}_{neutral}^{(i)} \mid seg\right) \quad for\ \ seg = 1, \ldots, r \qquad (6.27),$$

where m_{seg} represents the number of frames in that segment. A sample extracted posture set is illustrated in Figure 6.4. This quantized version of the posture feature will cause discontinuities when added to any continuous trajectory. To remedy this, we approximate the piece-wise feature using cubic smoothing splines. The smoothing spline $\boldsymbol{\Phi}_{posture,\,approx}$ for the quantized signal $\boldsymbol{\Phi}_{posture,\,segment}$, is calculated through Eq. 6.28 based on [220]:

$$\arg\min_{\boldsymbol{\Phi}_{posture,\,approx}} \left\{ z\sum_{i}\beta_i \left\| \boldsymbol{\Phi}_{posture,\,approx}^{(i)} - \boldsymbol{\Phi}_{posture,\,segment}^{(i)} \right\|^2 + (1-z)\sum_{i}\gamma_i \left\| \frac{\Delta^2 \boldsymbol{\Phi}_{posture,\,approx}^{(i)}}{\Delta i^2} \right\|^2 \right\} \qquad (6.28).$$

Here, γ is the roughness measure, for our application set to 1, $z \in [0, 1]$ is the smoothing parameter, for our application set to 0.01, and β is the weight equal to 1. Figure 6.4 illustrates the extraction and approximation process for the posture SF set. It is important to note that the posture features should be extracted only successive to extraction of the movement features.

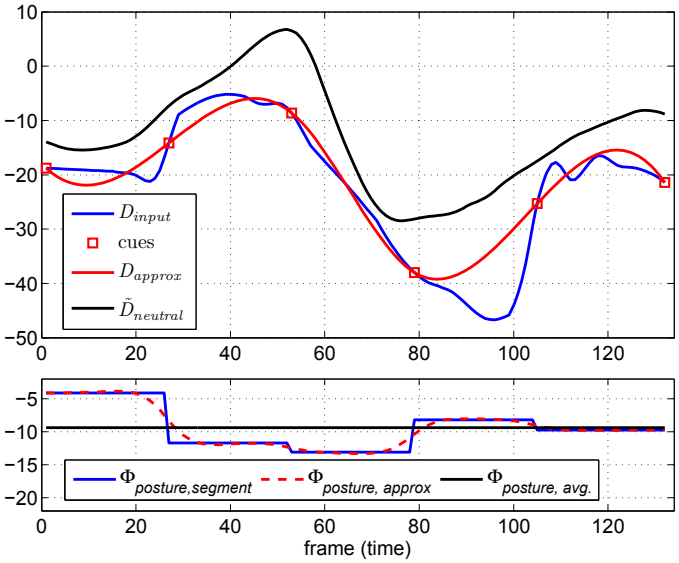


Figure 6.4. Extraction of posture SF from the input signal. Smoothing splines are used to approximate the feature.

6.3.4 Uniform Temporal Feature

Based on Eq. 6.13 we define $\Phi_{temporal}$ to describe the relationship between the temporal lengths of stylistic sequences with respect to neutral ones. We calculate this feature using:

$$\Phi_{temporal} = m_{input}/m_{neutral} \quad (6.29),$$

where m_{input} is the temporal length of the input and $m_{neutral}$ is the length of the corresponding neutral sequence. An important point to consider is that for this SF, m_{input}

and $m_{neutral}$ need to correspond to sequences that contain similar actions ($P_{input} = P_{neutral}$). This similarity should be true, both in terms of semantics as well as in terms of the number of actions in the sequences. For example, if $\boldsymbol{\mathcal{D}}_{input}$ is a walking sequence with two strides, $\boldsymbol{\mathcal{D}}_{neutral}$ must similarly contain walking with only two strides. This feature can also be characterized or described by the speed with which a particular stylistic action is performed with respect to the neutral version of that action.

Successive to calculating $\Phi_{temporal}$, UTW can be used to speed-match a neutral sequence with the stylistic input, hence transferring this feature. If $\Phi_{temporal} = 1$, the neutral sequence requires no temporal modification as it is equal in length to the stylistic sequence. For $\Phi_{temporal} < 1$, the neutral sequence needs to be compressed, and when $\Phi_{temporal} > 1$, it needs to be stretched. Examples of $\Phi_{temporal}$ are illustrated in Figure 6.5.

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

sequences non-uniformly, thus resolving the issue of non-uniform temporal features and allowing them to be extracted as movement and/or posture SFs.

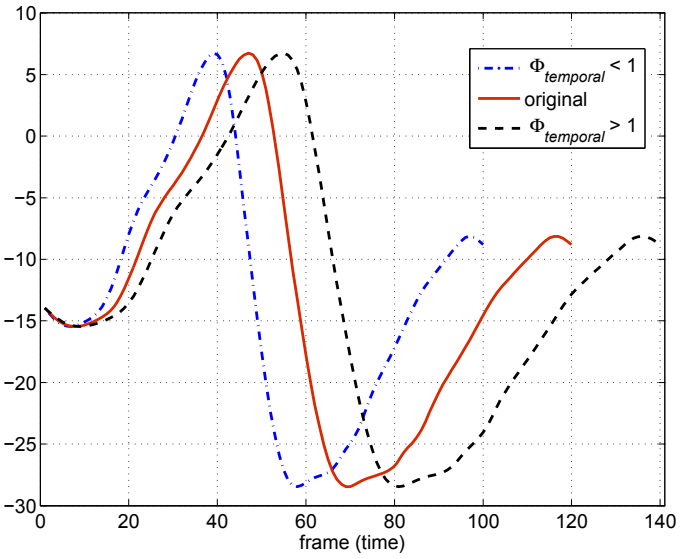


Figure 6.5. $\Phi_{temporal}$ for versions of the signal with different speeds.

6.4 Results and Discussion

We utilize sequences from the Carnegie Mellon University (CMU) dataset as well as our own recorded data (Carleton dataset). Details are provided in Appendix A. From the CMU dataset, sad, macho, drunk, and lavish walks are employed. Also, 5 neutral walks are used from which the best corresponding sequence is selected for each stylistic input walk. From our own dataset, tired and energetic walks, jumps, and runs are used as inputs, along with 5 neutral walks, 5 neutral jumps, and 5 neutral runs. Table 6.1 and Table 6.2 present the types of actions and STs used in this study. The former refers to sequences from the CMU dataset and the latter refers to those from the Carleton dataset.

The 5 neutral actions are represented only once in the tables. Following, the movement, posture, and uniform temporal features are extracted in the presented order. Furthermore, to illustrate the performance of the proposed method, style translation is carried out.

Table 6.1. The actions and datasets used from the CMU dataset.

	CMU dataset				
index	1	2	3	4	5
action	Walk	Walk	Walk	Walk	Walk
SF type	Neutral	Sad	Macho	Drunk	Lavish

Table 6.2. The actions and datasets used from the Carleton dataset.

	Carleton dataset								
index	6	7	8	9	10	11	12	13	14
action	Walk	Walk	Walk	Jump	Jump	Jump	Run	Run	Run
SF type	Neutral	Tired	Energetic	Neutral	Tired	Energetic	Neutral	Tired	Energetic

6.4.1 Extraction of Features

Figure 6.6 presents the average and standard errors of ω over DOFs of the sequences presented in Table 6.1 and Table 6.2. One-way analysis of variances, ANOVA, for the effect of different DOFs on the calculated ω, shows no significant effect at the $p < 0.05$ level despite the existing variations. Specifically, for the CMU samples, $F(75,228) = 1.2$, $p = 0.158$ and for Carleton samples, $F(51,260) = 1.31$, $p = 0.091$. One-way ANOVA for the effect of actions on calculated ω shows significance with $F(3,300) = 19.6$, $p < 0.0001$ for the CMU samples and $F(5,306) = 9.49$, $p < 0.0001$ for Carleton samples. This can be clearly observed in Figure 6.6 where, for example, energetic jump requires a significantly higher average ω compared to tired jump. This analysis indicates that while within a particular action, the DOF doesn't significantly impact ω, the action from which the features are being extracted, and the feature types themselves, do have a significant

influence on the optimum frequency values.

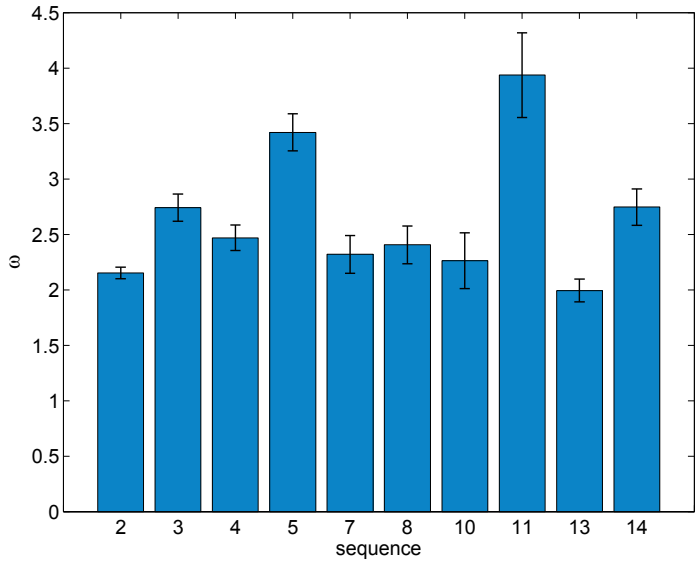


Figure 6.6. Average ω measured for different input sequences. Error bars represent standard errors over DOFs.

Table 6.3 presents the results for the calculated $\Phi_{temporal}$ values. Energetic sequences along with lavish walk, which are shorter (faster) than neutral, have resulted in $\Phi_{temporal} < 1$. For the rest of the samples, $\Phi_{temporal} > 1$, indicating longer lengths (slower) with respect to neutral. Generally, $\Phi_{temporal}$ is quite intuitive and simple to speculate without computation. Sequences that have SFs associated with low energy are often longer than neutral, thus $\Phi_{temporal} > 1$, while those attributed to higher energy are faster, and so $\Phi_{temporal} < 1$. However, for instances where exact ratios are required, Eq. 6.29 is utilized.

Table 6.3. Uniform temporal features for the test data.

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

6.4.2 Style Translation

Once the three SF components have been extracted from a stylistic input sequence, the features can be applied to a neutral sequence to produce a stylistic theme and change the action from neutral to stylistic. As discussed in Chapter 4, this process is called style translation [24]. Here, we utilize style translation to visualize the features extracted using our approach.

When manipulating motion data, synchronization is often lost [24]. As a result, out of tune motion of the feet causes an artifact in which the character seems to be skating rather than taking firm and solid steps. This is referred to as foot-skating [221] (or foot-sliding). Nevertheless, Eq. 6.24 calculates a global ω for the entire sequence. In other words, the process utilizes identical temporal cues for extraction of features from different DOFs. While this approach significantly reduces synchronization-related artifacts such as footskating, other artifacts are introduced as sub-optimal frequencies are utilized for some DOFs. As a result, some extracted SFs will be inaccurate. Specifically, as ω is calculated for the entire sequence, the sub-optimal set of cues, for some DOFs, might fall on specific extrema which are in fact part of the SF; whereas, if ω was calculated for that specific DOF, a different number of cues would have allowed for that extrema to be extracted as a feature. For this reason, we calculate ω for each joint separately and extract the features accordingly. The negative side effects (footskating) can be remedied through post-processing such as foot-plant-based techniques, which are forms of physical constraints [222]. Here, we first map the joint angle curves of the outputs onto the Cartesian space. Subsequently, the skate trajectory of the stance foot is calculated for the duration of that stance and subtracted from the motion, eliminating the artifact. Animation software such

as MotionBuilder (http://www.autodesk.com/products/motionbuilder/overview) also allow for cleanup of this artifact.

Video Clip C (https://www.youtube.com/watch?v=KBKUhWM9Fxw) and Figures 6.7 and 6.8 (both extracted from the video) present the style translation outputs for the sequences from CMU and Carleton datasets. Successful conversion of neutral actions to designated STs point to the accuracy of our method. The neutral actions are the inputs and the stylistic ones are the style translation outputs.

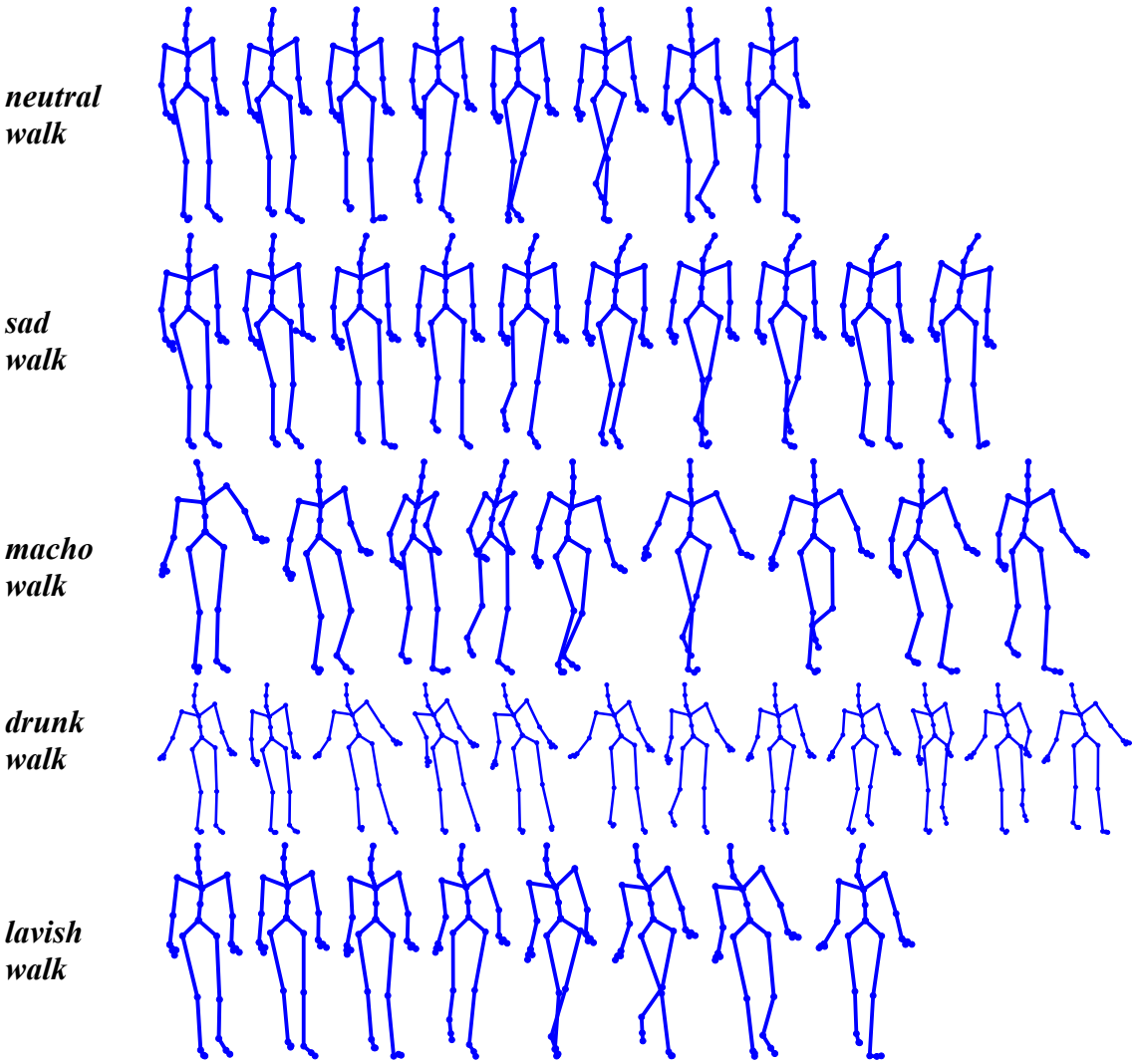


Figure 6.7. CMU dataset style translation outputs (from Video Clip C).

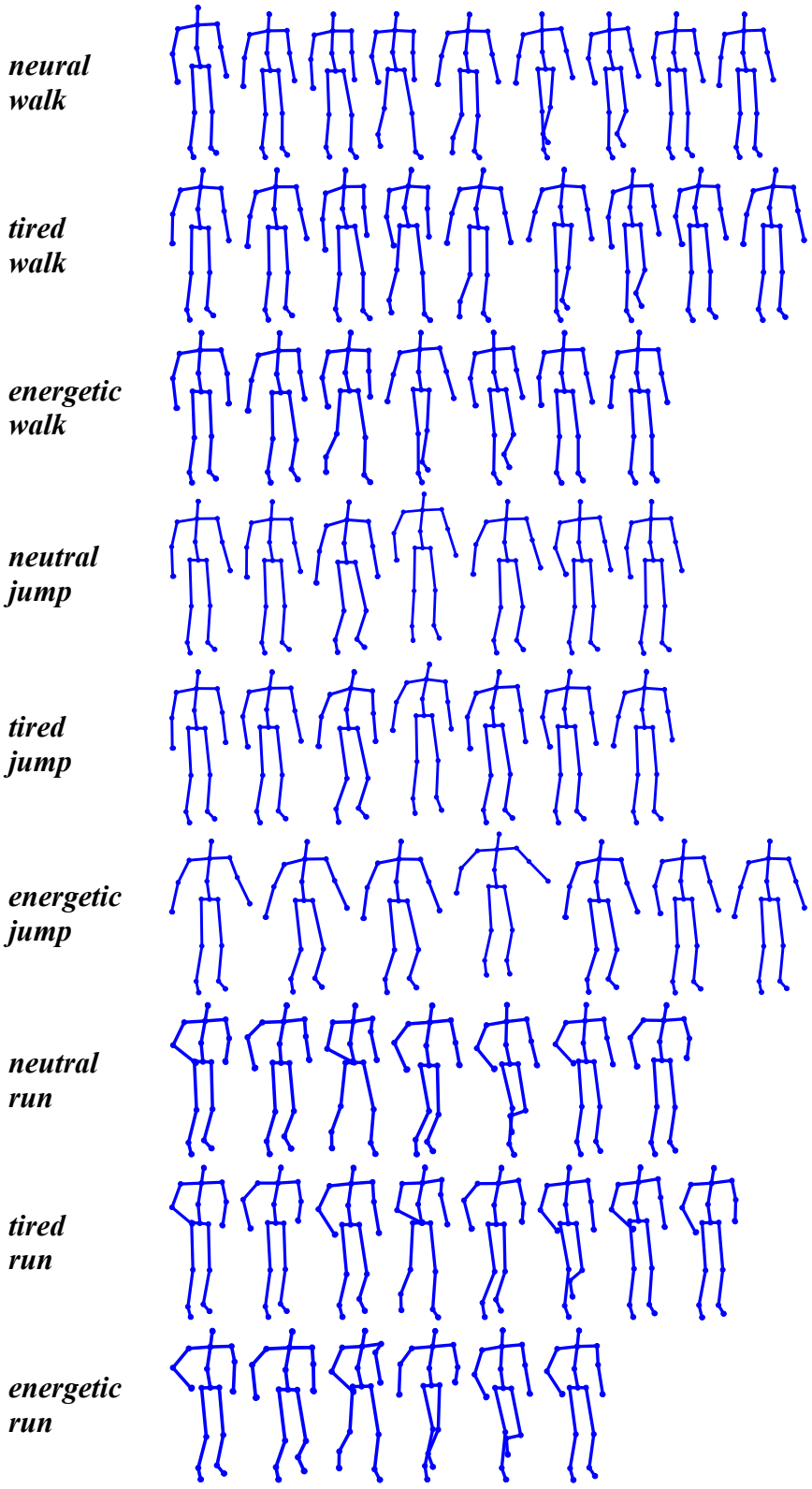


Figure 6.8. Carleton dataset style translation outputs (from Video Clip C).

To further evaluate the outputs, 10 participants, were asked to watch and provide feedback on the SFs that are perceived from the outputs. The average age of participants was 31.3, the standard deviation was 11.5, 6 were males, and 4 were females. A forced-choice questionnaire was used for this purpose. Sequences from the two datasets were presented separately and choices were divided based on the dataset. This was done for two reasons: (a) the skeleton structure for the two datasets is different which could influence the results had they been investigated together, (b) sad (from the CMU dataset) and tired (from the Carleton dataset) actions are very prone to misperception by participants. Such sequences are difficult to distinguish in real life as both contain features such as slower motion, decreased sway, and downward tilt in shoulders and head. Table 6.4 and Table 6.5 present the recognition rates for the style translation results. The former presents the results for the data from the CMU dataset while the latter are those of the Carleton dataset. The high recognition rates indicate accurate perceptual quality of the extracted features.

In general, the quality of our style translation outputs is on par with methods such as [24] or [23] where high quality stylistic sequences were generated. Our system performs well for different actions such as walking, jumping and running and a variety of styles such as sad, macho, drunk, lavish, tired, and energetic. However, compared to linear interpolation/extrapolation approaches such as [23] or learning systems such as [24], our method illustrates higher generalization capabilities where using a frequency range of ω = 2 to 8 Hz for temporal cues, the movement SFs can be extracted without a neutral reference. To extract the posture features, however, our system also needs a neutral reference. In addition, the uniform temporal feature can be speculated with an acceptable

accuracy.

Table 6.4. Recognition rates for the style translation outputs for the CMU dataset.

		Perceived				
		Sad	**Macho**	**Drunk**	**Lavish**	**Neutral**
Output	**Sad**	**0.8**	0.0	0.0	0.0	0.2
	Macho	0.0	**1.0**	0.0	0.0	0.0
	Drunk	0.0	0.0	**1.0**	0.0	0.0
	Lavish	0.0	0.0	0.0	**1.0**	0.0

Table 6.5. Recognition rates for the style translation outputs for the Carleton dataset.

		Perceived								
		Walk Tired	**Walk Energetic**	**Walk Neutral**	**Jump Tired**	**Jump Energetic**	**Jump Neutral**	**Run Tired**	**Run Energetic**	**Run Neutral**
Output	**Walk Tired**	**0.8**	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
	Walk Energetic	0.0	**0.9**	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	Jump Tired	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0	0.0	0.0
	Jump Energetic	0.0	0.0	0.0	0.0	**0.9**	0.1	0.0	0.0	0.0
	Run Tired	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0
	Run Energetic	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.7**	0.3

Unlike most existing techniques for stylistic motion control which are purely computational, our proposed method draws inspiration from the way biological motion is executed and perceived, namely, as three separate components of movement, posture, and time. Our system successfully differentiates and separately extracts these components. As a result, we believe our system can be used to study the physiology of affective and stylistic motion and contribute to psychological studies as well. The system can also be

used for multimedia and HCI applications such as animation and gesture-based interaction systems. To summarize, our system benefits from:

- Accurate extraction of features, and thus, high quality animation
- Separation of the three feature components namely movement, posture, and time
- Higher generalization compared to existing methods

Chapter 7.

Expert-driven Perceptual Shortcuts for Synthesis of Secondary Features

7.1 Introduction

Based on the review of the literature presented Chapter 2, we can conclude that the developed techniques for synthesis and editing of secondary features (SFs) are effective and valuable. Nevertheless, despite the accuracy of advanced computational approaches, they are often complex and computationally expensive. Furthermore, they seldom provide sufficient generalization as the systems need to be trained or configured using new data. Comprehensive studies on perceptual outcomes and implications of such

systems have also been mostly overlooked. On the other hand, although subjective rule-based methods are simple and intuitive, they are mainly difficult to mathematically model and subsequently apply to new data. Moreover, the relative perceptual significance and impact of different observed features has not been widely explored. Finally, for both categories, evaluating the scalability of the models can lead to more dynamic solutions.

To overcome these issues and in line with our approach in utilizing perceptually guided techniques, this chapter presents an expert-driven perceptual approach for generation of SFs. Our method is simple, intuitive, efficient, and the solution is scalable. To address the problem, we propose and utilize low-cost and controllable Gaussian radial basis functions (RBFs) as constructs of features in different categories of style and affect. We then develop a user interface that can be used to add multiple RBFs to motion sequences. Experienced animators are asked to use this interface and add up to 3 RBFs per degree of freedom (DOF) to a neutral walking sequence with the goal of generating stylistic sequences, namely, happy, sad, energetic, tired, feminine, and masculine walks. The neutral input sequence is synthesized through aligning and averaging multiple segmented neutral walks. This process ensures that the input is highly neutral and invariant of personal walking styles. The sets of RBF-based edits are recorded and analyzed and feature sets are computed. Perception feedback is subsequently collected regarding the computed feature sets. The study shows that the computed features are perceptually effective and accurate. The study also sheds light on how the mentioned motion variations are performed and perceived. Details regarding the use of posture vs. motion features in performing the actions, the most important body joint, and the inversion effect are realized. The relative significance of the features is embedded in the solution,

enabling for only few features to model affective/stylistic motion. This property enables the use of the most important features that model the variations, hence perceptual shortcuts. Accordingly, only a limited number of features are sufficient for correct perception ratings. Finally, our proposed system provides the possibility of applying weights to the derived features, generating a spectrum of intensities for affects and styles.

The contents of this chapter have been published as [223, 224].

7.2 System Overview

As the first step towards an expert-driven set of SFs, simple and controllable mathematical functions with which features can be modeled and collected are required. Accordingly, we first introduce the constructs of SFs for our framework in the following section. We subsequently validate the numerical and perceptual accuracy of the constructs in modeling different SFs. These constructs are also used in Chapter 8 in another framework that successfully classifies and translates SFs, further validating their accuracy.

The next step towards developing the system is a highly general motion sequence to be used as the input. For this purpose, the same average neutral walk created using the HDM05 dataset as described in Appendix A, is used as the input. To recap: multiple neutral walk cycles were manually segmented, aligned using CoTW, and averaged. The resulting sequence is a two-stride high quality and artifact-free walk which does not contain any personalized behavior, making it a very suitable representation of a neutral walk for perception studies. To further ensure the perceptual quality of this sequence, it is

perceptually validated in Section 7.6.1. Figure 7.1 presents the 17-joint body structure and spatial orientation of the model. Similar to Chapter 4, we use the semantic names of the joints rather than the exact names used in the dataset files. Figure 7.2, extracted from Video Clip D (https://www.youtube.com/watch?v=ffBxjScTDAo) illustrates frames from the sequence.

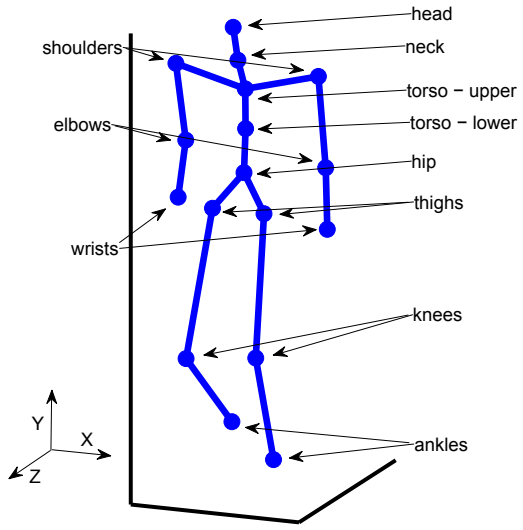


Figure 7.1. The human body model and its spatial orientation used in this study. The HDM05 body structure is slightly modified for this study.

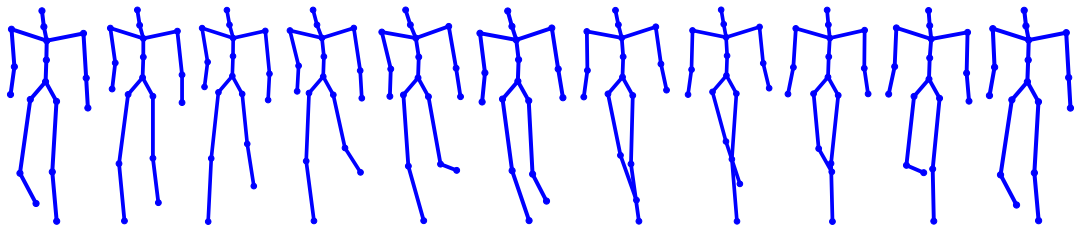


Figure 7.2. Frames from the average neutral sequence used as input (from Video Clip D).

To record perceptual SFs, a user interface is required, using which animators can modify the neutral input using the pre-defined set of mathematical functions. Accordingly, we

built a graphical user interface (GUI) in MATLAB. Experienced animators were asked to use the GUI and convert the neutral walk into stylistic/affective ones. The gathered data were summarized and analyzed. Using the summary of features, motion sequences with different SFs are produced. Finally, the generated results were evaluated through a second corpus of human subjects who are naïve towards motion studies. Figure 7.3 presents the schematic of the process in this chapter.

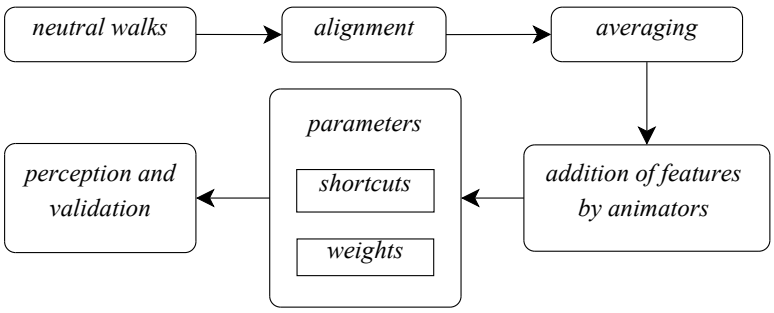


Figure 7.3. Schematic of the process used in this study.

7.3 RBFs as Constructs for SF

With existing animation and motion editing software tools such as Autodesk MotionBuilder (http://www.autodesk.com/products/motionbuilder/overview), Autodesk Maya (http://www.autodesk.com/products/autodesk-maya/overview), Autodesk 3DS Max (http://www.autodesk.com/products/autodesk-3ds-max/overview), and Blender (http://www.blender.org), animators often use ad-hoc and free-form means for changing existing motion trajectories. As a result, analysis and interpretation of the processes used by animators for altering motion are very difficult if not impossible. To overcome this obstacle, we propose the use of standard mathematical functions, namely Gaussian RBFs

as constructs for SFs. These functions are highly controllable and easy to analyze and study.

A radial function $\phi: \mathbb{R}^s \rightarrow \mathbb{R}$ is defined as a univariate function $\phi(r)$, where $r = \|t\|_2$, and $\|\cdot\|_2$ is a norm operator such as the Euclidean norm. Consequently, a Gaussian RBF is defined by:

$$\varphi(t; \mu, \sigma^2) = \phi(\|t - \mu\|_2) = exp\left\{\frac{-\|t - \mu\|_2^2}{2\sigma^2}\right\} \quad (7.1),$$

where μ is the mean and σ^2 is the variance. Accordingly, we can model the SF trajectory of the *i*th DOF with a weighted sum of M RBFs, resulting in:

$$\mathbf{\Phi}_i = \sum_{j=1}^{M} \alpha_j \varphi\left(t; \mu_j, \sigma_j^2\right) \quad (7.2).$$

where j denotes the RBF index and α is the amplitude or intensity of each RBF. Hence, the SF set for an m-frame long n-dimensional motion sequence is represented by $\{\mathbf{\Phi}_1, \mathbf{\Phi}_2, \dots, \mathbf{\Phi}_n\}^T{}_{n \times m}$ or by the $n \times M$ parameter set matrix:

$$\mathbf{\Pi} = \begin{bmatrix} \{\alpha, \mu, \sigma^2\}_1^1 & \cdots & \{\alpha, \mu, \sigma^2\}_M^1 \\ \vdots & \ddots & \vdots \\ \{\alpha, \mu, \sigma^2\}_1^n & \cdots & \{\alpha, \mu, \sigma^2\}_M^n \end{bmatrix} \quad (7.3)$$

To test the validity of our choice of constructs, we extracted the SF for happy, sad, energetic, tired, feminine, and masculine walks using linear subtracting a neutral walk from the stylistic sequences, successive to alignment using CoTW. The warping details were presented in Chapter 4. Our goal for this validation test is to determine whether the

SFs decomposed using RBFs are robust both numerically and perceptually. Figure 7.4 illustrates SFs of two of the trajectories from an energetic walk, modeled with a weighted sum of 5 Gaussian RBFs ($M = 5$). The linear least squares method has been used to decompose the trajectories into the basis functions. We see that the two trajectories are almost perfectly modeled using only 5 RBFs.

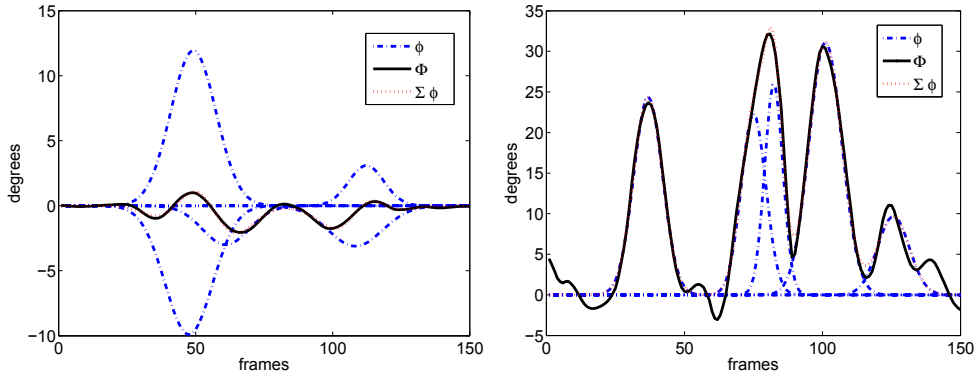


Figure 7.4. Two different SF trajectories of an energetic walk approximated using 5 RBFs.

By assigning different values for M, we can control the precision of the approximated trajectory and the amount of residues. Figure 7.5 illustrates the average RMSE/DOF vs. number of RBFs used to approximate the SF sets of happy, sad, energetic, tired, feminine, and masculine walking sequences. We have employed 1 to 8 RBFs to model the SF sets. This approximation can extend well beyond 8 and as expected the residues decrease as more RBFs are used and approximations become more accurate.

Perceptually, increasing the number of RBFs beyond a certain point is of little significance. To evaluate the perceived quality of SFs decomposed using RBFs, we extracted and modeled the SF sets of the 6 stylistic walks using $M = \{1, 3, 5, 7\}$ RBFs. The modeled SFs are then added back onto the neutral portion of the original sequences.

Each sequence was animated for 5 human subjects, 3 males and 2 females with an average age of 26.3 and standard deviation of 3.6. Table 7.1 presents the percentage of the audience who were able to correctly identify the secondary themes (STs) of the re-synthesized sequences when compared to the original sequences. Based on the results, only a few RBFs, as little as $M = 3$, is sufficient for approximating SFs.

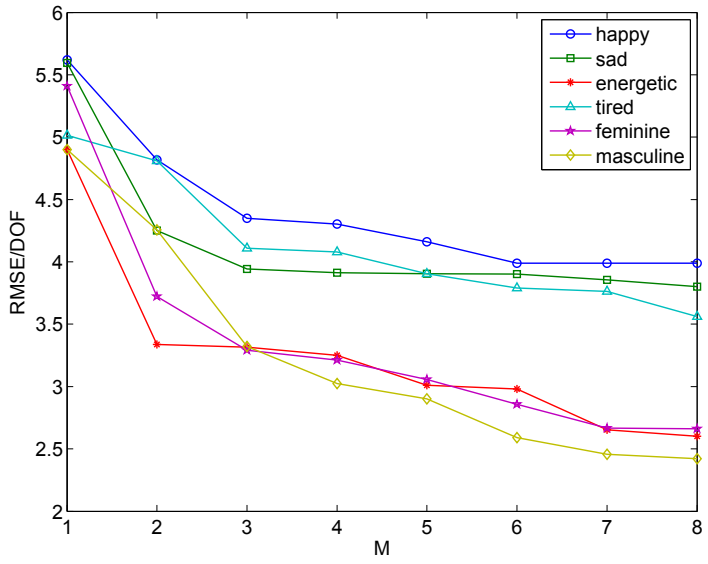


Figure 7.5. The average RMSE/DOF vs. number of RBFs used to approximate the SF sets of 15 happy, sad, young, old, energetic, and tired walking sequences.

Table 7.1. Residual error rates and audience identification results for approximated secondary themes.

	M = 1		*M = 3*		*M = 5*		*M = 7*	
	RMSE /DOF	*perc.*	*RMSE /DOF*	*perc.*	*RMSE /DOF*	*perc.*	*RMSE /DOF*	*perc.*
Happy	5.62	0.40	4.35	0.80	4.16	0.80	3.99	0.80
Sad	5.60	0.40	3.94	0.80	3.91	1.00	3.86	1.00
Energetic	4.91	0.60	3.31	1.00	3.01	1.00	2.65	1.00
Tired	5.09	0.60	4.01	0.80	3.90	0.80	3.75	1.00
Feminine	5.41	0.60	3.29	1.00	3.06	1.00	2.67	1.00
Masculine	4.90	0.40	3.32	0.80	2.91	1.00	2.46	1.00
Average	5.25	0.50	3.70	0.87	3.49	0.93	3.23	0.97

Combined numerical and perceptual accuracy of Gaussian RBFs in modeling the SFs make them perfect candidates for being used as constructs in our system. Moreover, in Chapter 8, we utilize Gaussian RBFs in a neural network setting for classification and translation of SFs. Accurate performance of these networks further validate suitability of these basis functions.

7.4 Interface for Feature Acquisition

Successive to introducing RBFs as building blocks for SFs, the aforementioned parameters (α, μ, σ^2) needed to be collected from animators. To facilitate this process, a user interface capable of generating the RBFs and adding them to different DOFs of motion in real-time was required. To the best of our knowledge, such a system did not exist at the time that this project was being conducted. Subsequently, we built a GUI for this purpose. For simple integration of the interface into the system and easier analysis of the results, the interface was developed in MATLAB using the *guide* functionality.

A snapshot of the GUI is presented in Figure 7.6. Using the system, users can browse and load motion capture data in the form of *bvh* files. The selected and loaded motion capture file can be animated. This component of the interface displays the original loaded sequences plus all the applied RBFs. A particular joint of the body can be selected. The selected joint is highlighted in red as opposed to all other joints which are displayed in blue, so that it could easily be distinguished by the user. Additionally, the axis, along which the selected joint is going to be modified also needs to be selected. If a set of RBFs have already been synthesized for a selected DOF (joint and axis), re-selecting that DOF

displays those RBFs for the user to see and modify if needed. Previously saved sets of RBFs can also be loaded and modified. For each DOF, up to 3 RBFs can be generated in Cartesian space. The three parameters for each RBF namely amplitude, location, and width (α, μ, σ^2), are assigned using sliders. Each RBF is interactively plotted as users make use of the sliders. To distinguish between the three RBFs, each is plotted with a different color. Making use of all three RBFs is not mandatory as 1, 2, or 3 of them can be used to modify a DOF, or the DOF can be left unchanged. Each set of generated RBFs can be summed to create a single curve which is displayed with a different color. The sum of RBFs can then be applied to the animated motion sequence. Upon completion of a task, the parameter sets for the generated RBFs can be saved.

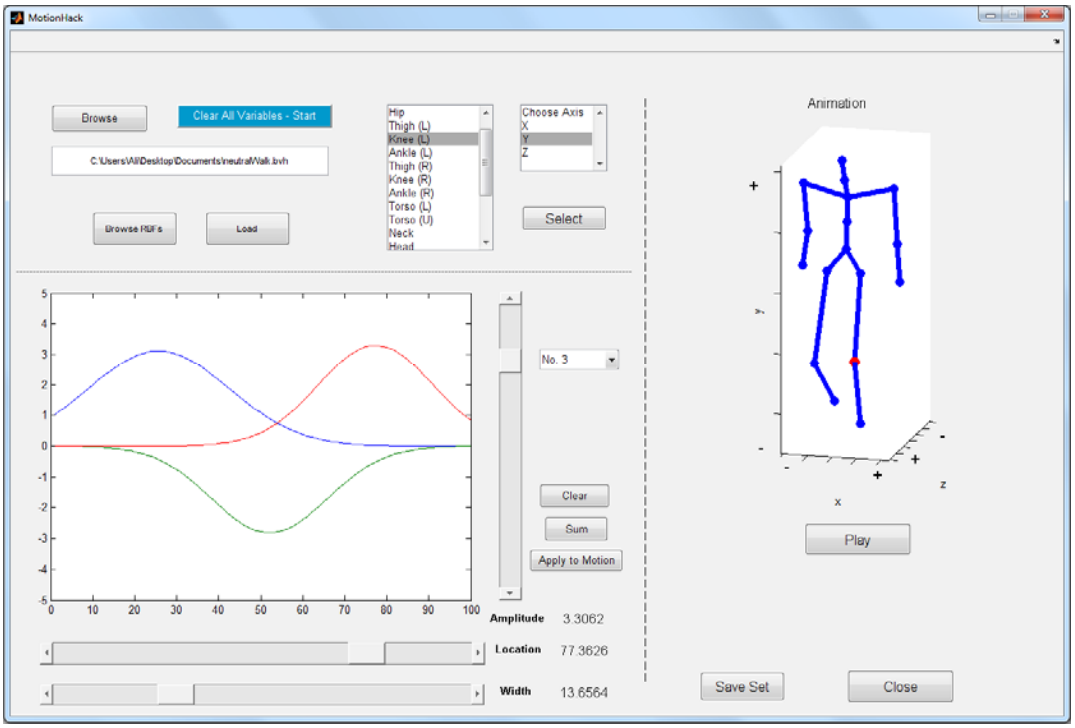


Figure 7.6. GUI used to generate SF by users. The interface is developed in MATLAB and enables loading of motion capture files, generation of RBFs, adding them to the sequence, and displaying the original or modified sequence.

7.5 Experimental Method

7.5.1 Participants and Materials

27 subjects in two groups participated in this study. For the experiments conducted with both groups, ethics approval was secured. The first group was composed of 11 participants who were experienced with animation. They were either graduate students with related experience or employees in the private sector, working for animation-related companies. Their mean age was 25.8 with a standard deviation of 4.2, 9 were males and 2 were females. They were provided with a paper-based description of the process and interface. Additional oral description was provided upon request. Their task was to convert the average neutral walk to happy, sad, energetic, tired, feminine, and masculine STs using the interface. They were first asked to practice with the interface to ensure proper utilization of its different functionalities. They were then asked to convert the loaded neutral walk into each of the mentioned STs using as many RBFs (up to 3) per DOF, and modify as many DOFs, as they felt required to complete the task. Upon conversion of the neutral walk to each of the STs, the added features were saved and the interface was reset. The order in which the STs were generated by different users was randomized to prevent arrangement side effects. The feature collection process was quite time consuming (between 90 to 120 minutes). Participants were compensated for their time. The second group of participants consisted of 16 individuals, 5 of whom were females and 11 were males. Their average age was 29.8 with a standard deviation of 10.9. They took part in providing perceptual feedback on the generated results for analyzing the different parameters involved in this study. They were naive towards human motion studies. A 7-point Likert forced-choice paper-based questionnaire was used. The

MATLAB based interface as well as the animated sequences were presented on a desktop computer with 3 GB of RAM, a 2.8 GHz processor, and a 23.6 inch 1080 HD LED screen.

7.5.2 Data Acquisition and Computation

Figure 7.7 illustrates the general process of utilizing the GUI for generation of SFs. The neutral average walk is first loaded into the GUI. Users then start by selecting a DOF (joint and axis) of the body that they think needs to be modified. Up to 3 RBFs are added to the selected DOF and the resulting animation is displayed. Once the modified motion of the DOF satisfies the animator, a new DOF is selected and the process is repeated. The animators can stop and save the parameters when they think synthesis of intended SF is achieved.

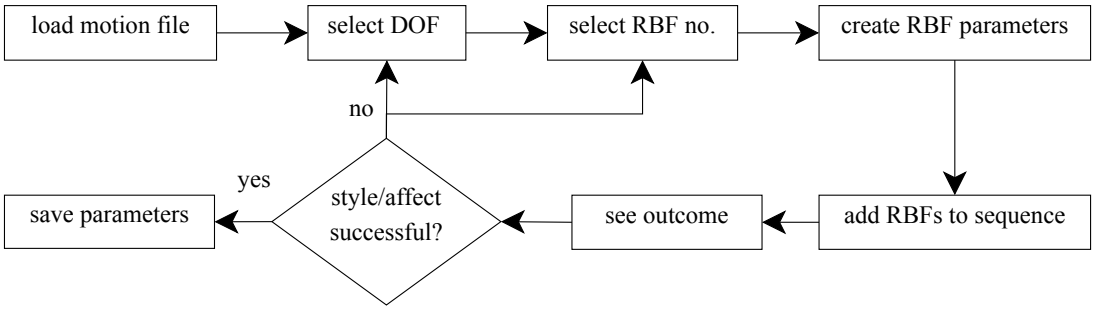


Figure 7.7. Process of using the interface for generating SF.

Based on Eq. 6.3, the feature set created by participant *i* for each SF type *j* is denoted by:

$$\mathbf{\Pi}_{i,j} = \begin{bmatrix} \{\alpha, \mu, \sigma^2\}_1^1 & \cdots & \{\alpha, \mu, \sigma^2\}_3^1 \\ \vdots & \ddots & \vdots \\ \{\alpha, \mu, \sigma^2\}_1^{54} & \cdots & \{\alpha, \mu, \sigma^2\}_3^{54} \end{bmatrix} \quad (7.4),$$

where $\{\alpha, \mu, \sigma^2\}_l^k$ are the parameters for the *l*th RBF used for the *k*th DOF. It is important

to note that some of the values in this matrix may be zero as some users may have chosen not to modify a particular DOF or not used all three RBFs. Subsequently, we sort each $\mathbf{\Pi}_{i,j}$ matrix for μ_l values in ascending format. In other words, the RBFs are temporally rearranged such that $\mu_1^k < \mu_2^k < \mu_3^k$. α_l^k and ${\sigma^2}_l^k$ are sorted along with associated μ_l^k values. For the *j*th SF, we calculate the sum of parameter matrices $\mathbf{\Pi}_{i,j}$ across the 11 participants and divide each element by number of animators (A) who made an edit using the corresponding RBF. This calculates the final feature matrix $\overline{\mathbf{\Pi}_j}$. In other words:

$$\overline{\mathbf{\Pi}_j} = \frac{1}{\boldsymbol{A}} \odot \sum_{i=1}^{11} \left(\mathbf{\Pi}_{i,j}\right)_{sorted} \tag{7.5},$$

where $\odot$ denotes element-wise multiplication and $\boldsymbol{A} = \begin{bmatrix} A_1^1 & \cdots & A_3^1 \\ \vdots & \ddots & \vdots \\ A_1^{54} & \cdots & A_3^{54} \end{bmatrix}$ is the animator recurrence matrix. $\frac{1}{\mathrm{A}}$ is an element-wise inversion of A. If an element of A is 0, meaning a particular DOF has not been edited, the corresponding element of $\frac{1}{\mathrm{A}}$ would be meaningless. Therefore, we substitute that element of $\frac{1}{\mathrm{A}}$ with 0. Using the recurrence matrix ensures that only modified DOFs and utilized RBFs are included in the averaging process, and not all values are divided by 11. This is done to preserve the features added to all DOFs regardless of the frequency of use. The calculated parameter matrix for each theme is used to analyze the type and purpose of different features added to different DOFs of the motion sequence.

7.6 Results

7.6.1 Validating the Input

We first validated the input sequence. Table 7.2 presents how participants perceived the average neutral walk. High recognition of the sequence as *neutral* points to high perceptual quality of the input. Moreover, the false classifications being distributed across different themes points to the absence of a particular dominant SFs in the input.

Table 7.2. Validation of the input neutral walk.

	Neutral	Happy	Sad	Energetic	Tired	Feminine	Masculine
% perceived	75.0	6.25	6.25	0.0	6.25	0.0	6.25

7.6.2 Distribution of Features across the Body

The animators were not asked to modify all DOFs of the motion sequence. As a result, only the DOFs perceived by the animators to be more essential towards generation of STs were modified. Figure 7.8 illustrates histograms of the number of animators with respect to different sections of the body.

We observe that except for sadness, in which all animators modified the shoulders and neck, no other ST type and section of the body drew that kind of consensus. For this theme, head, upper torso, elbows, and wrists followed. For generating the happy, energetic, and tired themes, shoulders drew the most attention. Wrists, knees, neck, head, and elbows followed. For energetic, elbows, wrists, knees, and head come next. In generating the tired theme, ankles, neck, head, and wrists come after the shoulders. For feminine, hip, thighs, and ankles are modified the most, followed by the knees and lower torso. Finally, for masculine SF, knees, ankles, shoulders and elbows drew the most

attention. Interestingly, no body parts of any of the STs were left unmodified.

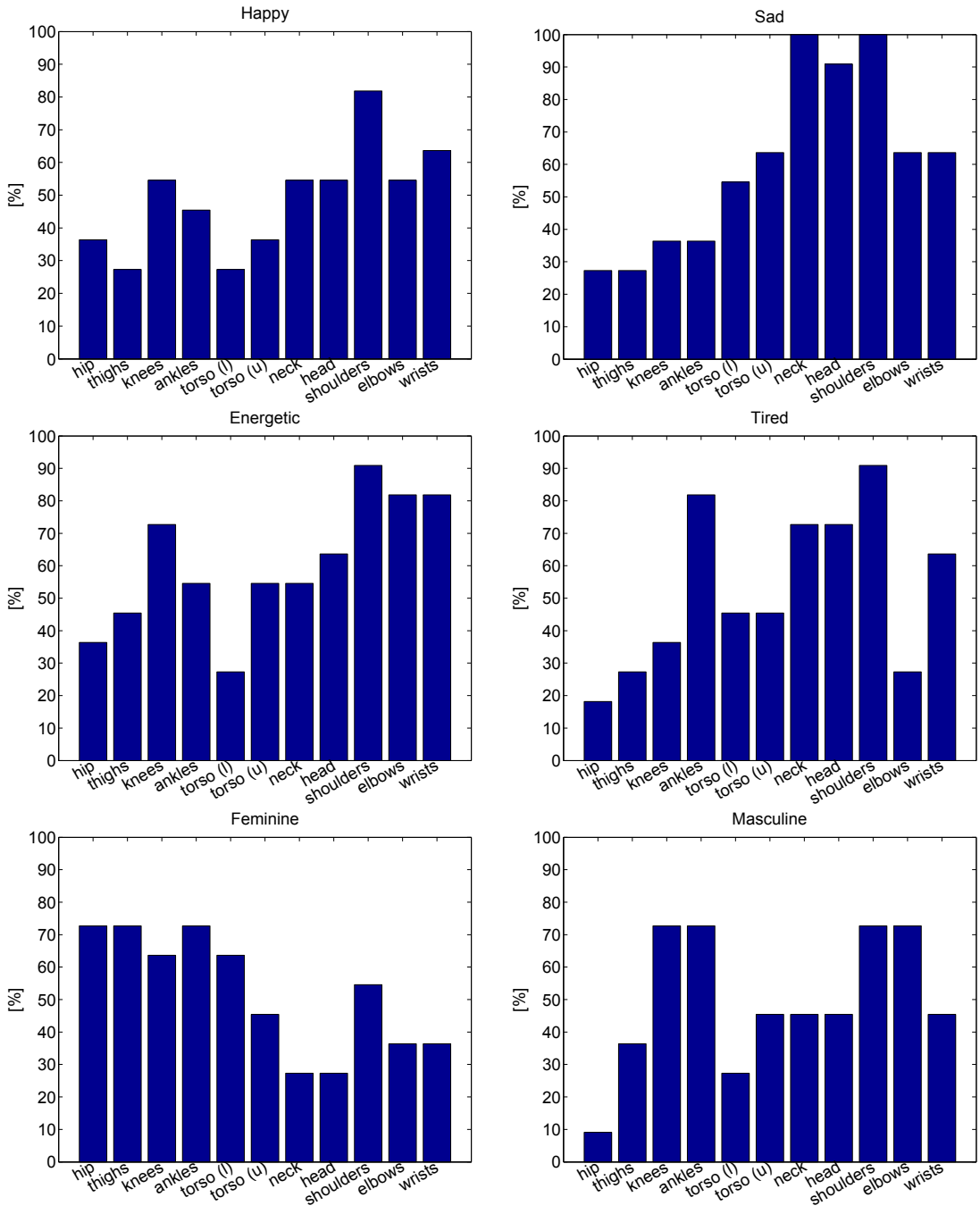


Figure 7.8. Number of animators that modified each body part.

For the six themes combined, one-way analysis of variances (ANOVA) indicates significant effect for body joints ($F(10,55) = 3.02$) at the $p < 0.01$ level indicating that

some body parts are modified by experienced animators more than others. Interestingly, the ST showed to be an insignificant factor ($F(5,60) = 0.62$), indicating similar attention is paid in generating different STs. This indicates that while within each theme, different joints require different amounts of modification, generally, the animators have not favored any particular theme over others. From a different perspective, we observe that for happy, sad, energetic, tired, and masculine the distribution of added RBFs per body part weighs in favor of upper-body regions, with sadness having the most upper-body to lower-body advantage. The only exception of the six categories is feminine in which the lower-body drew more attention compared to upper-body regions. In terms of left/right half of the body, animators chose to treat shoulders, elbows, wrists, thighs, knees, and ankles (which are composed of left and right sides) in symmetric fashion, meaning in all cases where these joints were altered, both left and right joints were modified.

7.6.3 Feature Properties

Common Shapes: The sets of RBFs created and used by the animators for modifying different DOFs of the motion can be categorized into 4 major types or shapes. Figure 7.9 illustrates these 4 commonly used features. Features type 1 and 2 are composed of a single RBF while feature types 3 and 4 are composed of 2. In the 2-RBF shapes, the blue and green curves show the individual RBFs while the red represents the sum. Feature types 1-1, 2-1, and 3-1 only contain $\alpha > 0$, types 1-2, 2-2, and 3-2 only contain $\alpha < 0$, and type 4-1 and 4-2 are composed of a combination of both $\alpha > 0$ and $\alpha < 0$. While different shapes were also observed in the results, such features were non-recurring and as outliers, we did not take them into account.

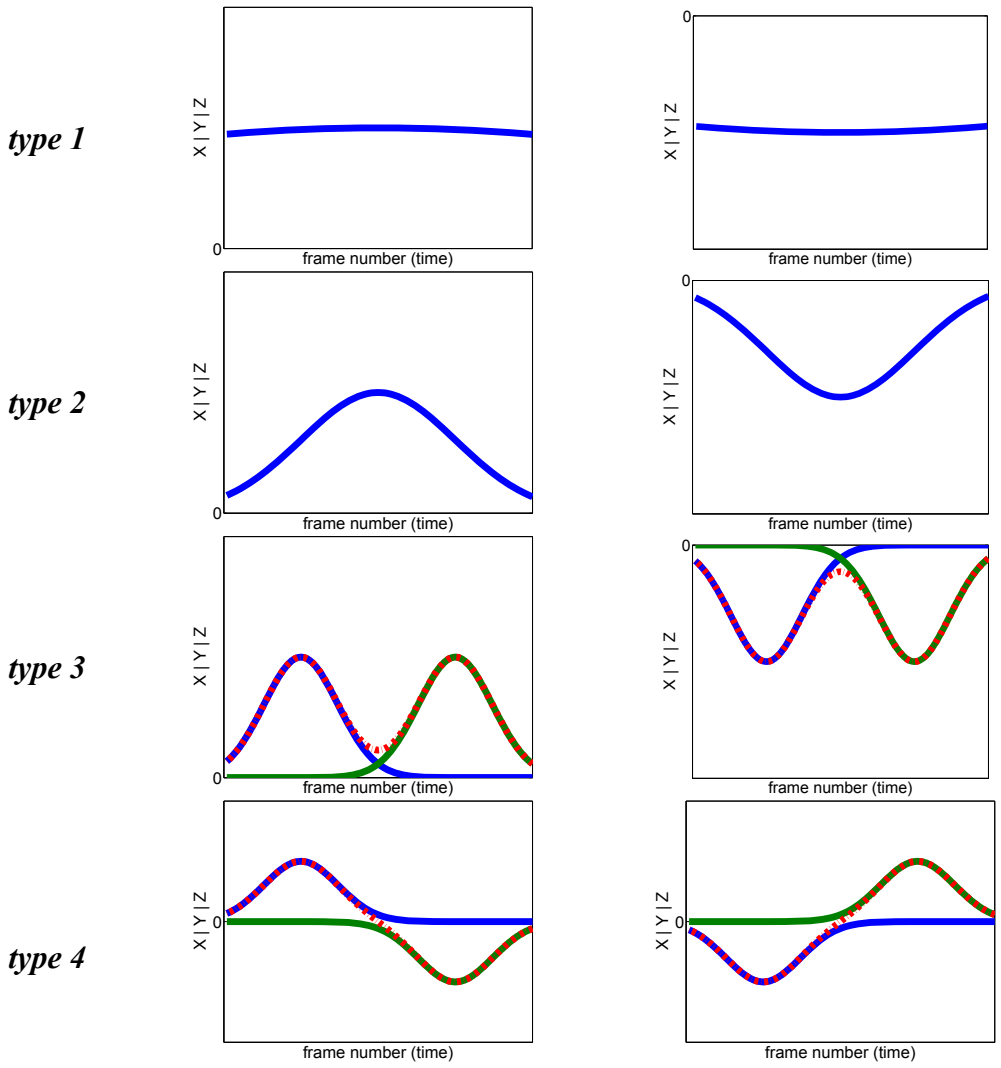


Figure 7.9. Commonly used feature shapes.

In general, feature type 1 is utilized to add a spatial offset to a motion trajectory of a particular DOF. These features often modify the body posture and not movement trajectories. An applied example of this is tilting down of the head. Feature type 2 is utilized when the animator intends to change the movement of a particular joint by increasing and subsequently decreasing it during a certain time frame, or vice versa. In other words, this feature can be used to add local maxima/minima where none exists, or to increase/decrease the amplitude of existing minima/maxima. For example, this feature

can be used to increase the movement of the right elbow. Similarly, feature type 3 is used to repeat this process at two different points. For example, in a 2 step walk, like the one being used as the input in this study, feature type 3 can be used to add two forward sways to the head, one with each step. Finally, feature type 4 is utilized to add opposing extrema to a trajectory. For example, it can be used to add sway to the hip joint where with the right step, the hip shows increased movement towards the left, and with the left step, it sways towards the right. As mentioned earlier, for shoulders, elbows, wrists, thighs, knees, and ankles where we have left and right joints, the two versions of a feature are often used. In other words, if feature type x-1 is used for a particular left joint, feature type x-2 (the negated version) is often used for the right joint, and vice versa. This is due to the symmetrical nature of walking and will not necessarily apply to other types of actions.

Posture vs. Movement: The use of feature type 1 vs. the other three types is of significance and worth further analysis. As described in Chapter 6, in addition to the uniform time feature, SFs are composed of two spatiotemporal components, namely posture and movement. To recap the notion, posture features are those that often stay unchanged through the sequence. In effect, they are changes to the initial posture of the body with which the motion is carried out. Movement features are the changes to the motion trajectories and vary throughout the sequence. The method in which the majority of DOFs are modified for a particular ST to be generated, can provide insight on how related features are presented in human motion. Table 7.3 presents the percentage of DOFs of the neutral walk that have been modified using posture features (type 1) vs. movement features (type 2, type 3, and type 3). We observe that sad and tired sequences

which are in fact quite similar are mostly modeled with posture features while the rest are mostly created by movement features. The maximum relative percentage of movement features is utilized for energetic while the maximum relative percentage of posture features is used for tired walk.

Table 7.3. Percentage of movement vs. posture based features used by animators to generate different themes.

	Happy	Sad	Energetic	Tired	Feminine	Masculine
% Movement	71.4	24.5	86.0	22.9	59.0	67.3
% Posture	28.6	75.5	14.0	87.1	41.0	32.7

Frequent Features: Table 7.4 presents the 10 most frequently synthesized features and their effect on the neutral input motion. As discussed, two general feature types, posture and movement, are used, which we refer to in the table as *tilted along* and *increased/decreased swing along* particular axes respectively.

Utilizing all of the features mentioned in Table 7.4, and other less frequently used ones that are not mentioned in this table, results in successful generation of the intended STs. However, we suggest that generating STs can be achieved using only few of these features, or in other words perceptual shortcuts.

Figure 7.10 illustrates the absolute values of the maximum of the sum of RBFs used to create the features in Table 7.4. The numbers are normalized by the height of the average actor from the input sequence. Normalization is carried out since the height of the walker can significantly affect the magnitude of added features. For example, increased arm swing for a tall actor is greater than that of a short one.

Table 7.4. Top 10 frequently used features for generation of the STs.

		ST					
		Happy	*Sad*	*Energetic*	*Tired*	*Feminine*	*Masculine*
Common features	*1*	Shoulders: increased swing along Z	Shoulders: tilted along –Y	Knees: increased swing along Y	Shoulders: tilted along –Y	Hip: increased swing along X	Shoulders: increased swing along Z
	2	Wrists: increased swing along Z	Head: tilted along –Y	Head: tilted along +Y	Head: tilted along –Y	Ankles: tilted, R along +X, L along –X	Knees: tilted, R along –X, L along +X
	3	Knees: increased swing along Y	Neck: tilted along –Y	Elbows: increased swing along X	Ankles: decreased swing along Z	Knees: tilted, R along +X, L along –X	Ankles: tilted, R along –X, L along +X
	4	Head: increased swing along X	Shoulders: decreased swing along Z	Elbows: increased swing along Z	Head: tilted along +Z	Torso L: increased swing along X	Elbows: tilted, R along –X, L along +X
	5	Wrists: increased swing along X	Wrists: decreased swing along Y	Shoulders: increased swing along Y	Wrists: tilted along –Y	Thighs: tilted, R along +X, L along –X	Ankles: increased swing along Y
	6	Hip: increased swing along X	Torso U: tilted along –Y	Shoulders: increased swing along Z	Ankles: decreased swing along Y	Torso U: increased swing along X	Knees: increased swing along Y
	7	Knees: increased swing along X	Head: tilted along +Z	Wrists: increased swing along Z	Neck: tilted along –Y	Thighs: increased swing along Y	Head: increased swing along Z
	8	Neck: tilted along +Y	Neck: tilted along +Z	Wrists: increased swing along X	Neck: tilted along +Z	Shoulders: tilted along –Y	Neck: increased swing along Z
	9	Head: tilted along +Y	Elbows: tilted along –Y	Thighs: increased swing along Z	Torso L: tilted along +Z	Elbows: increased swing along X	Shoulders: increased swing along Z
	10	Head: tilted along –Z	Hip: tilted along –Y	Thighs: increased swing along Y	Torso U: tilted along +Z	Wrists: increased swing along X	Elbows: increased swing along Z

Figure 7.10 shows that the most frequent features do not necessarily have the highest intensities. We observe that, increased wrists' swing along Z in happy, decreased ankles'

swing along Z in tired, and increased ankles' swing along Y in masculine, have the greatest intensities. This observation is logical since movement features are spatiotemporally (but not necessarily perceptually) dominant with respect to posture features. As a result, their alteration with the aim of creating the ST would require stronger RBFs. Moreover, the three mentioned features belong to wrists and ankles, which generally show the most spatiotemporal movement in walking [81]. The features described in Table 7.4 are computed based on the summary of edits (as described in Section 7.5.2). While the general shapes can be described based on the 4 illustrated categories (Figure 7.9), in rare instances, small variations (such as slight temporal shifts) are observed. The numbers in Figure 7.10 present the shapes (based on Figure 7.9) that most resemble the top 10 features of each theme. Moreover, 1 denotes that the feature is in the form of posture while 2, 3, and 4 indicate movement.

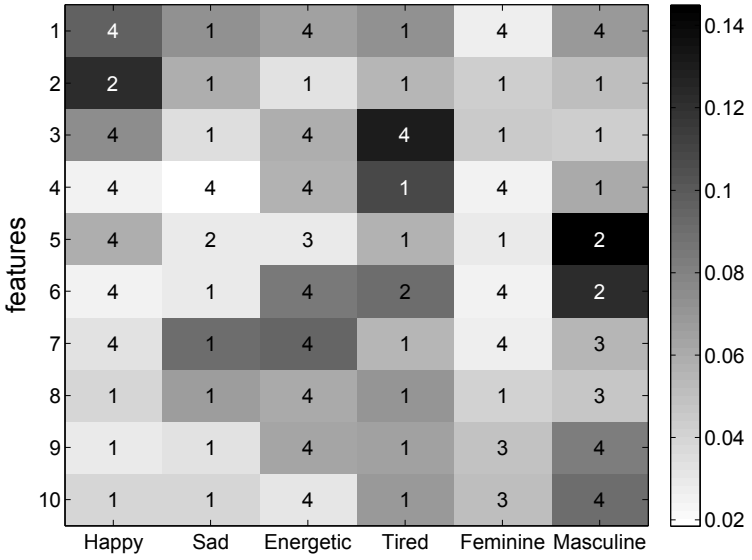


Figure 7.10. Normalized intensities of features from Table 7.4. The numbers in the cells represent the shape of feature based on Figure 7.9 (some numbers appear in white to maintain readability with respect to their dark backgrounds).

7.6.4 System Parameters and Perception

As mentioned in Section 7.5.1, to evaluate the performance and parameters of the proposed method, a second group of subjects participated in this study. The same group provided the information required for validating the input neutral sequence (Section 7.6.1). They were asked to rate the amount of affect or style in several displayed sequences that were generated using the system. The sequences were generated when 1, 4, 7, and 10 features were used. The weight parameter w from Eq. 5.9 was another variable in this experiment with the values of 0.4, 0.7, 1.0, and 1.3. Video Clip D illustrates the generated sequence when 10 features are used. The weight applied to these sequences is $w = 0.7$. Figure 7.11 presents frames from this video. Subjectively, high quality animation is achieved.

Figure 7.12 illustrates the perception results. Error bars represent standard errors ($SE = SD/\sqrt{sample\ size}$). Generally, a direct relation between the two parameters and the perceptual ratings of affect and style is observed.

We performed 2-factor ANOVA for weight × number of features. Based on the results presented in Table 7.5, for all themes, both weight and number of features show significant effect at the $p < 0.0001$ level as viewers perceive higher levels of affect or style as the number of features or the weights increase. There is no significant interaction between the two variables, indicating that the two variables do not affect one another. In other words, the weight variable influences perception with 3 features almost in the same way that it influences perception with the use of all 10 features. The same can be said about the effect of number of features on the weight variable.

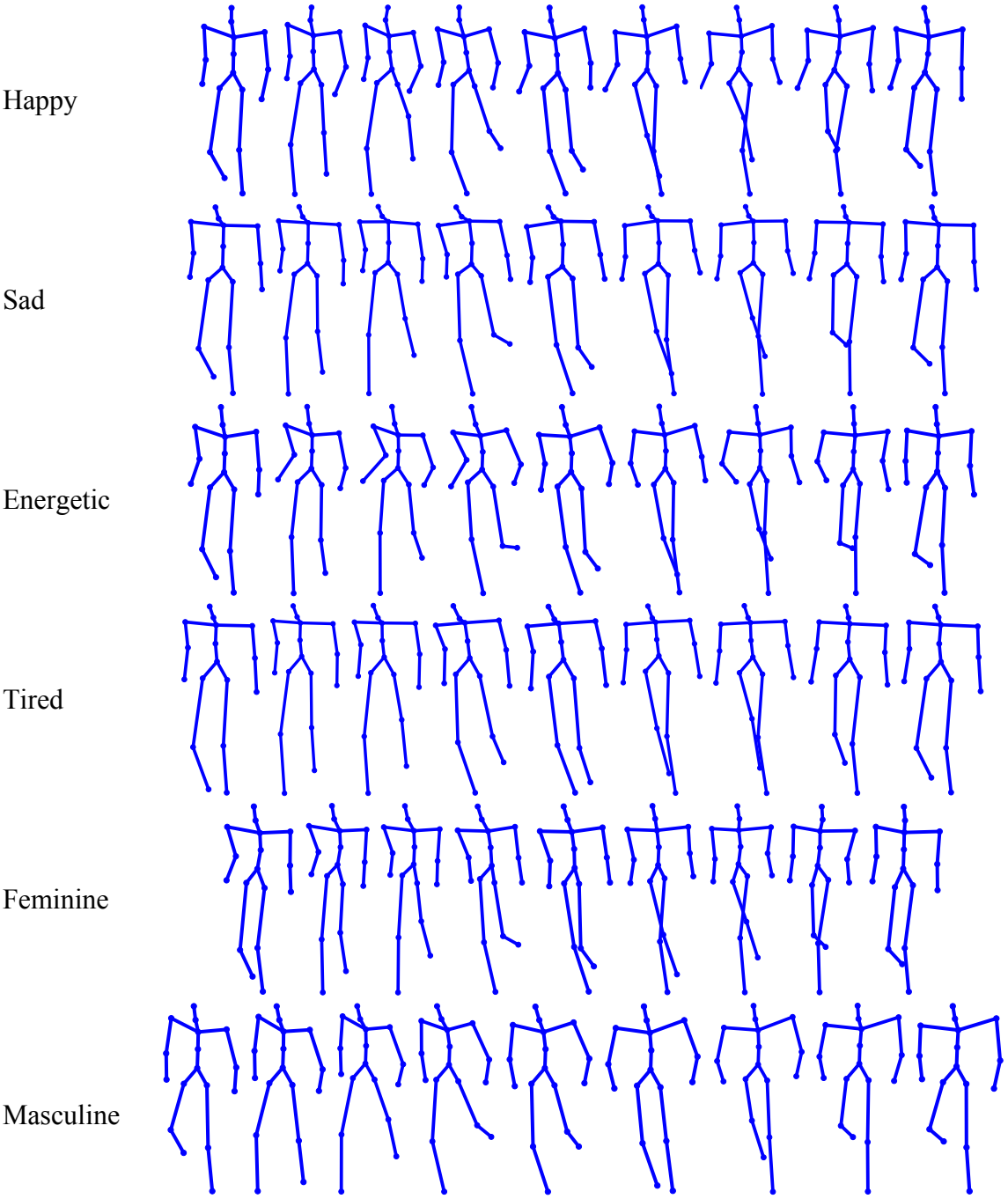


Figure 7.11. Frames from generated affective/stylistic sequences. 10 features are used and the applied weight is $w = 0.7$ (from Video Clip D).

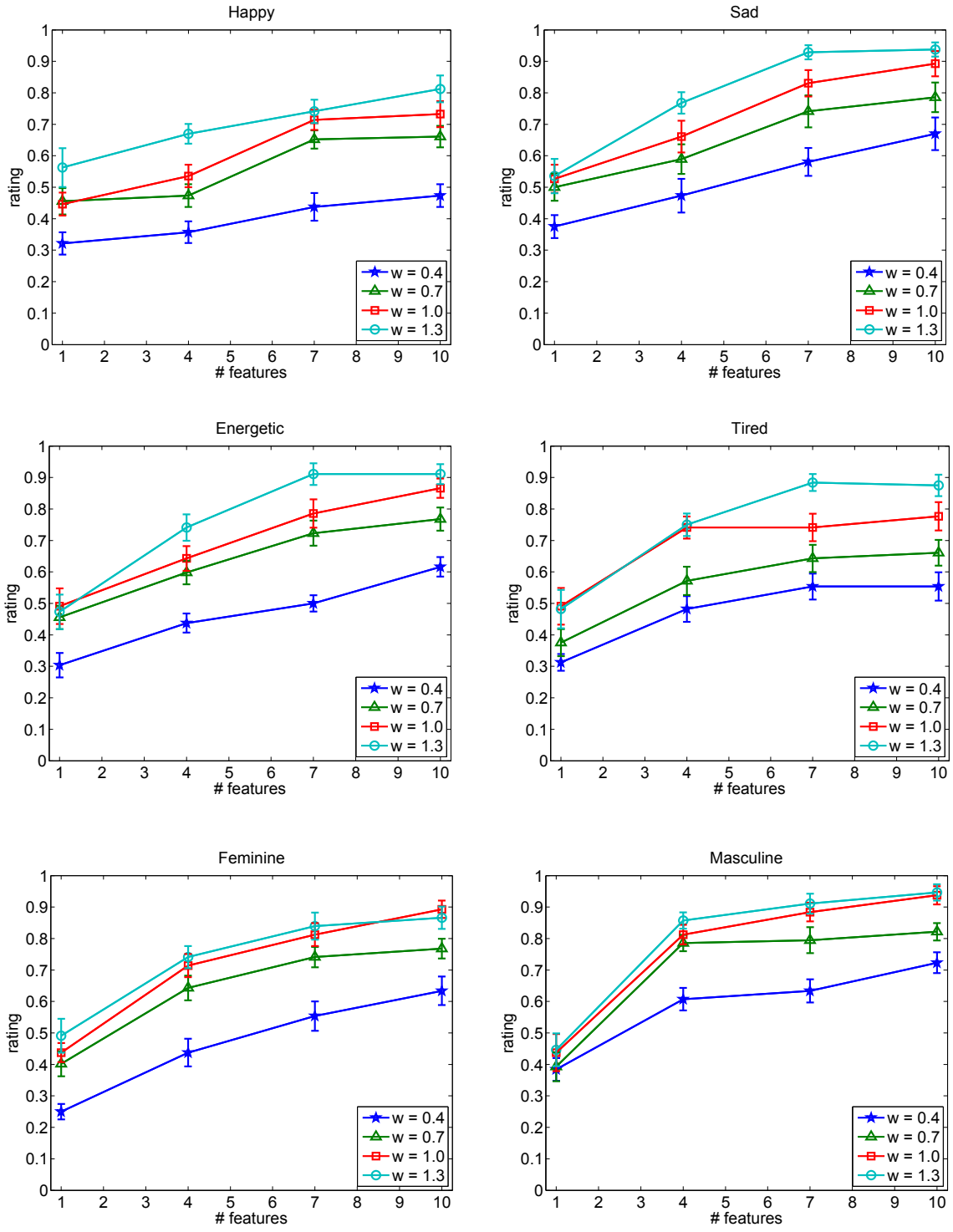


Figure 7.12. Perceptual ratings for the two variables, weight and number of features, used to generate affective/stylistic themes. Error bars represent standard errors.

Table 7.5. Two-way ANOVA results for weight and number of features as the two independent variables. *n.s.* denotes not significant. Weight and number of features show significant effect on perception while there is no interaction between the two.

		Factor					
		Weight		Number of features		Interaction	
		$F(3,240)$	$p <$	$F(3,240)$	$p <$	F(9,240)	$p <$
ST	Happy	29.38	0.0001	41.73	0.0001	0.82	n.s.
	Sad	48.33	0.0001	27.54	0.0001	0.73	n.s.
	Energetic	65.89	0.0001	42.01	0.0001	1.25	n.s.
	Tired	43.28	0.0001	33.17	0.0001	0.77	n.s.
	Feminine	83.33	0.0001	39.67	0.0001	0.36	n.s.
	Masculine	119.71	0.0001	24.79	0.0001	1.6	n.s.

7.6.5 Inversion

To further investigate the system, we tested negative w values. Surprisingly, we noticed the appearance of features belonging to the opposite themes. For example, when a negative weight was applied to happy features, the output sequences appeared as sad. Sequences obtained with $w = -0.7$ are illustrated in Video Clip D and Figure 7.14.

We further tested this concept by applying a weight of $w = -0.7$ to all the 10 features and collecting perceptual ratings. The audience were asked to select a rating between –7 to +7, with +7 denoting the maximum original theme, 0 denoting neutral, and –7 denoting the maximum opposite theme. The results are presented in Figure 7.13 where the audience perceived the opposite theme in all cases. For some themes, this inversion effect seemed to be stronger. This inversion effect was especially stronger in masculine, happy, and feminine compared to energetic and tired themes.

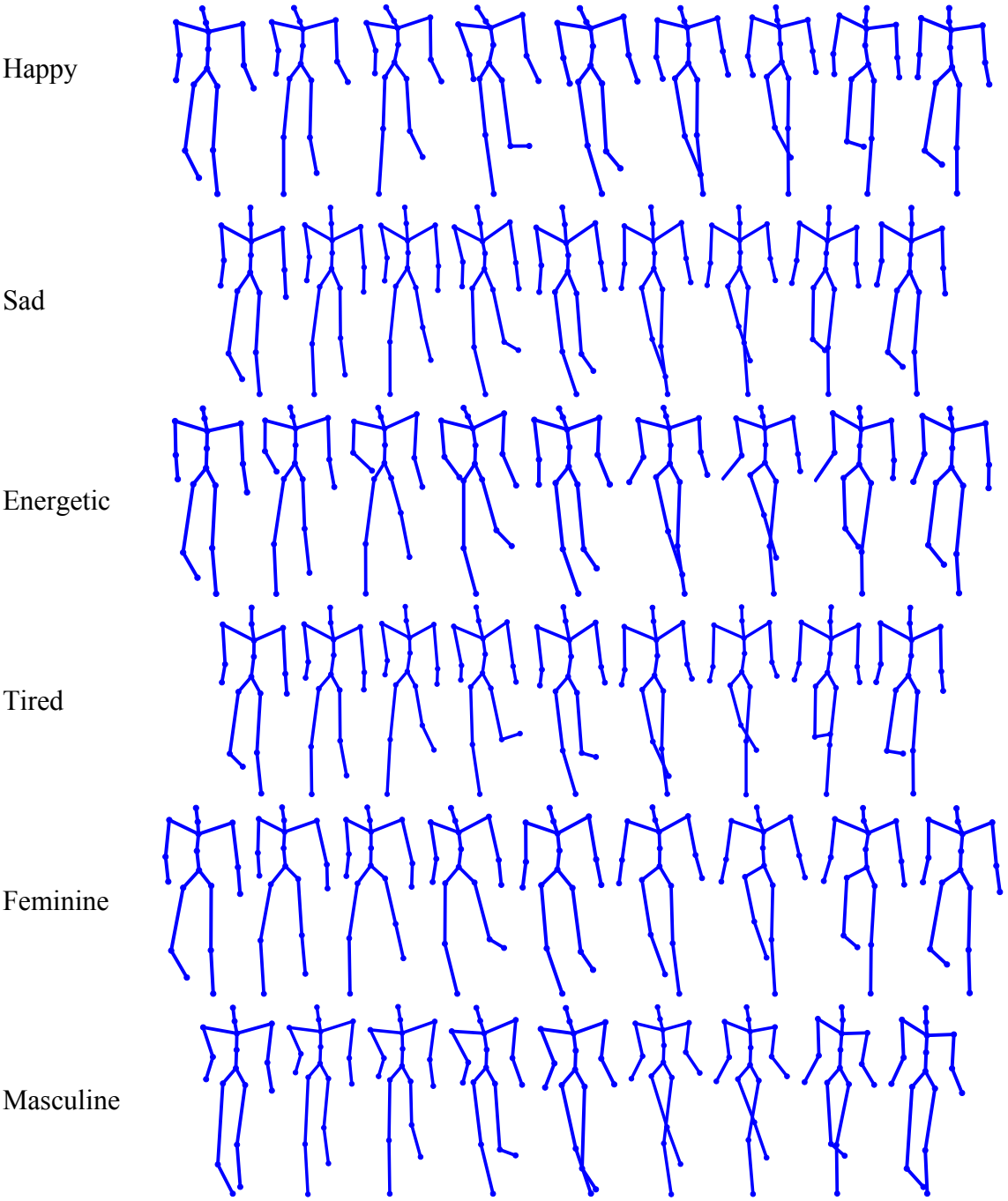


Figure 7.13. Frames from the sequences generated with 10 features and w = – 0.7 (from Video Clip D). Features belonging to opposite themes have appeared due to the negative weight factor.

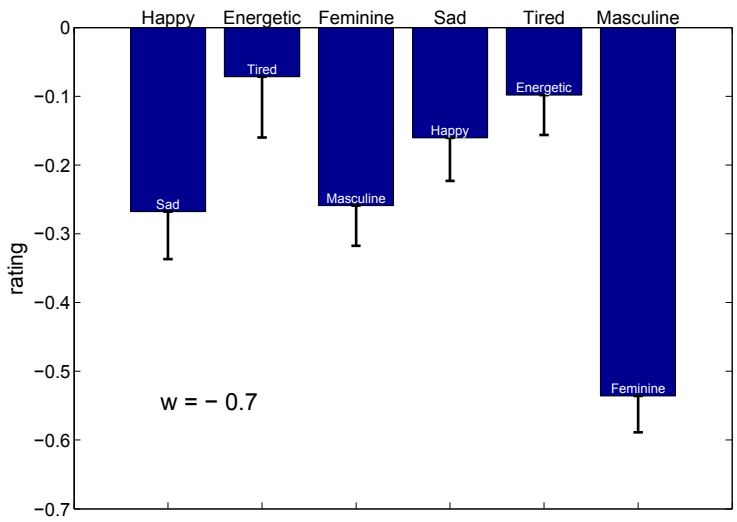


Figure 7.14. Perceptual ratings for negative weights show theme inversion.

7.6.6 Feedback

After data acquisition was concluded, the goal and methodology of the project was described for the animators and feedback was acquired. The animators were asked to rate, on a 7-point scale, the potential usefulness of findings and implications of this study in terms of motion perception and performance. They were also asked to rate the potential for an autonomous system developed based on this study. Our method scored an average 74.0% for the former and 75.3% for the latter. This feedback was especially valuable as the participants providing the feedback were experienced animators.

In regards to the interface itself, several animators mentioned that the RBF-based approach is “unusual”. This is because they were accustomed to free-form transformations for motion editing. The use of RBFs however, was not perceived as a limitation. It would have been possible to use and record free-form edits and subsequently model the features using a sum of Gaussian RBFs. This approach, however, would have approximated the features, and residues would have naturally been left

unaccounted for, whereas our approach yields results in the form of perfect RBFs. The GUI was simple to use as animators tuned each RBF using only three sliders. We encountered no problems during the data acquisition process after the short (~10 min.) training period. We had in fact carried out pilot studies and tests before the data acquisition and the system had gone through several rounds of improvement.

7.7 Discussion

7.7.1 Features and Perception

Towards analyzing the use of posture vs. movements features, we draw on the notion of activation and pleasantness in motion. To review this concept, we refer the reader to Russell's model or circumplex for affect, which is based on activation and pleasantness as its two axes [53]. This model only discusses affects and not energy or gender. However, we can assume that, despite energetic actions often being associated with pleasant moods, an exclusively energetic action does not necessarily vary on the pleasantness axis while being positive on the activation axis. The same can be said about a tired action where it is negative on the activation axis and does not convey any information in terms of pleasantness given a happy tired person can easily be realized. Similarly, for feminine and masculine, pleasantness does not play a role, and a particular judgment on activation is hard to make. However, due to exaggerated features, as well as being associated with faster gait cycles [64], slightly positive activation can be assumed for gender-related themes. Figure 7.15 illustrates our modified version of the model based on which the dominance of posture or movement with respect to one another can be estimated. In Table 7.3, we observed that happy, energetic, feminine, and masculine

variations are dominant in the use of movement features while sad and tired variations make more use of posture features. Accordingly, we can conclude that generally, themes with high activation are mostly generated using movement features whereas in themes that are low on activation, posture has a more crucial role. From a different standpoint, we can argue that activity requires movement. It is therefore logical for themes associated with higher activity to be associated with and represented by movement features. Where motion appears inactive, on the other hand, it is logical for the alternative type of feature, i.e. posture, to take the dominant role. This concept is shown in Figure 7.15.

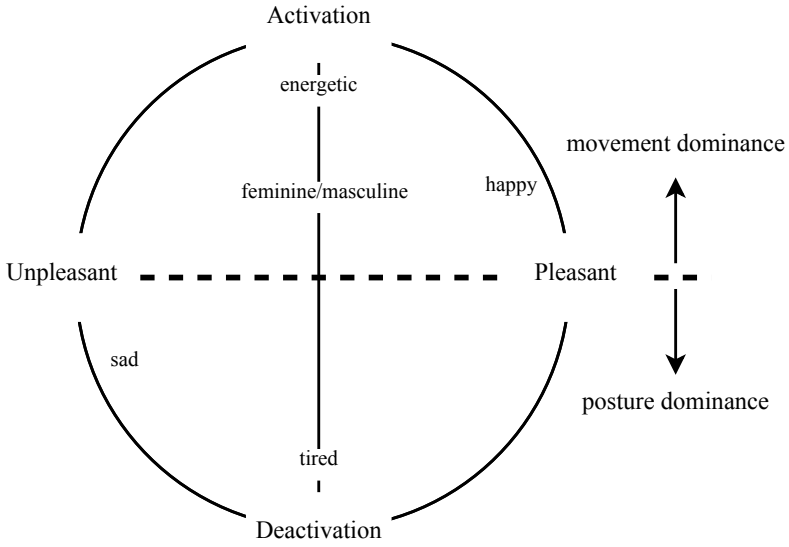


Figure 7.15. A modified version of Russell's model [40] that describes gender and energy related themes along with happiness/sadness.

Regarding the shapes of synthesized features, it is observed that all 4 shapes (Figure 7.9) contain some sort of symmetry. It should be noted that while Gaussian RBFs are symmetric in nature, asymmetric features could have been synthesized using a sum of

RBFs, and therefore could have been used if necessary. However, asymmetric features were almost never perceived required by the animators, perhaps due to the symmetric nature of the input, i.e. two-step walk cycle. It is also possible that the features were perceived as a combination of small asymmetric components plus dominant but symmetric ones, thus disregarding the asymmetric components of the features.

There have been many studies in the literature attempting to describe features responsible for presentation and perception of STs in motion. For example, it was observed in [81] that the following features are present in happiness: inclination of shoulders, lateral movements of hand/arm, arms being stretched out frontal, arms being crossed in front of chest, opening/closing of hands, and sideway positioning of back of hands, while for sadness: collapse of the upper body, downward bending of the head, sideways bending of the head, lateral movements of hand/arm, and sideway positioning of back of hands were reported. The study also suggests that in the happiness, more movement features are present compared to posture features. These features do not contradict our findings and proposed feature sets. Identical features in both happy and sad themes reported in the mentioned study, however, seem unusual and the number of reported features are low, especially as those related to legs and feet are ignored. Nevertheless, the latter is understandable, given the reports that suggest that emotions are mostly conveyed by the upper body [19]. We too have shown that features most frequently used by animators for generating happiness and sadness are mostly present in the upper body joints. In [225], it was illustrated that head inclination (posture) is central for sadness and that otherwise, movement features are dominant for emotions. In [79], regression analysis and PCA were used to analyze emotions in recorded sequences. It was reported that posture features are

more dominant in sadness compared to happiness. In terms of movement, for happiness, shoulders showed the most change followed by elbows, while for sadness, the same order was observed followed by the hips. Using PCA for happiness, elbows displayed the only significant change while for sadness, the knees were dominant followed by hips. For energy-related features, we were unable to find suitable studies that investigate motion features. For gender-related features, Troje illustrated that movement-related features are generally more critical compared to posture-related ones [73]. Increased hip sway in feminine and head sway in masculine walks have been suggested in [62]. It has been illustrated that body sway and lateral movements are dominant compared to posture changes and that increased sway of hip for feminine and shoulders for masculine are dominant features [74]. The analysis based on our method reconfirms and expands these affect and gender related studies.

In regards to the study on the impact of the number of perceptual shortcuts and w in Section 7.6.4, as the number of features increase, the perceptual ratings increase as well. In most cases, the increase in the ratings is more evident in the first few number of features and approaches an asymptote as the number of features grows. This asymptote corresponds to the maximum achievable average rating, in most cases approximately 0.9. The ratings are a bit lower in the happy theme with a maximum of approximately 0.8. The reason for this asymptote is the fact that the features are arranged in the order of most use, i.e. perceptual significance. As a result, higher order features are less important and convey less information regarding the themes. Similarly, the ratings increased with the increase of w. In most cases, the difference between $w = 0.4$ and $w = 0.7$ is greater than the difference between other consecutive pairs of weights, for example $w = 1.0$ and

w = 1.3. This property can be associated with the ratings of larger weights approaching the maximum possible value, i.e. 1.0. The 16 participants, however, indicated a decrease in visual quality for w = 1.3 and in some cases, even for w = 1.0. This is due to the fact that the group of animators sometimes exaggerated when asked to generate the features. As a result, we opt for weights of approximately 2/3 for animation purposes. For example, w = 0.7 was used in Video Clip D and Figure 7.11. In addition, for analysis of the inversion effect, the negative amount for the same weight was utilized.

Regarding the loss of visual quality for high values of w, an interesting observation was that as w approached values greater than 1, the audience stressed more loss of visual quality for happy, energetic, feminine, and even masculine themes compared to sad and tired. The reason for this can be the fact that most features used for the former themes are movement-based. Exaggerating movement features results in unbound and unnatural motion, which can become distracting and unappealing. Posture features, on the other hand, when exaggerated, do not seem as unappealing as movement features since they only bring about changes in the structure.

7.7.2 Time

A point to consider is that the synthesized affective/stylistic actions in this study were speed-matched with the neutral input. In other words, a time control unit was not embedded in the GUI for animators to use. Generally, there are two types of time features: uniform and non-uniform (also referred to as non-linear). We argued in Chapter 6 that non-uniform temporal features are manifested as changes in movement. Let us assume that we alter a motion trajectory (or sequence) such that its first half is linearly compressed by x frames and its second half is linearly stretched by x frames. While the

overall length of the resulting trajectory is preserved, there have been non-linear temporal modifications applied. These non-linear modifications can be interpreted as a spatiotemporal curve, i.e. movement feature. This version of time feature has been taken into account since movement changes in motion have been collected and analyzed. A uniform or linear time feature, on the other hand, is available when the entire trajectory is linearly stretched or compressed to achieve a new length. This type of time feature is simple to calculate or even estimate. Stylistic and affective variations of motion that are positive on the activation axis of the presented model in Figure 7.15, are often faster (shorter) compared to the neutral version, while lower activation motions are slower (longer). Moreover, it has been previously documented that viewers can easily recognize emotions of speed-matched affective sequences [79] as posture and movement features alone provide sufficient cues for perception. We can therefore conclude that the set of features derived in this study, or a subset of them, can successfully be employed to synthesize scalable affective/stylistic features, and speed alterations can subsequently be applied to the derived sequences using linear operations such as uniform time warping.

7.7.3 Inversion

The illustrated inversion effect can have significant implications for psychophysics as well as multimedia applications. Similar effects have been previously addressed in the literature. For example, Barclay et al. [64] illustrated that a feminine walker is perceived as male, and vice versa, when the stimuli is inverted. Our approach slightly differs, however, since: (*i*) the affect/style features alone are inverted rather than the entire sequence (along with the action); (*ii*) features of each joint are inverted along their local axis and not the global axis. While deriving a detailed and accurate neural or

psychological model that can describe this phenomenon requires in depth study of the brain functionality, we can speculate that the existence of *some* opposite features in opposing themes could be one of the possible explanations (see Table 7.4). This would especially seem sensible for opposing posture features, rather than movement features. For example, in energetic and tired themes, we have tilted head along +Y and tilted head along –Y respectively. These features will convert to one another should they be spatially inverted. For increased/decreased swing (movement) features, more exploration of the underlying reasons for the observed inversion effect is required.

7.7.4 Generalization

One of the benefits of the proposed set of features is that it uses notions of increased/decreased swing and tilt, which are descriptions derived from the mathematical RBF-based feature set. The use of only two general and time-independent features increases the possibility of generalizing the models to actions other than walking. Nevertheless, the precise values of the parameters (α, μ, σ^2) will most likely need adjustment. As an example, let us assume an existing neutral sequence of jumping jacks and consider applying the first three features for each theme. To convert the sequence to happy, increased shoulder swing along Z, increased wrist swing along Z, and increased knee swing along Y would all contribute towards accomplishing the goal. Similarly, tilted shoulders, head, and neck along –Y would contribute to sadness. Increased knee swing along Y, tilted head along +Y, and increased elbow swing along X would increase the perception of energy, while tilted shoulders and head along –Y and decreased ankle swing along Z would contribute to a low-energy theme. Finally, increased hip swing along X and tilted ankles and knees (right along +X and left along –X) would contribute

towards femininity while increased shoulder swing along Z and tilted knees and ankles (right along –X and left along +X) would increase masculinity. The exact location, width, and amplitudes of the RBFs need to be customized towards jumping jacks instead of walking. One should note, however, that these adapted features, despite contributing towards the particular theme, might be sub-optimal. This means that animators as well as a viewing audience might indicate a different and more optimal order for the arrangement of these features. Nevertheless, different action classes can easily be uploaded to the interface and animators can define the features necessary for creating designated variations. The collected data can then be summarized to produce a specialized feature set. For more complex actions, however, three RBFs may be insufficient, and so, it might be required to apply small changes to the interface.

7.8 Summary

The summary of the significances of our proposed method in this chapter, are as follows:

- There have been many studies on the subject of motion affect/style presentation and perception, which are mostly based on recorded sequences. Our method, on the other hand, is based on opinions of experienced animators, making the findings efficient while being very effective.

- To the best of our knowledge, ranking or relative significance for the reported features in affective/stylistic motion is not available in the literature to the depth presented in this chapter. This property allows for the most important features to be utilized for effective and efficient synthesis of affective/stylistic features.

- Our proposed system presents a mathematical and tunable basis for creating affect/style features and the model allows for scaling of the reported features.
- The proposed method and set of features are simple and intuitive.
- The proposed method is computationally inexpensive for practical utilization.
- Utilizing only a subset of the proposed set of features leads to high perception ratings.
- Our approach is simple to expand for other classes of affect/style as well as actions.
- Analysis of the findings adds to the existing body of knowledge on execution and perception of affect and style.
- Our model explores and partially describes the inversion effect. However, we believe further studies are required in this regard.

Chapter 8.

A Unified System for Recognition and Translation of Secondary Features

8.1 Introduction

In Chapters 6 and 7, we presented our proposed system for extraction of Secondary Features (SFs) followed by an expert-driven method for synthesis of SFs. However, based on the reviewed literature regarding different techniques for interpretation and synthesis of SFs (Chapter 2), to the best of our knowledge, there is no system which can carry out both classification and synthesis (or translation) of affect and style in motion. Such systems can be beneficial from different perspectives. For example, the system

setup can be utilized to retrieve and interpret STs and at the same time alter them if required. In this chapter, we tackle this problem and develop a unified system capable of performing both tasks. Our system shows very accurate classification performance, outperforming several other classifiers, and produces high quality style translated animation.

Classifiers that have been successfully utilized for recognition of motion features can often be reconfigured for generating features. For example, K-nearest neighbor (KNN), which has been used for motion recognition [226], can be used as a regression module capable of synthesizing motion features. Such an approach, however, will produce interpolated/extrapolated motion. Drawbacks of interpolation/extrapolation methods include lack of sufficient generalization. Other classifiers such as probabilistic models (such as Bayesian networks) have been utilized for both recognition [227] and synthesis [183], but a unified system is yet to be put forth. Modifying these existing frameworks to perform both tasks would most likely require significant re-structuring and tuning.

Artificial neural networks (ANNs), which are the core of the method proposed in this chapter, are viable candidates for a unified system due to their capacity in characterizing input-output relations between complex data. Previously, ANNs have been employed for classification of STs mostly in the form of multilayer perceptrons (MLPs). Nevertheless, two shortcomings can be noticed in previous ANN-based works. First, despite their high potential, low classification rates have been achieved. For example, the method used in [86] has an accuracy of 33% and the system proposed in [228] performs with an accuracy of 60%, 84%, and 87% for recognition of different features such as valence and arousal. Our previous research with resilient backpropagation neural networks resulted in average

88% and 78% accuracies for recognition of PTs and STs respectively [229, 230]. The second issue with previous ANN-based methods is that they have not been successfully employed for style translation. In this chapter, ensembles of neural networks are configured and trained successive to pre-processing of data. Multiple experimental results with high classification and style translation rates, as well as significant generalization capability, demonstrate the effectiveness of our approach.

The contents of this chapter have been published as [231, 232].

8.2 Pre-processing

In order to modify and adapt the data for use in the system, pre-processing is required. Two pre-processing steps are often carried out when dealing with motion data: time warping [233] and principal component analysis (PCA) [73]. The former is performed to temporally align the sequences while the latter is carried out for dimensionality reduction.

The CoTW method developed in Chapter 4 is used for aligning the data. As we described in Chapter 4, CoTW enables the selection of an optimum reference with respect to which all other sequences are warped. This functionality is especially beneficial, and is used for our purpose of training the classifiers. The reference sequence is automatically selected and the entire dataset consisting of both training and test data are warped accordingly.

Successive to alignment, PCA [234] is applied to the dataset, for dimensionality reduction as well as acquiring more distinct features. Periodic full-body motion contains a significant amount of redundancy in its many DOFs. As a result, PCA has been shown to

be a very effective means of reducing dimensionality in this type of data [73]. In other words, due to high correlation between different parts of the body, there is no need to utilize the entire dimensionality of the data. Therefore, we employ PCA and compute the principal components (PCs) for representing the sequences in the dataset. Thus, lower dimension classifiers can be used. Moreover, more distinct features often result in higher classification rates. For style translation, however, PCA will not be used since the entire sequence needs to be reconstructed and the output PCs would be difficult to directly interpret and animate.

Here, the PCA procedure is introduced based on [235]. As described in Appendix A, a motion matrix is formalized as $\boldsymbol{\mathcal{D}} = [\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \cdots \boldsymbol{\theta}_n]$ with n DOFs. The ith joint angle trajectory $\boldsymbol{\theta}_i$ with m frames is defined by $\boldsymbol{\theta}_i = \left\{\theta_i^{(t)} : t = 1, \cdots, m \in \mathbb{N}\right\}$. Accordingly, the motion matrix is zero-centered in DOF-wise fashion through:

$$\widehat{\boldsymbol{\mathcal{D}}} = \boldsymbol{\mathcal{D}} - \frac{1}{m}\mathbf{1}_{m\times m} \times \boldsymbol{\mathcal{D}} \tag{8.1}$$

Using the formulation above, we calculate the covariance matrix by:

$$\boldsymbol{\Gamma} = \widehat{\boldsymbol{\mathcal{D}}}\,\widehat{\boldsymbol{\mathcal{D}}}^T \tag{8.2}$$

The eigenvalues ($\boldsymbol{v}$) and eigenvectors ($\mathbf{Q}$) of $\boldsymbol{\Gamma}$ are then computed through solving:

$$(\boldsymbol{\Gamma} - \boldsymbol{v}\mathbf{I})\mathbf{Q} = \mathbf{0} \tag{8.3}$$

Subsequently, assuming $\mathbf{Q} = [\mathbf{q}_1\ \mathbf{q}_2\ \cdots \mathbf{q}_r]$ is the set of eigenvectors which correspond to the r largest eigenvalues, we obtain eigenmotion features by projecting $\boldsymbol{\mathcal{D}}$ onto the eigenmotion space using:

$$\mathbf{X} = \mathbf{Q}^T \boldsymbol{\mathcal{D}} \quad (8.4),$$

where the redundancy in the data is significantly eliminated. Figure 8.1 illustrates that for the sequences with different STs in our dataset around 94% of the sequence can be represented using only 6 PCs while for the HDM05 this rate can be achieved with only 3 PCs. This rate climbs to over 99% for 15 PCs. It is important to note that the joint angle matrix and the 3D displacement vector (root trajectory) are of different nature (degrees vs. meters). Therefore, the location vector must first be removed and PCA applied to the remaining joint angle matrix.

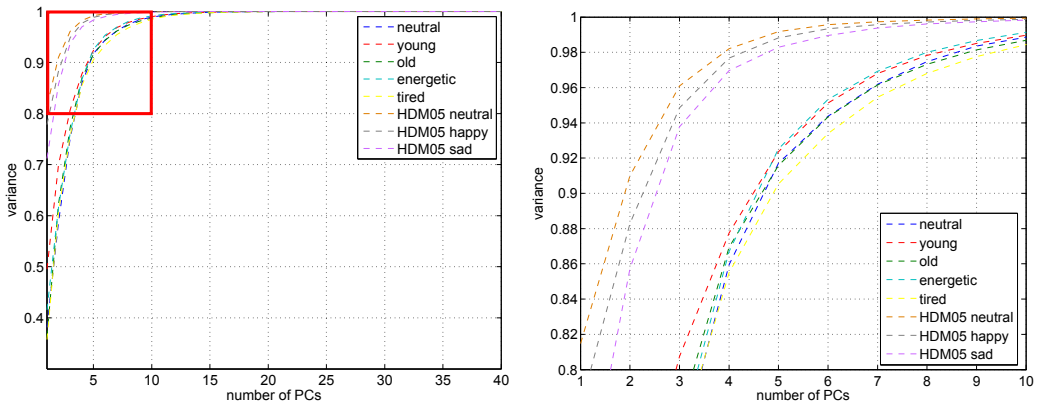
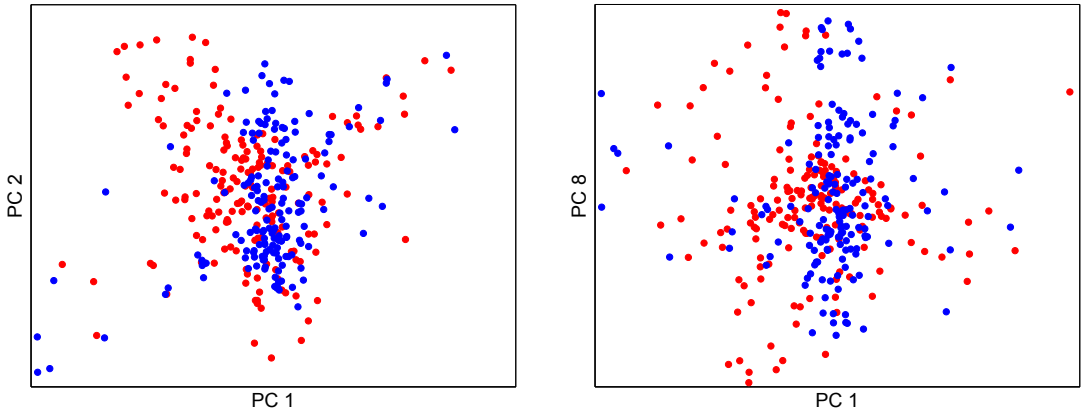


Figure 8.1. The amount of variance captured using PCA. Approximately 94% of the information is captured with the first 6 PCs for sequences from our dataset while the same amount is captured with the first 3 PCs for sequences from the HDM05.

In addition to dimensionality reduction, PCA results in features which are more distinctive and easier to classify. In Figure 8.2, we illustrate an energetic and a tired sequence in the PC subspace. In (a) we illustrate PC 1 vs. PC 2 in which blue and red clusters have emerged. While for action classification, the first few PCs would suffice, for

style recognition, higher PCs would also be informative as they contain smaller variations in the data which most likely correspond to SFs. Figure 8.2 (b) illustrates PC 1 vs. PC 8 for the same sequences where the two classes are recognizable in the subspace.



(a) (b)

Figure 8.2. Visualization of PC subspace for an energetic (red) and tired (blue) walking sequence. In (a), by using the first two PCs, distinct clusters are formed. Similarly, distinct clusters are visible in (b) where PCs 1 and 8 are employed. These higher order PCs are likely to be informative and beneficial for being used in ST-related classifiers.

8.3 System Setup

Generally, ANNs can be configured and trained in a variety of different ways. To the best of our knowledge, however, the impact of the type of ANNs and associated training methods, especially for human motion recognition has not been widely explored. In Chapter 7, we demonstrated that Gaussian RBFs are a powerful means for modeling SFs, from both computational and perceptual viewpoints. Accordingly, we take advantage of this observation and utilize Gaussian RBF neural networks (RBFNNs) for our system.

The proposed system has two modules: the first acts as a classifier while the second one is utilized for style translation. Both modules are composed of ensembles of Gaussian RBFNNs.

For classification, temporal alignment and PCA are carried out on the data as described in the previous section. Regarding the ANN configuration, one possible approach towards the system would be to employ single RBF networks for the entirety of the sequences (all DOFs). However, learning all DOFs of a sequence, which often contain phase shifts with respect to one another, is difficult and confusing for ANNs [236]. Therefore, we design an ensemble of networks, which includes a different and separate network for each DOF. For classification, each network classifies each DOF to the best of its ability, and accordingly, a subsystem classifies the entire motion sequence based on majority vote [237]. For translation, each DOF is transformed by a separate network. Furthermore, our experiments indicate that it is difficult for a single ANN, or even an ensemble of ANNs, to learn different themes for a given class of action. For example, when we experimented with an ensemble learning both young and old themes, we observed high confusion rates. This is because the weights corresponding to a young walk cannot accurately represent the old walk, resulting is high error rates, especially for blind test data. To overcome this issue, we generate a different ensemble for each ST: happy, sad, energetic, tired, young, and old.

For classification, coefficients of the PCs are used. Each ensemble learns the relationship between the coefficients of the first and 9 subsequent PCs of a particular theme. $\boldsymbol{Q}_{\tau}^{(1)}$ represents the vector of the first PC coefficients of the sequence with theme τ used as the

training input while $\boldsymbol{Q}_{\tau}^{(i)}$ is the vector of the ith PC coefficients of the same sequence used as target. Using the classifier following the training phase, $\boldsymbol{Q}_{x}^{(1)}$ of an unknown sequence is fed to all 6 thematic ensembles. Accordingly, each ensemble outputs $\boldsymbol{Q}_{y} = \left\{\boldsymbol{Q}_{y}^{(i)}; i = 2, \dots 10\right\}$ where y is the theme of the ensemble to which $\boldsymbol{Q}_{x}^{(1)}$ is fed. The ensemble that satisfies:

$$\arg\min_{y} \left\| \boldsymbol{Q}_{x}^{(i)} - \boldsymbol{Q}_{y}^{(i)} \right\|_{2} \tag{8.5},$$

for the majority of $i = \{2, \dots, 10\}$, determines the ST of the sequence. In other words, the ensemble that better predicts the higher order PCs based on the first PC is of the same ST as the input to which the first PC belongs. In rare instances where the voting component of the system reaches a tie,

$$\arg\min_{y} \left(\sum_{i=2}^{10} \left\| \boldsymbol{Q}_{x}^{(i)} - \boldsymbol{Q}_{y}^{(i)} \right\|_{2} \right) \tag{8.6},$$

determines the sequence. This means, in case the number of predicted PCs are equal for two or more ensembles, the ensemble that minimizes the total *amount* of difference in predictions determines the ST.

Translation of styles is carried out using a different ensemble of Gaussian RBFNNs. Similar to the classification module, we train this set of networks as 6 different non-connected anticipators, one for each ST. The networks do not use the PC vectors, but rather just the warped motion data. Each anticipator is composed of an individual neural network per DOF. The ensembles learn the relationship between a DOF of a neutral set

and the corresponding DOF of a stylistic set. In other words, the ith RBFNN is trained with $\boldsymbol{\theta}_{neutral,i}$ as the input and $\boldsymbol{\theta}_{stylistic,i}$ as the target. Consequently the translation system will have 6 ensembles each consisting of n RBFNNs that have learned how to transform a neutral motion trajectory to corresponding stylistic ones. In order to use the system for the purpose, a neutral sequence is fed to the ensemble with the designated ST and the generated output is the stylistic version, hence, style translation. Figure 8.3 presents the overall schematic of the classification/translation system.

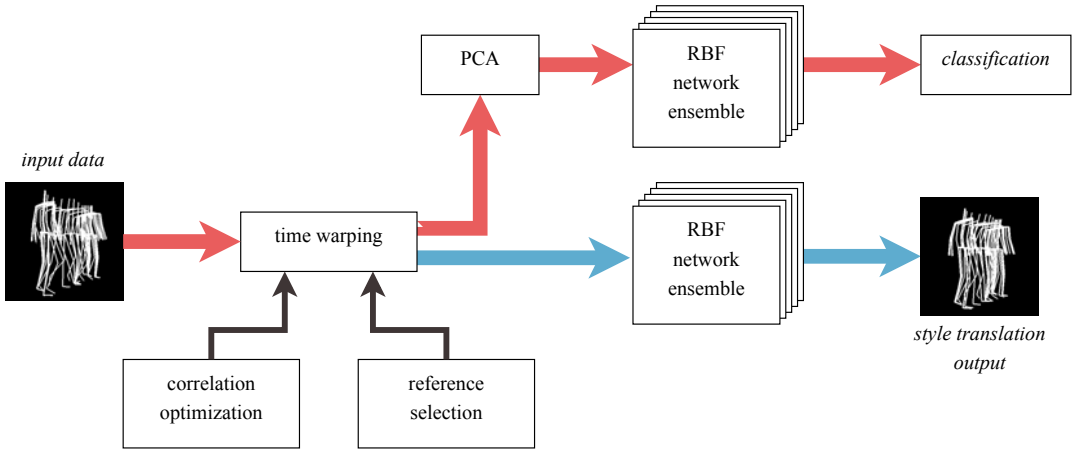


Figure 8.3. Overview of the classification and style translation system. The original data are first warped. PCA is applied when performing classification. The ensemble of RBF networks are then trained based on the two modes resulting in either classification or translation of STs.

8.4 Training

In Chapter 7, we illustrated that the SFs can be accurately modeled using a weighted sum of only a few RBFs. Computing the set of optimum parameters $\{\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$, however, is a challenging problem. In order to calculate the parameter set and employ them for

classification and synthesis of STs, RBF networks have been proposed and utilized [238]. The layout of a typical RBFNN is illustrated in Figure 8.4.

Based on the setup described earlier, for classification, the training input matrices for the ensemble corresponding to ST τ and trained with k samples can be arranged in the form of:

$$\boldsymbol{\Psi}_{input,\tau} = \left[\boldsymbol{Q}_{\tau,1}^{(1)}\ \boldsymbol{Q}_{\tau,2}^{(1)}\ \ldots \boldsymbol{Q}_{\tau,k}^{(1)}\right] \quad for\ network\ 2, \ldots, 10 \tag{8.7}.$$

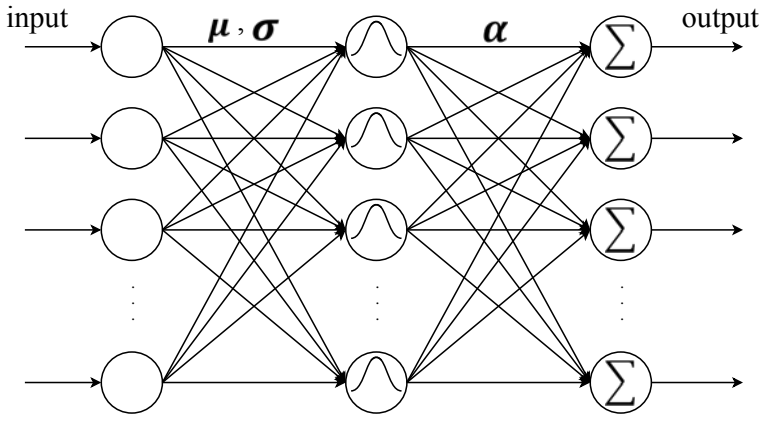


Figure 8.4. RBF network layout.

The target matrices will then be in the form of:

$$\boldsymbol{\Psi}_{target,\tau} = \left[\boldsymbol{Q}_{\tau,1}^{(i)}\ \boldsymbol{Q}_{\tau,2}^{(i)}\ \ldots \boldsymbol{Q}_{\tau,k}^{(i)}\right] \quad for\ network\ i = 2, \ldots, 10 \tag{8.8}.$$

Similarly, for the style translation module, the input matrices are represented by:

$$\boldsymbol{\Psi}_{input,\tau} = \left[\boldsymbol{\theta}_{neutral,i,1}\ \boldsymbol{\theta}_{neutral,i,2}\ \ldots \boldsymbol{\theta}_{neutral,i,k}\right] \quad for\ network\ i = 1, \ldots, n \tag{8.9},$$

where n is the DOF and k samples are used in training. The target matrices are denoted by:

$$\boldsymbol{\Psi}_{target,\tau} = \left[\boldsymbol{\theta}_{\tau,i,1}\ \boldsymbol{\theta}_{\tau,i,2}\ \ldots \boldsymbol{\theta}_{\tau,i,k}\right] \quad for\ network\ \ i = 1, \ldots, n \tag{8.10}.$$

To train the RBFNNs, different methods such as the least squares technique can be used [238]. Our system learns by iteratively adding RBF neurons, minimizing the sum of output residues. The width of the RBFs are fixed and manually assigned. While this may seem as a drawback at first, as there can be no limits to the number used RBFs, wider RBFs can be approximated by the sum of few thinner fixed-width RBFs (if needed). The number of RBF neurons are also manually assigned. The weights of the network are calculated by minimizing:

$$\left\|\boldsymbol{\Psi}_{target} - \boldsymbol{\alpha}\boldsymbol{\Psi}_{input}\right\| \tag{8.11},$$

where the solution is $\boldsymbol{\alpha} = \left(\boldsymbol{\Psi}_{input}^{T}\boldsymbol{\Psi}_{input}\right)^{-1}\boldsymbol{\Psi}_{input}^{T}\boldsymbol{\Psi}_{target}$. It should also be noted that there are alternative training methods for RBF networks which can be employed.

8.5 System Evaluation

Here we describe the method and criteria used to evaluate the system. For the classification module, 2 other classifiers are employed, namely KNN and SVM. Different variations of each of the two classifiers are used which are described in detail in the following sections. For evaluating the translation component of the system, 10 participants were asked to watch and provide feedback on the system outputs. The average age of participants was 31.3, the standard deviation was 11.5, 6 were males, and 4 were females. A paper-based forced-choice questionnaire is used. Ethics approval is secured. Sequences from each category (affect, energy, and age) were presented

separately since they contain very similar features that are indistinguishable even for real recorded motion sequences.

8.5.1 KNN Classifier

K-nearest neighbor is a supervised classifier [239]. The classifier is trained using a dataset of samples through indexing the dataset. A test sample is subsequently classified by measuring the proximity to *K* training samples of a particular class. Proximity is measured through distance functions such as Euclidean, city-block, and cosine. Other objective functions, for instance correlation, can be used. In this study, we utilize the four mentioned functions for validation. For m-dimensional test vector $\boldsymbol{\theta}_{test} = \left[\theta_{test}^{(1)}, \theta_{test}^{(2)}, \dots, \theta_{test}^{(m)}\right]$ and training vector $\boldsymbol{\theta}_{training} = \left[\theta_{training}^{(1)}, \theta_{training}^{(2)}, \dots, \theta_{training}^{(m)}\right]$, the Euclidean distance is measured by:

$$d_{Euclidean} = \left\|\boldsymbol{\theta}_{training} - \boldsymbol{\theta}_{test}\right\|_{\mathbf{2}} = \sqrt{\sum_{i=1}^{m}\left(\theta_{training}^{(i)} - \theta_{test}^{(i)}\right)^2} \tag{8.12}.$$

City-block is calculated through:

$$d_{city} = \left\|\boldsymbol{\theta}_{training} - \boldsymbol{\theta}_{test}\right\|_{\mathbf{1}} = \sum_{i=1}^{m}\left|\theta_{training}^{(i)} - \theta_{test}^{(i)}\right| \tag{8.13}.$$

Cosine distance is calculated by:

$$d_{cosine} = \frac{\boldsymbol{\theta}_{training} \cdot \boldsymbol{\theta}_{test}}{\left\|\boldsymbol{\theta}_{training}\right\| \left\|-\boldsymbol{\theta}_{test}\right\|} = \frac{\sum_{i=1}^{m}\left(\theta_{training}^{(i)}\theta_{test}^{(i)}\right)}{\sum_{i=1}^{m}\theta_{training}^{(i)}\sum_{i=1}^{m}\theta_{test}^{(i)}} \tag{8.14}.$$

Finally, correlation is measured using:

$$d_{correlation} = \rho\left(\boldsymbol{\theta}_{training}, \boldsymbol{\theta}_{test}\right) = \frac{\sum_{i=1}^{m}\left(\theta_{training}^{(i)} - \mu_{\boldsymbol{\theta}_{training}}\right)\left(\theta_{test}^{(i)} - \mu_{\boldsymbol{\theta}_{test}}\right)}{\sqrt{\sum_{i=1}^{m}\left(\theta_{training}^{(i)} - \mu_{\boldsymbol{\theta}_{training}}\right)\sum_{i=1}^{m}\left(\theta_{test}^{(i)} - \mu_{\boldsymbol{\theta}_{test}}\right)}} \quad (8.15).$$

Figure 8.5 illustrates classified test sample(s) using a KNN classifier where a Euclidean distance function is used (left). We observe that for $K = 2$, the test sample is classified as Class 1 while for $K = 5$, it is classified as Class 2. On the right, the test data are classified as Class 1 when correlation is used while being classified as Class 2 when Euclidean distance is employed.

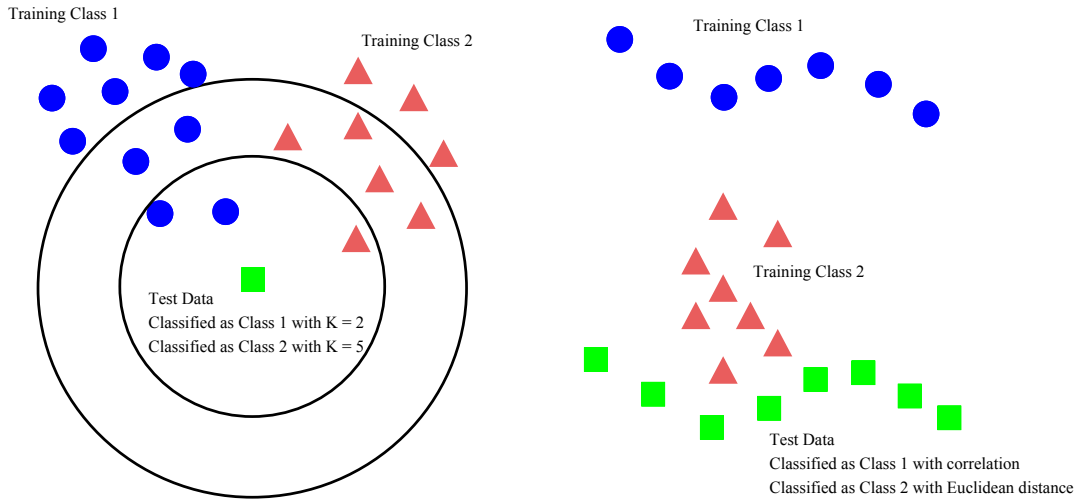


Figure 8.5. Performance of a KNN classifier with different parameters. Both K and the objective function can influence the classification outcome.

8.5.2 SVM Classifier

Support vector machines (SVMs), also referred to as support vector networks, are binary supervised classifiers [240]. A two-class training set is mapped onto a space and training is carried out by a characterizing a kernel or hyper-plane that best divides the two classes by maximizing the support vectors or distances between the leading training data and the

kernel. Accordingly, this classifier performs best when the data are linearly (or non-linearly based on the kernel) separable. Different kernels are employed in this classifier, for example, linear, quadratic, cubic, and RBF among others. In a space where the output is a function of time (t), the linear, quadratic, and cubic kernels are modeled as:

$$\kappa_{linear|quadratic|cubic} = \sum_{i=0}^{j} a_i t^i \quad for\ j = 1,2, or\ 3 \tag{8.16},$$

where $j = 1$ defines a linear kernel, $j = 2$ defines a quadratic kernel, and $j = 3$ yields a cubic kernel. a_i are the kernel parameters which are calculated through the training process. Finally, the Gaussian RBF kernel is defined by:

$$\kappa_{RBF} = a_1 . exp(a_2 \|t - a_3\|^2) \tag{8.17},$$

where, a_1, a_2, and a_3, are the kernel parameters.

Figure 8.6 presents a classified test sample using SVM with two different kernel functions, linear and quadratic. We observe that when a linear kernel is used, the test sample is classified as Class 2 while when a quadratic kernel is used, it is classified as Class 1.

8.6 Results and Discussion

In order to evaluate our proposed method, we utilize our own dataset along with the HDM05. Details about the datasets can be found in Appendix A. For the purpose of this chapter, our dataset consists of 75 walking sequences, 15 in each of the neutral, young,

old, tired, and energetic categories. The HDM05, after segmentation, contains 48 sequences, 16 in each of the neutral, happy, and sad classes. Other motion capture data such as those available in the Carnegie Mellon University (CMU) dataset (described in Appendix A) are publically available and accessible. However, in order for such data to be readily usable for ANN purposes, the data need to be very consistent. Ideally, the sequences need to be controlled cycles with similar model structures but performed multiple times. Furthermore, for the goal put forth in this research, i.e. classification/translation of ST, they need to be carried out with different ST types. To the best of our knowledge, at the time that we conducted this research, there were no publically available datasets from which multiple repeated and controlled affective/stylistic sequences (such as those described in the above paragraphs) could be derived; hence, the choice of the datasets.

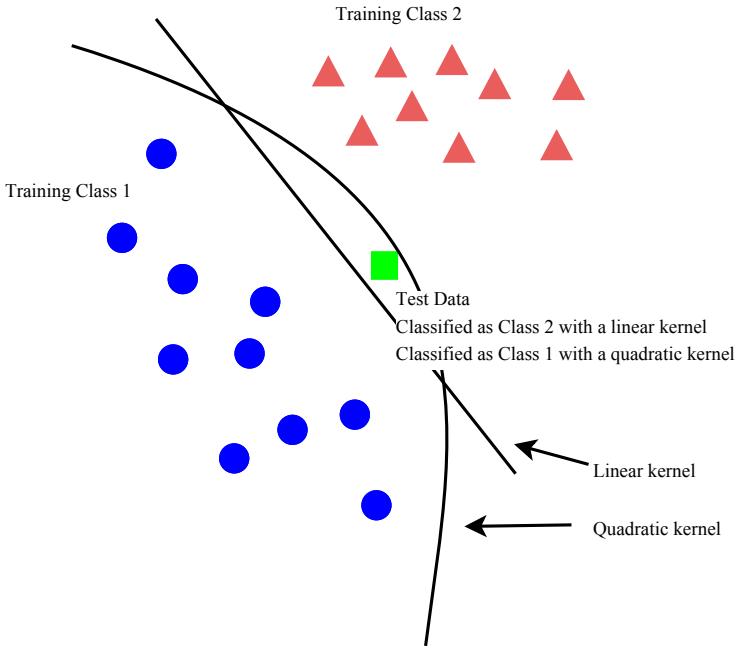


Figure 8.6. Performance of a SVM classifier. Different kernels result in different recognition outcomes.

The ANN ensembles are generated and trained in MATLAB as described in Sections 8.3 and 8.4. In general, the ensembles train well and expected learning curves are achieved. Figure 8.7 presents a sample learning curve from the system.

When classification is carried out, 15-fold and 16-fold cross validations are used for our dataset and the HDM05 dataset respectively. Naturally, styles such as happy, energetic, and young are often confused with one another even when observed by human subjects. Similarly, confusion rates for sad, tired, and old are quite high. It is therefore fair to expect that ANNs show high error rates should a 6-class system be used. Therefore, we use binary-classes for evaluation of the outputs and do not mix the affect, energy, and age related themes.

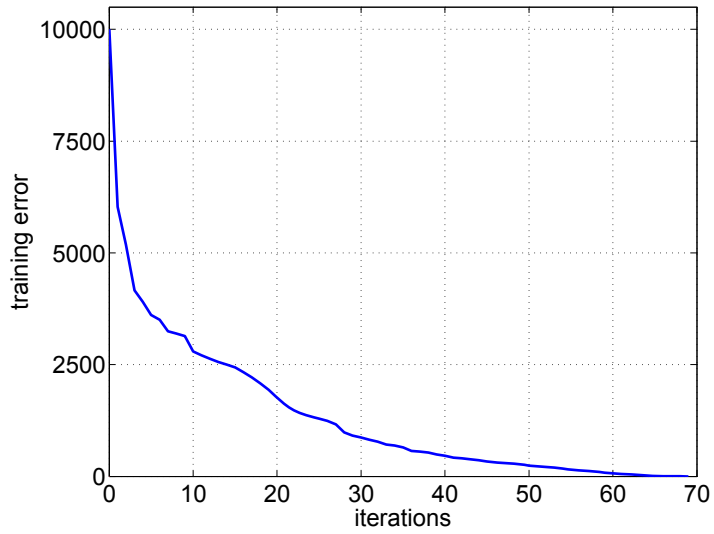


Figure 8.7. A sample RBFNN learning curve.

We evaluate the RBFNN by comparing its performance with KNN and SVM classifiers as described earlier. For KNN, we used the 4 different similarity functions described in

the previous section. $K = 1$ was used. Similarly, for SVM, we used the 4 different kernel functions described in the previous sections. Table 8.1 presents the results. Generally, for the three classifiers with different parameters and for different STs, KNN performs with an average accuracy of 85.8±10.0, SVM has a classification rate of 76.1±14.2, and the RBFNN performs with a classification rate of 93.5±4.0. The fact that KNN outperforms SVM indicates that the data are not simply separable using hyper-planes or kernels. Moreover, it entails that the training data are very intertwined and are more suitable for non-parametric separation. Therefore, the classification accuracy of our system is higher than KNN and SVM. Further evaluating the performance of KNN, we observe that different objective functions do not significantly affect recognition rates. One-way ANOVA indicates no significant effect at the $p < 0.05$ level for the objective function with $F(3,20) = 0.1$. Greater variations, however, are observed for the SVM with respect to the different kernels used, where the quadratic kernel shows the best performance. The effect of the kernel is significant at the $p < 0.05$ level with $F(3,20) = 4.83$.

Table 8.1. Classification rates for the KNN classifier with 4 different objective functions and the SVM classifier with 4 different kernels compared to RBFNN.

Classifier	***Parameters***	**Happy**	**Sad**	**Young**	**Old**	**Energetic**	**Tired**	**Average**
KNN	*Euclidean*	75.0	81.2	86.7	93.3	93.3	93.3	87.1±7.7
	City block	68.7	75.0	93.3	86.6	93.3	100.0	86.1±12.1
	Cosine	68.7	75.0	86.7	93.3	86.7	93.3	83.9±10.0
	Correlation	68.7	75.0	86.7	93.3	93.3	100.0	86.2±12.0
	Average	70.3 ±3.2	76.6 ±3.1	88.3 ±3.3	91.6 ±3.3	91.6 ±3.3	96.6 ±3.9	85.8±10.0
SVM	*Linear*	62.5	62.5	93.3	80.0	86.7	86.7	78.6±13.2
	Quadratic	68.7	75.0	86.7	100.0	93.3	86.7	85.1±11.6
	Cubic	68.7	68.7	80.0	100.0	80.0	80.0	79.6±11.4
	GRBF	56.2	50.0	80.0	60.0	60.0	60.0	61.0±10.1
	Average	64.0 ±6.0	64.0 ±10.7	85.0 ±6.4	85.0 ±19.1	80.0 ±14.4	78.3 ±12.6	76.1±14.2
RBFNN	*Gaussian*	87.5	93.8	93.3	93.3	100.0	93.3	93.5±4.0

In terms of the influence of ST on classification rates for KNN, one way ANOVA shows that the effect is significant at the $p < 0.0001$ level with $F(5,18) = 36.9$. This effect, however is not significant at the $p < 0.05$ level for SVM with $F(5,18) = 2.44$. The RBFNN shows the best rate for energetic while happy is classified with the least accuracy.

As described earlier, style translation is carried out using the ensemble of RBFNNs trained with neutral sequences as inputs and stylistic ones as outputs. The different parameters of the system in the training phase do not significantly influence the translation process. A sample trajectory transformed from neutral to energetic with different training termination rates ($\varepsilon = 0$ and $\varepsilon = 10$), number of RBFs ($M = 15$ and $M = 2$), and different RBF spreads ($\eta = 5$ and $\eta = 15$), show comparable results. This is shown in Figure 8.8 where the differences in the output trajectories are only a few degrees and thus insignificant.

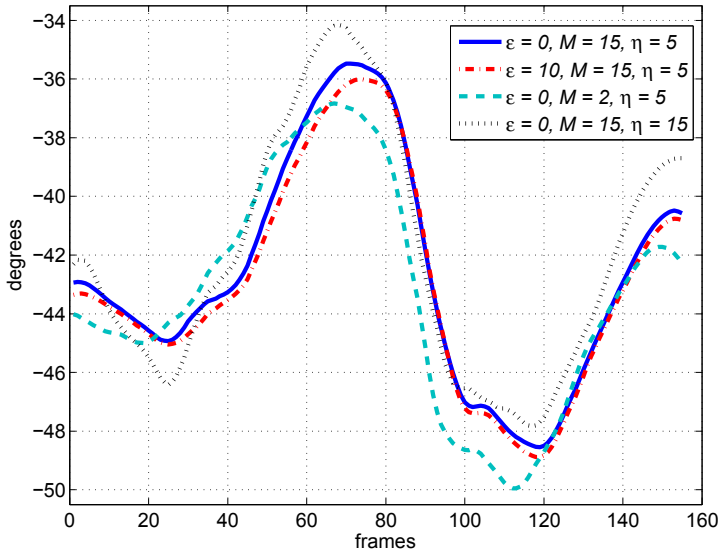


Figure 8.8. A sample style translation output with different network parameters.

Similarly, other DOFs of the input data show reasonable resilience towards the mentioned parameters. Nevertheless, the different number of RBF neurons impacts the translation performance for blind data (not been used in the training process). When many RBFs (more than 15) are utilized, over-fitting occurs. This issue manifests itself as unnatural motion in the output sequences. Our experiments show that 5 to 10 RBFs are sufficient for proper generalization of the problem.

Video Clip E (https://www.youtube.com/watch?v=lw0yU277Pg4) and the frames presented in Figure 8.9 illustrate the neutral input from the HDM05 and outputs obtained using the system. Increased lateral body sway and wider steps are visible in the sequences generated by the ensemble trained with happy target data. The ensemble trained with sad target data has generated features such as downward tilt in the neck, dropped shoulders, and smaller steps, which indicate successful translation to sadness.

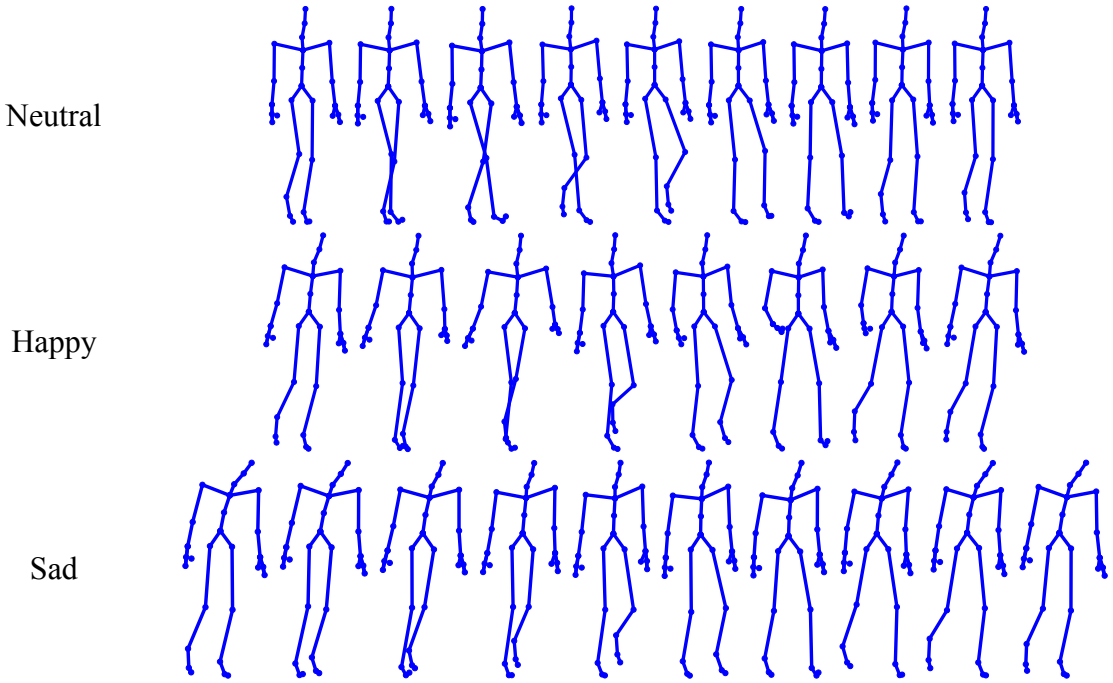


Figure 8.9. Neutral input from the HDM05 dataset and style translation outputs obtained using the RBFNN system (from Video Clip E).

Similarly, Video Clip E and the frames presented in Figure 8.10 illustrates poses from the input neutral walk from our dataset along with the stylistic sequences created using the system. Longer steps, more sway in shoulders, and an upright posture indicates successful translation of energetic and young STs. Smaller steps, decreased sway in limbs, and downward tilt in shoulders and head point to successful translation of tired and old STs. No post processing was carried out on the output data. Very little footskating indicates the accurate performance of our method.

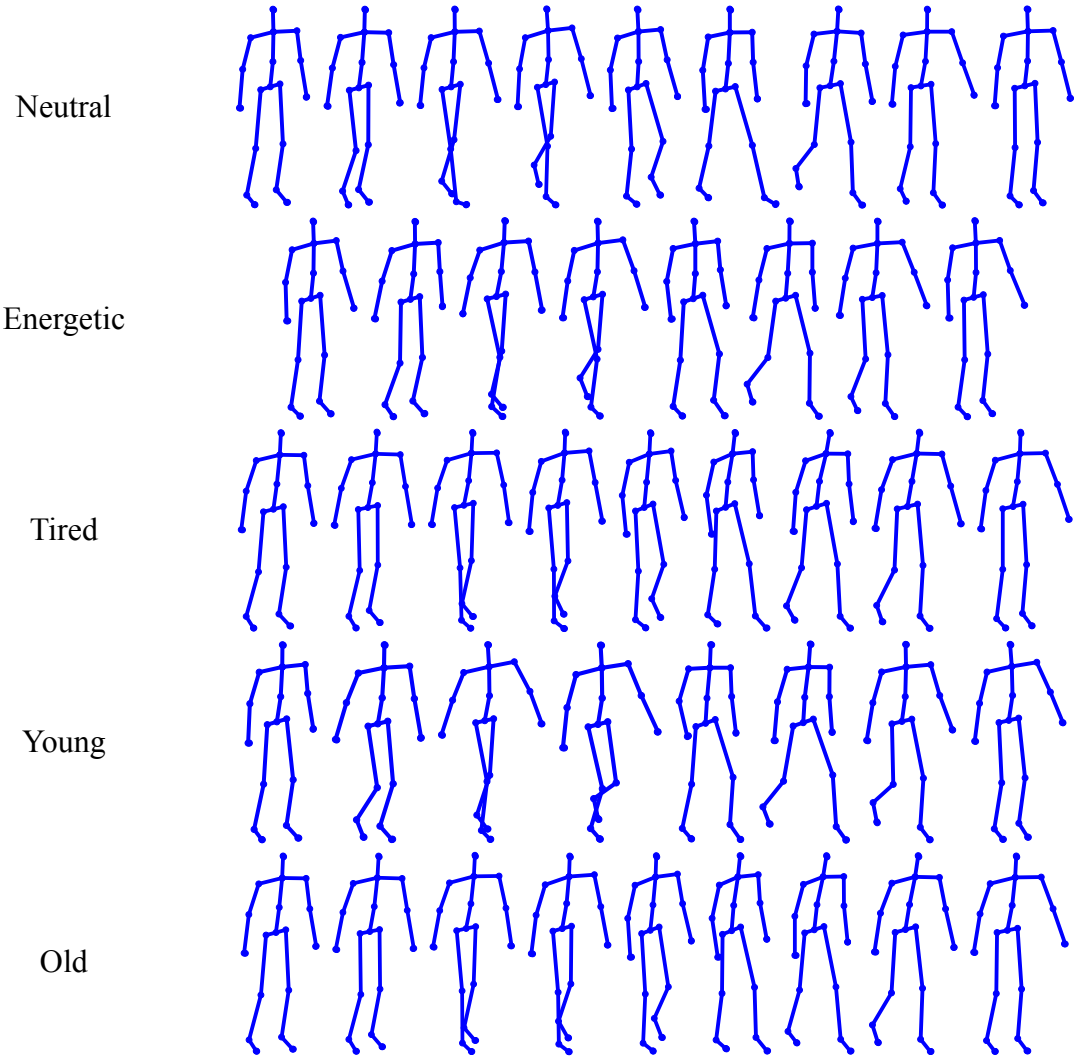


Figure 8.10. Neutral input from our dataset and style translation outputs obtained using the RBFNN system (from Video Clip E).

Based on the description in Section 5, user evaluation of the results were carried out to determine the performance of the style-translation outcome. Table 8.2 presents the correct perception rates. The results in general show very low confusion rates, denoting successful style translation. The ratio of the correctly perceived STs (true positive) range from 0.7 (happy and energetic) to 1.0 (old) while the maximum false positive is 0.2 (neutral).

Table 8.2. Successful perception rates for the style translation outputs.

		Perceived										
		Happy	**Sad**	**Neutral**		**Young**	**Old**	**Neutral**		**Energetic**	**Tired**	**Neutral**
Output	**Happy**	**0.7**	0.1	0.2	**Young**	**0.9**	0.0	0.1	**Energetic**	**0.7**	0.1	0.2
	Sad	0.0	**0.8**	0.2	**Old**	0.0	**1.0**	0.0	**Tired**	0.0	**0.9**	0.1

Further investigations show that even when actions outside the initial learning input class (neutral) are fed to the ensembles, style translation is carried out with adequate accuracy. For example, when an old walk is used as the input of a neutral-to-young ensemble, the output is transformed to the young style. Artifacts, however, are observed in some joints. Considering the pre-existence of un-related affect/style features in these input sequences, the relatively high quality outputs are very promising.

Chapter 9.

Towards Perceptual Validity in Animation of Motion

9.1 Introduction

We have so far established several methods for processing different aspects of secondary features (SFs). These processes include temporal alignment (Chapter 4), validation of incremental processing (Chapter 5), extraction (Chapter 6), synthesis (Chapter 7), and recognition/translation (Chapter 8). Overall, one of the general goals of the developed techniques is synthesis or altering features where required. The concept that we study in this chapter is to answer the following question: When and why should SFs be altered in

or added to motion sequences? In other words, what are the motivating factors for processing SF in motion and what are their relative degrees of significance?

To answer the abovementioned questions, we propose a general paradigm that describes different factors that can influence animation scenes. We call this paradigm, Perceptual Validity (PV). PV is a model based on the parameters that impact our perception of motion and aims at believability, acceptability, and perceptual (not necessarily format-related) attractiveness of the animated characters. Our PV paradigm associates various visual cues to viewers' perception according to the literature and a set of experiments, and establishes principles to help animators create content that is perceived as intended. The PV principles, which we propose, are particularly important as they can be used as the foundation of intelligent algorithms for procedural generation of believable character animation. Once the factors that influence perception of motion are recognized, the processing techniques developed in the previous chapters, as well as existing methods in the literature, can be utilized aim-fully.

The PV paradigm incorporates and explores three different components along with their internal and external relations. These factors are:

- context of the scene in which motion animation is depicted
- visual cues for primary themes (PTs)
- visual cues for secondary themes (STs)

Together, the relationships are categorized into four different components. Each component is described with tangible examples. A case study is carried out to validate

one of the components. We refrained from performing case studies on all components as they are mostly self-evident or backed up by the literature. Subsequently, a study is performed that describes the relative importance of each of the components and underlying elements. Finally, we compare and discuss our paradigm against Disney's 12 principles of animation [27].

While this dissertation focuses on body motion, the face has been known to convey a significant amount of STs [47] despite its small spatiotemporal significance compared to the rest of the body. As a result, we incorporate the face in our paradigm along with the body. This also shows the inclusiveness of our paradigm. The PV paradigm is presented following a background on the face domain.

The contents of this chapter have been published as [241, 242].

9.2 Incorporating Facial Features

Generally, the human face is known to convey a rich amount of affective and stylistic information [47]. As a result, a considerable amount of research has been done on creating personality-rich facial animation. The Facial Action Coding System (FACS) was one of the first systematic efforts to determine the smallest possible facial actions, or Action Units (AUs), and associate them to facial expressions [243]. Associating facial actions with personality requires a reasonably adequate personality model for the agent, and a thorough study of the effect of facial actions on the perception of personality. The latter, has not been done properly yet, but the former has been the subject of some recent works. A multi-layer personality model has been proposed [244], which is, more

precisely, a multi-layer behavioral model that includes layers of personality, mood, and emotions on top of each other. Every layer controls the one below it, and the facial actions and expressions are at the bottom. The model allows element of parameters at each level to individualize the agent. At the personality level, it utilizes the Big Five model with five parameters. The following observations can be made regarding this system:

- The problems associated with using the Big Five such as difficulty in visualization and correlation between dimensions
- Hierarchical dependence of personality and emotional states.

It has, in fact, been suggested that these should be treated independently [245]. Other proposed models follow similar notions [246, 247]. The latter uses a two-dimensional model similar to the one proposed in [248] for personality (called performatives) and also separates them from emotions as two independent components activating facial actions through a belief network. This two-dimensional model has been used to associate facial actions to personality dimensions, Dominance and Affiliation [49].

As mentioned earlier in Chapter 2, the location of common emotional states in a circumplex in 2D space with arousal and valence (activation and pleasantness) as the dimensions have been defined [53] and presented in Figure 2.2. Control was later suggested as the third parameter [249] while Uncertainty and Agency have been proposed in [250] and [251]. The significant visual cues that can be considered more important perceptual factors in animation have been studied in [252]. Finally, the temporal effects and also the issue of conflicting signals have been reviewed in [253].

9.3 Proposed Paradigm

The proposed paradigm is called Perceptual Validity since we believe abiding by the elements that it puts forth will ensure the perceptual quality of human motion animation. The paradigm takes into account PTs and STs in both face motion (FM) and body motion (BM) domains. Furthermore, the context in which the animation is presented is taken into account. The paradigm proposes 4 major components: association, contextual dependency, internal consistency, and external consistency, each of which is composed of several elements. In the following the different components and elements are described accompanied with examples.

9.3.1 Association

Whether for FM or BM, for both PTs and STs, it is imperative to display the correct visual cues based on the intentions of the animators. Therefore, the first step towards synthesizing a perceptually sound sequence of human motion is to recognize the set of visual cues (features) which will generate each possible PT/ST based on the requirements of the scene. The fashion in which the stimulus is visualized and displayed can significantly impact the two themes as perceived by the audience. We call the correct presentation and preservation of visual cues, *association*. In other words, association dictates that the visual motion cues displayed by the characters comply with the intentions of the animators in terms of PTs and STs.

While PTs are quite intuitive due to the effects in which different actions result in, the STs are not that simple to artificially generate or identify. In fact, there have been many publications on the visual cues responsible for perception of different STs for both FM

and BM. The visual cues corresponding to Ekman's basic set of emotions [254, 255] (as an accepted example for a subset of STs) for both FM and BM can be utilized. Examples of these visual cues are available in [225, 256].

In addition to correct utilization of visual cues, it is essential that the stimuli be presented such that the intended themes are not perceived differently. Preservation of the stimuli must occur both temporally and spatially. It is well known and demonstrated that spatial and/or temporal alterations in visual cues for both FM and BM will affect the perceived STs [7].

Another implication of this rule is that extra movements, such as those responsible for creating the famous foot-skate artifact [222] or tremor in the motion, must not appear in the motion. This is due to the fact that such movements in joints and other body parts might alter the present primary or secondary features or make them undetectable to the untrained eye.

Examples: It has been shown that face inversion results in difficulties in its recognition [257] and distortions in its features [258]. More particularly for effects of spatiotemporal alterations on face stimuli, it has been displayed that inversion of face stimuli reduces the accuracy emotion recognitions such as fear, anger, and disgust, and even results in sadness being identified as neutral by the audience [259]. Similar trends regarding importance of spatiotemporal preservations have been observed for BM stimuli in which, point light walkers are often used for perceptual studies. It has been shown that 1.6 to 2.7 seconds of motion are required for correct gender recognition of a walker [64]. Moreover, when stimuli are inverted, the gender is often recognized as the opposite, meaning for

upside down presentation of the point light walker videos, a male walker is more likely to be recognized as female, and vice-versa [64]. In another study, it was indicated that when hip movement is greater than the shoulder movement, the sequence appears as feminine to untrained eyes, and vice-versa [163, 164]. It was later shown that this "extra movement" is rather velocity than displacement [74]. This shows the necessity for temporal preservation of the stimuli. The aforementioned findings and many more similar investigations signify the importance of proper stimuli presentation.

9.3.2 Contextual Dependency

Performance and perception of PTs and STs performed by characters are heavily linked to the theme of the scene or the context [260, 261, 262]. It is safe to claim that the audience does not expect to see behaviours that contradict or are out of context. While in terms of physics of human behaviour, such inconsistencies are possible, the typical audience would not expect to experience such scenes, or at the very least would consider it to be odd. Accordingly, the second factor which we believe is essential in PV is *contextual dependency*. This component of the paradigm indicates that both primary and secondary themes for both body and face motion depend on and need to be consistent with the context of the scene. It should be noted that the need for contextual dependency is present for both themes. Some PTs are simply not acceptable with certain contexts and some may have global or cultural implications making them unacceptable at given scenarios. Also, for STs, the audience expects contextual reasons for expressive behaviours.

A point to consider in this regard is that contextual dependency is local with respect to the character. This means that for each character in the scene, the context might differ

based on the story, background, and agenda. Therefore, context only refers to the perspective of the character of interest. Should several characters be present in the scene, each will have his/her own local context.

Examples: An audience expects to see running, walking, jumping, and dribbling as primary actions during a basketball game. Displaying actions of dancing is therefore unexpected and improper, unless the local context validates the actions. An example of this local context could be a scene of goal celebration by the team members of the scoring team. For an example of Contextual Dependency for STs, we could mention that it is unorthodox to see scenes of clapping, whistling, and laughing at a funeral and such scenes will cause perceptual disbelief and distaste. Again, similar to PTs, local contexts can validate apparent dissociations.

9.3.3 Internal Consistency

A component of PV, which we call *internal consistency*, refers to PTs of the character being consistent with one another, while a similar consistency exists for STs. As both FM and BM domains need to be taken into account, many different scenarios can be considered. Namely, consistency needs to apply to the PTs of BM, the PTs of the FM, the STs of BM, and the STs of FM. Moreover, cross-internal-relations of FM and BM must be considered. This means that PTs of BM need to be in line with the PTs of FM and the STs of BM need to be consistent with the STs of FM. Models such as Russel's circumplex [53], as described earlier, can provide valuable guidelines for utilization of consistent STs in characters.

Examples: A viewer would not expect to see a person playing a guitar and kicking a

soccer ball at the same time (internal correspondence for primary themes). An example for internal consistency of STs, this time for FM, is that one would not expect to see the eyebrows squeezed and pressed together with anger and frustration while lips and cheeks display features associated with happiness, for example being raised.

9.3.4 External Consistency

In addition to internal consistency, the two themes, PT and ST, must be consistent with each other. This constitutes the final component for PV, which we call *external consistency*. Accordingly, PTs and STs in FM as well as PTs and STs in BM need to be consistent. Another implication of this component is cross-external-relations for FM and BM. In other words, in addition to the mentioned rule, PTs of FM need to be consistent with the STs of the BM while the STs of FM are consistent with the PTs of BM.

Examples: One would not expect to see a character jumping up and down when feeling sad or depressed and showing signs of such emotions. On the other hand, fast and energetic actions are more expected when dealing with STs such as excitement, happiness, or anxiety. Similar expectations are present for facial actions and expressions.

9.3.5 Summary

Table 9.1 presents the detailed statements of the paradigm. We observe that the model contains 4 major components. Association, contextual dependency, and external consistency are each composed of 4 elements while internal consistency consists of 6 elements. Together, these 18 elements form our proposed paradigm. Figure 9.1 illustrates a graphical representation of the proposed paradigm in which the arrows symbolize the notion of correspondence/consistency.

Table 9.1. Components and elements of the proposed paradigm.

No.	Component	Element
1	Association	Proper display of PT_{BM}
2		Proper display of PT_{FM}
3		Proper display of ST_{BM}
4		Proper display of ST_{FM}
5	Contextual Dependency	Dependence: PT_{BM} and *context*
6		Dependence: PT_{FM} and *context*
7		Dependence: ST_{BM} and *context*
8		Dependence: ST_{FM} and *context*
9	Internal Consistency	Consistence: PT_{BM} and PT_{BM}
10		Consistence: PT_{FM} and PT_{FM}
11		Consistence: ST_{BM} and ST_{BM}
12		Consistence: ST_{FM} and ST_{FM}
13		Consistence: PT_{BM} and PT_{FM}
14		Consistence: ST_{BM} and ST_{FM}
15	External Consistency	Consistence: PT_{FM} and ST_{FM}
16		Consistence: PT_{BM} and ST_{BM}
17		Consistence: PT_{FM} and ST_{BM}
18		Consistence: PT_{BM} and ST_{FM}

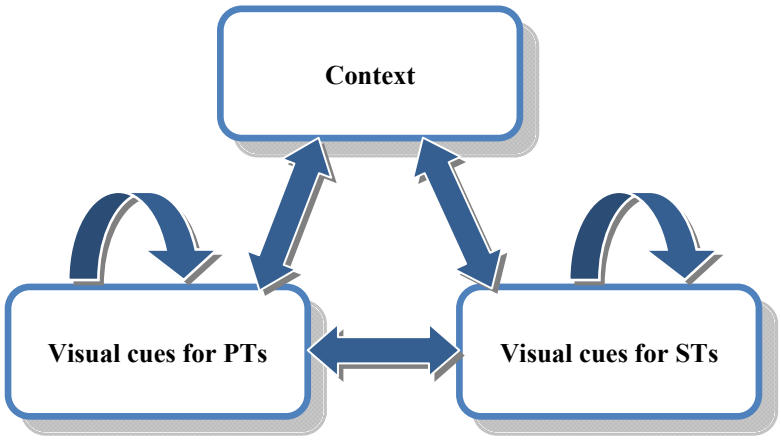


Figure 9.1. Graphical representation for the concept of Perceptual Validity.

9.4 Experiments and Results

In this section, we report a study aimed at determining the relative significance of the different components and elements presented in Table 9.1. Several participants were

asked for their opinions on the topic. As we believe that it is essential to incorporate the opinions of animators, half of the participants were selected with experience in motion animation. A case study is also provided that validates the concept of contextual dependency through a set of experiments in both FM and BM domains. We selected this component of the paradigm for validation since it is less self-evident (compared to internal consistency for example). Furthermore, there are also less supporting related work for this component.

9.4.1 Relative Significance

We believe it is important to determine the relative significance of each of the components and elements of the paradigm in both face and body domains. 22 participants were asked to provide their confidence levels regarding the importance and impact of each element on a 10-point Likert scale. Detailed description of each of the elements in Table 9.1 in addition to supporting examples were read for the participants. Initially, definitions and examples of the terms: bodily actions (PT_{BM}), bodily expressions (ST_{BM}), facial actions (PT_{FM}), facial expressions (ST_{FM}), and context were provided. Then, the descriptions and examples of the elements of the paradigm as described in Section 9.3 were read to the participants. 11 of the participants were experienced with animation of human motion, and 11 were inexperienced towards animation of human motion. The experienced participants had a mean age of M = 25.8 and SE = 4.2. They were either employees of animation studios or graduate students with experience in the field of animation, 9 of whom were males and 2 were females. The naïve participants had a mean age of M = 30.8 and SE = 12.7, 7 of whom were males and 4 were females. The necessary ethics approval was secured. No compensation was provided for the

participants' time.

Figure 9.2 illustrates the mean and standard errors (SEs) of both experienced and naïve participants for each of the PV elements mentioned in Table 9.1. The two important points that should be considered in this figure are experienced participants alone and the experienced and naïve participants together. In both cases, consistence of PTs between body and face has been perceived as the least important. Interestingly, the most important element as perceived by experienced participants is the internal consistency of PTs for FM, while the naïve participants perceived consistence of PTs in the body as the most important. Internal consistency of STs in the face as well as proper display of SFs in the face have also been perceived as very important by both experienced and naïve participants.

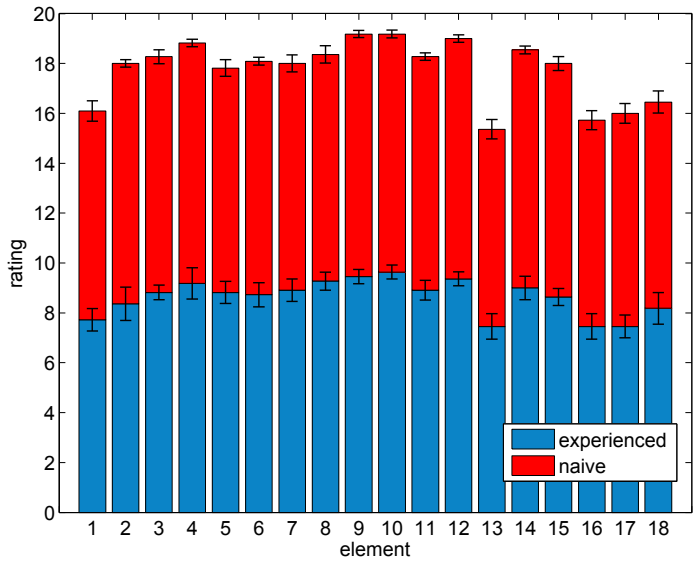


Figure 9.2. Mean and standard errors of the ratings provided by experienced and naïve participants for each element of Table 9.1.

One-way analysis of variances (ANOVA) for the ratings of the 18 elements of the paradigm shows that for experienced participants, there is significant effect at $p < 0.005$ ($F(17,197) = 2.42$). Similarly for the naïve participants, the element poses a significant effect at $p < 0.0001$ ($F(17,197) = 3.95$). For the two groups of participants together, similar significant effect is observed at $p < 0.0001$ with $F(17,395) = 5.17$.

To further evaluate the relative significance of the four components, we averaged the elements for each of the main components of PV. Figure 9.3 illustrates the average and standard errors of the results for the two participant groups.

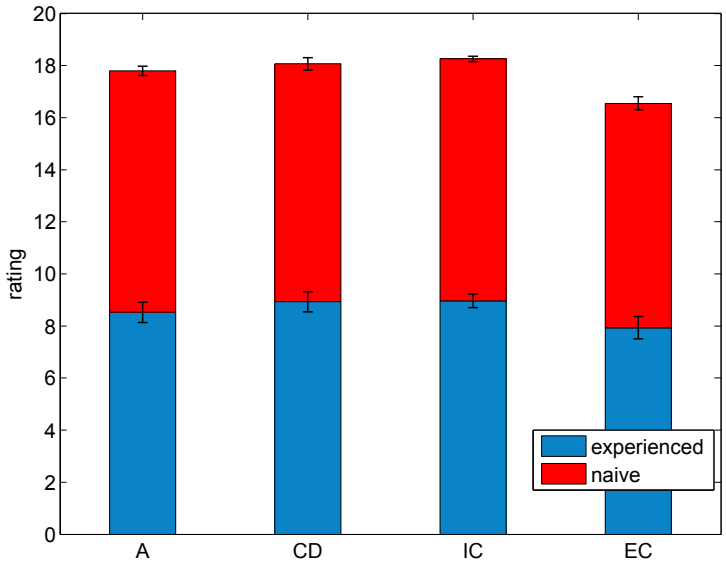


Figure 9.3. Mean and standard errors of the ratings of experienced and naïve participants for each component of the paradigm. Components are in the order presented in Table 9.1.

One-way ANOVA indicates neither experienced nor naïve participants perceive a components to be significantly different ($F(3,43) = 1.7$, $p = 0.183$ for experienced and $F(3,43) = 2.45$, $p = 0.077$ for naïve). However, one-way ANOVA on both ratings

together, indicates that the two groups together do perceive a significant difference at the $p < 0.05$ level with $F(3,87) = 3.34$. Both experienced and all participants perceive external consistency as the least important while internal consistency is perceived as the most significant.

9.4.2 Case Study

The goal of this section is to provide a case study, validating the concept of contextual dependency for both FM and BM. The stimuli for the FM experiments were created using iFACE, a 3D facial animation software, developed by Arya and DiPaola [263].The BM stimuli were from the Carleton dataset described in Appendix A. Alterations to the BM data, where needed, were applied in MATLAB and the data were shown in a combinational point light and stick-figure form. A total of 23 naïve participants, 14 males and 9 females, with a mean age of $M = 21.6$ and $SE = 4.0$ took part in this experiment. The necessary ethics approval was secured. No compensation was provided for the participants' time. A set of descriptions regarding the context of the scene was provided to the participants. The stimuli were then presented on a 14.6 inch LCD laptop screen and participants were asked to rate the validity of animation to represent the context provided, on a 10-point Likert scale.

Two sets of experiments were carried out one for FM and one BM. Each experiment consisted of three cases. Case 1 dealt with slight primary actions and neutral secondary themes, case 2 investigated the effect of unexpected primary actions, and case 3 was regarding the effect of unexpected secondary themes in the animation.

In experiment 1, the scenario illustrated a travel agent describing different flight

schedules for a vacation to a tropical resort to a customer. The customer was happy with the offers and planning to take the trip. Then, three sets of stimuli were presented to the participants, all belonging to a white female agent. In the first case, the agent did not show any particular facial actions or expressions (neutral), only mild nodding occurred. The second case showed fast single eyebrow rises and no particular expressions. Finally, in the third case expressions of sadness were displayed. Figure 9.4 illustrates these facial expressions.

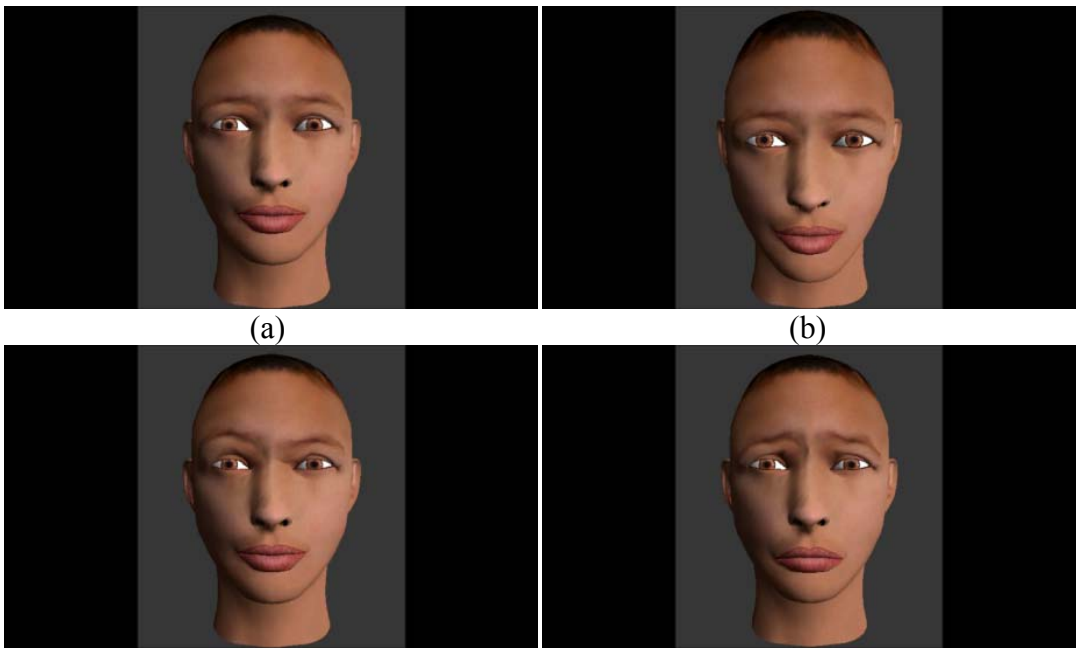


Figure 9.4. Frames from the video used in experiment 1 for testing contextual dependency: (a) presents a neutral face, (b) is captured from case 1 showing neutral face with slight nodding, (c) is captured from case 2, illustrating fast single eyebrow rising, and (d) from case 3, showing expressions of sadness.

In experiment 2, it was described that the same agent from experiment 1 printed the ticket (after being purchased by the customer) and walked to the printer in the other side of the room to pick it up. Case 1 of this experiment showed a neutral female walk while in case

2 a regular female walk was presented during which the agent threw random punches in the air, and in case 3, a tired female walk was presented. Figure 9.5 illustrates frames from the walking stimuli.

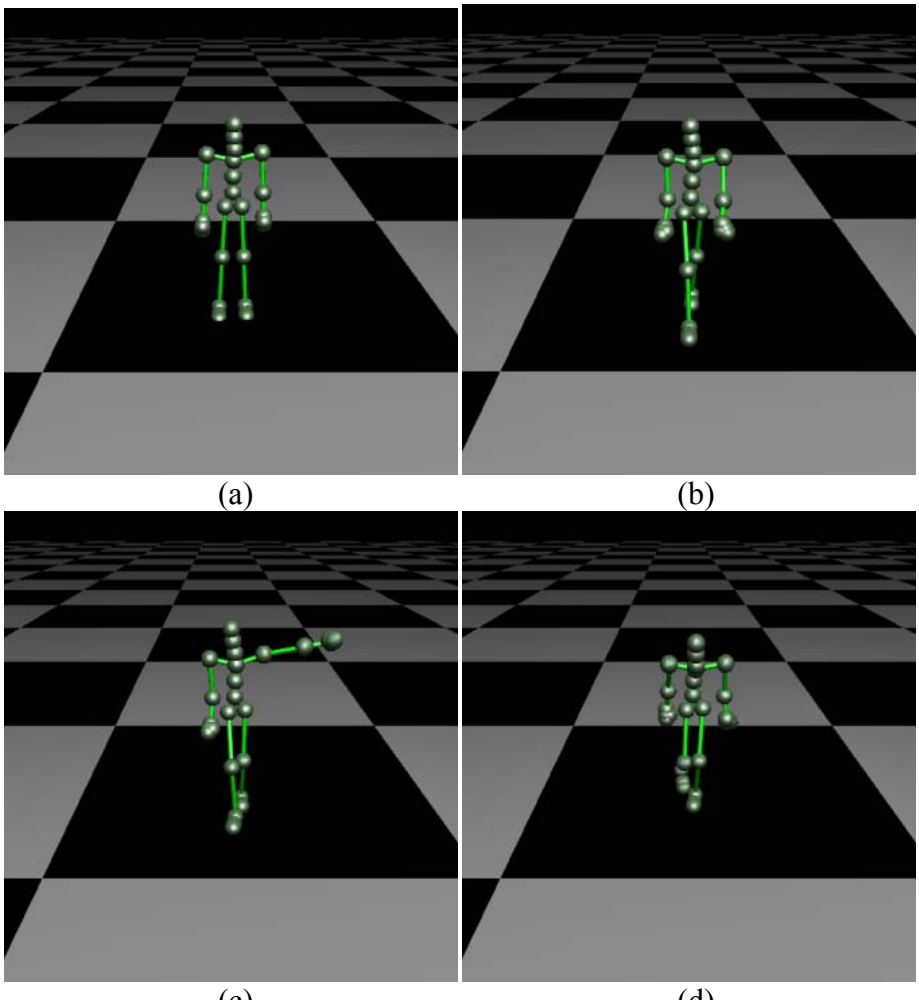


Figure 9.5. Frames from the video used in experiment 2 for testing contextual dependency: (a) presents a neutral stance, (b) is captured from case 1 showing neutral walk, (c) is captured from case 2, illustrating fast normal walk and sudden punches on the way, and (d) is from case 3, showing very tired walking.

The results for the experiments are illustrated in Figure 9.6 where the mean ratings are displayed and error bars represent the standard errors. One-way ANOVA shows that

there was a significant effect for contextual dependency in FM domain at the $p < 0.001$ level ($F(2,66) = 48.9$). A similar effect was observed in BM domain at $p < 0.001$ ($F(2,66) = 61.28$). It is generally observed that for case 1, where contextual dependency has been taken into consideration, in both FM and BM domains, approval rates are quite high. For case 2 where dependency has been disregarded for the PT, ratings drop in both domains. For BM, however, the drop rate is slightly higher than FM. Similarly, in case 3, where dependency has been disregarded for the ST, the ratings drop in both domains. The drop rate for case 3 in general is less than that of case 2. Also, a higher drop is observed for FM.

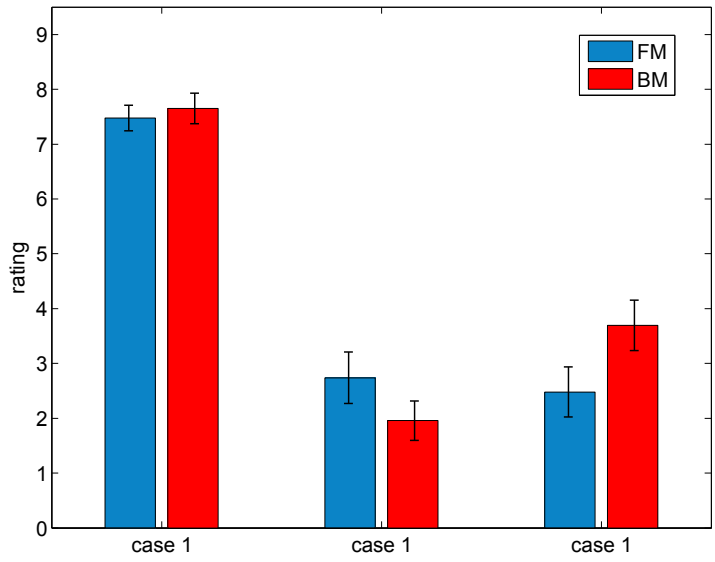


Figure 9.6. Mean ratings and standard errors for contextual dependency experiments in FM and BM domain. Case 1 represents full dependency while in case 2, the component has not been considered for the PT, and in case 3, the component has not been considered for the ST.

9.5 Discussion

9.5.1 Relative Significance

The study on the relative impacts of the elements of our proposed paradigm showed that as expected, different elements and components maintain different significance levels, the most important of which (according to experienced participants) is consistency between PFs in FM. Moreover, consistency of SFs in FM and proper display (association) of SFs in FM have received higher ratings with respect to others. This is in accordance to earlier studies emphasizing the significant impact of facial features [264]. Moreover, it is interestingly observed that the least important element is internal consistency between PFs of FM and BM while the importance of internal consistency between SFs of FM and BM is quite high. This indicates that the face and body can perform independently in terms of actions, but not emotions and expressions. In addition, proper display of PFs in BM is not perceived as significantly important (compared to that of SFs), which points to proper display of SFs being generally more important. This could be due to the fact that SFs, by nature, are spatiotemporally smaller compared to PFs, and slightest alterations can change the expression, while for primary actions, there is more room for variation (spatiotemporally). In other words, the perceptual message of PFs can be conveyed even when they are not visually sound while the same cannot be said about SFs. Finally, internal consistency is perceived as the most important component while having the least important element, namely, consistence of the PF of BM with PF of FM. This indicates that aside for consistence of PFs for FM and BM, other elements of this component are very vital in character animation.

The 11 experienced participants, when exposed to the paradigm and asked about its potential applications for animators, provided an average 8.3/10 approval rating with *SE* = 1.5. This indicates that the model as a whole is perceived to be useful and helpful by people with experience in animation of human motion. The animators have generally felt that the paradigm along with its ranking of elements and components can provide a valuable set of guidelines on when and how to modify and process motion animation.

9.5.2 Case Study

In order to validate all the different possible elements of the paradigm, numerous experiments would be required. In this chapter, we provided a case study in support of contextual dependency. Association has been widely studied in the literature, and the two consistency laws are quite intuitive and self-evident. Our experiments in both FM and BM domains illustrated significant effect for contextual dependency. This means in both FM and BM, and for both PT and ST, disregarding contextual dependency significantly reduced the validity of the sequences for the participants. In case 1, where the original sequences were displayed, BM shows higher ratings. This can be due to the fact that the BM stimuli were acquired using a motion capture system while the FM stimuli were synthesized using artificial CG techniques. In case 2, where unexpected primary actions were displayed, ratings for both FM and BM significantly dropped. This is quite expected based on the model and follows the proposed paradigm. Unexpected movements with no particular explanation for the actions reduce the validity significantly. Written feedback by the participants showed that some sort of reason for punching or sudden eyebrow raise must be provided in order for the sequences to be valid. Display of excess movements for BM, on the other hand, reduced the validity slightly more than FM. This can be due to

many reasons, but written feedback by the participants provided an interesting insight into this trend. Most participants who, in case 2, provided higher ratings for BM, associated the sudden eyebrow raising of the FM stimuli to some sort of nervous tick. This contradicts the findings of the comparative impact study (where primary themes of FM are perceived as more important), but it should be noted that the difference between corresponding elements of Figure 9.2 are not significant. In case 3 where unexplained secondary actions were displayed, the FM ratings almost stayed the same (compared to case 2). Therefore, a sad face, according to most participants, is not at all valid when applied to a travel agent about to sell a ticket. The BM ratings, however, showed an approximate 16% increase in validity. Written feedback from participants related this increase to the probability of the agent being tired from a long working day (or being old in some cases). STs of FM being more important than STs of BM, is in accordance with the comparative impact study.

An important point to consider is that when creating synthetic animation content, the animator might intentionally disregard some aspects of PV due to implementation of a particular style or a specific agenda or role of that animated character. In such cases, one should not expect to precisely witness the estimated impacts of the elements of the proposed model.

9.5.3 Disney's Principles for Animation

In order to further analyze our proposed paradigm, we compare and map the components of PV to Disney's principles of animation [27]. The goal of this analysis is to examine how the concepts of PV correspond to Disney's set of 12 principles and vice versa. It should be noted, however, that Disney's principles recount for general animation and are

not limited to human motion unlike our model which only discusses FM and BM. Table 9.2 shows the correlation between the two paradigms in human motion domain.

Table 9.2. Mapping Disney's set of principles for animation with components of PV. A: Association; CD: Contextual Dependency; EC: External Consistency; IC: Internal Consistency.

	Disney Principles		PV Components			
No.	**Principle**	**Summary/Implications**	**A**	**CD**	**IC**	**EC**
1	Squash and stretch	Volume of squashed or stretched object is constant	✓			
2	Anticipation	Specific movements are anticipated based on physical properties; Focus on an object about to be subjected to force		✓		
3	Staging	Presenting an action, mood, personality, or expression clearly	✓	✓		
4	Straight ahead action and pose to pose	Animation process: "straight ahead action" and "pose to pose"	✓			
5	Follow through and overlapping action	Parts of bodies (with degrees of freedom) will move after the body has stopped; Different parts of a body can move with different rates	✓			
6	Slow in and slow out	A body needs time to accelerate and decelerate	✓			
7	Arcs	Natural actions often follow a path of a trajectory in arch format	✓			
8	Secondary action	While a character performs a main action, it can perform secondary (smaller) supporting actions			✓	✓
9	Timing	Correct timing (number of frames per action or second) results in realistic scenes	✓			
10	Exaggeration	Realism vs. style: Disney preferred realism but in a bit wilder form	✓	✓		
11	Solid drawing	Animation in 3D with weight and volume	✓			
12	Appeal	Characters being charismatic and interesting, even if not necessarily sympathetic	✓	✓		

Many of the Disney's principles are concerned with correct presentation and preservation of animation cues as well as physical laws. As a result, they can be related to association. Principle 1, squash and stretch, is concerned with weight and volume, hence physical

characteristics. Thus, it can be related to association where correct depiction of motion cues is proposed, regardless of whether the cues belong to PTs or STs. Often regarded as the most important principle, it resonates with the first component of our paradigm. The fourth principle is solely concerned with animation techniques and notions such as volume, size, proportions, etc. As a result, this principle is only similar to association. Principle 5, follow through and overlapping action, suggests that some degrees of freedom (DOFs) of a moving body keep moving when that body comes to a stop, referring to physical kinetic laws. As this principle strictly considers physical rules, we can only relate it to association in the framework of human motion. Similar to principle 5, the sixth principle called slow in and slow out, is concerned with acceleration and decelerations of bodies, referring to physical laws yet again. Consequently, we map this principle to association. The seventh principle, arcs, suggests that natural motion is generally arc-like. While this property is difficult to map onto our paradigm, we can only distinctly relate it to association where naturally and correct appearing features and cues are discussed. The ninth principle, timing, suggests that correct frame rates and timing results in realistic scenes. As this law is directly in regards to preservation and correct presentation of animation content, it can be mapped to association. Principle 11, solid drawing, again emphasizes robust visualization, taking into account volume, weight, solidity, and other factors in 3D. As a result, similar to several other principles, this is mapped onto association.

Anticipation, which is the second principle, puts focus on bodies or objects in the scene. Naturally, this principle is developed through the events that lead to the particular scene, hence, context. Therefore, we relate this principle to contextual dependency.

The third principle called staging directly relates to association as it dictates correct and clear depiction of PTs and STs. Moreover, this principle demands the themes to be clearly displayed based on the "story", which in turn dictates the context. Therefore, the principle also related to contextual dependency.

Principle 8 is called secondary action. This principle dictates that a main action is accompanied by a series of less significant yet "supporting" actions. While the terminology for this principle does not clarify whether STs are considered a form of secondary actions, it is safe to assume that this in fact is the case. As a result, this principle captures both consistency principles namely internal consistency and external consistency. In other words, the general idea behind this principle can be regarded as action cues (regardless whether primary or secondary) being in support of one another and not contradicting each other.

Exaggeration, the tenth principle, prohibits extreme visual distortion of content. Suggesting that animation should be depicted realistically, we can conclude that preservation and presentation of cues in natural format is preferred through this principle. Thus, the principle closely relates to association.

Finally, the twelfth principle called appeal, suggests that characters should be charismatic and interesting, even if they are not necessarily sympathetic. Here the concept of appeal can refer to correct and accurate content from a visual and form-related point of view, hence association. However, it is has been suggested that characters can significantly appeal to the audience based on the story-line, events, decisions, and actions. Therefore, this principle relates to contextual dependency of the PV paradigm as well.

As shown above, most of Disney's animation principles are concerned with appealing and accurate presentation and format of animation content, hence PV's association. The proposed paradigm, however, is more general and focused on consistency of cues with one another as well as context. While PV is by no means intended to replace Disney's set of principles, we believe it can be a valuable addition to the existing set of guidelines for animators, specifically for those in the field of motion animation.

9.6 Summary

In this chapter, we proposed an empirical paradigm based on which the methods presented in previous chapters can be utilized. The goal of the paradigm is to provide a set of guidelines for animators employing human motion. We suggest that taking into account the 18 elements of the paradigm can ensure *perceptual validity* in animated motion scenes. Specifically, the paradigm suggests that visual cues that compose actions (primary and secondary) need to be spatiotemporally and perceptually accurate. Moreover, PTs and STs need to be consistent with the context of the animation scene. Finally, the STs and PTs need to be consistent, both internally and towards one-another.

In addition to the PV paradigm, other factors affect how appealing an animated human motion sequence is perceived. Studying the overall quality of animation and defining measures for that are beyond the scope of this dissertation, and can include a variety of spatial and temporal factors. Presence of noise is one of these factors, for which a vast amount of research has been carried out (to recognize and eliminate different types of noise). A survey can be consulted at [265]. Temporal and spatial resolutions, rendering

techniques, texturing, and lighting are among other parameters which can have significant effect on visual and format-related video quality. These issues are often addressed in terms of both hardware and software. Style and target audience are also among parameters that need to be considered alongside visual and format-related quality and within a broader study.

An important conclusion that can be drawn from the proposed model and supporting arguments is that for animators to create characters that behave in a perceptually valid manner, human motion content needs to be carefully controlled. Alterations based on context, style, and other intentions, may need to be applied to pre-recorded or pre-animated sequences. While cues for PTs often have a more significant spatiotemporal presence, cues for STs are even more important from a perceptual standpoint. Thus, accurate and perceptually guided processing of STs taking into account the notion of PV in motion is an essential step towards creating the perfect animation.

Chapter 10.

Concluding Remarks

10.1 Research Conclusions

In this dissertation, we proposed that style and affect, or more broadly, secondary themes (STs) and corresponding secondary features (SFs), are critical in animation of human motion. These features are the foundation of achieving *personality* in animated characters and lead to *natural* and *believable* motion sequences.

As we humans are the target audience of animated motion scenes presented in different media such as animated movies, digital games, and virtual worlds, accurate and valid perception of synthetic or recorded STs in motion are of critical importance. Hence, we believe that perceptually guided motion processing procedures should lead the way in

different aspects of human motion especially STs. The procedures for which we proposed such methods include time warping for temporal alignment (which is utilized in almost every motion processing system), extraction of SFs, synthesis of SFs, unified recognition/translation of SFs, and a set of guidelines for employing these systems to achieve higher perceptual accuracy.

10.2 Summary of Contributions

In different chapters of this dissertation, we developed several perceptually guided and accurate systems for different types of motion processing. Following, we present a summary of our contributions and advantages of the developed systems:

- A new time warping method for motion:
    - Our method uses correlation as the basis of characterizing alignment which we showed to be more accurate and suitable compared to distance, which is mostly used as a determinant for alignment.
    - The proposed method outperforms several other techniques in terms of alignment and also produces smooth animation with low distortion.
    - The time warping method benefits from high spatial and temporal customizability which can be beneficial for different applications.
    - When multiple sequences are being processed, the optimum sequence which would demand the least warping from other sequences to achieve

alignment can be automatically computed using our system.

- We validated the notion of additivity in perception of affect from limb motion. In this study we showed that the sums of perceptual intensities of affect for single limbs are highly correlated with perceptual intensities of affect from multiple limbs, hence, additivity.

- Extraction of SFs:

    - A model was proposed and formalized that describes the relationship between PTs and STs.

    - A system was developed that extracts SFs as three separate components: posture, movement, and time (speed).

    - Our method performs in spatiotemporal Cartesian or joint angle spaces, and does not explore latent spaces which often require additional processing for interpreting the results.

    - Through style translation, we showed the high precision of the extracted SFs.

- Expert-driven perceptual shortcuts for SFs:

    - Gaussian radial basis functions (RBFs) were introduced and validated as accurate means of modeling SFs.

    - We developed an interface using which SFs can be produced and added to

motion sequences using Gaussian RBFs.

- Animators converted a neutral walk to happy, sad, tired, energetic, feminine, and masculine walks. The added functions were recorded and analyzed, using which, a set of SFs were derived for each theme.
- A separate group of participants validated the derived sets of features.
- We showed that only using a few of the features can convey the intended themes.
- The intensities of the features are controllable (scalable) making our approach highly dynamic.
- A variety of different properties regarding execution and perception of affective and stylistic motion such as the percentage of posture-related features vs. movement-related features was revealed.
- The inversion effect, meaning perception of opposite themes with imposed negative weights to the features, was incorporated and addressed through the system.

- A unified system for recognition and translation of SFs:
  - Inspired by how humans are capable of performing both, we developed a system that can achieve both with minimal need for adjustments.
  - Ensembles of Gaussian RBF neural networks were used for the system.

- The system outperforms several types of k-nearest neighbor (KNN) and support vector machine (SVM) classifiers.
- The system was capable of converting neutral inputs to stylistic outputs. The style translation results were validated through audience perceptual feedback.

- An empirical paradigm for perceptual validity:
  - We proposed a paradigm that incorporates different components regarding accurate preservation and presentation of cues in motion, consistency of existing features with the presented context, inner-consistency of actions and SFs, and inter-consistency of actions with SFs.
  - Both body motion (BM) and face motion (FM) is incorporated into the paradigm.
  - We suggest that animators can use this paradigm as a reference for different factors that need to be taken into account to achieve perceptually valid motion sequences.
  - A case study validated one of the components of the paradigm.
  - Animators and naïve participants were asked to provide a ranking for different elements of the paradigm, according to which, the relative significance of different elements was revealed and discussed.
  - By mapping the set of proposed components and elements of the

paradigm with Disney's principles of animation, we showed that while there is interpretive overlap between the two, they explore different realms, as Disney's is more focused on general animation while PV explores details of motion animation. We believe the two can be used in complementary fashion alongside each other for animation purposes.

10.3 Future Directions

There are enhancements that the proposed systems can benefit from. As our methods such as CoTW (Chapter 4), spline-based extraction of features (Chapter 6), and training of ensembles of RBF neural networks (Chapter 8), utilize many motion data with high DOFs and long temporal lengths, and often utilize iterative processes, computational costs and runtime were quite high. This is despite the fact that measures such as dynamic programming have been taken to increase speed. Therefore, speedup techniques should be explored in the future. These methods include utilizing state of the art software and hardware, implementing the systems on GPUs, implementing the systems using lower level programming languages such as C/C++ or Java, and exploring more efficient programming algorithms and techniques.

The study on additivity in perception of affect from limb motion, presented in Chapter 5, can be expanded to provide a more accurate model for describing the proposed relationship between single and multiple limbs in perception of STs. A detailed study aiming at the underlying reasons for the lack of perfect additivity can be conducted. Finally, the data required to expand the study to other classes of affect and style can be

recorded and variations of the experiments can be designed and carried out.

The extracted features using our spline-based method presented in Chapter 6, can be compared to the sets of developed features based on the study in Chapter 7. Such a study can reveal interesting conclusions with respect to execution and perception of SFs in motion.

The unified system for recognition and transfer of SFs in Chapter 8 can be enhanced to achieve more generalization. A system capable of accurately transferring learned features from a particular action (for example walking) onto another action class (such as jumping or running) can be an extremely valuable asset to the field. Developing unified systems based on other classifiers such as hidden Markov models (HMMs) as well as the simple classifiers used for recognition such as KNN and SVM can also prove valuable for evaluating the proposed approach.

With respect to the different proposed systems for extraction and synthesis/translation of features, expanding the data to other classes of affect and style such as angry, nervous, young, old, unhealthy, etc. can be a valuable step. Furthermore, incorporating combinational themes for affect-gender-energy can be an interesting extension. For example, actions performed as happy-tired or feminine-sad can be considered, which can introduce a wide range of obstacles. Overcoming these obstacles can lead to more accurate and practical systems and possible development enhanced machine learning techniques.

The proposed PV paradigm can be used in a practical game or animation engine for synthesis or control of animated motion. User studies can respectively be carried out to

better evaluate and perhaps update the model. Such studies can lead to a great deal of new questions in perceptual motion processing which can ultimately further enhance the state of the art.

Finally, for different methods presented in this dissertation, opinions of experienced animators can play a more significant role. Animators can and should be consulted regarding our systems and findings and expert-knowledge can be more central. Developing practical tools and incorporating opinions of animators, we believe, can be a valuable addition to the existing research.

Appendix A

Datasets and Data Structure

A.1 Motion Capture Process

Motion capture is the process of recording biological motion. Different methods and sensor types have been developed among which inertial sensors [266], accelerometer sensors [267], magnetic markers [268], marker-less optical [269], and marker-based optical [270] have been widely used. A review on the history of vision-based motion capture can be found at [271, 272] and a survey on recent advances in the field is available at [8]. The motion capture data used in this dissertation have been recorded using optical marker-based systems.

Optical marker-based motion capture systems record motion sequences in 3D space.

Typically, in such systems, 40-50 light-reflective markers are placed on a bodysuit worn by the performer whose motion will be recorded. These markers are tracked using 6-12 cameras which emit infrared. The cameras have very high temporal (up to 240 Hz) and spatial (less than 1 mm) resolutions. The locations of the markers recorded by all the cameras are processed through the system software and a 3D body model is generated along with a motion matrix that contains the locations of a number of virtual markers or joints in 3D space. The virtual markers generally overlap real body joints. Figure A.1 illustrates a motion capture process along with the model. Here, the body model in the figure is composed of 18 virtual joints.

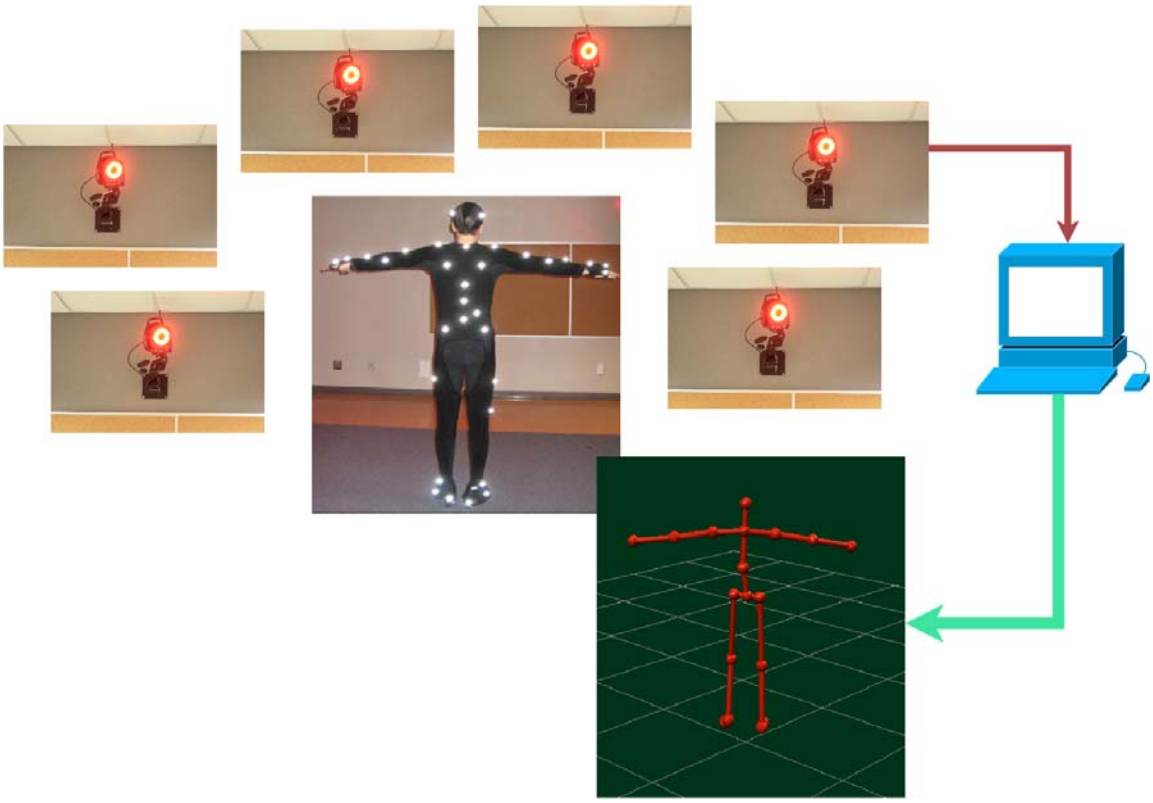


Figure A.1. Schematic of a motion capture process. Light reflective markers are attached to the body suit and tracked using cameras that emit infrared. The computer software then computes the articulated motion model and matrix using the data received from the cameras.

A.2 Data Representation

A motion sequence can be characterized through the location of body joints at each frame or instance of time, or through other means such as joint angles. The data used in this dissertation are in the latter form. In this representation, the orientation of each joint is represented by three rotation angles with respect to a local set of axis on the parent joint. A motion sequence can subsequently be represented by a number of consecutive postures variable with time. In other words,

$$\boldsymbol{\mathcal{D}} = [\boldsymbol{p}^{(1)}\ \boldsymbol{p}^{(2)} \dots \boldsymbol{p}^{(m)}]^T \quad (A.1),$$

where, $\boldsymbol{p}$ represents each posture and m is the number of frames ($m \in \mathbb{N}$). In turn, each posture can be represented by a finite number of joints or virtual makers. Therefore, $\boldsymbol{p} \in \mathbb{R}^{3l}$ where l is the number of joints representing each posture in three dimensions ($l \in \mathbb{N}$). The ith posture is consequently be represented by:

$$\boldsymbol{p}^{(i)} = [\theta_{1,x}^{(i)}\ \theta_{1,y}^{(i)}\ \theta_{1,z}^{(i)} \dots \theta_{l,x}^{(i)}\ \theta_{l,y}^{(i)}\ \theta_{l,z}^{(i)}] \quad (A.2),$$

where, θ is the joint angle, with $\theta \in \mathbb{R}$ and $0 \leq \theta < 360$. Accordingly, the joint angle trajectory of the jth degree of freedom (DOF) is described by:

$$\boldsymbol{\theta}_j = \left[\theta_j^{(1)}\ \theta_j^{(2)} \cdots\ \theta_j^{(m)}\right]^T \quad (A.3),$$

An additional displacement vector characterizes the root joint. This root joint, often set as the hip joint, locates the character in Cartesian space. Therefore, the displacement vector is defined by $\boldsymbol{d} = \left[\boldsymbol{d}_x\ \boldsymbol{d}_y\ \boldsymbol{d}_z\right]$, where:

$$\boldsymbol{d}_x = \left[d_x^{(1)}\ d_x^{(2)} \cdots\ d_x^{(m)}\right]^T \quad (A.4),$$

and $\boldsymbol{d} \in \mathbb{R}^{3m}$. Based on this definition, a motion sequence, in addition to the Eq. A.1 representation, can be formalized by:

$$\boldsymbol{\mathcal{D}} = [\boldsymbol{d}\ \boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \cdots \boldsymbol{\theta}_n] \quad (A.5),$$

with the motion matrix has $n + 3$ DOFs (for l-joint model, $n = 3l$).

A.3 Datasets

Three different datasets are used in this dissertation, namely the Carnegie Mellon University motion dataset (referred to as the CMU dataset), the HDM05 dataset from Max Planck Institute for Computer Science, and the Carleton University dataset.

The CMU dataset (http://mocap.cs.cmu.edu) is perhaps the most complete and widely used dataset of motion capture files. It contains over 2600 sequences in a variety of different actions, activities, interactions, and styles. The *bvh* (Biovision hierarchy) versions of the data that are widely used for motion studies are available at: https://sites.google.com/a/cgspeed.com/cgspeed/motion-capture/cmu-bvh-conversion.

The *bvh* files contain joint structures, joint angle trajectories, and the root joint location trajectory (the latter two are combined in the form of $\boldsymbol{\mathcal{D}}$ as described above). To the best of our knowledge, details such as those regarding the actors are not disclosed. A description on the capture process and marker format is available at http://mocap.cs.cmu.edu/info.php.

The HDM05 dataset (http://www.mpi-inf.mpg.de/resources/HDM05) is also used in our research. Several types of actions, including walking, is carried out by 5 actors with different classes of affect. A total of approximately 1500 sequences are available in this dataset. Neutral versions of the actions are also available. Details regarding the joint structure, capture process, available sequences, and actors can be found in [270].

Finally, we recorded our own dataset, which we refer to as Carleton dataset. Recording for this dataset is ongoing. So far, 5 actors have performed several types of actions with different styles such as feminine, masculine, energetic, and tired, among others. Recording is carried out via a Vicon MX40 motion capture system. Around 100 long motion sequences each consisting of multiple actions and styles have been recorded. Different joint structures have been used for different types of applications.

Appendix B

Questionnaires

Here, we present a sample questionnaire used to conduct the user studies described in this dissertation. The consent form read and signed by each participant is first presented, followed by samples of question. As there is a great amount of overlap in the questions posed for the studies in different chapters, only samples are provided. For example, when the audience are asked to select the affect or style of the animated walker, for different chapters, different combinations of answers are used, whereas in this appendix, only one sample is mentioned. Similarly for rating the amount of perceived affect/style, a variety of different sequences and themes are presented and asked to be ranked, where, in this appendix, only one example is presented.

B.1 Participant Informed Consent Form

You have been solicited as a research participant for our project entitled:

Perception of Human Motion and Styles

The research is being conducted by:

- Dr. Ali Arya, School of Information Technology, Carleton University, Ottawa, Canada, arya@carleton.ca
- S. Ali Etemad, Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada, etemad@sce.carleton.ca

Purpose

This project aims at understanding the way human viewers perceive characteristics such as emotion, gender, and energy in human motion. Your answers in this questionnaire will be used to analyze how participants perceive different style of human motion.

Task Requirements

- You will be asked to watch videos of human motion and answer questions accordingly.
- Each action will be replayed until the participant decides to answer the question.
- The information regarding your age, gender, and department might also be used in this study.

Dissemination

This research is a part of a project supervised by Dr. Ali Arya (School of Information Technology) and conducted by S. Ali Etemad (Ph.D. Candidate, Department of Systems

and Computer Engineering) towards a Ph.D. degree at Carleton University. The result of this research will be used in Mr. Etemad's dissertation and might also be published and/or presented in conferences and/or journals, as well as grant applications.

Anonymity/Confidentiality

You may choose to provide your names and emails or choose to remain anonymous. If you do provide your name and email, however, they will not be used or published in the research as they will only be kept for possible future verification.

Risks

There are no known risks associated with this activity.

Right to Withdraw

As a participant, you may withdraw at any time for any reason.

Compensation (only used for the study conducted in Chapter 7)

Participants will be compensated with a 10$ gift card for their time.

Ethics Approval

This research has been reviewed and cleared by the Carleton University Research Ethics Board (REB) and questions and concerns can be addressed to the REB chair.

Research Ethics Board:

Professor Antonio Gualtieri, Chair

Research Ethics Board

Carleton University Research Office

Carleton University

1125 Colonel By Drive

Ottawa, Ontario K1S 5B6

Tel: 613-520-2517 E-mail: ethics@carleton.ca

Your signature below indicates that you have read the above and voluntarily agree to participate. If you have any questions, please ask them before signing.

***** I have read and understand the above information *****

Participant's name: _____

Phone number: _____

Email: _____

Gender (mandatory): _____

Age (mandatory): _____

Program of Study: _____

Year of Study: _____

Nationality: _____

Signature (mandatory): _____

Date (mandatory): _____

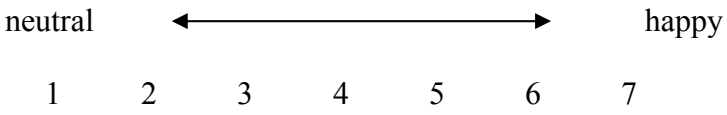
B.2 Question Type I

What is the style of the displayed sequence?

neutral happy sad energetic tired feminine masculine

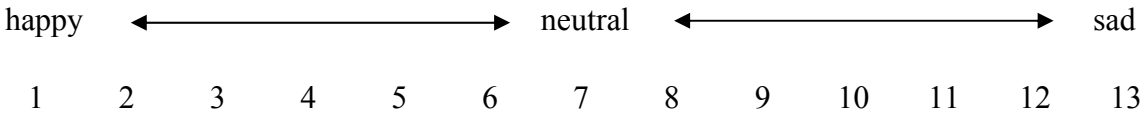
B.3 Question Type II

Rank each of the following sequences from 1 to 7 in terms of Happiness, 1 meaning it is completely Neutral, 7 meaning it is very Happy.



B.4 Question Type III

Rank the following sequence from 1 to 13 in terms of Happy-Sad: 1 meaning it is Happy, 7 meaning it is Neutral, and 13 meaning it is Sad.



B.5 Question Type IV

1- Do you have experience in animation?

Yes_____ No_____

If yes, what is your level of expertise?

Professional_____ Experienced_____ Working knowledge_____ Beginner_____

2- Do you have experience in motion capture?

Yes_____ No_____

If yes, what is your level of expertise?

Professional_____ Experienced_____ Working knowledge_____ Beginner_____

3- In this research, participants will add curves (in the form of Gaussian functions shown in Figure B.1) to the motion of different joints on the body in order to convert a neutral walk into different types of walk for example, happy, sad, feminine, masculine, tired, and energetic.

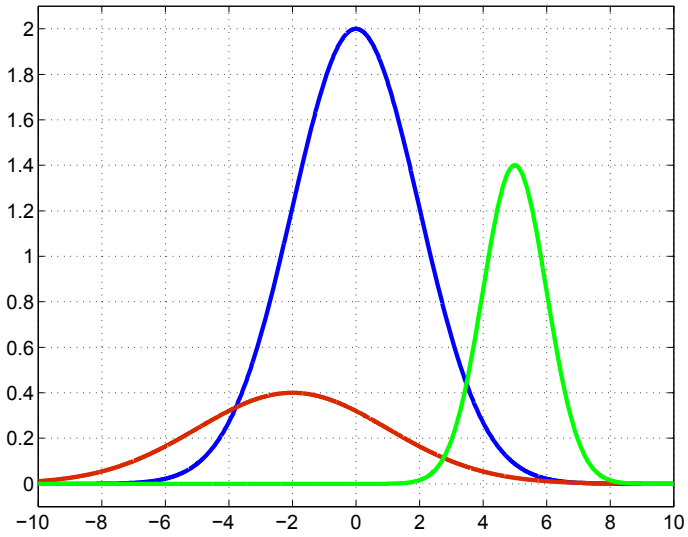


Figure B.1. Gaussian functions.

The added functions will be studied and used to determine how people perceive different types of motion.

Now please proceed with the experiment:

The goal is to convert the neutral sequence to happy, sad, tired, energetic, feminine, masculine

i) load the neutral walk

ii) clear all variables and start

iii) play the sequence (the sequence plays for three cycles)

iv) you can load a set of functions that you produced earlier (optional)

v) select a joint and axis (you can choose to add functions to all of the joints and axis or a

subset of them)

vi) create the functions using the sliders

vii) generate the sum of the functions

viii) apply them to the motion

ix) play the sequence

x) you can go back to your set of functions and modify them if necessary

xi) save the functions when comfortable with the outcome

xii) go back to (*i*) and redo the process for new style

B.6 Question Type V

Please answer the following question:

Proper display of bodily actions features is critical for generating perceptually valid animation:

Strongly disagree ⟷ Strongly agree

1 2 3 4 5 6 7

B.7 Question Type VI

Please read the following description about a research project:

In this research, participants will add curves (in the form of Gaussian functions shown below) to the motion of different joints on the body in order to convert a neutral walk into different types of walk for example, happy, sad, feminine, masculine, tired, and energetic. The added functions will be studied and used to determine how people perceive different

types of motion.

a) Does this research make sense to you?

Yes_____ No_____

b) Do you think the findings of this research can be useful for an animator who constantly works with human motion?

Yes_____ No_____

If yes, how helpful do you think this it is (5 is the highest)?

1 2 3 4 5

c) Do you think the tools developed as a result of this study can be useful for an animator who constantly works with human motion?

Yes_____ No_____

If yes, how helpful do you think the tools can be (5 is the highest)?

1 2 3 4 5

Appendix C

Implementation

The programs and functions that were used for this dissertation were implemented in MATLAB on a computer with a 64-bit Windows operating system, an AMD Athlon 2.80 GHz dual core processor, and 3.00 GB of RAM.

For loading the data, a routine capable of parsing and saving the motion matrix in MATLAB's workspace was used. The routine also enables saving the motion files as numeric matrices or *bvh* files in designated system folders. A separate routine capable of displaying motion data, in the form of pint-light and stick-figure, for users either prior or after use in the systems was developed. The routine also enables rendering of the motion data. The displayed sequences and images of animation poses such as Figure 5.15 were rendered directly using this routine. The function provides customization capabilities

such as selecting the frame rate, joint and bone colors, angle of view, etc. Extensions (file formats) of rendered frames can also be selected by users.

A variety of different pre and post-processing routines were implemented for this study. These routines include segmentation and saving long sequences as several smaller sequences, low pass filtering for noise removal, automatic removal of idle-frames (in which the character illustrates no motion prior or after performing a main action), foot-skate cleanup, and many more.

Some of the systems that were developed can be widely beneficial for researchers in the field. For example:

- our proposed time warping technique (Chapter 4),
- the method developed for spatiotemporal extraction of secondary features in three separate movement, posture, and time components (Chapter 6),
- the interface for collecting expert-knowledge for generating different variations of an input sequence (Chapter 7),
- the proposed set of features for expert-driven synthesis of secondary features (Chapter 7),
- the neural network system for recognition and translation of secondary features (Chapter 8).

Additionally, other routines such as the following can be very helpful for researchers in the field:

- data load-and-parse,
- save data and *bvh*,
- motion display,
- motion rendering,
- pre-processing: segmentation, noise filtering, removal of idle-frames,
- post-processing: foot-skate cleanup.

We intend to release these resources for public use in the future.

Generally, implementation of the different systems was carried out with no particular problem. The only issue that we stumbled in few of the systems was runtime. Three main factors contributed to the high runtimes:

- number of sequences required for configuration, training, and testing the systems,
- high dimensionality (degrees of freedom and number of frames) of the data,
- considerable computational demand of the implemented systems such iterative processes that were frequently used (for example in the time warping method in Chapter 4 or training of the neural networks in Chapter 8).

These factors are intrinsic in most motion processing studies and difficult to overcome. Nevertheless, we were able to achieve reasonable processing speeds by employing efficient algorithms. For example, instead of a typical iterative search for the time

warping method presented in Chapter 4, we employed dynamic programming, which is significantly more efficient. Furthermore, for iterative loops, parallel-computing methods such as the *parfor* MATLAB function (parallel distribution of the "for" loop load across different cores) was used. As mentioned in Chapter 10, for future work, the systems, or at least some of the routines that are frequently used in different developed systems, can be implemented in C/C++ and compiled. This will significantly reduce runtime. Naturally, more powerful hardware such as the main processor, number of cores, and RAM can aid towards further reducing the runtimes. Finally, GPU implementation of the entire systems or some of the related subroutines can be explored as another possible solution.

References

[1] R. Parent, Computer Animation: Algorithms and Techniques, Morgan-Kaufmann, 2001.

[2] W. T. Freeman, D. B. Anderson, P. Beardsley, C. N. Dodge, M. Roth, C. D. Weissman, W. S. Yerazunis, H. Kage, K. Kyuma, Y. Miyake and K. I. Tanaka, "Computer vision for interactive computer graphics," *IEEE Computer Graphics and Applications,* vol. 18, no. 3, pp. 42-53, 1998.

[3] F. Farhadi-Niaki, S. A. Etemad and A. Arya, "Design and usability analysis of gesture-based control for common desktop tasks," *LNCS 8007, Proceedings of the 15th International Conference on Human-Computer Interaction International,* vol. 5, pp. 215-224, 2013.

[4] L. Jarmon, T. Traphagan, M. Mayrath and A. Trivedi, "Virtual world teaching, experiential learning, and assessment: An interdisciplinary communication course in Second Life," *Computers & Education,* vol. 53, no. 1, pp. 169-182, 2009.

[5] S. A. Etemad, N. Sepehri Boroujeni and A. Arya, "On the environmental impacts of virtual technologies," *Proceedings of the International Conference on Environmental Pollution and Remediation,* p. 88, 2011.

[6] R. M. Baecker, "Picture-driven animation," *Proceedings of the 1969 ACM spring joint computer conference,* pp. 273-288, 1969.

[7] T. W. Calvert, J. Chapman and A. Patla, "Aspects of the kinematic simulation of human movement," *IEEE Computer Graphics and Applications,* vol. 2, no. 9, pp. 41-50, 1982.

[8] T. Moeslund, A. Hilton and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding,* vol. 104, no. 2-3, pp. 90-126, 2006.

[9] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing,* vol. 28, no. 6, pp. 976-990, 2010.

[10] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien and D. Ramanan (Eds.), "Computational studies of human motion: Part 1, tracking and motion synthesis," *Foundations and Trends in Computer Graphics and Vision,* vol. 1, no. 2-3, pp. 77-254, 2005.

[11] B. Rosenhahn, R. Klette and D. Metaxas (Eds.), "Human motion: Understanding, modelling, capture, and animation," *Computational Imaging and Vision,* vol. 36, 2008.

[12] A. Arya, S. DiPaola and A. Parush, "Perceptually valid facial expressions for character-based applications," *International Journal of Computer Games Technology,* vol. 2009, p. 462315, 2009.

[13] R. McDonnell, F. Newell and C. O'Sullivan, "Smooth movers: Perceptually guided human motion simulation," *Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 259-269, 2007.

[14] S. A. Etemad and A. Arya, "Modeling and transformation of 3D human motion," *Proceedings of the 5th International Conference on Computer Graphics Theory and Applications,* pp. 307-315, 2010.

[15] C. Darwin, The expression of the emotions in man and animals, London: John

Murray, 1872.

[16] R. Blake and M. Shiffrar, "Perception of human motion," *Annual Reviews of Psychology,* vol. 58, pp. 47-73, 2007.

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

[18] F. E. Pollick, "The features people use to recognize human movement style," in *Gesture-Based Communication in Human-Computer Interaction*, Berlin Heidelberg, Springer, 2004, pp. 10-19.

[19] A. Normoyle, F. Liu, M. Kapadia, N. I. Badler and S. Jörg, "The effect of posture and dynamics on the perception of emotion," *Proceedings of the ACM Symposium on Applied Perception,* pp. 91-98, 2013.

[20] J. Min, H. Liu and J. Chai, "Synthesis and editing of personalized stylistic human motion," *Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games,* pp. 39-46, 2010.

[21] G. Liu, Z. Pan and Z. Lin, "Style subspaces for character animation," *Computer Animation and Virtual Worlds,* vol. 19, no. 3-4, pp. 199-209, 2008.

[22] R. Poppe, "Vision-based human motion analysis: An overview," *Computer Vision and Image Understanding,* vol. 108, no. 1-2, pp. 4-18, 2007.

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

[24] E. Hsu, K. Pulli and J. Popović, "Style translation for human motion," *ACM Transactions on Graphics,* vol. 24, no. 3, pp. 1082-1089, 2005.

[25] M. Brand and A. Hertzmann, "Style machines," *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, ACM Press/Addison-Wesley Publishing Co.,* pp. 183-192, 2000.

[26] C. Jones, Chuck Amuck: The life and times of an animated cartoonist, New York: Farrar, Straus & Giroux, 1989.

[27] F. Thomas and O. Johnston, The illusion of life: Disney animation, Bdd Promotional Book Co., 1981.

[28] E. F. Miller, David Hume, Essays, Moral, Political, and Literary, ed., 1985.

[29] M. Alam and J. S. Dover, "On beauty: Evolution, psychosocial considerations, and surgical enhancement," *Archives of Dermatology,* vol. 137, no. 6, pp. 795-807, 2001.

[30] M. Neff and E. Fiume, "Aesthetic edits for character animation," *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 239-244, 2003.

[31] M. Neff and E. Fiume, "AER: aesthetic exploration and refinement for expressive character animation," *Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 161-170, 2005.

[32] M. Neff and E. Fiume, "Methods for exploring expressive stance," *Graphical Models,* vol. 68, no. 2, pp. 133-157, 2006.

[33] K. Aizawa, K. Kodama and A. Kubota, "Producing object-based special effects by

fusing multiple differently focused images," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 10, no. 2, pp. 323-330, 2000.

[34] A. Jhala and M. R. Young, "Cinematic visual discourse: Representation, generation, and evaluation," *IEEE Transactions on Computational Intelligence and AI in Games,* vol. 2, no. 2, pp. 69-81, 2010.

[35] C. Browne and F. Maire, "Evolutionary game design," *IEEE Transactions on Computational Intelligence and AI in Games,* vol. 2, no. 1, pp. 1-16, 2010.

[36] S. Colton, "Creativity versus the perception of creativity in computational systems," *Proceedings of the AAAI Spring Symposium on Creative Systems,* pp. 14-20, 2008.

[37] C. Pedersen, J. Togelius and G. N. Yannakakis, "Modeling player experience for content creation," *IEEE Transactions on Computational Intelligence and AI in Games,* vol. 2, no. 1, pp. 54-67, 2010.

[38] G. N. Yannakakis and J. Hallam, "Real-time game adaptation for optimizing player satisfaction," *IEEE Transactions on Computational Intelligence and AI in Games,* vol. 1, no. 2, pp. 121-133, 2009.

[39] R. McGloin, K. L. Nowak, S. C. Stiffano and G. M. Flynn, "The effect of avatar perception on attributions of source and text credibility," *Temple University, Philidelphia, PA,* 2010.

[40] P. S. Reitsma and N. S. Pollard, "Perceptual metrics for character animation: sensitivity to errors in ballistic motion," *ACM Transactions on Graphics,* vol. 22, no. 3, pp. 537-542, 2003.

[41] L. Ren, A. Patrick, A. A. Efros, J. K. Hodgins and J. M. Rehg, "A data-driven approach to quantifying natural human motion," *ACM Transactions on Graphics,* vol. 24, no. 3, pp. 1090-1097, 2005.

[42] J. K. Tang, H. Leung, T. Komura and H. P. Shum, "Emulating human perception of motion similarity," *Computer Animation and Virtual Worlds,* vol. 19, no. 3-4, pp. 211-221, 2008.

[43] S. Chikkerur, V. Sundaram, M. Reisslein and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting,* vol. 57, no. 2, pp. 165-182, 2011.

[44] J. K. Hodgins, J. F. O'Brien and J. Tumblin, "Perception of human motion with different geometric models," *IEEE Transactions on Visualization and Computer Graphics,* vol. 4, no. 4, pp. 307-316, 1998.

[45] M. Cavazza, R. Earnshaw, N. Magnenat-Thalmann and D. Thalmann, "Motion control of virtual humans," *IEEE Computer Graphics and Applications,* vol. 18, no. 5, pp. 24-31, 1998.

[46] P. Hemp, "Avatar-based marketing," *Harvard Business Review,* pp. 48-57, 2006.

[47] A. J. Calder and A. W. Young, "Understanding the recognition of facial identity and facial expression," *Nature Reviews Neuroscience,* vol. 6, pp. 641-651, 2005.

[48] K. Isbister and C. Nass, "Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics," *International Journal of Human-Computer Studies,* vol. 53, no. 2, pp. 251-267, 2000.

[49] A. Arya, L. N. Jefferies, J. T. Enns and S. DiPaola, "Facial actions as visual cues

for personality," *Computer Animation and Virtual Worlds,* vol. 17, no. 3-4, pp. 371-382, 2006.

[50] J. Bates, "The role of emotion in believable agents," *Communications of the ACM,* vol. 37, no. 7, pp. 122-125, 1994.

[51] M. Paleari and C. L. Lisetti, "Toward multimodal fusion of affective cues," *Proceedings of the 1st ACM International Workshop on Human-Centered Multimedia,* pp. 99-108, 2006.

[52] C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer and K. Alvarez, "Developing multimodal intelligent affective interfaces for tele-home health care," *International Journal of Human-Computer Studies,* vol. 59, no. 1, pp. 245-255, 2003.

[53] J. Posner, J. A. Russell and B. S. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology,* vol. 17, no. 3, pp. 715-734, 2005.

[54] F. Simion, L. Regolin and H. Bulf, "A predisposition for biological motion in the newborn baby," *Proceedings of the National Academy of Sciences,* vol. 105, no. 2, pp. 809-813, 2008.

[55] R. Fox and C. McDaniel, "The perception of biological motion by human infants," *Science,* vol. 218, no. 4571, pp. 486-487, 1982.

[56] B. I. Bertenthal, "Infants' perception of biomechanical motions: Intrinsic image and knowledge-based constraints," *Visual perception and cognition in infancy,* pp. 175-214, 1993.

[57] N. F. Troje and C. Westhoff, “The inversion effect in biological motion perception: Evidence for a “life detector”?,” *Current Biology,* vol. 16, no. 8, pp. 821-824, 2006.

[58] V. T. Inman and H. D. Eberhart, “The major determinants in normal and pathological gait,” *The Journal of Bone & Joint Surgery,* vol. 3, no. 543-558, p. 35, 1953.

[59] M. P. Murray, A. B. Drought and R. C. Kory, “Walking patterns of normal men,” *The Journal of Bone & Joint Surgery,* vol. 46, no. 2, pp. 335-360, 1964.

[60] M. P. Murray, “Gait as a total pattern of movement: Including a bibliography on gait,” *American Journal of Physical Medicine & Rehabilitation,* vol. 46, no. 1, pp. 290-333, 1967.

[61] J. Napier, “The antiquity of human walking,” *Scientific American,* vol. 216, no. 4, p. 56, 1967.

[62] M. P. Murray, R. C. Kory and S. B. Sepic, “Walking patterns of normal women,” *Archives of physical medicine and Rehabilitation,* vol. 51, no. 11, p. 637, 1970.

[63] G. Johansson, “Visual perception of biological motion and a model for its analysis,” *Perception & psychophysics,* vol. 14, no. 2, pp. 201-211, 1973.

[64] C. D. Barclay, J. E. Cutting and L. T. Kozlowski, “Temporal and spatial factors in gait perception that influence gender recognition,” *Perception & Psychophysics,* vol. 23, no. 2, pp. 145-152, 1978.

[65] W. H. Dittrich, “Action categories and the perception of biological motion,” *Perception,* vol. 22, no. 1, pp. 15-22, 1993.

[66] B. I. Bertenthal and J. Pinto, “Global processing of biological motions,”

Psychological science, vol. 5, no. 4, pp. 221-225, 1994.

[67] R. McDonnell, S. Jörg, J. K. Hodgins, F. Newell and C. O'sullivan, "Evaluating the effect of motion and body shape on the perceived sex of virtual characters," *ACM Transactions on Applied Perception,* vol. 5, no. 4, p. 20, 2009.

[68] J. E. Cutting and L. T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bulletin of the psychonomic society,* vol. 9, no. 5, pp. 353-356, 1977.

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

[70] D. Jokisch, I. Daum and N. F. Troje, "Self recognition versus recognition of others by biological motion: Viewpoint-dependent effects," *Perception,* vol. 35, pp. 911-920, 2006.

[71] L. T. Kozlowski and J. E. Cutting, "Recognizing the sex of a walker from a dynamic point-light display," *Perception & Psychophysics,* vol. 21, no. 6, pp. 575-580, 1977.

[72] D. R. Saunders, D. K. Williamson and N. F. Troje, "Gaze patterns during perception of direction and gender from biological motion," *Journal of Vision,* vol. 10, no. 11, p. 9, 2010.

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

[74] G. Mather and L. Murdoch, "Gender discrimination in biological motion displays based on dynamic cues," *Proceedings of the Royal Society of London. Series B: Biological Sciences,* vol. 258, no. 1353, pp. 273-279, 1994.

[75] W. H. Dittrich, T. Troscianko, S. E. Lea and D. Morgan, "Perception of emotion from dynamic point-light displays represented in dance," *Perception,* vol. 25, no. 6, pp. 727-738, 1996.

[76] T. J. Clarke, M. F. Bradshaw, D. T. Field, S. E. Hampson and D. Rose, "The perception of emotion from body movement in point-light displays of interpersonal dialogue," *Perception,* vol. 34, no. 10, pp. 1171-1180, 2005.

[77] F. E. Pollick, H. M. Paterson, A. Bruderlin and A. J. Sanford, "Perceiving affect from arm movement," *Cognition,* vol. 82, no. 2, pp. B51-B61, 2001.

[78] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell and A. W. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," *Perception,* vol. 33, pp. 717-746, 2004.

[79] C. L. Roether, L. Omlor, A. Christensen and M. A. Giese, "Critical features for the perception of emotion from gait," *Journal of Vision,* vol. 9, no. 6, p. 15, 2009.

[80] M. Thrasher, M. D. van der Zwaag, N. Bianchi-Berthouze and J. H. Westerink, "Mood recognition based on upper body posture and movement features," in *LNCS 6974 (Proceedings of Affective Computing and Intelligent Interaction)*, Berlin Heidelberg, Springer, 2011, pp. 377-386.

[81] H. G. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology,* vol. 28, no. 6, pp. 879-896, 1998.

[82] H. M. Patterson, F. E. Pollick and A. J. Sanford, "The role of velocity in affect discrimination," *Proceedings of the Twenty-Third Annual Conference of the Cognitve Science Society,* pp. 756-761, 2001.

[83] J. M. Montepare, S. B. Goldstein and A. Clausen, "The identification of emotions from gait information," *Journal of Nonverbal Behavior,* vol. 11, no. 1, pp. 33-42, 1987.

[84] A. Barliya, L. Omlor, M. A. Giese, A. Berthoz and T. Flash, "Expression of emotion in the kinematics of locomotion," *Experimental Brain Research,* vol. 225, no. 2, pp. 159-176, 2013.

[85] E. Crane and M. Gross, "Motion capture and emotion: Affect detection in whole body movement," *Affective Computing and Intelligent Interaction,* pp. 95-101, 2007.

[86] F. E. Pollick, V. Lestou, J. Ryu and S. B. Cho, "Estimating the efficiency of recognizing gender and affect from biological motion," *Vision Research,* vol. 42, no. 20, pp. 2345-2355, 2002.

[87] L. Wang, W. Hu and T. Tan, "Recent developments in human motion analysis," *Pattern Recognition,* vol. 36, no. 3, pp. 585-601, 2003.

[88] O. Arikan and L. Ikemoto, Computational Studies of Human Motion: Tracking and Motion Synthesis, Now Publishers Inc, 2006.

[89] P. Slinger, S. A. Etemad and A. Arya, "Intelligent toolkit for procedural animation of human behaviours," *Proceedings of the ACM Future Play,* pp. 27-28, 2009.

[90] K. Amaya, A. Bruderlin and T. Calvert, "Emotion from motion," *Graphics*

interface, pp. 222-229, 1996.

[91] A. W. C. Fu, E. Keogh, L. Y. Lau, C. A. Ratanamahatana and R. C. W. Wong, "Scaling and time warping in time series querying," *The VLDB Journal—The International Journal on Very Large Data Bases,* vol. 17, no. 4, pp. 899-921, 2008.

[92] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 26, no. 1, pp. 43-49, 1978.

[93] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," *1st SIAM International Conference on Data Mining,* 2001.

[94] L. Rabiner and B. H. Juang, Fundamentals of speech recognition, Prentice Hall, 1993.

[95] C. Myers, L. Rabiner and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 28, no. 6, pp. 623-635, 1980.

[96] D. Lemire, "Faster retrieval with a two-pass dynamic-time-warping lower bound," *Pattern Recognition,* vol. 42, no. 9, pp. 2169-2180, 2009.

[97] F. Zhou and F. Torre, "Generalized time warping for multi-modal alignment of human motion," *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1282-1289, 2012.

[98] A. Bruderlin and L. Williams, "Motion signal processing," *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques,* pp. 97-104, 1995.

[99] A. Witkin and Z. Popovic, "Motion warping," *Proceedings of the 22nd ACM annual conference on Computer graphics and interactive techniques,* pp. 105-108, 1995.

[100] M. Gleicher, "Retargetting motion to new characters," *Proceedings of the 25th Annual ACM Conference on Computer Graphics and Interactive Techniques,* pp. 33-42, 1998.

[101] L. Kovar and M. Gleicher, "Flexible automatic motion blending with registration curves," *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation ,* pp. 214-224, 2003.

[102] L. Kovar and M. Gleicher, "Automated extraction and parameterization of motions in large data sets," *ACM Transactions on Graphics,* vol. 23, no. 3, pp. 559-568, 2004.

[103] M. Müller, T. Röder and M. Clausen, "Efficient content-based retrieval of motion capture data," *ACM Transactions on Graphics,* vol. 24, no. 3, pp. 677-685, 2005.

[104] M. Müller and T. Röder, "Motion templates for automatic classification and retrieval of motion capture data," *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 137-146, 2006.

[105] F. Zhou, F. Torre and J. K. Hodgins, "Aligned cluster analysis for temporal segmentation of human motion," *Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition,* pp. 1-7, 2008.

[106] M. Kim, K. Hyun, J. Kim and J. Lee, "Synchronized multi-character motion editing," *ACM Transactions on Graphics,* vol. 28, no. 3, p. 79, 2009.

[107] M. Raptis, D. Kirovski and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 147-156, 2011.

[108] G. Cimen, H. Ilhan, T. Capin and H. Gurcay, "Classification of human motion based on affective state descriptors," *Computer Animation and Virtual Worlds,* vol. 24, no. 3-4, pp. 355-363, 2013.

[109] A. Heloir, N. Courty, S. Gibet and F. Multon, "Temporal alignment of communicative gesture sequences," *Computer Animation and Virtual Worlds,* vol. 17, no. 3-4, pp. 347-357, 2006.

[110] E. Hsu, M. da Silva and J. Popović, "Guided time warping for motion editing," *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 45-52, 2007.

[111] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," *Advances in Neural Information Processing Systems,* pp. 2286-2294, 2009.

[112] Y. Caspi and M. Irani, "Aligning non-overlapping sequences," *International Journal of Computer Vision,* vol. 48, no. 1, pp. 39-51, 2002.

[113] Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 11, pp. 1409-1424, 2002.

[114] F. L. Pádua, R. L. Carceroni, G. A. Santos and K. N. Kutulakos, "Linear sequence-to-sequence alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 2, pp. 304-320, 2010.

[115] I. N. Junejo, E. Dexter, I. Laptev and P. Pérez, "View-independent action recognition from temporal self-similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 1, pp. 172-185, 2011.

[116] C. Lu and M. Mandal, "A robust technique for motion-based video sequences temporal alignment," *IEEE Transactions on Multimedia,* vol. 15, no. 1, pp. 70-82, 2013.

[117] W. Lao, J. Han and P. H. N. de With, "Automatic video-based human motion analyzer for consumer surveillance system," *IEEE Transactions on Consumer Electronics,* vol. 55, no. 2, pp. 591-598, 2009.

[118] G. S. Pingali, Y. Jean and I. Carlbom, "Real time tracking for enhanced tennis broadcasts," *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition,* pp. 260-265, 1998.

[119] A. D. Wilson, "Robust computer vision-based detection of pinching for one and two-handed gesture input," *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology,* pp. 255-258, 2006.

[120] F. Farhadi-Niaki, J. Gerroir, A. Arya, S. A. Etemad, R. Laganière, P. Payeur and R. Biddle, "Usability study of static/dynamic gestures and haptic input as interfaces to 3D games," *Proceedings of the 6th International Conference on Advances in Computer-Human Interactions,* pp. 315-323, 2013.

[121] J. Chan, H. Leung and H. Poizner, "Correlation among joint motions allows classification of Parkinsonian versus normal 3-D reaching," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 18, no. 2, pp. 142-149, 2010.

[122] M. Livne, L. Sigal, N. F. Troje and D. J. Fleet, "Human attributes from 3D pose tracking," *Computer Vision and Image Understanding,* vol. 116, no. 5, pp. 648-660, 2012.

[123] G. W. Taylor, G. E. Hinton and S. T. Roweis, "Modeling human motion using binary latent variables," *Advances in neural information processing systems,* pp. 1345-1352, 2006.

[124] S. A. Etemad, P. Payeur and A. Arya, "Automatic temporal location and classification of human actions based on optical features," *Proceedings of the 2nd International IEEE Congress on Image and Signal Processing,* pp. 1-5, 2009.

[125] J. Barbič, A. Safonova, J. Y. Pan, C. Faloutsos, J. K. Hodgins and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," *Proceedings of Graphics Interface,* pp. 185-194, 2004.

[126] X. Zhao, Y. Fu and Y. Liu, "Human motion tracking by temporal-spatial local Gaussian process experts," *IEEE Transactions on Image Processing,* vol. 20, no. 4, pp. 1141-1151, 2011.

[127] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," *Proceedings of the European Conference on Computer Vision,* pp. 494-507, 2010.

[128] Q. Cai and J. K. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 21, no. 11, pp. 1241-1247, 1999.

[129] C. Faloutsos, M. Ranganathan and Y. Manolopoulos, "Fast subsequence matching

in time-series databases," *ACM,* vol. 23, no. 2, pp. 419-429, 1994.

[130] E. Keogh, T. Palpanas, V. B. Zordan, D. Gunopulos and M. Cardle, "Indexing large human-motion databases," *Proceedings of the 30th International Conference on Very Large Data Bases,* vol. 30, pp. 780-791, 2004.

[131] F. Liu, Y. Zhuang, F. Wu and Y. Pan, "3D motion retrieval with motion index tree," *Computer Vision and Image Understanding,* vol. 92, no. 2, pp. 265-284, 2003.

[132] Z. Deng, Q. Gu and Q. Li, "Perceptually consistent example-based human motion retrieval," *Proceedings of the 2009 Symposium on Interactive 3D Graphics and Games,* pp. 191-198, 2009.

[133] B. Krüger, J. Tautges, A. Weber and A. Zinke, "Fast local and global similarity searches in large motion capture databases," *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 1-10, 2010.

[134] M. Kapadia, I. K. Chiang, T. Thomas, N. I. Badler and J. T. Kider Jr., "Efficient motion retrieval in large motion databases," *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games,* pp. 19-28, 2013.

[135] M. Müller, A. Baak and H. P. Seidel, "Efficient and robust annotation of motion capture data," *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 17-26, 2009.

[136] J. Liu, S. Ali and M. Shah, "Recognizing human actions using multiple features," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[137] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[138] H. Yu, G. M. Sun, W. X. Song and X. Li, "Human motion recognition based on neural network," *Proceedings of the 2005 International IEEE Conference on Communications, Circuits and Systems,* vol. 2, pp. 979-982, 2005.

[139] S. A. Etemad and A. Arya, "Recognition and synthesis of 3D human motion with personalized variations," *Proceedings of the IEEE International Conference on Multimedia Computing and Systems,* pp. 106-111, 2009.

[140] H. Ren and G. Xu, "Human action recognition with primitive-based coupled-HMM," *Proceedings of the 16th International Conference on Pattern Recognition,* vol. 2, pp. 494-498, 2002.

[141] C. S. Chan and H. Liu, "Fuzzy qualitative human motion analysis," *IEEE Transactions on Fuzzy Systems,* vol. 17, no. 4, pp. 851-862, 2009.

[142] M. Karg, K. Kuhnlenz and M. Buss, "Recognition of affect based on gait patterns," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 40, no. 4, pp. 1050-1061, 2010.

[143] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 2, pp. 288-303, 2010.

[144] N. F. Troje, "Retrieving information from human movement patterns," in *Understanding events: How humans see, represent, and act on events*, 2008, pp.

308-334.

[145] X. Sun, M. Chen and A. Hauptmann, "Action recognition via local descriptors and holistic features," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 58-65, 2009.

[146] R. Laban, The mastery of movement, Plays, Inc., 1971.

[147] V. Maletic, Body, space, expression: The development of Rudolf Laban's movement and dance concepts, New York: Mouton de Gruyte, 1987.

[148] D. Chi, M. Costa, L. Zhao and N. Badler, "The EMOTE model for effort and shape," *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques,* pp. 173-182, 2000.

[149] T. Yu, X. Shen, Q. Li and W. Geng, "Motion retrieval based on movement notation language," *Computer Animation and Virtual Worlds,* vol. 16, no. 3-4, pp. 273-282, 2005.

[150] S. Okajima, Y. Wakayama and Y. Okada, "Human motion retrieval system based on LMA features using interactive evolutionary computation method," in *Innovations in Intelligent Machines–2*, Berlin Heidelberg, Springer, 2012, pp. 117-130.

[151] D. Bouchard and N. Badler, "Semantic segmentation of motion capture using laban movement analysis," in *Intelligent Virtual Agents, LNCS 4722*, Berlin Heidelberg, Springer, 2007, pp. 37-44.

[152] L. Zhao and N. I. Badler, "Acquiring and validating motion qualities from live limb gestures," *Graphical Models,* vol. 67, no. 1, pp. 1-16, 2005.

[153] L. Torresani, P. Hackney and C. Bregler, "Learning motion style synthesis from perceptual observations," *Advances in Neural Information Processing Systems,* pp. 1393-1400, 2006.

[154] J. Lee, J. Chai, P. S. Reitsma, J. K. Hodgins and N. S. Pollard, "Interactive control of avatars animated with human motion data," *ACM Transactions on Graphics,* vol. 21, no. 3, pp. 491-500, 2002.

[155] Z. Popovic, "Editing dynamic properties of captured human motion," *Proceedings of the IEEE International Conference on Robotics and Automation,* pp. 670-675, 2000.

[156] S. A. Etemad and G. A. Wainer, "DEVS-based modeling of a human motion data synthesis system," *Proceedings of the ACM/SCS Summer Computer Simulation Conference,* pp. 469-474, 2010.

[157] A. C. Fang and N. S. Pollard, "Efficient synthesis of physically valid human motion," *ACM Transactions on Graphics,* vol. 22, no. 3, pp. 417-426, 2003.

[158] O. Arikan, D. A. Forsyth and J. F. O'Brien, "Motion synthesis from annotations," *ACM Transactions on Graphics,* vol. 22, no. 3, pp. 402-408, 2003.

[159] R. Tomovic and R. B. McGhee, "A finite state approach to the synthesis of bioengineering control systems," *IEEE Transactions on Human Factors in Electronics,* vol. 2, pp. 65-69, 1966.

[160] M. Vukobratovic and D. Juricic, "Contribution to the synthesis of biped gait," *IEEE Transactions on Biomedical Engineering,* vol. 1, pp. 1-6, 1969.

[161] M. A. Townsend and A. Seireg, "The synthesis of bipedal locomotion," *Journal of*

biomechanics, vol. 5, no. 1, pp. 71-83, 1972.

[162] M. A. Townsend and A. A. Seireg, "Effect of model complexity and gait criteria on the synthesis of bipedal locomotion," *IEEE Transactions on Biomedical Engineering,* vol. 6, pp. 433-444, 1973.

[163] J. E. Cutting, "A program to generate synthetic walkers as dynamic point-light displays," *Behavior Research Methods,* vol. 10, no. 1, pp. 91-94, 1978.

[164] J. E. Cutting, "Generation of synthetic male and female walkers through manipulation of a biomechanical invariant," *Perception,* vol. 7, no. 4, pp. 393-405, 1978.

[165] Y. Y. Tsai, W. C. Lin, K. B. Cheng, J. Lee and T. Y. Lee, "Real-time physics-based 3d biped character animation using an inverted pendulum model," *IEEE Transactions on Visualization and Computer Graphics,* vol. 16, no. 2, pp. 325-337, 2010.

[166] A. Bruderlin and T. W. Calvert, "Goal-directed, dynamic animation of human walking," *Proceedings of the ACM 16th Annual Conference on Computer Graphics and Interactive Techniques,* pp. 233-242, 1989.

[167] A. Bruderlin and T. Calvert, "Knowledge-driven, interactive animation of human running," *Proceedings of Graphics Interface,* pp. 213-221, 1996.

[168] K. Perlin, "Real time responsive animation with personality," *IEEE Transactions on Visualization and Computer Graphics,* vol. 1, no. 1, pp. 5-15, 1995.

[169] K. Perlin and A. Goldberg, "Improv: A system for scripting interactive actors in virtual worlds," *Proceedings of the 23rd ACM Annual Conference on Computer*

Graphics and Interactive Techniques, pp. 205-216, 1996.

[170] K. Pullen and C. Bregler, "Motion capture assisted animation: Texturing and synthesis," *ACM Transactions on Graphics,* vol. 21, no. 3, pp. 501-508, 2002.

[171] R. Urtasun, P. Glardon, R. Boulic, D. Thalmann and P. Fua, "Style-based motion synthesis," *Computer Graphics Forum,* vol. 23, no. 4, pp. 799-812, 2004.

[172] A. Shapiro, Y. Cao and P. Faloutsos, "Style components," *Proceedings of Graphics Interface,* pp. 33-39, 2006.

[173] A. Ahmed, F. Mokhtarian and A. Hilton, "Parametric motion blending through wavelet analysis," *Eurographics,* pp. 347-353, 2001.

[174] R. Boulic, N. Magnenat-Thalmann and D. Thalmann, "A global human walking model with real-time kinematic personification," *The visual computer,* vol. 6, no. 6, pp. 344-358, 1990.

[175] K. Grochow, S. L. Martin, A. Hertzmann and Z. Popović, "Style-based inverse kinematics," *ACM Transactions on Graphics,* vol. 23, no. 3, pp. 522-531, 2004.

[176] S. Coros, P. Beaudoin and M. van de Panne, "Generalized biped walking control," *ACM Transactions on Graphics,* vol. 29, no. 4, p. 130, 2010.

[177] Z. Popović and A. Witkin, "Physically based motion transformation," *Proceedings of the 26th ACM Annual Conference on Computer Graphics and Interactive Techniques,* pp. 11-20, 1999.

[178] X. Wei, J. Min and J. Chai, "Physically valid statistical models for human motion generation," *ACM Transactions on Graphics,* vol. 30, no. 3, p. 19, 2011.

[179] A. Safonova, J. K. Hodgins and N. S. Pollard, "Synthesizing physically realistic

human motion in low-dimensional, behavior-specific spaces," *ACM Transactions on Graphics,* vol. 23, no. 3, pp. 514-521, 2004.

[180] L. Kovar, M. Gleicher and F. Pighin, "Motion graphs," *ACM Transactions on Graphics,* vol. 21, no. 3, pp. 473-482, 2002.

[181] C. K. Liu and Z. Popović, "Synthesis of complex dynamic character motion from simple animations," *ACM Transactions on Graphics,* vol. 21, no. 3, pp. 408-416, 2002.

[182] S. J. Lee and Z. Popović, "Learning behavior styles with inverse reinforcement learning," *ACM Transactions on Graphics,* vol. 29, no. 4, p. 122, 2010.

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

[184] O. Arikan and D. A. Forsyth, "Interactive motion generation from examples," *ACM Transactions on Graphics,* vol. 21, no. 3, pp. 483-490, 2002.

[185] J. Lasseter, "Principles of traditional animation applied to 3D computer animation," *ACM Siggraph Computer Graphics,* vol. 21, no. 4, pp. 35-44, 1987.

[186] S. Werda, W. Mahdi and A. B. Hamadou, "Lip localization and viseme classification for visual speech recognition," *International Journal of Computing & Information Sciences,* vol. 5, no. 1, pp. 62-75, 2007.

[187] G. Tomasi, F. van den Berg and C. Andersson, "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data," *Journal of Chemometrics,* vol. 18, no. 5, pp. 231-241, 2004.

[188] P. O. Gray, Psychology, 5th ed., New York: Worth, 2006.

[189] J. M. Wolfe, K. R. Kluender, D. M. Levi, L. M. Bartoshuk, R. S. Herz, R. L. Klatzky and S. J. Lederman, Sensation and Perception, 2nd ed., Sinauer Associates, 2008.

[190] N. P. V. Nielsen, J. M. Carstensen and J. Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping," *Journal of Chromatography A,* vol. 805, no. 1, pp. 17-35, 1998.

[191] V. Pravdova, B. Walczak and D. L. Massart, "A comparison of two algorithms for warping of analytical signals," *Analytica Chimica Acta,* vol. 456, no. 1, pp. 77-92, 2002.

[192] G. Tomasi, "Practical and computational aspects in chemometric data analysis," *Ph.D. Thesis, The Royal Veterinary and Agricultural University,* 2006.

[193] T. Skov, F. van den Berg, G. Tomasi and R. Bro, "Automated alignment of chromatographic data," *Journal of Chemometrics,* vol. 20, no. 11-12, pp. 484-497, 2006.

[194] S. Wang, J. Yao, J. Liu, N. Petrick, R. L. Van Uitert, S. Periaswamy and R. M. Summers, "Registration of prone and supine CT colonography scans using correlation optimized warping and canonical correlation analysis," *Medical Physics,* vol. 36, pp. 5595-5603, 2009.

[195] S. A. Etemad and A. Arya, "A customizable time warping method for motion alignment," *Proceedings of the 7th IEEE International Conference on Semantic*

Computing, pp. 387-388, 2013.

[196] S. A. Etemad and A. Arya, "Correlation optimized time warping for motion," *submitted.*

[197] J. Wang and B. Bodenheimer, "Synthesis and evaluation of linear motion transitions," *ACM Transactions on Graphics,* vol. 27, no. 1, p. 1, 2008.

[198] J. M. Wang, D. J. Fleet and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, no. 2, pp. 283-298, 2008.

[199] D. Ormoneit, H. Sidenbladh, M. J. Black and T. Hastie, "Learning and tracking cyclic human motion," *Advances in Neural Information Processing Systems,* pp. 894-900, 2001.

[200] M. Pražák, L. Hoyet and C. O'Sullivan, "Perceptual evaluation of footskate cleanup," *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 287-294, 2011.

[201] C. Bouillaguet, H. C. Chen, C. M. Cheng, T. Chou, R. Niederhagen, A. Shamir and B. Y. Yang, "Fast exhaustive search for polynomial systems in F2," in *Cryptographic Hardware and Embedded Systems*, Berlin Heidelberg, Springer, 2010, pp. 203-218.

[202] G. P. Bingham, "Scaling judgments of lifted weight: Lifter size and the role of the standard," *Ecological Psychology,* vol. 5, no. 1, pp. 31-64, 1993.

[203] S. A. Etemad, A. Arya and A. Parush, "Spatial perceptual weights of energy-related features in animation of human motion," *Proceedings of Computer Graphics*

International, p. S15, 2011.

[204] T. Pozzo, C. Papaxanthis, J. L. Petit, N. Schweighofer and N. Stucchi, "Kinematic features of movement tunes perception and action coupling," *Behavioural Brain Research,* vol. 169, no. 1, pp. 75-82, 2006.

[205] J. T. Todd, "Perception of gait," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 9, no. 1, pp. 31-42, 1983.

[206] A. Bruderlin, C. G. Teo and T. Calvert, "Procedural movement for articulated figure animation," *Computers & Graphics,* vol. 18, no. 4, pp. 453-461, 1994.

[207] S. A. Etemad, A. Arya and A. Parush, "Additivity in perception of affect from limb motion," *Neuroscience Letters,* vol. 558, pp. 132-136, 2014.

[208] S. A. Etemad, A. Arya and A. Parush, "Spatial perceptual weights of secondary variants in human motion for multimedia applications," *Technical Report, Carleton University, SCE-13-01*, 2013.

[209] Z. Zhang and N. F. Troje, "View-independent person identification from human gait," *Neurocomputing,* vol. 69, pp. 250-256, 2005.

[210] D. Jokisch and N. F. Troje, "Biological motion as a cue for the perception of size," *Journal of Vision,* vol. 3, no. 4, pp. 252-264, 2003.

[211] D. Bernhardt and P. Robinson, "Detecting affect from nonstylized body motions," *Proceedings of the International Conference on Affective Computation and Intelligent Interaction, LNCS 4738,* pp. 59-70, 2007.

[212] M. M. Gross, E. A. Crane and B. L. Fredrickson, "Effort-shape and kinematic assessment of bodily expression of emotion during gait," *Human movement*

science, vol. 31, no. 1, pp. 202-221, 2012.

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

[214] S. A. Etemad and A. Arya, "Separation and extraction of energy variants from human motion using temporal minimization," *Proceedings of the IEEE Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems,* pp. 73-77, 2011.

[215] S. A. Etemad and A. Arya, "Extraction of secondary features from gait trajectories using spatiotemporal spline optimization," *Technical Report, Carleton University, SCE-11-05,* 2011.

[216] A. H. Guest, Labanotation: the system of analyzing and recording movement, Psychology Press, 2005.

[217] M. Unuma, K. Anjyo and R. Takeuchi, "Fourier principles for emotion-based human figure animation," *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques,* pp. 91-96, 1995.

[218] S. Taheri, V. Patel and R. Chellappa, "Component-based recognition of faces and facial expressions," *IEEE Transactions on Affective Computing,* 2013, 10.1109/T-AFFC.2013.28.

[219] R. H. Bartels, J. C. Beatty and B. A. Barsky, An introduction to splines for use in computer graphics and geometric modelling, San Francisco: Morgan Kaufmann, 1998.

[220] C. De Boor, A practical guide to splines, New York: Springer-Verlag, 1978.

[221] E. Lyard and N. Magnenat-Thalmann, "A simple footskate removal method for virtual reality applications," *The Visual Computer,* vol. 23, no. 9-11, pp. 689-695, 2007.

[222] L. Kovar, J. Schreiner and M. Gleicher, "Footskate cleanup for motion capture editing," *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pp. 97-104, 2002.

[223] S. A. Etemad and A. Arya, "Expert-driven features for style and affect in motion: Synthesis, analysis, and perception," *submitted.*

[224] S. A. Etemad and A. Arya, "Mining expert-driven models for affective motion," *ACM CHI Conference on Human Factors in Computing Systems (Workshop on Gesture-based Interaction Design: Communication and Cognition),* 2014.

[225] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of Nonverbal Behavior,* vol. 28, no. 2, pp. 117-139, 2004.

[226] C. Shan, S. Gong and P. W. McOwan, "Beyond facial expressions: Learning human emotion from body gestures," *Proceedings of the British Machine Vision Conference,* pp. 1-10, 2007.

[227] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications,* vol. 30, no. 4, pp. 1334-1345, 2007.

[228] A. Kleinsmith, N. Bianchi-Berthouze and A. Steed, "Automatic recognition of non-

acted affective postures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 41, no. 4, pp. 1027-1038, 2011.

[229] S. A. Etemad and A. Arya, "3D human action recognition and style transformation using resilient back-propagation neural networks," *Proceedings of IEEE International Conference on Intelligent Computing and Intelligent Systems,* pp. 296-301, 2009.

[230] "Segmentation and classification of human actions and actor characteristics with 3D motion data," *International Journal of Artificial Intelligence & Applications,* vol. 3, no. 4, pp. 65-81, 2012.

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

[232] S. A. Etemad and A. Arya, "Motion style translation with radial basis function networks," *International Conference on Multimedia and Human Computer Interaction,* p. 36, 2013.

[233] E. Hsu, K. Pulli and J. Popović, "Style Translation for Human Motion," *ACM Transactions on Graphics,* vol. 24, no. 3, pp. 1082-1089, 2005.

[234] I. T. Jolliffe, Principal Component Analysis, John Wiley & Sons, Ltd., 2005.

[235] M. J. Er, S. Wu, J. Lu and H. L. Toh, "Face recognition with radial basis function (RBF) neural networks," *IEEE Transactions on Neural Networks,* vol. 13, no. 3, pp. 697-710, 2002.

[236] H. Lee, S. Hong and E. Kim, "Neural network ensemble with probabilistic fusion

and its application to gait recognition," *Neurocomputing,* vol. 72, no. 7-9, pp. 1557-1564, 2009.

[237] L. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 2, pp. 281-286, 2002.

[238] E. J. Hartman, J. D. Keeler and J. M. Kowalski, "Layered neural networks with Gaussian hidden units as universal approximations," *Neural Computation Summer,* vol. 2, no. 2, pp. 210-215, 1990.

[239] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician,* vol. 46, no. 3, pp. 175-185, 1992.

[240] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.

[241] S. A. Etemad, A. Arya, A. Parush and S. DiPaola, "Perceptual validity in animation of human motion," *submitted.*

[242] S. A. Etemad and A. Arya, "Perceptually valid motion for avatars," *Proceedings of the International Conference on Multimedia and Human Computer Interaction,* p. 72, 2013.

[243] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[244] S. Kshirsagar and N. Magnenat-Thalmann, "A multilayer personality model," *Proceedings of the 2nd international symposium on Smart graphics,* pp. 107-115, 2002.

[245] D. Rousseau and B. Hayes-Roth, "Interacting with personality-rich characters,"

Report No. KSL 97, p. 6, 1997.

[246] A. Egges, S. Kshirsagar and N. Magnenat-Thalmann, "A model for personality and emotion simulation," in *Knowledge-based intelligent information and engineering systems*, Springer Berlin Heidelberg, 2003, pp. 453-461.

[247] C. Pelachaud and M. Bilvi, "Computational model of believable conversational agents," in *Communication in Multiagent Systems*, Springer Berlin Heidelberg, 2003, pp. 300-317.

[248] J. S. Wiggins, P. Trapnell and N. Phillips, "Psychometric and geometric characteristics of the Revised Interpersonal Adjective Scales (IAS-R)," *Multivariate Behavioral Research,* vol. 23, no. 4, pp. 517-530, 1988.

[249] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states," *Genetic, Social, and General Psychology Monographs,* vol. 121, pp. 339-361, 1995.

[250] L. Z. Tiedens and S. Linton, "Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing," *Journal of Personality and Social Psychology,* vol. 81, no. 6, pp. 973-988, 2001.

[251] J. A. Ruth, F. F. Brunel and C. C. Otnes, "Linking thoughts to feelings: investigating cognitive appraisals and consumption emotions in a mixed-emotions context," *Journal of the Academy of Marketing Science,* vol. 30, no. 1, pp. 44-58, 2002.

[252] C. Wallraven, M. Breidt, D. W. Cunningham and H. H. Bülthoff, "Evaluating the perceptual realism of animated facial expressions," *ACM Transactions on Applied*

Perception, vol. 4, no. 4, p. 4, 2008.

[253] C. Pelachaud, "Modelling multimodal expression of emotion in a virtual agent," *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 364, no. 1535, pp. 3539-3548, 2009.

[254] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," in *Nonverbal Communication, Interaction, and Gesture*, pp. 57-106.

[255] T. Dalgleish, Handbook of cognition and emotion, Chichester, UK: Wiley, 1999.

[256] P. Ekman and W. V. Friesen, Pictures of facial affect, Consulting Psychologists Press, 1975.

[257] R. K. Yin, "Looking at upside-down faces," *Journal of Experimental Psychology,* vol. 81, no. 1, pp. 141-145, 1969.

[258] J. C. Bartlett and J. Searcy, "Inversion and configuration of faces," *Cognitive Psychology,* vol. 25, no. 3, pp. 281-316, 1993.

[259] S. J. McKelvie, "Emotional expression in upside-down faces: Evidence for configurational and componential processing," *British Journal of Social Psychology,* vol. 34, no. 3, pp. 325-334, 1995.

[260] R. L. Birwhisttell, Kinesics and context: Essays on body motion communication, University of Pennsylvania Press, 1970.

[261] J. F. Cohn, "Foundations of human computing: facial expression and emotion," *Proceedings of the 8th International Conference on Multimodal Interfaces,* pp. 233-238, 2006.

[262] M. Pantic, A. Pentland, A. Nijholt and T. Huang, "Human computing and machine understanding of human behavior: a survey," in *Artifical Intelligence for Human Computing*, Springer Berlin Heidelberg, 2007, pp. 47-71.

[263] A. Arya and S. DiPaola, "Multi-space behavioural model for face-based affective social agents," *EURASIP Journal on Image and Video Processing, Special Issue on Facial Image Processing,* 2007.

[264] P. Ekman, "Facial expression and emotion," *American Psychologist,* vol. 48, no. 4, p. 384, 1993.

[265] A. Buades, B. Coll and J. M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation,* vol. 4, no. 2, pp. 490-530, 2005.

[266] T. Cloete and C. Scheffer, "Benchmarking of a full-body inertial motion capture system for clinical gait analysis," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* pp. 4579-4582, 2008.

[267] A. Whitehead, N. Crampton, K. Fox and H. Johnston, "Sensor networks as video game input devices," *Proceedings of the ACM conference on Future Play,* pp. 38-45, 2007.

[268] S. Yabukami, H. Kikuchi, M. Yamaguchi, K. I. Arai, K. Takahashi, A. Itagaki and N. Wako, "Motion capture system of magnetic markers using three-axial magnetic field sensor," *IEEE Transactions on Magnetics,* vol. 36, no. 5, pp. 3646-3648, 2000.

[269] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall and H. P. Seidel, "Markerless motion capture with unsynchronized moving cameras," *IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 224-231, 2009.

[270] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger and A. Weber, "Documentation mocap database HDM05," *Technical Report, No. CG-2007-2,* 2007.

[271] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Proceedings of the IEEE Nonrigid and Articulated Motion Workshop,* pp. 90-102, 1997.

[272] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding,* vol. 81, no. 3, pp. 231-268, 2001.