1-1-2015

Experimental Analysis of Motion Style

Aline Normoyle

*University of Pennsylvania, alinen@seas.upenn.edu*

Follow this and additional works at: [http://repository.upenn.edu/edissertations](http://repository.upenn.edu/edissertations)

Part of the Computer Sciences Commons

**Recommended Citation**

Normoyle, Aline, "Experimental Analysis of Motion Style" (2015). Publicly Accessible Penn Dissertations. 1920. [http://repository.upenn.edu/edissertations/1920](http://repository.upenn.edu/edissertations/1920)

This paper is posted at ScholarlyCommons. [http://repository.upenn.edu/edissertations/1920](http://repository.upenn.edu/edissertations/1920)

For more information, please contact libraryrepository@pobox.upenn.edu.
Experimental Analysis of Motion Style

Abstract
For interactive 3D applications and games, the ability to edit, combine, and adapt motion capture is an important technique for reducing labor, project complexity, and storage overhead. Stylistic motion capture, in particular, has unique challenges since editing can alter the original intent of the actor -- sometimes with poor consequences as viewers can pick up subtle differences in body language. Thus, this thesis aims to enhance motion capture techniques, particularly when it is important that the stylistic intent of a motion is not unintentionally altered, when we might need to quantify the stylistic nature of the motion, and when better tools can help naive users works with large stylistic motion sets. In our first experiments, we investigate how changes introduced by motion editing might alter the emotional content of a motion. We find that emotions were mostly conveyed through the upper body, that the perceived intensity of an emotion can be reduced by blending with a neutral motion, and that posture changes can alter the perceived emotion but subtle changes in dynamics only alter the intensity. In our next experiments, we investigate whether fundamental numerical differences exist between neutral and stylistic motions of a single motion category and find that neutral walks had significantly lower torques than stylistic motions. Next, we model the variation over our stylistic walks using PCA by decoupling each walking example motion into an average and stylistic part such that only the stylistic part is transformed to PCA space. Although our use of PCA for motion modeling is not novel, our decision to decouple motions is a simple enhancement that makes the use of the PCA model produce consistently good output for blending, clustering, searching, and browsing. In our final experiments, we investigate whether using a minimap, created from our PCA model, to browse stylistic motions helps non-expert users search and edit styles. We find that 82\% of participants prefer the map to a label-based interface and that when not under time pressure, users performed equally well with the map, although using fewer mouse clicks and producing more varied results.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Computer and Information Science

First Advisor
Norman I. Badler

Keywords
Character animation, Motion capture

Subject Categories
Computer Sciences

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1920
EXPERIMENTAL ANALYSIS OF MOTION STYLE

Aline Normoyle

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania in Partial
Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

Graduate Group Chairperson

-----------------------------

Norman I. Badler, Professor,
Computer and Information Science

Lyle Ungar, Professor,
Computer and Information Science

Dissertation Committee

Stephen H. Lane,
Associate Professor of Practice,
Computer and Information Science

Jessica Hodgins, Professor,
Carnegie Mellon University
(External Committee Member)

Ladislav Kavan, Assistant Professor,
Computer and Information Science

Lyle Ungar, Professor,
Computer and Information Science
EXPERIMENTAL ANALYSIS OF MOTION STYLE

COPYRIGHT

2015

Aline Normoyle
For interactive 3D applications and games, the ability to edit, combine, and adapt motion capture is an important technique for reducing labor, project complexity, and storage overhead. Stylistic motion capture, in particular, has unique challenges since editing can alter the original intent of the actor – sometimes with poor consequences as viewers can pick up subtle differences in body language. Thus, this thesis aims to enhance motion capture techniques, particularly when it is important that the stylistic intent of a motion is not unintentionally altered, when we might need to quantify the stylistic nature of the motion, and when better tools can help naive users work with large stylistic motion sets. In our first experiments, we investigate how changes introduced by motion editing might alter the emotional content of a motion. We find that emotions were mostly conveyed through the upper body, that the perceived intensity of an emotion can be reduced by blending with a neutral motion, and that posture changes can alter the perceived emotion but subtle changes in dynamics only alter the intensity. In our next experiments, we investigate whether fundamental numerical differences exist between neutral and stylistic motions of a single motion category and find that neutral walks had significantly lower torques than stylistic motions. Next, we model the variation over our stylistic walks using PCA by decoupling each walking example motion into an average and stylistic part such that only the stylistic part is transformed to PCA space. Although our use of PCA for motion modeling is not novel, our decision to decouple motions is a simple enhancement that makes the use of the PCA model produce consistently good output for blending, clustering, searching, and browsing. In our final experiments, we investigate whether using a minimap, created from our PCA model, to browse stylistic
motions helps non-expert users search and edit styles. We find that 82% of participants prefer the map to a label-based interface and that when not under time pressure, users performed equally well with the map, although using fewer mouse clicks and producing more varied results.
Contents

Abstract iii

Contents v

List of Tables viii

List of Figures x

1 Introduction 1

2 Background 6

  2.1 What is motion style? 6
  2.2 Style editing 7
  2.3 Motion Decomposition 8
  2.4 Physically-based motion editing 8
  2.5 Statistically-based motion editing 10
  2.6 Remarks and summary 13

3 Effects of motion editing on emotional style 16

  3.1 Introduction 16
  3.2 Emotion recognition 18
    3.2.1 Impact of Motion Editing 19
  3.3 Experiment 1: Emotion Recognition and Movement Analysis 20
List of Tables

2.1 Examples shown to work well for different approaches to creating, editing, and modeling motion styles. .......................... 14

2.2 Strengths and weaknesses of different approaches to creating, editing, and modeling styles. .......................... 15

3.1 Characteristic frame for each emotion from the clips with the best recognition rates. ............................................. 17

3.2 Confusion matrix from our first experiment. Entries show the percentage of times participants chose each emotion in a forced choice experiment. The displayed emotions are listed on the left, the selection of the viewer at the top. ............................................. 23

4.1 Examples: Motions are trimmed to have the matching foot contacts. Top, a dancing jig walk. Bottom, the average of 10 neutral walking motions. 42
4.2 Speeds for different stylistic examples. Colored plots showing the joint speeds for every joint for each frame. The x-axis represents frames and the y axis represents each joint. Brighter areas represent faster speeds. The upper body is stored in the top of each image and the lower body is shown lower. For example, in the dinosaur and marching styles, the upper body does not move much. Old man exhibits very little movement whereas jaunty and elated has a lot of movement over the whole body. Depressed has much slower movements than elated or jaunty. The top, left-most image is a neutral walk computed as the average walk over 10 neutral examples.

4.3 Anthropometrics. Limb masses as percentages of total body mass. COM positions as percentages of limb length. When a segment is composed of multiple joints, such as the trunk and the hands, we divide the mass between the joints. Parameters for linear regression curves for estimating segment density from body density. Adapted from Winter (2009).
List of Figures

3.1 Emotion recognition from our first experiment. Each column summarizes the percentage of correct responses for each emotion. The red middle band indicates the median value, the box top and bottom shows the first and third percentiles (75% and 25%), and the top and bottom whiskers show the maximum and minimum recognition percentage. Crosshairs show outliers. ................................................................. 23

3.2 Histograms of rotational speeds (radians/second) for the major joints of our character: root, left and right hips, knees, ankles, spine, shoulders, elbows, wrists, and neck. The x axis shows bins corresponding to speeds of 1, 2, 4, 8, and 16 radians/second respectively. The y axis shows counts for each speed across all motions of each category, normalized to range from 0 to 1 based on the maximum bin size. Our results are consistent with previous published research which states that anger and happiness tend to have larger and faster joint movements, whereas fear and sadness tend to have the least joint movement. For our motions, surprise and disgust lie somewhere in between these two extremes. ...................... 25

3.3 Joint amplitudes. Our captured motions are consistent with published research which states that happiness and anger have higher amplitudes whereas sadness and fear have lower amplitudes. ...................... 26
3.4 Stimuli examples from our second experiment in which we hid either the head, the lower body, or the upper body. .......................... 27

3.5 Error rates for each condition (left) and for each emotion (right). Emotions were recognized significantly less often when the upper body was hidden (NU: no upper body) than when it was visible (OR: original, NH: no head, NL: no lower body). Error bars represent one standard error of the mean in all graphs. .................................................. 30

3.6 Intensities for each condition (left) and for each emotion (right). Emotions were rated to have a lower intensity on a scale from 1 to 5 when the upper body was occluded. Also, anger, fear, and happiness were rated to have a significantly higher intensity than disgust, sadness, and surprise. 31

3.7 Body blend (BB) condition. The upper body joints are blended with a neutral motion having the arms at the side. The above example shows the result of blending 50% of the original motion with 50% of the neutral motion. Joint rotations are represented using quaternions and blended with slerp. This condition changes both pose and velocity. ............... 33

3.8 Stimuli examples from body blend (BB) conditions. Original motions appear in the first column. The second column shows BB25, which retains 75% of the original motion. The third row shows BB50, which retains 50% of the original motion. As the poses moved towards neutral, the perceived intensity of the emotion is decreased. ......................... 34

3.9 Stimuli examples from the offset condition. Original poses appear in the first column. Modified poses appear in the second column. .............. 36

3.10 Error rates for each alteration (left) and for all emotions and alterations (right). Emotions were recognized significantly less often when blended to 50% with a neutral motion or when an offset was added. Error bars represent one standard error of the mean in all graphs. ................. 37
3.11 Intensity for each alteration (left) and for all emotions and alterations (right). Emotions were rated to have a lower intensity on a scale from 1 to 5 when blended to 50% with a neutral motion or when time warped.

4.1 Segment density versus body density. Given the body density $X$, we estimate the corresponding limb density $Y$, where $Y = MX + B$. These lines are based on the slope $M$ and intercept $B$ terms from Table 4.3. Adapted from Winter (2009).

4.2 Volumes and centers of mass. We use public anthropometrics data to estimate limb masses and moments of inertia. The red volumes display the relative sizes of masses for each joint. The blue dots display the center of mass.

4.3 Sum torques. Histograms of the sum of squared torques (log scale) for each frame. Each row corresponds to a motion. The x-axis contains counts for each frame having the same range of sum of torques. The color of the bin represents the quantity of frames, with the largest bins corresponding to white dots. Neutral motions (bottom) tend to have lower sum of squared torques.

4.4 Power. Histograms of the sum of squared power each frame. The largest bins correspond to white dots. Neutral motions tend to have lower power, but the lowest power motions are the ones with the shortest, slowest movements (old man).

4.5 Energy. Histograms of the changes in energy each frame. The largest bins correspond to white dots. Slowest motions tend to have the smallest scores with this metric.

4.6 Torque changes. Histograms of the torque force changes for each frame. The largest bins correspond to white dots. Faster motions tend to have higher values.
4.7 Ground reaction force changes. Histograms of the contact forces changes for each frame. The largest bins correspond to white dots. 57

5.1 Process for converting input motions to PCA coordinates. We first compute the average motion across all input samples and then use that average to decompose each input motion into an average part and a stylistic, or offset, part. We then only model the motion offsets in PCA space. 60

5.2 Rotational artifacts which result if we do not preprocesses the motion set to only model offsets. 62

5.3 Joint distributions for problematic exponential map components. Left, exponential map component distributions from the original motions. Right, exponential map component distributions from the offset motions. Modeling the stylistic offsets in PCA space avoids rotational artifacts. The exponential map distributions (joint angle components) of the offset motions are centered and have smaller magnitudes than modeling the original joint angles. 64

5.4 Right, process for converting from PCA coordinates to a motion. We convert from PCA coordinates to an offset motion and then add back to the average motion. We model the timing in PCA space, but not the root position, so to complete the reconstruction, we use a treadmill approach which sets the root positions based on foot contact velocity. 65

5.5 A single map point represents a large family of motions. Above, random motions generated from the same map point. The user can generate 5 random examples at a time. Clicking on a thumbnail animation (from the right) will preview the motion on the center character. 67
5.6 Using search to repair motions. In this example, blending violates joint limits. Left, unnaturally bent joints in an extrapolated walking example. Right, a nearby motion in PCA space with valid joint limits, found using CMA-ES ................................................................. 68

5.7 Using search to repair motions. Blending can produce motions with self-intersections. Left, a poor motion with extreme self-intersections; bounding spheres used for collision checking are shown in the inset; Right, an improved motion without intersections found using CMA-ES. .... 69

5.8 The map of the first two PCA components reveals clusters in the data: crouching motions, upright walks, and wide-stance walks with bent arms. 70

5.9 A single map point represents a large family of motions. Above, different motions created using different browsing heuristics. The red dot indicates the current map position. The first two map components correspond to first 2 PCA coordinates. The zero heuristic sets all remaining coordinates to zero. The closest heuristic sets remaining components to the closest input example, which in this case is a limp. The weights heuristic sets the remaining components to a weighted sum of nearby input motions. .... 71

6.1 Screenshot showing the interactive minimap with two bookmarks, labeled 1 and 2. Blue dots represent the original motion examples. Green dots represent bookmarks. The red dot represents the current motion shown on the character. We implement this map using a principal component analysis (PCA) transformation of the original stylistic motion set, which automatically clusters similar motions together and transitions smoothly between styles. ................................................................. 75

6.2 Using bookmarks to blend between specific motions. From left to right, linear blend, trilinear blend, and a 5-way blend. ............................ 78
6.3 Blend UI used in our Untimed experiment. The Select UI from our Timed experiment looks very similar, with buttons replacing the ability to set blend weights.

6.4 Left, Durations. Participants spent significantly more time creating an angry motion than a happy or confident motion. Error bars represent standard errors. Right, Mouse clicks. The map requires significantly fewer mouse clicks than the blend interface.

6.5 Analysis of dissimilarity of created motions between UI type and Question. The responses for the map display much more variety. Several differences are significant: The silly motions created with the map are significantly more varied than every other category except for frightened motions created with the map. Not surprisingly, the normal motions show the least variety for both the map and the blend interface. The frightened motions created with the map are also significantly more varied than the happy, normal, and angry motions created with the blend interface (highlighted in the diagram).
Chapter 1

Introduction

For games and other interactive applications, the ability to generate variations on motion capture is an important technique for reducing the labor, complexity, and storage requirements associated with animating a very large cast of interactive characters. Relying solely on motion capture for this purpose is extremely difficult:

• The number of motions needed for a typical game is huge (typically 80+ hours of interactions with the player, for hundreds of characters);

• It is difficult to anticipate all data needs up front;

• The amount of storage required becomes enormous;

• It is expensive to maintain the facilities and staff needed to organize actors, capture, and then clean their performance data; and

• The turn-around for acquiring new motions is slow.

Editing and procedural animation mitigates these problems. The ability to compute new motions automatically from existing performances means that there is no longer a need to capture many small variations on the same asset. It also gives animators and game designers direct control of the characters for the purposes of prototyping and
tweaking character performances. Additionally, if it is possible to compute motions on-the-fly, the runtime memory storage requirements are reduced and the need to anticipate motion requirements up front becomes less.

For these reasons and because style cannot be easily simulated or generated procedurally, methods for working specifically with stylistic motion capture are valuable. In this context, we define “style” broadly as the manner in which something is done, so a style model might alter a captured performance so that a character gestures happily, drunkenly, nervously, or sleepily in response to a player and while handling dynamic changes in the environment. Such techniques enable virtual characters (particularly, tertiary characters who would otherwise be too expensive to be configured with nuanced performances) to display compelling body language in a wide-range of situations.

Many interesting projects have proposed techniques for generating new motion styles from existing motion capture. Three primary approaches are editing/decomposition approaches, physically-based approaches, and statistically-based approaches. Examples from the first category include the ability to add noise (Perlin, 1995; Bodenheimer et al., 1999), modify signal curves through blending and warping (Witkin & Popovic, 1995; Amaya et al., 1996; Rose et al., 1998; Chi et al., 2000; Wang et al., 2006; Neff & Kim, 2009), and decompose and edit signal curves (Unuma et al., 1995; Bruderlin & Williams, 1995; Vasilescu, 2002; Urtasun et al., 2004a; Shapiro et al., 2006). These approaches tend to be manual or semi-automatic and facilitate off-line motion creation with artistic supervision. Alternatively, physically-based approaches can automatically change the physical characteristics of an input motion, such as its gravity, character weight distribution, and foot contacts. Physically-based approaches are ideal for dynamically altering motions to react to impacts (Kokkevis et al., 1996; Zordan & Hodgins, 2002; Zordan et al., 2005; Metoyer et al., 2008), and adapting to new terrains and physiology (Popović & Witkin, 1999; Komura & Shinagawa, 2001; Liu et al.; Tak & Ko, 2005; da Silva et al., 2008). By definition however, a physically-based approach
cannot model how mental states, emotions, individual preferences, or aesthetics affect movement. Statistical methods fill this gap by learning these relationships from examples. Statistical approaches can synthesize individual variations (Chien & Liu, 2006; Lau et al., 2009; Ma et al., 2010; Min et al., 2010), dances (Brand & Hertzmann, 2000; Li et al., 2002), and caricatured movements (Hsu et al., 2005; Taylor & Hinton, 2009; Pullen & Bregler, 2002). However, unlike physically-based approaches, which tend to generalize well, signal-based, decomposition, and statistical approaches are limited. They often can only be used for off-line motion synthesis because synthesized motions usually require foot-skate cleanup, removal of self-intersections and other repairs due to the fact that physics are not modeled. Also, because these models depend on matching similar poses and velocities, they are limited in their ability to generalize. A style model learned from a walking motion cannot be applied to a run or idling motion because their poses and dynamics are too different. In summary, physically-based approaches cannot solve all the problems of style while non-physically-based approaches generally often cannot be done robustly at runtime, are limited by the example motions available, and can require a lot of storage. In Chapter 2, we discuss the strengths and weaknesses of different approaches to motion editing in more detail.

In the remainder of this thesis, we investigate problems related to working with stylistic motion capture for body language. Although motion capture has the potential to faithfully reproduce an actor’s intent, how we edit and represent this captured data can potentially alter the intended performance. Viewers are sensitive to errors in body language and can easily misinterpret the performance if the transfer is poor. In Chapter 3 we look at which aspects of body language are important for conveying emotions with two goals in mind. The first goal is to understand how changes introduced by motion editing might alter the emotional content of a motion, so we can ensure that important aspects of a performance are preserved. The second goal is to gain insight on how we may edit captured motion to change its emotional content, further increasing its
reusability. In the first study, we captured an actor performing vignettes displaying happiness, sadness, anger, fear, disgust, and surprise and transferred his performance to a virtual character. Then, we identified which of the animated performances had the best recognition among viewers. In the second study, we focused only the best-recognized animations to determine what parts of the body were most important for conveying the emotion. In the third study, we systematically altered the poses and dynamics for our best recognized animations. For our captured motions, we found that emotions are mostly conveyed through the upper body, that the perceived intensity of an emotion can be reduced by blending with a neutral motion, and that posture changes can alter the perceived emotion but subtle changes in dynamics only alter the intensity.

Both the findings in Chapter 3 and previous work consistently find correlations between emotions and dynamics. For example, happy and angry movements are associated with higher joint velocities (Wallbott, 1998; Coulson, 2004; Roether et al., 2009; Normoyle et al., 2013). We perform an analysis in Chapter 4 to determine if there are similar fundamental differences between neutral and stylistic motion. Unlike in Chapter 3, where we study emotional vignettes, we restrict our analysis to a single category of motion, walks, but include a wider variety of styles (due to emotions, moods, aesthetics, personality, and caricatures). Using this new dataset, we compare the energy and naturalness between neutral and stylistic examples through various metrics estimated with inverse dynamics. Overall, we find that our neutral walks have significantly lower torques than stylistic motions, but that our other metrics did not show consistent patterns. Assuming this results holds for other categories of motion, this result could be used to infer neutral motions from stylistic ones, which is potentially useful for editing and categorizing motion.

The analysis in Chapter 4 helps us understand what non-stylistic motions have in common, but not how stylistic examples differ. Furthermore, Chapter 4 focused on dynamics, rather than pose. In Chapter 5, we describe how we model the variations
between our stylistic walks using PCA and show how this model can be used for blending, sampling, searching, clustering and browsing stylistic motions in real-time. Although our use of PCA for motion modeling is not novel, our decision to decouple motions and only model the motion offset is a simple enhancement that makes the use of the PCA model produce consistently good output.

In Chapter 6, we investigate whether a minimap which clusters similar motions together helps users browse styles. Enhancing browsing capabilities has the potential to help users find motions that are difficult to describe with queries, e.g. when the user might not already have an idea of what is available or when labels in a motion set might not match a user’s expectations. We build our minimap using the PCA model from Chapter 5 which allows users to interactively preview styles by dragging the mouse in the map and to set bookmarks to annotate the map and single out motions for editing. To quantify the strengths and weaknesses of the minimap, we design two user studies which compare our map interface to a traditional label-based interface. In both studies, users perform two tasks: the first task asks users to find a motion with a given semantic quality, such as happy; the second task asks users to reproduce a walking animation. Our first study gave participants ample time to finish each task whereas our second study had a strict 45 second time limit. Most of our participants were non-expert users, rating their experience with character animation as lower than 3 on a 7-point scale. We found that for the Untimed experiment, users were equally satisfied with both the map and label interfaces, but were able to create and find motions using significantly fewer mouse clicks and exhibiting more creative results using the map. For the Timed experiment, both interfaces were equally poor at helping people find motions with a given semantic quality; however, the label-based interface was better for finding a matching motion example. Overall, most participants greatly preferred the map (14 out of 17 participants, 82%), enjoying the ability to quickly scrub through the entire motion set. The creation of better tools for working with large motion sets again helps us leverage captured data
by helping users understand and edit captured clips.
Chapter 2

Background

2.1 What is motion style?

Motion Style is a frequently used term in character animation, but what do researchers mean when they say “style”? In some works, style is used very broadly to refer to an input motion (da Silva et al., 2008) or input motion set (Grochow et al., 2004), but more commonly, style is used to imply variations on a specific theme, such as the individual variations between people performing the same motion (Vasilescu, 2002; Lau et al., 2009; Chen et al., 2010; Ma et al., 2010) or the different ways one can perform the same behavior (such as a sad walk versus a nervous walk). Amaya et al. (1996) used the term “primary motion” to talk about a behavior category such as walking and the term “secondary motion” to describe the manner of walking, such as sadly. Rose et al. (1998) used a grammar analogy for describing variation in movement: parameterized motions were called “verbs” and the parameters for controlling them were called “adverbs”. Brand & Hertzmann (2000) explicitly mention “stylistic variations on a basic motor theme” as their focus, considering walking and running as two styles of moving forward. Liu et al. defined style numerically in terms of the relative “stiffnesses” of each joint, and also considered walking and running different styles of the same behavior. The work of
Torresani et al. (2006) defined style as the combination of content (the main action) and style (“the particular way that action is performed”).

In our work, we define style as the manner in which something is done. In particular, the following overview will focus on algorithms for transforming a given motion so that it can convey new meanings, for example, how to transform a neutral walk into a sneaky walk, a sad walk, an injured walk, a silly walk, or Charlie Chaplin’s walk?

2.2 Style editing

Early research with motion capture techniques indirectly addressed style. For example, Perlin (1995) introduced the use of Perlin noise for adding variation to motion, which was capable of giving the impression of different individuals, or could be layered onto existing motions to create a rhythmic nervous or agitated effect. Similarly, Bodenheimer et al. (1999) proposed a noise function based on biomechanical principles for the purposes of adding natural-looking detail to cyclic animations. Also working on joint curves, Wang et al. (2006) proposed a simple signal editing technique which exaggerated changes in joint accelerations to produce cartoonish, exaggerated joint movements.

Early research also demonstrated the power of blending for creating arbitrary combinations of styles. For example, Witkin & Popovic (1995) proposed “Motion Warping” for editing individual poses to create new motions, including those that may convey a new style, such as a sad walk or a limping walk. Amaya et al. (1996) investigated an “emotion transform”, based on joint speeds and amplitudes, capable of warping from a neutral motion to an emotional version of the same behavior. Rose et al. (1998) proposed a simple, robust, real-time technique which blended motions to support arbitrary style combinations. For example, walks of different speeds, different genders, and different emotions could be combined to create male and female characters capable of walking
with any emotion and any speed. Chi et al. (2000) defined Laban parameterizations for motion which allowed a user to intuitively change the shapes and timing of gestures.

### 2.3 Motion Decomposition

Motion decomposition has also been used for the synthesis of new stylistic motions, particularly emotions, aesthetics, and mental states. A user decomposes the joint curves into components which can then be transferred and modified between motions. For example, Unuma et al. (1995) used Fourier decomposition to change a normal walk into a brisk walk. Bruderlin & Williams (1995) used Laplacian band filtering to decompose motion signals which could be altered and combined to create effects such as nervousness. Vasilescu (2002) used N-mode SVD to decompose motions into content and individual style. The user could then transfer individual styles between motions (demonstrated by transitioning a style from a walking input to a climbing stairs output). Urtasun et al. (2004a) used principal component analysis (PCA) to change the speeds of walkers. Similarly, Shapiro et al. (2006) investigated independent component analysis (ICA), a dimensionality reduction technique similar to PCA, for creating “Style Components” which a user can interactively transfer between motions.

Decomposition, motion warping, blending, and additive noise techniques tend to be manual or semi-automatic, to allow a user to preview and guide changes. For example, Bruderlin & Williams (1995) and Shapiro et al. (2006) provided editing tools which allowed a user to preview and tweak motions interactively. Additionally, these approaches operate in joint curve space. Thus, it is generally necessary to work with similar categories of motion (such as walking or dancing), to time warp motions to align similar poses, and to post-process the results to repair foot sliding and other artifacts due to violations of physical correctness.
2.4 Physically-based motion editing

Physically-based approaches modify the dynamics of an input motion based on a physics simulation. These approaches are ideal for realistically responding to impacts, changing gaits, estimating the effects of different body types (height, weight, and mass distribution) or the effect of carrying different weights. Physically-based editing typically consists of three steps. In the first, we estimate the dynamics which likely produced the input motion (since the input motion is kinematic, this includes estimating the mass properties of the character). In the second, the user changes the physical properties of the physical simulation, for example, the positions of foot contacts, the body type, or gravity. Running the simulation with these new constraints produces a new motion which is mapped back to the character.

Early work in this area focused on using physics to realistically respond to impacts. For example, Kokkevis et al. (1996) estimated the forces and torques on each DOF curve so that they could be modified by new external forces. Their system could realistically knock down a soldier. Zordan & Hodgins (2002) designed a proportional derivative (PD) controller capable of maintaining balance while tracking an input motion. When a hit occurred, the stiffness and damping terms were modified so that the character reacted realistically to the impact. Similarly, Zordan et al. (2005) used rag doll physics to realistically fall after an impact. A fast search algorithm was then used to find good kinematic motion clips to blend back into after the fall. Metoyer et al. (2008) extended physical reactions to impacts with an anticipatory (flinching) motion model.

Popović & Witkin (1999) was one of the first to use physics for general motion editing. Their approach used a simplified character representation to model and edit the dynamics of an input motion. Two years later, Komura & Shinagawa (2001) used a similar approach to estimate muscle activations from the input motion. The muscle model could then be used to create new motions that looked fatigued or injured. The model could also be used to estimate movements for different body types, such as
These early approaches worked best for highly dynamic motions, e.g. where most of the motion is the result of passive physics forces, such as the center of mass during a jump. Liu et al. proposed a modified formulation capable of editing low-dynamic motions like walks, where the human is initiating movement purposefully (instead of passively). Using their formulation, Liu et al. was able to edit motions to give the appearance of holding heavy objects, walking up and down inclines, and walking with springy shoes or a limp.

These approaches are suitable for synthesizing new motions offline and although they reduce the amount of data which needs to be captured, they still require developers to anticipate which motions will be needed beforehand. da Silva et al. (2008) proposed a real-time system capable of maintaining balance while following an input motion. They were able to generate walking motions in the styles of marching, sideways steps, and stomping which could then interact realistically in new environments. Jain et al. (2009) proposed a real-time technique based on solving small, quick optimization problems which allowed a character to maintain balance using features in the environment, such as the railings in a gondola.

Other research has looked at editing dynamics directly. For example, Sok et al. (2010) investigated the use of “normalized dynamics” which can propagate changes in momentum and force forward in time and Neff & Kim (2009) parameterized input motions in terms of the center of mass, ankle and wrist positions, and pelvic orientation. Tak & Ko (2005) fit a physical model to the motion, but used a constraint state estimation problem based on a Kalman filter framework instead of a constrained optimization formulation.

In physically-based systems, style most often refers to an input motion whose style is preserved when changing the dynamics (Popović & Witkin, 1999; Liu et al.; da Silva et al., 2008). For example, the use of the abstracted model by Popović & Witkin
(1999) ensured that the detailed movements in the original model were not lost. Liu et al. learned the parameters for their physical model from the input motion and then propagated them to the result. Thus, if the input was a sad walk with a “carry briefcase” constraint, the output would be a sad person carrying a briefcase. Similarly, da Silva et al. (2008) tracked the input motion with their physical model and so maintained the flavor of the input motion.

2.5 Statistically-based motion editing

Although physically-based methods can robustly edit motions to account for different body types, physical impacts and world/body constraints, by definition, they do not model how mental states, personality, emotions, or aesthetics affect movement. Statistical models strive to learn these relationships through examples. Because these techniques are probabilistic, they can also capture the natural variability in human movements. Once the model is learned, it can be used to create new stylistic motions, to recognize styles, or in some cases, to add and edit styles. Also depending on the nature of the examples and the technique, statistical approaches are capable of modeling both an individual style, e.g. their “movement signature” (Laban, 1971), or a category of movement, such as sneaky walks and different styles of dance.

Statistical modeling approaches typically operate in joint curve space (e.g. in terms of rotations). As a result, a meaningful style model must be built from motions belonging to a similar motion category (such as walking or dancing). Before learning the model, input motion clips additionally need to be time warped so that similar poses are aligned. A final consideration is that generated motions will often require post processing to fix artifacts such as foot sliding. This results from the lack of a physical model to ensure that synthesized motions are also physically correct.

Many statistical models have been proposed for generating and translating stylistic
movements. Most approaches are based on a generative model, which is capable of producing new motions which are statistically similar to the input examples. In one of the earliest works of this kind, Brand & Hertzmann (2000) used a hierarchical hidden Markov model (HMM), called a “Style Machine”, to model categories of movement, such as dance or ambulatory movement. A low level HMM modeled the primary structure of the motion (e.g., foot steps) and a top level HMM modeled various styles. Li et al. (2002) also focused on dance. Their method built a generative model based on the concept of “motion textures”. A low level model, called a “textron” modeled short clips of motion with a linear dynamic system while a high-level model, based on a transition matrix, modeled choreographies. Torresani et al. (2006) used linear regression with Laban effort features to produce different styles of dance. Taylor & Hinton (2009) used a conditional restricted Boltzmann machine to learn different styles of walking. Min et al. (2010) learned a generative model of both individual variation and categorical style using a multilinear model. Other previous work used Bayesian networks to model individual variations across motions of the same type, such as cheering, walking, side-stepping, or swimming (Lau et al., 2009; Ma et al., 2010).

Statistical style models have also been used to detect similar motions, to bias motion editing, or to add detail to a sparse set of keyframes. For example, Ren et al. (2005) compared the use of mixture of Gaussians (MoG) models, HMMs, and switching linear dynamic systems (SLDS) for detecting unnatural motions. Pullen & Bregler (2002) used a model based on a given motion set to add detail to user-created keyframes. The authors approximated correlations between the degrees of freedom with an input motion set and then used these correspondences to fill in missing trajectory information. In later work, Grochow et al. (2004) proposed style-based IK for biasing IK poses such that they are similar to an input motion set (modeled using a Gaussian Process Latent Variable Model (GP-LVM)). The authors demonstrate how style-based IK could be used to fill in missing information, for example, in trajectory keyframing or predicting the
positions of missing markers during motion capture.

Other generative approaches focus on modeling a particular person’s patterns of movement. For example, Ashida et al. (2001) incorporated actions into the upper body (head touches, body touches, head scratches, body scratches, yawning, looking around, looking down, and gesturing) to support a more realistic heterogeneity of pedestrians for crowd simulation. The distribution of upper body motions were modeled statistically-based on observed data. Neff et al. (2008) built a model of an individual’s gesturing style from annotated video data. The model was then used to generate conversational animations having the same distribution of corresponding gestures. Lee & Popović (2010) used inverse reinforcement learning to model how a player avoids obstacles while traversing a virtual environment. Chen et al. (2010) used a HMM to model the transitions between different attacks to capture an individual’s fighting technique.

An alternative approach is to build a model for translating between styles. In this case, the model maps how an input style relates to an output style. For example, Hsu et al. (2005) used a time-invariant linear model to map between two styles, determined by at least two input motions. Similarly, Xia et al. (2015) used local regression models to learn mappings between input and output motions without the need for motion annotations. Once learned, the linear models from these works can convert other examples from the input motion style into the output motion style. Analogously, Ikemoto et al. (2009) used a Gaussian process model to learn motion edits which could then be applied to a entire motion. Wu et al. (2006) investigated an online method for transferring style (defined in terms of joint angle statistics) from one motion onto a new motion. In the same spirit, the work of Min et al. (2010), described above, showed that their generative model could be used to edit and transfer styles.


### 2.6 Remarks and summary

We segment approaches to motion style into three primary categories: physically based approaches, statistical modeling, and signal editing. Each approach has its strengths and weaknesses and new techniques increasingly combine aspects from each, for example, building statistical models which include physics (Wei et al., 2011), or physically-based models which include optimization (Hämäläinen et al., 2014). In general, however, different approaches tend to excel for producing certain subsets of styles (See Table 2.1 for a listing). For example, physics-based approaches are best for responding to impacts, jumping, and changing body properties whereas statistics-based models remain best for aesthetic styles, such as personality, moods, and dance. For a practical application, such as games, a mix of motion capture (or keyframed animation) can be dynamically swapped with a physical approach whenever appropriate. A weakness of statistical approaches is that they are only as good as the data fed into them and sometimes, the model does not add functionality over the original motion set, which can be adapted in real-time using simple techniques such as blending and IK. For example, a statistical model of a sneaky walk may not be easier to work with in terms of speed, memory, and simple artist pipeline, than simply using the motion directly. Not surprisingly, the advantages of statistical methods only appear when datasets become larger. In these cases, statistics-based methods aid generating stylistic variations and can facilitate quick search and more naturalistic editing. In Table 2.2, we summarize pros and cons for each approach.
<table>
<thead>
<tr>
<th>Method</th>
<th>Demonstrated examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physically based approaches</td>
<td>changing foot contacts to create new gaits (limps, creatures, zig-zag steps, changing gravity), carrying heavy weights, changes in body mass, responding to hits and impacts, falls, changing shoes, modifying jumps, kicks, spins and flips, fatigue and injury effects by modifying the forces and movements of individual joint, grabbing railings to maintain balance, wall climbing</td>
</tr>
<tr>
<td>Statistical modeling approaches</td>
<td>dance, sneaking, hobbling, stealthy walking, walking at different speeds, jaunty walking, stride walking, side-ways stepping, cheering, swimming, gesturing, following paths and avoiding obstacles, limping, crouching, proud walking, marching, catwalking, dinosaur walking, gangly walking, kickboxing, drag queen walk</td>
</tr>
<tr>
<td>Signal editing approaches</td>
<td>emotions (happy, sad, angry, afraid, grief-stricken), cartoonish motions, sneaking, runway walking, gesturing, walking with different moods (briskness, tiredness, shivering, clueless, tired, delirious, determined, frenzied, ashamed, bored, goofy), speed, stair climbing and descending, elderly kickboxing</td>
</tr>
</tbody>
</table>

**Table 2.1:** Examples shown to work well for different approaches to creating, editing, and modeling motion styles.
<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physically based approaches</td>
<td>• No need of examples from the desired style</td>
<td>• Cannot model how mental states, personality, aesthetics, individuality, etc. affect movement</td>
</tr>
<tr>
<td></td>
<td>• Increasing support for real-time</td>
<td>• Solving few new motions can be slow</td>
</tr>
<tr>
<td></td>
<td>• Guarantees a physically correct result</td>
<td>• Some formulations may not always converge to a desired minimum, e.g. we may not always get a solution</td>
</tr>
<tr>
<td></td>
<td>• Dynamics can be changed for any input motion, not just categorically similar ones</td>
<td></td>
</tr>
<tr>
<td>Statistical modeling approaches</td>
<td>• Flexibility: a model could be built from any consistent motion category</td>
<td>• No guarantees on physical correctness</td>
</tr>
<tr>
<td></td>
<td>• Model is learned automatically</td>
<td>• Often requires post processing</td>
</tr>
<tr>
<td></td>
<td>• Domain knowledge of underlying processes not needed</td>
<td>• Style can only be transferred between similar motions</td>
</tr>
<tr>
<td></td>
<td>• Generating new motions from the learned model is often fast</td>
<td>• Some models (such as kernel regression) can require the full original motion set to be stored</td>
</tr>
<tr>
<td></td>
<td>• Can model both individual and categorical differences in motion</td>
<td>• Offline computation, so the result can be checked</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Requires examples of the desired style</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Limited by the motions in your database</td>
</tr>
<tr>
<td>Signal editing approaches</td>
<td>• Intuitive user interfaces</td>
<td>• Some methods require examples in the desired style</td>
</tr>
<tr>
<td></td>
<td>• Fast: animators can see changes immediately</td>
<td>• Requires post processing</td>
</tr>
<tr>
<td></td>
<td>• Many styles can be modeled and transferred (with sufficient trial and error)</td>
<td>• No guarantees on physical correctness</td>
</tr>
<tr>
<td></td>
<td>• Style examples might not always needed</td>
<td>• Style can only be transferred between similar motions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Often needs to be offline, so the result can be checked</td>
</tr>
</tbody>
</table>

**Table 2.2:** Strengths and weaknesses of different approaches to creating, editing, and modeling styles.
Chapter 3

Effects of motion editing on emotional style

3.1 Introduction

To engage us in a movie or game, virtual characters need the ability to convey emotions in a convincing way. This goal is most often achieved by capturing human performances directly with motion capture, as this technology faithfully records the nuances of an actor’s performance. However, one limitation of motion capture is that it does not inherently adapt to new situations. Thus, extensive research has explored how to increase the reusability and flexibility of motion capture clips, leading to numerous techniques in common use today, such as inverse kinematics, interpolation, blending, retargeting, morphing, move trees, motion graphs, overlays, and splicing. Further extensive research, mostly in psychology, has explored how we perceive emotions. However, little research has looked at the perceptual effects of motion editing techniques.

In this chapter, we investigate which aspects of body language are important for conveying emotions with two goals in mind. The first goal is to understand how changes introduced by motion editing might alter the emotional content of a motion, so we
can ensure that important aspects of a performance are preserved. The second goal is to gain insight on how we may edit captured motion to change its emotional content, further increasing its reusability. We study six basic emotions (Ekman, 1992) shown to be readily recognized across cultures: anger, disgust, fear, happiness, sadness, and surprise (Figure 3.1). The motion of a character’s body can effectively express all basic emotions (Atkinson et al., 2004), including its context and intensity, and is the focus of our study.

Many motion editing techniques affect part of the body, change poses, or modify the motion dynamics. Thus, we performed three experiments to gain a better understanding of the perception of emotions based on posture and dynamics. We recorded an actor giving ten short performances of each basic emotion, which were mapped to a generic humanoid virtual character. In our first experiment, we analyse this large set of clips to search for differences in posture and dynamics between emotions and to compare the perception of our stimuli to previous work. We also establish a baseline set of twelve animation clips (two for each emotion) having the highest recognition rates among viewers. As some editing techniques affect only specific joints of the body, our second experiment determines what part of the body (either head, upper body and head, or lower body) conveys the emotion most strongly. As previous work has established a relationship between pose, dynamics, and the perception of emotion (Wallbott, 1998; Pollick et al., 2001; Atkinson et al., 2004; Coulson, 2004; Roether et al., 2009), our third experiment systematically alters the poses and joint velocities. Our experiments yield
the following observations:

- We confirm several findings on the perception of emotions based on a large and varied set of clips. For example, we confirm that happy and angry movements have higher velocities and greater joint amplitudes whereas sadness has slower joint velocities and smaller amplitudes.
- Emotions are mostly conveyed through the upper body.
- The perceived intensity of an emotion can be reduced by blending with a neutral motion.
- We find that posture changes can alter the perceived emotion and its intensity while changes in dynamics only alter the intensity.

3.2 Emotion recognition

Extensive research in psychology (Roether et al., 2009) aims to understand the perceptual significance of body motion for conveying emotions. Coulson (2004) showed that happiness and sadness are clearly recognizable while disgust is harder to discern. He also found that surprise and fear were harder to discern from purely static poses. Wallbott (1998) observed a relationship between emotion type and posture characteristics, showing that differences in emotion can be partly explained by the dimension of activation, where activation refers to velocities, accelerations, and jerk. Atkinson et al. (2004) showed that emotion could be recognized from body motion, with exaggerations in motion increasing the recognition accuracy. Roether et al. (2009) observed that elbow and hip flexion were important attributes for anger and fear, while head inclination was important for recognizing sadness in motions. Sawada et al. (2003) found that dancers varied the speed, force, and directness of their arm movements when conveying
joy, sadness, and anger. Other work by Ennis & Egges (2012) observed that negative emotions are better recognizable.

The recognition of emotions is mostly consistent across cultures and geometries. Kleinsmith et al. (2006) conducted a study to evaluate the cultural differences in the perception of emotion from static body postures, observing moderate similarity across cultures. Pasch & Poppe (2007) evaluated the importance of the realism of the stimuli on the perception of emotion, demonstrating that high realism did not always conform to an increase in agreement of the emotional content. McDonnell et al. (2008) investigated the role of body shape on the perception of emotion and found that emotion identification is largely robust to change in body shape.

Several previous studies focus on specific motion categories such as gait (Crane & Gross, 2007; Roether et al., 2009), knocking and drinking (Pollick et al., 2001), and dance movements (Sawada et al., 2003), whereas this study uses non-restricted actor portrayals, similarly to other previous work (Wallbott, 1998; Atkinson et al., 2004; McDonnell et al., 2008). This choice was made for two reasons. First, we did not want to inadvertently restrict the motions to parts of the body. For example, if we based our experiments on knocking, we might bias the upper body to display most of the emotion. Second, because several related studies have already performed systematic comparisons between motions of the same type, we wanted to investigate how their findings compared to our varied motion set.

### 3.2.1 Impact of Motion Editing

Avoiding unwanted artifacts during motion editing is an important issue in computer animation. Ren et al. (2005) presented a data-driven approach to quantifying naturalness in human motion, which they evaluated against edited and keyframed animations. Ryall et al. (2012) and Reitsma & Pollard (2003) observed viewers’ sensitivity to timing
changes made to walking and ballistic motions and found that participants were more sensitive to time warping when slow motions were made faster than when fast motions were made slower. Safonova & Hodgins (2005) analysed the effect of interpolation on physical correctness and found that such operations can create unrealistic trajectories for the center of mass or create unrealistic contacts.

3.3 Experiment 1: Emotion Recognition and Movement Analysis

In our first experiment, we determine the recognition rates achievable with our clips. We use the results to select well recognized motions for the subsequent experiments, to compare our stimuli to previous studies, and to validate existing findings against our diverse motion set.

3.3.1 Stimuli creation

We invited an experienced stage actor to give ten short performances of each of the six emotions: anger, disgust, fear, happiness, sadness, and surprise (1 actor × 6 emotions × 10 portrayals = 60 animation clips). The actor was asked to convey each emotion as convincingly as possible using his entire body and no vocal cues. He was also told that his face and hands would be blurred. Given these instructions, the actor improvised ten entirely different portrayals of each emotion, some of which were repeated until the actor was satisfied with his performance. For example when performing anger, the actor’s performances ranged from tantrums, to yelling and pointing, and finally to head shaking and staring (to show contained rage, or annoyance). In general, our actor portrayed each emotion using stereotypical movements. To show disgust, he most often turned
away, made sweeping away gestures with his hands, or acted sick to his stomach. To show fear, he would back away, place his hands up in front of the body, or freeze in place. To show sadness and grief, he would cross his arms, hang his head, or cover his face with his hands. To show happiness, he cheered, jumped for joy, bowed, and twirled. To show surprise, he performed startled jumps, followed by relief. Although such performances may be considered staged exaggerations of natural emotions, they are typical of the types of movements commonly captured for entertainment. Previous research has shown that exaggerated expressions of emotion are more readily recognizable by viewers (Atkinson et al., 2004). As a result, the recognition rates for our stimuli might be higher than for more realistic displays of emotion.

Our actor’s body motions were recorded with a 12-camera optical Vicon system and post-processed in Vicon Nexus and Autodesk MotionBuilder. A standard skeleton with 22 joints was used. We then created a skinned character adapted to the actor's skeleton.

As we did not record facial or finger motions, we box blur the face and hands of the virtual character similar to McDonnell et al. (2008) to ensure that the motionless, unnatural-looking faces or fingers would not distract the viewer. We kept the blurred area as small as possible, hiding the facial features and the finger motions of the character but still showing the general orientation of the head and the hands. The box blur algorithm was implemented as a Maya plugin.

The resulting character was rendered in a neutral environment. A few clips had to be excluded from the experiment because they included too many occluded markers to be cleaned accurately, or resulted in clips that were too short. In total, we obtained 55 animated clips (11 anger, 9 disgust, 7 fear, 10 happiness, 9 sadness, and 9 surprise), each between 2 and 10 seconds long at 24 fps.
3.3.2 Method

Fifteen participants (8M, 7F) between 17 and 53 (mean 25.6) watched all 55 clips. All participants were naïve to the purpose of the experiment and had normal or corrected to normal vision. After each clip, they were asked to specify which emotion they thought was conveyed in the video with a forced-choice between anger, disgust, fear, happiness, sadness, and surprise. Although forced-choice questionnaires have the drawback of restricting the responses from our participants and could potentially inflate recognition rates (Frank & Stennett, 2001), we chose this format to be able to compare our study to related work, which mostly uses forced-choice.

As the goal of this experiment was to find how well our stimuli conveyed the basic emotions, participants could perform the study at their own pace and view each clip as often as they wanted. They were sent a link to the study and were allowed to view the stimuli on their own computers and to take breaks. The clips were presented in a different random order for each participant. Three example questions were given to participants before beginning the experiment, so that participants were familiar with the rendering style and question format before starting.

After all 55 clips had been viewed and an emotion selected for each clip, participants were asked to watch them a second time and to rate the intensity and energy of the emotion on a scale from 1 (not intense/low exertion) to 5 (very intense/high exertion). Definitions of intensity – “How deeply the person feels the emotion” – and energy – “The level of exertion and vigor of the person’s movement” – were displayed on screen.

The entire experiment took between 30 and 45 minutes to complete. Participants had the option to perform the study in our lab, where they were compensated with food and refreshments, or they could choose to perform the study elsewhere without compensation.
3.3.3 Results

On average across all clips, 62.4% of the clips were recognized correctly. Figure 3.1 summarizes our results. In general, we see that happiness, anger, fear, and surprise were recognized best whereas disgust and sadness were recognized least. Anger was generally very well recognized with the exception of two motions which had only one correct response each. The two poorly recognized motions were subtle anger expressions (“contained rage”) which did not translate well to the character without facial animation. Many participants mis-categorized these anger motions as sadness. In general, clips with low recognition rates fell into two categories: participants had no agreement on the displayed emotion, suggesting that they were merely guessing, or there was high agreement among participants for the wrong emotion. Most clips with low scores were in the first category. Examples from the second category included a sadness clip where our actor was waving his hands (to gesture “go away”), which was chosen as anger or happiness, and a happiness clip which was predominately chosen as anger.

Disgust was recognized least well in our experiment (although two disgust motions had high recognition rates). Disgust is known to be a less readily recognized emotion and our result is consistent with a large body of work (Ekman, 1992; Atkinson et al., 2004). However, sadness is typically recognized at a higher rate. Our hypothesis regarding this finding is that the actor often tried to show grief and distress, where the body typically moves more than in a depressed individual. Without facial capture, this subtlety was lost. In the confusion matrix in Table 3.2, we see that disgust was most often confused with sadness. With 12.5% of the total selections, disgust was also selected less often than any other emotion (all emotions were displayed equally often: 16.7% of the time).

From this experiment, we choose the two animated performances with the highest recognition rates to use for all subsequent experiments: two anger motions each with 100% correct recognition; two disgust motions (recognition rates 80% and 93%); two fear motions (93% and 100%); two happy motions (93% and 100%); two sad motions
Figure 3.1: Emotion recognition from our first experiment. Each column summarizes the percentage of correct responses for each emotion. The red middle band indicates the median value, the box top and bottom shows the first and third percentiles (75% and 25%), and the top and bottom whiskers show the maximum and minimum recognition percentage. Crosshairs show outliers.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Happiness</th>
<th>Sadness</th>
<th>Anger</th>
<th>Disgust</th>
<th>Fear</th>
<th>Surprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td><strong>68.0</strong></td>
<td>3.3</td>
<td>24.7</td>
<td>0.7</td>
<td>0.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Sadness</td>
<td>13.3</td>
<td><strong>50.4</strong></td>
<td>10.4</td>
<td>8.1</td>
<td>12.6</td>
<td>5.2</td>
</tr>
<tr>
<td>Anger</td>
<td>5.6</td>
<td>12.7</td>
<td><strong>72.1</strong></td>
<td>6.7</td>
<td>0.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Disgust</td>
<td>9.6</td>
<td>20.7</td>
<td>7.4</td>
<td><strong>46.7</strong></td>
<td>11.1</td>
<td>4.4</td>
</tr>
<tr>
<td>Fear</td>
<td>0.0</td>
<td>6.7</td>
<td>1.0</td>
<td>9.5</td>
<td><strong>71.4</strong></td>
<td>11.4</td>
</tr>
<tr>
<td>Surprise</td>
<td>5.2</td>
<td>9.6</td>
<td>8.1</td>
<td>3.0</td>
<td>8.9</td>
<td><strong>65.2</strong></td>
</tr>
</tbody>
</table>

Table 3.2: Confusion matrix from our first experiment. Entries show the percentage of times participants chose each emotion in a forced choice experiment. The displayed emotions are listed on the left, the selection of the viewer at the top.
(80% and 86%); and two surprise motions (100% and 87%). In case of a tie, the clips that looked least similar to each other were selected.

We also analyse the pose and velocities of our motion clips and compare our findings to previous studies. Figure 3.2 shows histograms of the rotational speeds for the major animated joints of our character, namely the root (pelvis), left and right hips, knees, ankles, spine, shoulders, elbows, wrists, and neck. To compute angular velocities, we first compute quaternion rates using 5-point central differencing and then convert the quaternion rate to an angular velocity vector according to (Diebel, 2006). Our motions are consistent with previous research which states that anger and happiness tend to have larger and faster joint movements, whereas fear and sadness tend to have smaller and slower joint movements (Roether et al., 2009).

Figure 3.3 compares the amplitudes for the head, shoulders, and elbows (amplitude is defined as the difference between the max joint angle and min joint angle for each motion category). Our findings are consistent with previous research which states that happiness and anger have higher amplitudes whereas sadness and fear have lower amplitudes. Existing research is less clear regarding disgust and surprise. For our dataset, surprise shared amplitude characteristics with fear, but had greater elbow movement. Disgust had high elbow movement but low head and shoulder movement.

Lastly, we looked at modal and average flexion angles, defined as the angle between limbs. Specifically, previous research describes reduced head angle for sad walking and increased elbow angle for fearful and angry walking (Roether et al., 2009). However, our motion set did not produce convincingly consistent results. Most joint angle distributions were not normally distributed. Based on histograms of joint angle, both sad and disgust motions had modal head angles of 160 (where 180 corresponds to looking straight forward and 90 degrees corresponds to looking straight down) whereas all others had modal head angles of 170 degrees. Elbow angle was greatest for disgust and fear (110 degrees, where 180 corresponds to a fully flexed arm), second largest for sadness and
Figure 3.2: Histograms of rotational speeds (radians/second) for the major joints of our character: root, left and right hips, knees, ankles, spine, shoulders, elbows, wrists, and neck. The x axis shows bins corresponding to speeds of 1, 2, 4, 8, and 16 radians/second respectively. The y axis shows counts for each speed across all motions of each category, normalized to range from 0 to 1 based on the maximum bin size. Our results are consistent with previous published research which states that anger and happiness tend to have larger and faster joint movements, whereas fear and sadness tend to have the least joint movement. For our motions, surprise and disgust lie somewhere in between these two extremes.
anger (150 degrees), and smallest for surprise and happy (170 degrees). To understand these results, we note that many of our sad clips had the hands at the face, and several of our disgust motions huddled the arms into the body. Many of the anger motions contained punching and swinging gestures.

We averaged the ratings of each participant over all clips of the same emotion and used a repeated measures ANOVA to determine that there were significant differences in intensity ($F(5,70)=43.1, p < 0.001$) and energy ($F(5,70)=91.4, p < 0.001$) for the different emotions. Post-hoc Tukey tests were used to determine that the intensities and energies of happy and angry were significantly higher than the other emotions; that all emotions except fear were significantly higher in intensity than sadness; and that all emotions were significantly higher in energy than sadness. The correlation coefficient between intensity and energy was 0.65.

### 3.4 Experiment 2: Partial Occlusions

Many motion editing operations can be applied to parts of a virtual character’s body. For example, inverse kinematics and overlays for reaching usually just affect the upper
body whereas ground clamping techniques to adapt walking motions usually just affect the lower body. Therefore, in our second experiment, we determine which parts of the body are important in conveying emotions.

### 3.4.1 Stimuli

For each of the six emotions, we chose the two clips with the highest recognition rates from Experiment 1 (see Section 3.3.3). We then occluded different parts of the body: the head motion (or NH for “No Head motion”), the lower body motion (NL), and the upper body motion (NU). The unaltered motion is labeled OR for “original”. We did not alter the root motion for any of the conditions to avoid very unnatural motions that could affect the ratings in unintended ways (see video for examples). To occlude the body parts, we erase all motion from the considered part and cover it with a nondescript flat cuboid that we attached to the character (see Figure 3.4). We obtain 6 Emotions x 2 Clips x 4 Occlusion types = 48 different clips.
3.4.2 Method

Sixteen participants who were not involved in the previous experiment watched all of the clips in small groups of 1–3 participants on a large projection screen in a seminar room. As the aim of this experiment is not to determine the highest possible recognition rate of each clip but to investigate differences between several partial occlusions, we chose a faster pace for this experiment. Participants viewed a clip once. Then they had a total of six seconds to specify the perceived emotion in a forced-choice between the six basic emotions and then the perceived intensity of that emotion on a scale from 1 to 5 similar to Experiment 1. After four seconds, a sound was played together with the number of the next clip to alert participants to look at the screen again. Then the next clip started. Participants were asked to watch each clip in its full length. As in the previous experiment, participants were first shown sample questions so they understood the question format and rendering style before beginning the experiment.

Although our pilots showed that six seconds was very short, participants were able to follow the instructions after a short training phase. However, we decided not to ask to rate the energy of the clips as we found that participants were not able to effectively distinguish between intensity and energy in such a short time. Our second reason for the very fast pace was that participants watched the same motions with different occlusions. Once a non-occluded animation has been viewed, it is possible for a participant to recognize that animation in subsequent clips and to infer the perceived emotion and intensities. The fast pace of our experiment did not give participants time to think about the motions. Based on questions and conversations in the debriefing, we assume that many participants started to recognize some of the motions towards the end of the experiment.

Before starting each experiment, we showed participants four training clips at the same pace as the experiment. The training clips were chosen from the unused clips in the first experiment. A short break to answer any questions ensured that participants
understood the instructions. The participants viewed all 48 clips in random order. After a short break, they viewed all 48 clips again in a different random order. The full experiment took about 25 minutes to complete and participants were rewarded with $5.

### 3.4.3 Results

#### Emotion recognition

Three participants either did not follow the instructions or checked the boxes in an illegible manner. Their answers had to be discarded, leaving 13 participants in our analysis. For the emotion recognition, we computed the error rates for each participant, emotion, and occlusion type by averaging over the two clips and two repetitions. We then performed a repeated measures ANOVA with the within-subject factors Occlusion type and Emotion. We used Newman-Keuls post-hoc tests to determine the origin of the significant effects.

We found a main effect of Occlusion type with $F(3, 36) = 62.8, p < 0.001$, due to the fact that the condition where the upper body was hidden had significantly higher error rates (lower recognition rates) than the other three occlusion types. This effect was surprisingly distinct as can be seen in Figure 3.5 (left). There were no significant differences between the other three occlusion conditions.

As expected, we also found a main effect of Emotion ($F(5,60) = 5.3, p < 0.001$), meaning that the different emotions were not recognized equally well. Fear was recognized best on average and significantly better than all other emotions except happiness (see Figure 3.5, right). Sadness had the lowest recognition rate (or highest error rate), which differed significantly from fear and happiness and reflects the already lower recognition rate of the original clips.

Furthermore, there is an interaction effect between Occlusion type and Emotion ($F(15, 180) = 7.5, p < 0.001$), which can be traced back to three causes: First, fear was
the only emotion where the error rates remained the same in all occlusion conditions. For all other emotions, hiding the upper body resulted in significantly higher recognition errors than all or most (for anger) other conditions. Second, sadness with an occluded head (NH) was recognized least of all OR, NH, and NL motions and significantly less well than eight other motions. Third, anger with occluded lower body (NL) was recognized second least of all OR, NH, and NL motions, leading to significant differences with three of them.

**Intensities**

To analyse intensity, we had to discard the results of one more participant as those had not been reported correctly, reducing the participant count to 12. Similarly to emotion recognition, we computed the averages for each participant, emotion, and occlusion type over the two clips and two repetitions, performed a repeated measures ANOVA with the within-subject factors *Occlusion type* and *Emotion*, and used Newman-Keuls post-hoc tests to determine the origin of the significant effects.

Here again, we found a main effect of Occlusion type ($F(3,33) = 29.5, p < 0.001$) based on a significant difference between the clips where the upper body was occluded (NU) and the clips in the three other occlusion conditions. The intensity was rated
Figure 3.6: Intensities for each condition (left) and for each emotion (right). Emotions were rated to have a lower intensity on a scale from 1 to 5 when the upper body was occluded. Also, anger, fear, and happiness were rated to have a significantly higher intensity than disgust, sadness, and surprise.

significantly lower for NU (see Figure 3.6, left). There was also a main effect of Emotion with $F(5, 55) = 8.3, p < 0.001$. The post-hoc test showed that the intensity ratings were split into two groups: Anger, Fear, and Happiness were rated to have higher intensities than Disgust, Sadness, and Surprise (see Figure 3.6, right). There were no significant differences within each group but each combination of ratings across groups was significantly different from each other. Finally, there is an interaction effect between Occlusion type and Emotion ($F(15, 165) = 3.5, p < 0.001$), which, however, does not add new insights. The posthoc test reveals that for each individual emotion, the intensities are rated significantly lower when the upper body is occluded than in the three other occlusion types, whereas there are no significant differences between those three.

Discussion

We infer that the upper body is crucial for the perception of emotions. The lower body or the head alone were not relevant in our set of clips to recognize the emotion. The irrelevance of the head for all emotions except sadness could have been due to our blurring of the head in the baseline clips: when we occluded the head entirely, there was no considerable impact. Alternatively, this finding might have be due to the head being
unimportant for recognizing the emotion in nearly all our clips. However, the relatively high error rate for sadness when the head was occluded complies with previous work that head motion is particularly important for displaying sadness.

However for the lower body, which was not blurred, the emotion could be effectively conveyed through lower body motions, for example through the kicking motion for anger, the running away motion for fear, or the jump for surprise. Interestingly, differences between the occluded recognition rates were smallest for two of the emotions that displayed very distinct lower body motions, namely fear and anger. Because we decided to hide the lower body motion but leave the root (pelvis) motion unaltered, viewers may have inferred the lower body motions based on the movements of the upper body. We also considered several other options, such as deleting the root motion or replacing the lower body motion with a neutral lower body motion. However, these options drastically changed the full body motions instead of just hiding parts of the body and were therefore discarded.

The intensity ratings largely mirror the recognition rates: when the error rates were higher, the intensity was judged lower. This is not surprising: participants might recognize an emotion less clearly when its intensity is low or might attribute a low intensity to a motion when they are unsure which emotion it represents.

### 3.5 Experiment 3: Posture and Dynamics

Previous research suggests that velocity, accelerations, and jerk (defined as the time derivative of acceleration) are important factors in emotional body language along with pose (Roether et al., 2009). Because motion editing procedures such as interpolation and blending change both the pose and dynamics, we investigate these effects. We hypothesize that small scale changes might affect the emotion intensity, while large scale changes might affect whether the emotion is recognized correctly.
3.5.1 Stimuli

We filtered the major joint curves of our best-recognized motions to produce changes to either the poses, the velocities, or both. For this experiment, we created four conditions: two conditions (BB25, BB50) in which we change poses and velocities by blending the upper body with a neutral posture, one condition (DTW) in which we change the timing but not the poses through dynamic time warping, and one condition (OFF) where we change the poses but not the timing by setting constant offsets to either the shoulders, elbows, or head.

Our two body blend conditions (BB25, BB50) blend the joints of the upper body, from the spine and upwards (Figure 3.7 and 3.8). The upper body was chosen because the previous experiment showed it to be the most relevant for the perception of emotions. To compute each blend condition, each frame (where a frame consists of the rotations for each joint) of the original motion is blended with a neutral pose having the arms down at the side. BB25 blends 75% of the original motion with 25% of the neutral pose. BB50 blends 50% of the original motion with 50% of the neutral pose.

Our dynamic time warping condition (DTW) modifies the timing of the motion such that no joint velocity is higher than a given maximum. We choose our maximum...
Figure 3.8: Stimuli examples from body blend (BB) conditions. Original motions appear in the first column. The second column shows BB25, which retains 75% of the original motion. The third row shows BB50, which retains 50% of the original motion. As the poses moved towards neutral, the perceived intensity of the emotion is decreased.
value separately for each emotion as 200% of the average speed of the fastest moving joint. If the speed of all joints for a set of frames does not exceed the maximum value, those frames remain unchanged. Given a maximum speed, dynamic time warping is performed by computing new times for each frame and then resampling the motion curves at the original framerate. Specifically, if a frame originally occurred at time $t$ and had its fastest joint $i$ moving at $v > v_{\text{max}}$, we adjust the time for this frame so that it occurs at $t + v/v_{\text{max}} \Delta t$, where $\Delta t$ is $1$/framerate. The curve is resampled by interpolating between the original poses.

Our offset condition (OFF) modifies the poses without changing the timing. Offsets were specified manually for either the shoulders, elbows, or spine and neck by specifying an offset pose $q^{\text{user}}$ for a single reference frame $q(\hat{t})$ of each original motion (Figure 3.9). From the offset and reference pose, we compute an offset rotation $q^{\text{offset}}$ which is then applied to all frames.

$$q^{\text{offset}} = (q(\hat{t}))^{-1} q^{\text{user}}$$

$$q_i^{\text{new}}(t) = q^{\text{offset}} q_i(t)$$

For our offset condition, we added offsets for each of our twelve clips (two per emotion). We created three motions with changed elbows, three motions with altered shoulders, three motions which modified the neck and spine upwards, and three motions where the neck and spine went downwards.

We apply these four posture and velocity conditions — BB25, BB50, DTW, and OFF — to the two clips with the best recognition rates for each emotion (and keeping the original motion OR) to obtain $6 \text{ Emotions} \times 2 \text{ Clips} \times 5 \text{ Alterations} = 60$ different clips.
Figure 3.9: Stimuli examples from the offset condition. Original poses appear in the first column. Modified poses appear in the second column.
3.5.2 Method

We used the same fast paced method as experiment 2 (see Section 3.4.2). Seventeen naïve participants, who were not involved in any of the previous experiments, took part in experiment 3, which took less than 30 minutes to perform. As before, they were rewarded with $5. As before, participants were shown sample questions before the experiment so they were familiar with the question format and rendering style before starting.

3.5.3 Results and Discussion

One participant with unclear answers had to be excluded, leaving 16 participants in the analysis. As in Experiment 2, we computed the averages for each participant, emotion, and Alteration type (OR, BB25, BB50, DTW, and OFF) over the two clips and two repetitions, performed a repeated measures ANOVA with the within-subject factors Alteration and Emotion, and used Newman-Keuls post-hoc tests to determine the origin of the significant effects.

Emotion recognition

As expected, we found a main effect of Alteration with $F(4,60) = 5.1, p < 0.01$ (see Figure 3.10, left). The error rates for the motions blended to 50% with the neutral motion (BB50) and the ones with offsets (OFF) were recognized significantly less well than the unmodified ones (OR). There were no significant differences between the recognition rates of BB25, the time-warped motion (DTW), and the original condition. However, the difference between the conditions OFF and DTW was significant. We also found a main effect of Emotion ($F(5,75) = 6.1, p < 0.001$) due to the sadness motion being recognized at a significantly lower rate than all of the other emotions, which restates a result we found throughout the whole study.
Finally, there is an interaction effect between Alteration and Emotion with \( F(20, 300) = 3.8, p < 0.001 \), mainly due to the offset and 50% neutral blended sadness clips (OFF and BB50) having significantly higher error rates than any other combination of Emotion and Alteration (see Figure 3.10, right). We found that those two combinations (sadness OFF, and sadness BB50) are also the origin for the main effects of Emotion and Alteration. The differences between the alterations of the other emotions were not significant.

**Intensities**

As expected, alterations changed the perceived intensities of our clips. The perceived intensities of the clips with the alterations BB50 and DTW were significantly reduced with a main effect of Alteration \( F(4, 60) = 5.1, p < 0.01 \), see Figure 3.11, left). As before, the clips with different emotions were also rated as having different intensities (main effect of Emotion with \( F(5.75) = 13.2, p < 0.001 \)), with anger and fear having the highest intensities, and sadness the lowest.

Finally, the interaction effect between Emotion and Alteration with \( F(20, 300) = 2.5, p < 0.001 \) showed that our modifications had a different effect depending on the emotion. For anger, time warping (DTW) significantly reduced the perceived intensity, whereas for sadness, blending (BB50) and adding an offset (OFF) reduced the intensity.
Figure 3.11: Intensity for each alteration (left) and for all emotions and alterations (right). Emotions were rated to have a lower intensity on a scale from 1 to 5 when blended to 50% with a neutral motion or when time warped.

significantly (see Figure 3.11, right).

3.5.4 Discussion

We found that motion editing techniques can affect the recognition of emotions and its perceived intensity. The conditions BB50 and OFF, which both modify the posture, influenced the emotion recognition. The conditions BB50 and DTW, which both modify the timing, lead to a lower perceived intensity. From these results, we might infer that posture is a strong indicator of the type of emotion while timing and dynamics contribute to its perceived intensity. However based on the interaction effects, these effects are not equally strong for all emotions. The decreasing intensity for OR, BB25, and BB50 in Figure 3.11, left, suggests that the average intensity of a motion can be decreased by blending that motion with a neutral motion.

Participants were not able to determine the emotion sadness in the alterations OFF and BB50 as well as for the other emotions and alterations. This could be due to one of the OFF clips and both BB50 clips changing the orientation of the head, which was shown to be crucial in Experiment 2. Not surprisingly, the perceived intensity also decreased for those two cases where the emotion was not well recognized.
3.6 Remarks and summary

In this chapter, we investigated how changes to captured motion clips, such as those which commonly occur through motion editing, might alter the recognition and perceived intensity of an emotional performance. Rather than look at categories of motion, such as gait, we study a varied set of emotion clips. From these, we learn that the upper body motion is most crucial for the recognition of emotions, that changes to posture can change the perceived motion type whereas changes to dynamics can change the perceived intensity, and that the perceived intensity of an emotion can be reduced by blending with a neutral motion. However, these results do not apply equally well to all motions and emotions, and future work should try to understand these differences. For example, our findings regarding the upper body may primarily be due to the fact that the movement of the legs is restricted to maintain balance.

Overall these findings might motivate one to take care when splicing and using IK to control the upper body, since such changes can affect emotion recognition and reduce the motion’s perceived intensity. When blending major joints, such as the head, one might use smaller blend weights so that emotional content is not diluted.

Both the findings in this chapter and previous work consistently find correlations between emotions and dynamics. For example, happy and angry movements are associated with higher joint velocities. In this next chapter, we will investigate whether similar fundamental differences exist between neutral and stylistic motions. Unlike this chapter, where we use unconstrained vignettes, the next chapter will focus on a single motion category, walking, but analyze a broader set of styles in addition to emotions.
Chapter 4

Numerical analysis of stylistic walks

In this chapter, we investigate whether fundamental numerical differences exist between stylistic and neutral motions. We focus on a single category of motion, walking, for a wide variety of styles representing emotions, age, personality, and caricatured performance. For these walking examples, we compute numerical metrics using inverse dynamics aimed at quantifying the energy and naturalness of each motion example. Specifically, our goal is to test whether

- Neutral motions are more energy efficient than stylized motions.
- Neutral motions are more natural than stylized motions.

Understanding the numerical differences between stylistic and neutral motions can help ensure that important characteristics are not altered during motion editing. Additionally, a good numerical definition can inform good features for categorizing motion as well as enhance editing techniques by allowing us to infer the neutral counterpart of any stylistic input. For example, we saw in Chapter 3 that blending with a neutral motion could alter the intensity of an emotional performance, and in the work by Hsu et al. (2005) we have a model which relies on having neutral-stylistic pairs. Over all the metrics we tested, the sum of torques appears to be the most correlated with our neutral motion examples.
4.1 Method

We analyze over 55 examples of stylized and neutral walking motions based on motion capture data. We also add one neutral motion example which is the average of each neutral example. Each motion was recorded at 120 fps, trimmed to contain a single walking cycle, smoothed, annotated with ground contact information, and then retargeted to the same character. To ensure that comparisons between motions are meaningful, each motion contains matching contact changes (single foot, double stance, single foot, double stance). Metrics are then estimated from motion dynamics, based on a biologically-based anthropometric model and inverse dynamics. To avoid problems at the boundaries when computing finite differences, we additionally padded motions at the beginning and end. Figure 4.1 shows two examples of time-aligned walks with matching contacts. The primary datasets used in our experiments were

- **Subject 120, Stylized motions** (CMU Graphics Lab): locomotion in gorilla, sneaky, neutral, robot, and zombie styles.

- **Subject 137, Stylized motions** (CMU Graphics Lab): picking up, waiting, and walking in cat, chicken, dinosaur, drunk, gangly teen, graceful lady, elderly, sexy lady, strong man, and neutral styles.

- **Subject 142, Stylized motions** (CMU Graphics Lab): walks in childish, clumsy, cool, depressed, elated, elderly, happy, joyful, lavish, marching, hurt knee, relaxed, rushed, sad, scared, sexy, shy, singing, and sneaking styles.

4.1.1 Inverse dynamics

Our input motions are kinematic and we estimate forces and torques with inverse dynamics using the standard Newton-Euler two-pass algorithm (Craig, 1989). First, we estimate the local velocities and accelerations of the root and joint angles using
finite differences. Then, we propagate velocities and accelerations from the root to each end effector. Finally, we propagate forces from the end effectors to the root to get the aggregate forces acting on the character.

Local velocities and accelerations are computed using 5-point central differencing. We pad the motions at the beginning and end to handle poor estimates at the boundaries. For angle rates, we compute central differences of each quaternion and then convert these quaternion rates to angular velocities and accelerations according to Diebel (2006). Table 4.2 visualizes the differences in speeds across different stylistic examples. These results are consistent with previous research which shows that happy motions have faster velocities and accelerations than sad motions. For example, in Table 4.2, elated and happy have higher velocities and accelerations than depressed. Also note that some stylistic examples, such as march and dinosaur have extremely little upper body motion compared to average (the average example is on the top left).

**Outward pass: computing velocity and acceleration**

In the outward pass, we compute angular and translational velocities for the root
Table 4.2: Speeds for different stylistic examples. Colored plots showing the joint speeds for every joint for each frame. The x-axis represents frames and the y axis represents each joint. Brighter areas represent faster speeds. The upper body is stored in the top of each image and the lower body is shown lower. For example, in the dinosaur and marching styles, the upper body does not move much. Old man exhibits very little movement whereas jaunty and elated has a lot of movement over the whole body. Depressed has much slower movements than elated or jaunty. The top, left-most image is a neutral walk computed as the average walk over 10 neutral examples.
and joint angles. Angular velocity and acceleration of joint \( i \) are computed recursively like so

\[
\omega^i = R^i_{i-1} \omega^{i-1}_{i-1} + \dot{\theta}^i \\
\dot{\omega}^i = R^i_{i-1} \dot{\omega}^{i-1}_{i-1} + R^i_{i-1} \omega^{i-1}_{i-1} \times \dot{\theta}^i \dot{\theta}^i
\]

where the subscript indicates the \( i \)-th joint and the superscript represents the coordinate frame of reference. In other words, the above calculations are performed in the local coordinate system of the \( i \)-th joint. \( R^i_{i-1} \) is a transformation matrix which converts from the parent’s coordinate frame to the joint’s reference frame.

Translational velocity and acceleration of joint \( i + 1 \) are computed similarly like so

\[
v^i_{i+1} = R^i_{i-1} v^{i-1}_i + \omega^i \times r^i_{i,i+1} \\
\dot{v}^i_{i+1} = R^i_{i-1} \dot{v}^{i-1}_i + \dot{\omega}^i \times r^i_{i,i+1} + \omega^i \times \omega^i \times r^i_{i,i+1}
\]

where \( R^i_{i-1} \) is a transformation matrix which converts from the joint’s coordinate frame to its parent’s frame. To initialize the outward computations, we set \( v^0_0 = 0 \), \( \dot{\theta}^0_0 = 0 \), and \( \ddot{\omega}^0_0 = 0 \). The linear acceleration is set to \( \dot{v}^0_0 = (0, -9.8, 0) \) to account for gravity.

Inward pass: computing forces and torques

For the inward pass, we propagate forces from the end effectors to the root. The
inertial forces are given by

\[ F^i_i = m^i_i \dot{v}^i_{cm_i} \]
\[ N^i_i = I^i_{cm_i} \dot{\omega}^i_i + \omega^i_i \times I^i_{cm_i} \omega^i_i \]

where the subscript indicates the \( i \)-th joint and the superscript represents the coordinate frame of reference. In other words, the above calculations are performed in the local coordinate system of the \( i \)-th joint. \( v^i_{cm_i} \) is the linear velocity about the center of mass, \( \omega \) is the angular velocity and \( I^i_{cm_i} \) is the moment of inertia around the center of mass. The reaction forces are given by

\[ f^i_i = R^i_{i+1} f^{i+1} + F^i_i \]
\[ \tau^i_i = r^{i}_{cm_i} \times F^i_i + r^{i}_{i-1,i+1} \times R^i_{i+1} f^{i+1} + R^i_{i+1} n^{i+1} + N^i_i \]

where \( R^i_{i+1} \) is a transformation matrix which converts from the child's coordinate frame to the current joint's reference frame.

The above calculations depend on having a model of masses and moments of inertia for each joint. In the next section, we describe the anthropometrics model we use.

**Anthropometric model**

We use published anthropometrics data to estimate masses and moments of inertia for the body. First, we infer height based on the distance from foot to head. From the height, we can estimate the mass. We use information from the National Institutes of Health (2015) to fit a regression line relating height \( X \) to weight \( Y \)

\[ Y = 73.0766 \times X - 61.7542 \]
Joint masses are estimated as a percentage of total body mass. When a segment is composed of multiple joints, such as the trunk and the hands, we divide the mass between the joints. Similarly, the center of masses for each joint can be estimated from the joint lengths. In Table 4.3, we list the fractions used to distribute the total mass among each joint and the offset distances as a percent of limb length used to compute the COM positions for each joint.

Given height and mass, we then estimate the whole body density $\rho$ using the model from Contini (1972)

$$\rho = 0.69 + 0.9 \times \frac{h}{w^{1/3}}$$

Using the total body density, we can estimate the densities of each of the joints. We use regression lines adapted from Winter (2009) and plotted in Figure 4.1. Each regression line maps the total body density to each limb density using the slopes and intercept terms listed in Table 4.3.

The shape of our joints are estimated with cylinders. The cylinder volume can be computed from the limb density $V = m/\rho$. The cylinder radius can then be computed from the volume and joint length $r_j = \sqrt{\frac{m_j}{\rho + \pi s L_j}}$ where $r_j$, $m_j$, and $L_j$ is the radius, mass, and length of joint $j$ respectively. Moments of inertia are based on cylinders, using the parallel axis theorem to align the COM. The resulting body model is visualized in

<table>
<thead>
<tr>
<th>Segment</th>
<th>Mass</th>
<th>COM</th>
<th>Density B (y-intercept)</th>
<th>Density M (slope)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand</td>
<td>0.006</td>
<td>0.506</td>
<td>-0.44</td>
<td>1.5</td>
</tr>
<tr>
<td>Forearm</td>
<td>0.016</td>
<td>0.430</td>
<td>-0.0675</td>
<td>1.125</td>
</tr>
<tr>
<td>Upper arm</td>
<td>0.028</td>
<td>0.436</td>
<td>0.4225</td>
<td>0.625</td>
</tr>
<tr>
<td>Foot</td>
<td>0.0145</td>
<td>0.5</td>
<td>0.3933</td>
<td>0.6667</td>
</tr>
<tr>
<td>Shank</td>
<td>0.0465</td>
<td>0.433</td>
<td>0.555</td>
<td>0.5</td>
</tr>
<tr>
<td>Thigh</td>
<td>0.1</td>
<td>0.433</td>
<td>0.3533</td>
<td>0.6667</td>
</tr>
<tr>
<td>Head</td>
<td>0.081</td>
<td>0.5</td>
<td>1.11</td>
<td>0</td>
</tr>
<tr>
<td>Trunk</td>
<td>0.497</td>
<td>0.5</td>
<td>1.03</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 4.3: Anthropometrics. Limb masses as percentages of total body mass. COM positions as percentages of limb length. When a segment is composed of multiple joints, such as the trunk and the hands, we divide the mass between the joints. Parameters for linear regression curves for estimating segment density from body density. Adapted from Winter (2009).
Figure 4.1: Segment density versus body density. Given the body density $X$, we estimate the corresponding limb density $Y$, where $Y = MX + B$. These lines are based on the slope $M$ and intercept $B$ terms from Table 4.3. Adapted from Winter (2009).
Figure 4.2: Volumes and centers of mass. We use public anthropometrics data to estimate limb masses and moments of inertia. The red volumes display the relative sizes of masses for each joint. The blue dots display the center of mass.

Figure 4.2.

4.2 Metrics: Energy efficiency

Our metrics are taken from de novo motion synthesis research, which create character animations from scratch. This research has repeatedly shown that low energy solutions tend to look natural. Our hypothesis is that these low energy motions also correlate with neutral, non-stylistic motions. Specifically, objective functions have investigated the results of minimizing torque (Popović & Witkin, 1999; Fang & Pollard, 2003; Wang et al., 2009, 2012), weighted accelerations (Fang & Pollard, 2003), metabolic energy
(Anderson & Pandy, 2001; Wang et al., 2009), or work (Tan et al., 2011).

In the following sections, we describe how our examples compare for different measurements of energy. Most of the distributions in the following sections are not usually normally distributed – some are bimodal, lop-sided, or uniform – nor do they have equal variance. Thus, we use a Welch’s anova test, which does not assume equal variances, to compare the significances of means between walking examples. When significant effects are present, we use a Games-Howell post hoc test to find their reasons.

4.2.1 Squared torques

Minimizing the sum of squared torques is a widely used objective function in animation. For our walking motion set, we compute

\[ \sum_{j>0}^{M} \tau_j^2 \]

where M is the number of joints in our skeleton, excluding the root. Figure 4.3 shows histograms for sum of squared torques (log scale) for each frame over our stylistic walking set. The histogram shows that our neutral examples (based on labels during capture) tend to have lower torques. For our histograms, the X axis represents log sum of torques values and the Y axis are counts. Brighter dots represent greater numbers of samples for the given torque value. We verify that the differences between means is significant (F(59,2861.89) = 4821.97, p < 0.0001). A Games-Howell post hoc test confirms that most motions have significantly higher sum of torques than our neutral motions (at the 0.05 significance level). Our lowest scoring walk, having mean is 5.9, is labeled neutral and has non-significant differences with only 3 other motions: 2 are labeled neutral and one is labeled sexy.
Figure 4.3: Sum torques. Histograms of the sum of squared torques (log scale) for each frame. Each row corresponds to a motion. The x-axis contains counts for each frame having the same range of sum of torques. The color of the bin represents the quantity of frames, with the largest bins corresponding to white dots. Neutral motions (bottom) tend to have lower sum of squared torques.
4.2.2 Power and work

Power estimates the work performed by each limb (in Watts, where Watts = Newton-meters * radians/second). Negative power corresponds to muscles absorbing mechanical energy whereas positive power corresponds to muscles generating energy (Winter, 2009). Muscle power is estimated as the product of the net muscle moment and angular velocity

\[ \sum_{j>0} (\tau_j \cdot \omega_j)^2 \]

where \( M \) is the number of joints in our skeleton (not including the root). Power may be positive or negative. Positive power indicates that muscles are generating mechanical energy to induce movement in the limbs. Negative power indicates that muscles are absorbing mechanical energy from the limbs (Winter, 2009).

Histograms of power for our motion set are shown in Figure 4.4. Motions have significant differences in power (\( F(59, 2863.08) = 483.69, p < 0.0001 \)), but we do not see a strong a pattern for our neutral examples as we do for torques. Neutral motions tend to be lower, but our lowest energy motions are old man, sad, and neutral.

Work is the integration of power over time (in Joules, where 1 Joule = 1 Watt/second). In the human body, the muscles are the only source of mechanical energy generation (Winter, 2009). Positive work occurs during concentric contraction, when the joint moment acts in the same direction as movement. Negative work is done during eccentric contraction of the muscles, where the joint moment acts in the opposite direction to the movement. Positive work represents the amount of energy transferred from the muscles to the limbs; whereas negative work indicates a flow of energy from the limbs to the muscles (Winter, 2009).

\[ \sum_{j>0} (\tau_j \cdot \omega_j \Delta t)^2 \]
M is the number of joints in our skeleton. $\Delta t$ is the duration of each frame, e.g 1/framesecond. $\omega_j$ is the angular velocity. Work can be either positive or negative, depending on whether the force and velocity are aligned. Not surprisingly, the results for work are similar to power.

### 4.2.3 Energy

The changes in kinetic and potential energy (also called internal work) has also been used to estimate the amount of work performed by the muscles (Burdett et al., 1983). We compute

$$W_{int} = \Delta E_k - mg \Delta z$$

where $E_k$ is the kinetic energy, computed as follows for each frame

$$\sum_j \frac{1}{2} \left( m_j v_j^2 + \omega_j \cdot I_j \omega_j \right)$$

where $m_j$ and $I_j$ are the mass and moment of inertia for joint $j$, $v_j$ is the linear velocity of the COM of joint $j$, $\omega_j$ is the angular velocity of joint $j$ and $M$ is the number of joints in our skeleton. This metric makes a simplifying assumption that negative and positive work take the same amount of metabolic energy. Nicolas et al. (2006) used this metric to estimate walking motions from skeletons using IK and warping. Figure 4.5 shows histograms of energy for our motion set. Motions have significant differences in energy ($F(59,2864.83) = 473.337, p < 0.0001$), but we do not see a strong a pattern for our neutral examples as we do for torques (although our neutral motions tend to be lower). The lowest energy motions are *old man*, *scared*, and *sad*. 
Figure 4.4: Power. Histograms of the sum of squared power each frame. The largest bins correspond to white dots. Neutral motions tend to have lower power, but the lowest power motions are the ones with the shortest, slowest movements (old man).
Figure 4.5: Energy. Histograms of the changes in energy each frame. The largest bins correspond to white dots. Slowest motions tends to have the smallest scores with this metric.
4.3 Metrics: Naturalness

Previous work in de novo character motion synthesis also define constraints to synthesize more natural-looking movements, such as smoothness (Fang & Pollard, 2003), small angular momentum about COM (Wang et al., 2009), and a stable head (Wang et al., 2009; Wampler & Popović, 2009). Because our examples are from motion capture, are naturalness metrics do not uncover strong differences between our stylistic and neutral examples. In particular, all our input motions had little head movement and small angular momentum about the COM in the vertical direction. Below, we show the results of two smoothness metrics measured in terms of changes in torques and changes in ground reaction forces.

Torques changes

Smooth motions also tend to have small changes in torque. Minimum torque changes have been shown to predict gently curved arm trajectories (Uno et al., 1989). We compute

\[ \frac{1}{N-1} \sum_{i} (\tau_{root}(t_i) - \tau_{root}(t_{i+1}))/dt \]

Figure 4.6 shows histograms for our motion set using these heuristics. Most motions did not have large changes in torques. The largest changes occurred with elated, surprised, and joy. These differences were significantly different from most styles and all our neutral motions (F(59, 2860.55) = 97.06, p < 0.00001).

Contact changes

Adding constraints for smooth contact force changes have been shown to aid synthe-
Figure 4.6: Torque changes. Histograms of the torque force changes for each frame. The largest bins correspond to white dots. Faster motions tend to have higher values.
sizing smooth motions (Fang & Pollard, 2003). We compute

\[ \sum_{i=1}^{m-1} \frac{f_c(t_i) - f_c(t_{i+1})}{d t} \]

where \( f_c \) is the ground contact force, to avoid jerking at foot contacts. Figure 4.7 shows histograms for our motion set using these heuristics. Most motions did not have large contact changes, with the exception of dance, elated, happy, joy, and march. These differences were significantly different from most styles and all our neutral motions (\( F(59, 2860.53) = 312.034, p < 0.00001 \)).

### 4.4 Remarks and summary

In Chapter 3, we saw that pose and dynamics are important factors for emotional body language. In this chapter, we investigate whether similar differences exist between neutral and stylistic motions. Specifically, we show how neutral and stylistic motions compare using several metrics related to energy and naturalness, and find that neutral motions had significantly lower sums of torques than stylistic motions, even considering slow-moving styles such as sadness. We did not find significant differences in our metrics for naturalness, such as with changes in torque, or changes in ground reaction forces.

There are several limitations with this approach. First, because our motions are kinematic, the metrics we compute are approximate and in some cases, may be extremely noisy even after smoothing (too much smoothing can also alter the motion too much). Additionally, our classifications for neutral and stylistic are based on annotations during the original motion capture and thus may not effectively convey that style to all users. Although numerical comparisons are useful, ultimately perception is what determines style.

Future work should check whether this trend is consistent for different stylist motion datasets. Assuming this trend is consistent, we could use this finding to detect neutral
Figure 4.7: Ground reaction force changes. Histograms of the contact forces changes for each frame. The largest bins correspond to white dots.
motions or to infer neutral motion from stylistic ones. Other future work might look into the factors which cause an animal to move sub-optimally. In some cases, the reasons are aesthetic, as in dance, or even socio-cultural, such as in a swagger or geisha’s walk. But in other cases, the style might have a physiological basis which could potentially be simulated, such as with pride, shame, sadness, or fear. For example, Tracy & Matsumoto (2008) found that both sighted and blind people have the same body language in response to winning or defeat, suggesting that the origins of such body language is biological.

In this chapter, we focus on the dynamics between our stylistic walking examples. In the next chapter, we analyze the poses by modeling the variation between walks using PCA. This formulation will be useful for blending, searching, sampling, and browsing our stylistic motion set.
Chapter 5

Modeling stylistic variation with PCA

In this chapter, we describe how we model the variations between the stylistic walks using PCA. Our approach differs slightly from typical PCA modeling approaches in computer animation in that we first compute the average motion across all our examples and then use this average to decouple each input into a stylistic and non-sylistic, or offset, part. We then only transform the offset part in PCA space (Figure 5.1). In the resulting space, the average motion is at coordinate zero, with similar motions clustered together into three primary groups around this average. The resulting PCA space can be used to robustly blend between examples, sample new randomized styles, and quickly search the space to avoid degenerate motions. Because we can reconstruct the root positions from foot contacts, we also avoid problems with foot skate.

The use of PCA for working with animation data is a very popular and widely used tool in animation. For example, Alexa & Mueller (2000) used eigenposes for animating shapes and Shin & Lee (2006) allowed users to modify and sketch eigenpose curves to edit motions. Alternatively, Urtasun et al. (2004a) used eigenmotions to retarget gaits from video to 3D walking animations; Tilmanne & Dutoit (2010) combined a PCA model with a Gaussian distribution to produce within-gaits variation; and Min et al. (2010) used eigenmotions to model the space of poses for their statistical model. As mentioned in Chapter 2,
other decomposition, transformation, and modeling techniques might also been used, such as Fourier decomposition (Unuma et al., 1995), Laplacian band filtering (Bruderlin & Williams, 1995), velocity curve transfer (Amaya et al., 1996), independent component analysis (Shapiro et al., 2006), principal geodesics analysis (Tournier et al., 2009), and gaussian processes (Grochow et al., 2004; Ikemoto et al., 2009; Wei et al., 2011). Our choice of PCA model using eigenmotions allows us to model the variations between each input motion in a straightforward way. By modeling the stylistic offsets from the average motion, the resulting PCA space simultaneously clusters similar motions together and supports smooth blending between arbitrary subsets of PCA coordinates. Using PCA for clustering was also used by Barbič et al. (2004) to segment a motion set into different behaviors.

5.1 From motion space to PCA space

To ensure our PCA space generates consistently good motions, we first compute an average motion across all input examples and then decouple each motion into an average and stylistic, or offset part. We then transform the offset motions to PCA space in the standard way, treating each motion offset as a N-tuple vector. Figure 5.1 summarizes this process.

5.1.1 Motion representation

Our input motion set consists of 55 stylistic, looping walking motions (the same ones introduced in Chapter 4, now preprocessed to seamlessly loop). Each motion consists of four contact phases: a right foot contact, double support, left foot contact, double support. We reduce each input motion into a set of 17 keyframes, 4 for each contact phase with a repeat of the first key to maintain a smooth looping motion. For each motion, the contact keys are taken at the same proportion of the phase (namely
Figure 5.1: Process for converting input motions to PCA coordinates. We first compute the average motion across all input samples and then use that average to decompose each input motion into an average part and a stylistic, or offset, part. We then only model the motion offsets in PCA space.

at 0%, 25%, 50%, and 75%) so that poses are time-aligned. Each key is represented as a tuple (t,q) where t is the time in seconds since the last keyframe and q is a pose consisting of joint rotations in exponential map coordinates (although quaternions also work, the exponential map only requires 3 components rather than 4). The frame time is computed as the time relative to the previous frame, with the first keyframe having $t=0$. Because PCA is sensitive to relative scales between variables, storing the time in seconds is important because it ensures that values have appropriate magnitude to our rotations. This approach to modeling time supports varying walking frequencies and asymmetric gaits, such as limps.

Each input motion $M_i$ is converted into a single row vector of keys with size $\text{NumKeys} \times (1 + \text{NumJoints} \times 3) = 1853$. Motions are reconstructed with spherical interpolation using splines according to Shoemake (1985). After preprocessing (Section 5.1.2), the final motion matrix passed to PCA contains the entire motion set with each row corresponding to an example and each column representing a keyframe feature.
5.1.2 Preprocessing

To avoid artifacts when working with the PCA space (Figure 5.2), we propose computing the average motion across the motion set and then decomposing each input motion into an average and stylistic part, so that only the stylistic offsets from the average motion are modeled in PCA space.

Figure 5.2: Rotational artifacts which result if we do not preprocess the motion set to only model offsets.

To compute the mean motion, we compute average keys by averaging each time offset and averaging each keyframe pose (recall that each motion has 17 keyframes taken at corresponding moments at each contact phase). Computing an average motion consists of computing average rotations for every joint in the pose. Special care must
be taken when averaging these rotations (a heuristic approach based on checking dot products between quaternions does not work with a large number of rotations – it is often impossible to guarantee that all dot products will have the same sign). We use the technique based on SVD decomposition described in Curtis et al. (1993), which provides an analytic solution for the rotation which minimizes the distance to all other rotations. Specifically, the average rotation matrix $R_{\text{ave}}$ is given by

$$R_{\text{ave}} = UV^T$$

where $U \Sigma V^T = \sum R_i$. Because we work with the exponential map and quaternions, this operation converts to rotation matrices, adds them, performs the SVD decomposition to find $U$ and $V$, and then converts the final result back to the exponential map. An alternate approach based on quaternions is given in Johnson (1995).

Once computed, the poses of the mean motion $\bar{M}$ are subtracted from each input motion $M_i$ to compute an offset motion $M_{\text{off}}^i$. Subtraction is performed by finding the rotational distance between joint angles.

$$q_{i_{\text{off}}} = q_i^{-1} \bar{q}_i$$

When computing offsets, care should be made to compute the smaller of the two angles between orientations, which can be done in the typical way by ensuring that the dot product between $q_i$ and the corresponding average angle $\bar{q}_i$ have the same sign.

Modeling the offset motions in PCA space, rather than the original motions $M_i$, avoids rotation artifacts when exploring the space. When we try modeling the input set directly, the distributions of some of the joint components are bimodal – either due to large angle changes between keys or due to angle representations being non-unique (e.g. angle changes of either -10 or 350 degrees represent the same rotation). Modeling the offsets improves the PCA model because it results in approximately normal joint angle
distributions and smaller joint magnitudes. Arbitrary blending between components is very robust as a result. In Figure 5.2, we show how the left wrist and both ankles display unnatural twists when we do not decouple motions first. The corresponding problematic joint distributions are shown in Figure 5.3 both before and after decoupling the motion set.

Figure 5.3: Joint distributions for problematic exponential map components. Left, exponential map component distributions from the original motions. Right, exponential map component distributions from the offset motions. Modeling the stylistic offsets in PCA space avoids rotational artifacts. The exponential map distributions (joint angle components) of the offset motions are centered and have smaller magnitudes than modeling the original joint angles.
5.1.3 PCA transformation

Given a set of offset motions $M_i^{off}$, we transform the dataset to PCA space in the usual way. We treat each offset motion as a 1853 feature tuple to compute the mean $\bar{M}^{off} = \frac{1}{N} \sum M_i^{off}$ and covariance matrix before performing the eigenvalue decomposition. We then keep the largest $K$ components so we can maintain 99% of the variance (requires 38 components for our example set).

The reconstructed motions in PCA space are thus

$$M_i^a \approx \bar{M}^{off} + \sum_{i=1}^{K} \alpha_{ik} e_k$$

where $\alpha_{ik}$ is the $k$th coordinate of the $i$th PCA component vector $\alpha_i$ which corresponds to $M_i$ in PCA space. $M_i^a$ is an offset motion, so complete reconstruction requires adding back the motion mean and reconstructing the root position.

5.2 From PCA space to motion space

To return to motion space from PCA space, we simply reverse the construction process: we transform our PCA coordinate to an offset motion, add back the average, and then reconstruct the root positions based on foot contacts. Figure 5.4 summarizes this process.

Specifically, given a PCA coordinate $\alpha_i$, we first add the PCA mean $\bar{M}^{off}$ as a tuple. Then we add back the average motion $\bar{M}$ rotationally like so

$$q_i = \bar{q}_i (q_i^{off})^{-1}$$

Now that we have our poses for $M_i$, we recompute the full motion using quaternion spherical interpolation (Shoemake (1985)) based on the reconstructed time offsets. The root position is computed from foot contacts using a treadmill approach, e.g. the velocity
Figure 5.4: Right, process for converting from PCA coordinates to a motion. We convert from PCA coordinates to an offset motion and then add back to the average motion. We model the timing in PCA space, but not the root position, so to complete the reconstruction, we use a treadmill approach which sets the root positions based on foot contact velocity.

of the root is the inverse of the velocity of the feet in contact. IK is used to prevent foot sliding in the middle of a contact phase.

5.3 Editing in PCA space

The PCA space supports blending arbitrary coordinates together, sampling to generate randomized styles, exaggerating styles through extrapolation, and efficient search for avoiding motions with self-collisions and bad joint angles.

5.3.1 Blending

PCA coordinates can be linearly combined together to blend motions.

$$\alpha_m = t \cdot \alpha_a + (1 - t) \cdot \alpha_b$$

where $t \in [0, 1]$ and $\alpha_a$ and $\alpha_b$ are the styles to blend. The resulting blends are nearly identical to blending the corresponding motion examples in motion space.
5.3.2 Dampening and exaggeration

Because every motion is the offset from the average, we can exaggerate a style by scaling the PCA coordinate.

\[ \alpha_m = t \alpha_a \]

where \( t \) is a positive scalar and \( \alpha_a \) is the style to exaggerate. When \( t \) is less than 1.0, the style is dampened. When \( t \) is greater than 1, the style is exaggerated. This process is equivalent to blending the current style with the average motion, similarly to the neutral blend conditions described in Chapter 3. Scaling the coordinate can be thought as changing the distance from the average in the direction of the style \( \alpha_a \).

5.3.3 Sampling

Random motions can be created by sampling from a multivariate distribution corresponding the PCA space, \( p(\alpha) = N(\mu, \Sigma) \) where \( \mu \) is a vector of zeros and \( \Sigma \) is a diagonal matrix of eigenvalues \( \sigma \) corresponding to our PCA space. To generate a random PCA component \( \alpha_{ik} \), we sample each coordinate from a normal distribution \( N(0, \sigma_k) \) where \( \sigma_k \) is the \( k \)th eigenvalue corresponding to the \( k \)th eigenmotion.

In Figure 5.5, we show several randomly generated motion examples which share the same first two PCA components. Randomized examples are shown in thumbnail animations on the right. Each map point represents a family of potentially valid motions.

5.3.4 Avoiding degenerate motions

Unconstrained blends and random sampling easily produce artifacts in the resulting motion, particularly self-collisions and violated joint rotations. Many techniques, such as optimization and direct post-processing using IK, could be used to directly repair these errors. Figure 5.6 shows the effect of detecting and repairing bad joint angles. Figure 5.7 shows the effect of detecting and repairing self-intersections (based on...
Figure 5.5: A single map point represents a large family of motions. Above, random motions generated from the same map point. The user can generate 5 random examples at a time. Clicking on a thumbnail animation (from the right) will preview the motion on the center character.

Bounding spheres. In these examples, we use Covariance Matrix Adaptation Evolution Strategy (CMA-ES, Hansen (2005)) to search for nearby PCA coordinates which do not violate constraints, based on the implementation by Hansen (2015). So that our routine returns more quickly, we give it a maximum number of iterations to search. We initialize the algorithm’s search space using the current, broken PCA coordinate as the starting point and the initial search standard deviations as the eigenvalues corresponding to our PCA basis. Each potential new example is scored according to

$$\min \sum_{i} w_i (\alpha_{i}^{\text{current}} - \alpha_{i}^{\text{x}}) + a (\sum \text{penetration distances} + \sum \text{joint limit distances})$$
where the first term is a weighted sum of eigencomponents such that components near
to the current component are scored higher. The second term represents our constraints.
If a constraint is satisfied, the values are zero. Otherwise, the values are the sum of
penetration distances and sum of joint violation distances. The weight $w_i$ are set to
the corresponding eigenvalues if that value is greater than 1.0 and 1.0 otherwise. The
constant term, $a$, weights constraint violations to be more important than staying close
to the starting point (set to 10000 in our experiments).
5.4 Browsing styles in PCA space

The first two PCA components, visualized on a 2D plot, creates a minimap of the motion set. On this map, the average motion is at coordinate (0,0) and similar motions are clustered together into three primary groups. Figure 5.8 displays the minimap and highlights each primary cluster: crouching motions, upright walks, and wide-stance walks with bent elbows. Blue dots represent the input motion examples. This map is equivalent to performing a classical multidimensional scaling of the motion examples using Euclidean distance (The Mahalanobis distance could also be used – the map looks similar for this dataset).

![Figure 5.8: The map of the first two PCA components reveals clusters in the data: crouching motions, upright walks, and wide-stance walks with bent arms.](image)

The minimap organizes the large input set of potential motions to blend spatially. However, the map is only 2 dimensional and doesn’t convey the full variability of potential motions that are available at that coordinate (recall, we use 38 PCA components). For example, Figure 5.9 shows several examples corresponding to the same map point. Thus, each map point represents a family of potential motions and several heuristics for choosing the remaining components could be used for exploring the space. In the last
section, we discuss randomly generating PCA components, but we also looked at the following three heuristics.

Figure 5.9: A single map point represents a large family of motions. Above, different motions created using different browsing heuristics. The red dot indicates the current map position. The first two map components correspond to first 2 PCA coordinates. The zero heuristic sets all remaining coordinates to zero. The closest heuristic sets remaining components to the closest input example, which in this case is a limp. The weights heuristic sets the remaining components to a weighted sum of nearby input motions.

**Closest heuristic**

This heuristic selects the remaining components based on the closest example point. To avoid snapping as the closest example changes, we blend the closest components
closest with the user’s current map coordinate \( \alpha_m \).

\[
\alpha_m = a\alpha_{\text{closest}} + (1 - a)\alpha_m
\]

where \( a \) is a scalar blending weight between 0 and 1. This heuristic allows the user to see each input example as they move around the map.

**Zero heuristic**

In this simple heuristic, we set the remaining PCA components to zero. This choice equates to choosing the average values for the remaining components.

**Weights heuristic**

This heuristic is similar to the method proposed by Urtasun et al. (2004b) to generate motions from new examples, where weights are inversely proportional the distance. As the user drags the current map point \( \alpha_m \), we compute the 2D distances (e.g. based only on the first 2 PCA components) between the user’s point and each example \( \alpha_i \). The weights are then the normalized inverse of these distances.

\[
w_i = \frac{\text{Distance}_2D(\alpha_m, \alpha_i)^{-1}}{\sum_j^N \text{Distance}_2D(\alpha_m, \alpha_j)^{-1}}
\]

Once weights are calculated, we compute a full PCA coordinate as a weighted sum over all the examples. \( \alpha_m = \sum w_i \alpha_i \). Because this calculation takes into account every input motion, the results end up looking very similar to the zero heuristic. To address this, the number of closest motions used for blending should be limited.
5.5 Remarks and summary

In this chapter, we construct a PCA model for the stylistic motion set introduced in Chapter 4. In PCA space, we can robustly blend, browse styles, sample randomized styles, and search the space to avoid invalid motions. From an analysis perspective, the model gives insight into how the postures of each motion cluster together, which we can easily see from visualizing the first 2 PCA components on a minimap. Although our use of PCA for motion modeling is not novel, our decision to decouple motions and only model the motion offset is a simple enhancement that makes the use of the PCA model produce consistently good output. In the next chapter, we investigate whether using a 2-dimensional minimap consisting of the first 2 PCA components helps users browse and combine styles.
Chapter 6

Browsing styles with a minimap

In this chapter, we investigate whether a minimap which clusters similar motions together helps users browse styles. As motion sets grow in size, they become increasingly difficult to browse. Many motion databases rely on queries to return subsets of motions to the user, but for stylistic motions, strictly query-based systems are difficult: the user must already have an idea of what is available in the motion set and the labels need to match the user’s expectations. For quantitative labels, such as walking speed or step length, labels are clear. However, for qualitative labels, such as 'happy', 'jaunty', or 'energetic', labels can be misleading. For example, the same walk might be labeled 'happy', 'jaunty', or 'energetic', depending on the annotator.

Given examples from a motion category, it is unclear how best to present the motion set to users so that they can browse the available variations effectively. Inspired by minimaps frequently used in games, we investigate organizing motions with a minimap based on the first 2 PCA components, as described in Chapter 5. With the map, users can interactively browse the styles which are immediately displayed on the character. Users can also annotate the map with bookmarks, which can either be used to remember the locations of motions on the map, or to be singled out for blending together to create new styles. A screenshot of this interface is shown in Figure 6.1.
We demonstrate the minimap idea using the same motion set from Chapter 5 consisting of 55 stylistic, looping walks. To quantify the strengths and weaknesses of the minimap, we design two user studies which compare our map interface to a traditional label-based interface. In both studies, users perform two tasks: the first task asks users to find a motion with a given semantic quality, such as happy; the second task asks users to reproduce a walking animation. In the first study, called our Untimed experiment, we asked participants to perform each task using both interfaces with ample time. In the second study, called the Timed experiment, we asked participants to perform each task with a strict, short time limit (45 seconds). Most of our participants were non-expert users, rating their experience with character animation as lower than 3 on a 7-point scale. We found that for the Untimed experiment, users were equally satisfied with both the map and label interfaces, but were able to create and find motions using significantly fewer mouse clicks and exhibiting more creative results using the map. For the Timed experiment, both interfaces were equally poor at helping people find motions with a given semantic quality; however, the label-based interface was better for finding a matching motion example. Overall, most participants greatly preferred the map (14 out of 17 participants, 82%), enjoying the ability to quickly scrub through the entire motion set.

6.1 Motion search and browsing

Finding motions from large motion datasets typically relies on the user to make queries. Queries might be semantic, based on user demonstration, based on finding similar examples or based on motion attributes. For example, semantic queries could consist simply of searching for motion annotations, as with the CMU mocap database (CMU, 2015), or could parse natural language to try to automatically stitch motions together (Arikan et al., 2003; Min & Chai, 2012). Demonstration-based queries include sketch-
Figure 6.1: Screenshot showing the interactive minimap with two bookmarks, labeled 1 and 2. Blue dots represent the original motion examples. Green dots represent bookmarks. The red dot represents the current motion shown on the character. We implement this map using a principal component analysis (PCA) transformation of the original stylistic motion set, which automatically clusters similar motions together and transitions smoothly between styles.
based interfaces, where a user’s drawings are converted into queries into a motion set (Thorne et al., 2004; Choi et al., 2012) and performance-based interfaces, where a user demonstrates the motion they want such as with a puppet (Numaguchi et al., 2011) or gestures (Chai & Hodgins, 2005). Example and attribute-based queries try to find similar sets of motions. For example, given a motion, the algorithm might search for others that have similar joint rotations and marker positions (Kovar & Gleicher, 2004; Meng et al., 2008; Liu et al., 2005; Krüger et al., 2010), similar motion encodings (Müller & Röder, 2006), or similar weighted PCA-space coordinates (Forbes & Fiume, 2005). In an attribute-based query, a user might search for high-level motion characteristics, such as body contact, energy, balance, and shape (Kapadia et al., 2013). The above techniques can be very efficient, but finding good motions relies on the user to formulate good queries, and it is unclear how to best organize and display the results of a query to a user. Our PCA-based minimap investigates one way in which a query resulting in many stylistic possibilities might be presented to the user.

Less research has explored how to help users browse a set of motions when they do not have a clear idea of what they are looking for. Sakamoto et al. (2004) displayed groups of motions side by side, each on their own character, as a motion map. Yasumuro et al. (2008) organized returned queries in a tree structure. Choi et al. (2012) created stickfigure images to summarize each motion (and later use for querying). Other works have focused solely on summarizing an entire motion with a single image (Assa et al., 2005; Lee et al., 2012). To our knowledge, no research has compared motion display methods to determine how they affect users’ ability to quickly browse. Although we do not investigate multiple display strategies, we do compare two display methods: our map versus a text-based list.

Browsing our PCA model for styles can also be posed as a parameter selection problem. In other words, how should we choose blend weights (our parameters) so that the user can find a suitable motion? In a seminal work, Marks et al. (1997) introduced
an interface called a “design gallery” which summarized potential parameter settings using both a high-level 2D map of the parameter space along with thumbnails showing the effect of different choices. In our interface, the sample animations in the map can be interactively explored and explicitly and immediately blended.

### 6.2 Interactive minimap

Our minimap gives users an interactive, visual way to explore a motion set that does not rely on semantic labels. The user gains a sense of the types of motions available by dragging the cursor over the minimap. The user can bookmark interesting motions to annotate the map or to set aside for editing and blending.

The implementation of our minimap is the PCA model based on motion offsets (Chapter 5). Furthermore, with careful preprocessing of the motion set (Section 5.1.2), the PCA space robustly supports direct blending of PCA coordinates, which allows a user to smoothly transition between examples when exploring.

Recall from Chapter 5 that the minimap displays the entire motion set such that similar motions are clustered together (Figure 5.8) Blue dots represent the input motion examples. The red dot represents the location of the current motion shown on the character. This map is equivalent to performing a classical multidimensional scaling of the motion examples using Euclidean distance and reveals three primary walking clusters: crouching motions, upright walks, and wide-stance walks with bent elbows. Because our primary goal is to support browsing and blending, we use the closest heuristic (Section 5.4) for browsing.

\[
    \alpha_m = a\alpha_{\text{closest}} + (1 - a)\alpha_m
\]

where \( a \) is a blending scalar between 0 and 1 and \( \alpha_m \) is the current map coordinate. This heuristic smoothly transitions between the original styles while browsing.
6.2.1 Bookmarking and blending

Free exploration using the minimap is good for giving an overview of the types of styles available in the motion space. However, it does not provide landmarks to give the different areas of the map context. We give users this functionality through bookmarking. Each bookmark is displayed with an animated thumbnail as shown in the screenshot in Figure 6.1. Additionally, these bookmarks can also be blended together to create new styles. In Figure 6.2, we show how two bookmarks can be used to linearly interpolate between two PCA coordinates; and three or more bookmarks can be used to blend using generalized barycentric coordinates. The blend weights are computed using only the 2D map coordinates, but the resulting motion is computed using the PCA coordinates, e.g. $\sum w_i \alpha_i$.

![Figure 6.2: Using bookmarks to blend between specific motions. From left to right, linear blend, trilinear blend, and a 5-way blend.](image)

6.3 Evaluation

We perform two user experiments to evaluate whether the map helps users find stylistic walks more quickly and whether the map helps users find motions which they might not find otherwise.

Our first experiment, called the Untimed experiment, tests users’ ability to perform tasks with varying difficulty. Our second experiment, called the Timed experiment, tests users’ ability to find motions quickly. Both experiments ask participants to perform two
different tasks with each interface: a description task, where users must find a motion having a requested quality, such as happy; and an example task, where users must find a matching walk for a given example.

For each experiment, we compared the map interface against a straightforward label-based interface (Figure 6.3). The goal of this comparison was to compare the strengths of a text-based interface to one that organized motions spatially. Thus, we deliberately had minimal labels for the map interface. The Untimed experiment shows the motion label on mouse over, and the Timed experiment showed no labels at all, only numbers for each bookmark. Note that the labels for each motion were taken from the original motion capture session and may not always be a good description of the animation. We also considered using a professional tool, such as Motion Builder, for comparison, but decided against it to both better support novice participants as well as keep our UI comparison simple.

In both experiments, participants had no prior experience with the interface or the motion set. Some participants also had no experience with computer graphics. Both UIs had capabilities for scrolling through motions, pausing and scrubbing through the animation, and moving the camera. Both experiments used the same motion set (55 stylistic walks plus the average), a large set for participants to work with over a short period. The interfaces used in both experiments had identical blending capabilities, e.g. any blend that can be found with the blend bar could be found with the map. However, the map allowed extrapolations outside the convex hull of the motion set.

In both experiments, we ask participants how satisfied they were with the result and how easily they found the motion. We also recorded the amount of time they spent on each question, the number of mouse clicks, and the animation that they created or found. In the following sections, we answer

- Are users more satisfied with their motions using either interface?
- Are motions easier to find with either interface?
• Which interface is easier to work with?

• Are there differences in the types of motions people find with each interface?

• Do users find motions more quickly with either interface?

We choose blending as an editing task not just because the PCA space readily supports it but also because this task quickly becomes very difficult when the set of potential blend motions becomes large. To create a blend, the user must continually selects motions for blending and then determines desired weights manually through trial and error. Often, it is not intuitive how arbitrary combinations of input motions will look when they are blended together.

6.3.1 Experiment 1: Untimed tasks

In this experiment, participants had ample time (5 minutes) to search and create motions for each task. After 5 minutes, a dialog would appear to ask them to finish and then grant them an additional minute. Participants were told of the time limit to start and shown the time remaining. The map interface consisted of the interactive minimap and bookmarking interface described in Section 6.2. The label interface, which we will call the Blend Bar interface, consisted of a simple list of labels and associated blend weights for each (Figure 6.3).

For the Description Task, we ask participants to find or create a walk with a specific quality: happy, non-stylistic, confident, silly, frightened, and angry. The first two questions (happy, non-stylistic) directly corresponded to several examples in the motion set, although there may be more than one possibility and the database labels may not match exactly. The second two (confident, silly) are subjective, but readily found. The final two (frightened, angry) don’t readily exist in the database, but might be obtained with blending.
For the Example Task, we ask participants to reproduce a motion example. Three of the example motion examples existed in the motion set (medium difficulty task), and three motion examples consisted of blends of two motions (hard difficulty task). For each question, participants were shown the example in the top left corner of the experiment interface.

We used a within subject design where participants use both the blend and map interfaces to perform each task. Both the order in which participants use each interface and the ordering of the questions were randomized.

Method

We recruited 9 volunteers for this experiment: 7 male and 2 female aged 17 to 27. On average, participants rated their level of experience in character animations as 2.67 out of 7.

Participants were first given a tutorial to practice with the controls and gain familiarity with the motion set. They were then asked for their age, gender, and to rate their level of experience with character animation on a 7 point Likert scale.
The description task (12 questions, 6 for each interface) was completed first. We asked participants “to find stylistic walks using the given interface.” After each question, we asked “How well does the walk convey X” and “How easy was it to find a walk with quality X”, where X was a given characteristic, such as silly. Participants could enter comments if desired and then press a button to continue when ready.

The example task (12 questions, 6 for each interface) was performed second. Participants were asked “to replicate examples of stylistic walks using the given interface. Some motions will exist ’as is’ from the motion set whereas others will require combining two styles.” After each question, we asked “How satisfied are you with how well your walk matches the example?” and “How easy was it to find a match?” on a 7 point Likert scale. Again, participants could enter comments optionally and click a button when they were ready to continue.

After completing both tasks for each interface, we asked participants which interface they liked overall; what they liked and disliked about the map; what they liked and disliked about the blend bar; and if they could combine elements of both, what would they do. This study took an hour on average to complete.

Results

From exit surveys, 8 out of 9 participants preferred the map. Participants liked that the map was [sic] “very visual”, “easier to determine blend ratio”, “clumps by similarity makes it easy to go through variations”, “easier to find unlisted animations”, “easy to try”, “previews of walks [e.g bookmarks] so I don’t forget”. Criticisms included “takes a little bit of time to get familiar with the mood of a region”, “less precise than the blend bar”, and “can’t see a point name directly”. Many participants requested more labels for the map and adding controls so that blend weights could be set directly. Many people wished they could find a motion with the map and then refine with the blend bar.

We perform a 2-way repeated measures ANOVA with factors UI type and Question to
analyze satisfaction, easy to find, durations, and mouse clicks. UI type was either the map interface (M) or the blend bar interface (B). When significant effects were present, we used Tukey’s honestly significant difference procedure to determine their reasons.

**Satisfaction**

There were no significant differences in UI type for satisfaction for either task. However, there was a main effect of Question for both tasks (example: $F(5,96) = 3.48, p < 0.01$; description: $F(5,96)=4.91, p < 0.001$). For the description task, participants were significantly more satisfied with their responses for happy and confident over frightened and angry. For the example task, participants were significantly more satisfied that they found the chicken motion than they were for one of the blends and gangly teen. Overall, these scores are consistent with our intended difficulty levels, but we cannot say that users are more satisfied with the motions they create with either interface.

**Easy to find**

There is no significant differences in UI type for “easy to find” for either task. However, there was a main effect of Question for both tasks (example: $F(5,96) = 5.02, p < 0.001$; description: $F(5,96)=8.13, p < 0.001$). For the description task, they found happy, normal-looking, and confident significantly easier to find than frightened and angry. For the example task, they found chicken significantly easier to find than every other motion example except for sexy. These results are consistent with our intended difficulty levels, but we do not find that either interface makes some motions easier to find.
Figure 6.4: Left, Durations. Participants spent significantly more time creating an angry motion than a happy or confident motion. Error bars represent standard errors. Right, Mouse clicks. The map requires significantly fewer mouse clicks than the blend interface.

**Duration**

On average, participants spent 94.6 seconds on each description task and 89.6 seconds on each example task. There was no significant differences in UI type for duration for either task. However, there was a main effect of Question for the description task (F(5,96)=3.22, p < 0.05, Figure 6.4). Participants spent significantly more time creating an angry motion than a happy or confident motion. These findings are again consistent with our intended difficulty levels.

**Number of mouse clicks**

The map interface required significantly fewer mouse clicks than the blend interface for both tasks (example: F(1,96) = 4.39, p < 0.05; description: F(1,96) = 5.05, p < 0.05, Figure 6.4). The map was designed to require fewer mouse clicks, since we intended users to search and blend primarily by dragging the mouse. This result verifies this design choice and adds some support, although not definitive, that the map may be easeir to use.

**Accuracy of matched motions**
For the example task, we measure how closely participants’ animations match the example, based on distances between joint angles, i.e.,

\[
\text{Distance}(q_1, q_2) = \cos(2 \times \text{Dot}(q_1, q_2)^2 - 1)
\]

where \(q_1, q_2\) are the joint quaternions to be compared. The accuracy score of a participant’s response is the sum of angle distances for each joint for each keyframe. There was also no effect of UI type in how well participants were able to match the examples, but there was a main effect of Question (\(F(5,96) = 2.8, p < 0.05\)) with the chicken motion example being significantly easier than every other motion except sexy. These results are consistent with the ratings for “easy to find”.

**Analysis of created motions**

For the description task, we measure the similarity of the animations created by each participant, measured as the distance to the average animation over all participants. The average is computed as in section 5.1.2 and the similarity score is the sum of joint angle differences, as above.

There is a main effect of UI type (\(F(5,96)=41.86, p < 0.001\)), question (\(F(5, 96)=6.66, p < 0.001\)), and interaction effects (\(F(5,96)=2.47, p < 0.05\)) when analyzing similarity. The motions created with the blend interface tended to be more similar than those created with the map (Figure 6.5). By question, the most extreme example occurs with silly which had significantly more variation than happy, normal-looking, confident, and angry. The most homogeneous example was for the normal-looking style which was significantly different from silly and frightened.

There are several possible explanations for this result. The blend interface’s labels sometimes influenced people’s choices. Additionally, the simplest way to create a style
Figure 6.5: Analysis of dissimilarity of created motions between UI type and Question. The responses for the map display much more variety. Several differences are significant: The silly motions created with the map are significantly more varied than every other category except for frightened motions created with the map. Not surprisingly, the normal motions show the least variety for both the map and the blend interface. The frightened motions created with the map are also significantly more varied than the happy, normal, and angry motions created with the blend interface (highlighted in the diagram).

was to set a single blend weight to one whereas the map encouraged approximate choices and blending. An other factor is that the map allowed some extrapolation, which users might choose for their silly motion (1 participant chose an extreme backwards lean using the map). Extreme and inexact choices both increased the spread using the map. However, visual inspection of the created motions support that people tended to be more creative using the map as well. Overall, this result supports our intent that the map encourages exploration and aids users in finding and selecting motions which they might not readily find just with labels.
6.3.2 Experiment 2: Timed tasks

In this experiment, participants had a strict time limit (45 seconds) to search for a motion. If the participant ran out of time, the experiment would automatically progress to the next screen. Participants were aware of the time limit and shown the time remaining during the experiment. The map interface consisted of the interactive minimap and bookmarking interface described in Section 6.2, but modified to snap to map points by default. For comparison, we created a simple text-based interface consisting of a list of buttons which applied the corresponding style when clicked.

Unlike the Untimed experiment, the tasks for the Timed experiment are intended to be straight-forward search and find tasks (no blends). For the description task, we ask participants to find a walk with qualities that are readily found in the dataset: sneaky, non-stylistic (Normal), confident, silly, tired, and cautious. The first two questions (sneaky, non-stylistic) have numerous examples in the input set, so users need to find their subjective best one.

For the example task, we ask participants to search and find a motion which exists in the motion set. For each question, participants were shown the example in the top left corner. To make the example task easier, we added a feature to re-align the walk cycle of the example motion with the user’s motion whenever the selection changed. As before, we use a within subject design where participants use both the blend and map interfaces to perform each task, with both UI type and questions randomized.

Method

We recruited 8 volunteers to participate: 6 male and 2 female aged 16 to 39. On average, participants rated their level of experience in character animations as 2.13 out of 7.

Participants again were allowed to gain familiarity with the interfaces and motion set before starting the experiment. They were then asked to provide demographic
information and their experience with character animation.

The description task (12 questions, 6 for each interface) was performed first. Participants were asked “to find stylistic walks using the given interface. Select the best walk which you believe conveys the desired stylistic characteristic.” After each question, we asked “How satisfied are you that you found the best walk for X?” and “How easy was it to find a walk for X?” on a 7 point Likert scale, where X was a given characteristic, such as silly. A value of 1 indicated that a user could not find a style. Participants could enter comments if desired and then press a button to continue when ready.

The example task (12 questions, 6 for each interface) was performed second. Participants were asked “to find matching stylistic walks using the given interface. Each example already exists in the motion set.” After each question, we asked “How satisfied are you that your walk matches the example?” and “How easy was it to find a match?” on a 7 point Likert scale. A value of 1 indicated that a user could not find a style. Again, participants could enter comments optionally and click a button when they were ready to continue.

After completing both tasks for each interface, we again asked participants which interface they liked overall; what they liked and disliked about the map; what they liked and disliked about the blend bar; and if they could combine elements of both, what would they do. This study took a half hour on average to complete.

Results

From exit surveys, 6 out of 8 participants preferred the map. Positive comments for the map included “much nicer sorting of motions”, “I really liked a lot that I didn’t have to read through everything”, “bookmarking was helpful”, “I felt like I could blend more easily”, “freedom of choice”, and “very easy to scrub through motions”. Common negative comments were “there isn’t a way of knowing a priori what the motions are.”, “hard to determine where to start”, and “I couldn’t tell where I was”. The most common
request was to include more labels for the map, either with axis labels, cluster labels, or mouse over labels.

We perform a 2-way repeated measures ANOVA with factors UI type and Question to analyze satisfaction, ease of search, duration, and mouse clicks. UI type was either the map interface (M) or select interface (S). When significant effects were present, we used Tukey’s honestly significant difference procedure to determine their reasons.

There were no significant results for the description task. People had an equally hard time with both UI types. Based on comments from participants, they often ran out of time trying to see if they could improve on their current motion.

For the example task, people were able to find motions more easily and quickly with the select interface. There were main effects of UI type and Question at the 0.05 significance level and a nearly significant interaction effect ($p = 0.053$). Participants found motions significantly easier with the select interface ($F(1,84) = 5.58, p < 0.05$) than the map interface. In particular, two questions were much harder to find with the map than with the select interface, likely because the motion was hidden in a crowded map area.

Correspondingly, participants took significantly less time to find a motion using the select interface. There were main effects of UI type ($F(1,84) = 4.19, p < 0.05$) and Question ($F(5,84) = 3.16, p < 0.05$), with one question being significantly harder than the two easiest questions.

Thus, we conclude that the label-based interface helps participants find motions more quickly, but it relies on the labels being properly descriptive (or learned by the user). Although people seem to like the smooth browsing, it takes time to preview the available motions. This result might be different with people who have time to learn where motions are located on the map; however, in these experiments, we deliberately worked with people having no experience with the motion set.
6.4 Remarks and summary

This chapter demonstrates the potential for a map based interface to aid novice users while performing character animation tasks. Although PCA coordinates are not typically intuitive, our participants had no trouble navigating the space of motions due to the fact that changes could immediately be seen on the character. The PCA space simultaneously clusters similar motions together and supports smoothly blending between PCA coordinates. Our user studies demonstrate that novice users easily understand the map and can use it to create new styles from the input motion set. Based on the Untimed experiment, the map provides a compact representation of the walking set which performed equally well as a simple, label-based interface in terms of satisfaction and ease of search. However, the map requires less “clicking around” and encourages participants to explore and find motions that they might not easily find otherwise. Although we see that neither interface performed well for the description task when there was a strict time limit, we did see that the label interface allowed users to find matching motions more quickly. A likely explanation for this result is that the map requires more time to scrub through the motion set to find a match. Despite this, many participants preferred scrubbing through the motion set to clicking through examples. Future work will look at how to best combine the benefits of the map and label interfaces.

Specifically, the interface tested in this chapter is still very limited. For example, overlapping motion points are hard to work with and although we implemented a zoom capability, participants did not use it. Future work will address this issue by testing alternate layouts, such as those aligned in a grid or circle, which avoid overlap. Furthermore, we purposefully used minimal labels, which made the map harder to work with. Many participants requested labels in their subjective evaluation of the interface. Labels would help users understand the structure of the space and could furthermore be customizable to ensure they are meaningful to each user.

Our participants were primarily novice users and this may have influenced the
positive reception of the tool, due to the tool's simplicity. Enhancements would be necessary before the map would likely be useful for professional work. Overall, the map view is good for browsing but lacks precision. A good analogy for a complete interface is the color picker widget from tools such as Photoshop. Similarly, we could create a motion picker widget where users quickly browse for a rough estimate of the motion they want using the map, but then fine-tune their choice by directly specifying numerical values.

It is not clear whether displaying motions by similarity aids browsing, especially when motions are a similar category, such as walking. For example in their 2001 study, Rodden et al. (2001) investigated whether organizing large sets of images by similarity aided designers. They found that although clustering by similarity was generally useful, it also had the drawback of adjacent examples “blending” together and becoming easily lost and overlooked. They speculate that a randomized ordering may be better when the user is not looking for a specific example.

Previous research has proposed numerous interfaces for showing motion examples to users, but no work has yet been done to analyze what interfaces are best for performing specific tasks, such as browsing and motion editing. For example, how does the spatial arrangement of the map affect someone’s ability to remember where motions are located? What is the best way to present motions to a user (e.g., one at a time, side by side, or superimposed into a single image)? Future work should perform side by side comparisons of these interfaces in the context of performing specific tasks. As motion sets continue to grow, there is an increased need for tools that help people search, explore, and customize such sets.
Chapter 7

Discussion

This work aims to understand and model stylistic motion capture so that it can be modified and adapted to new situations through editing, decomposition, physics-based methods, and statistics-based methods. In Chapter 2, we categorize the primary ways in which motion capture is edited to so that it can be adapted to new situations: through editing, decomposition, physics-based methods, and statistics-based methods. Different approaches are best suited to different categories of styles and differ in their abilities to generalize and work in real-time.

In Chapter 3, we investigated how changes to captured motion clips, such as those described in Chapter 2, might alter the recognition and perceived intensity of an emotional performance. Rather than look at categories of motion, such as gait, we study a varied set of emotion clips captured from an actor. From perceptual experiments which systematically altered the poses and dynamics of each performance, we learn that changes to posture can change the perceived motion type whereas changes to dynamics can change the perceived intensity, and that the perceived intensity of an emotion can be reduced by blending with a neutral motion. These findings motivate one to take care when editing not to modify such qualities if the original emotion must be preserved. Through occluding parts of the body, we also learned that some joints are crucial for
conveying some emotions, for example, the head and neck are important for sadness. We also speculate that emotions are primarily expressed through the free joints of the body, for example, the movement of the legs is restricted while standing because they must help the actor maintain balance. Thus, when editing emotional content, some joints can be modified more than others without diluting the performance. Further research in this area could develop automated methods for preserving pose and dynamics during editing, or for purposely changing dynamics to alter the emotional content.

In Chapter 4, we tested a series of metrics aimed at quantifying the energy and naturalness of stylistic motions. Unlike in Chapter 3, we limit this analysis to categories of motions, specifically walking, in our case. Overall, we find that neutral motions had significantly lower sums of torques than stylistic motions, even considering slow-moving styles such as sadness. We did not find significant differences in our metrics for naturalness, such as with those considering smoothness. Future work would need to check whether this trend is consistent for different motion categories. Assuming this trend is consistent, we could use this finding to detect neutral motions or to infer neutral motion from stylistic ones (useful for motion editing and categorization). For example, in Chapter 3 we saw how blending with a neutral motion could reduce the intensity of an emotional motion. However, in that experiment, we only blended the upper body with a neutral motion. Using a technique which solved for a whole body neutral motion would potentially allow us to dampen (or exaggerate) any input motion. Other future work might look into the factors which cause an animal to move sub-optimally. In some cases, the reasons are aesthetic, as in dance, or even socio-cultural, such as in a swagger or geisha’s walk. But in other cases, the style might have a physiological basis which could potentially be simulated, such as with pride, shame, sadness, or fear. For example, Tracy & Matsumoto (2008) found that both sighted and blind people have the same body language in response to winning and losing.

In Chapter 5, we model the variations between our stylistic walks based on poses
rather than metrics. Our approach decouples each input example into an average and stylistic part. We then only model the stylistic part in PCA space. Although our use of PCA for motion modeling is not novel, our decision to decouple motions and only model the motion offset is a simple enhancement that makes the use of the PCA model produce consistently good output for blending, clustering, searching, and browsing our stylistic examples.

In Chapter 6, we use the PCA space to create an interactive minimap for browsing the stylistic walking set and creating blends. We found that for the Untimed experiment, users were equally satisfied with both the map and label interfaces, but were able to create and find motions using significantly fewer mouse clicks and exhibiting more creative results using the map. For the Timed experiment, both interfaces were equally poor at helping people find motions with a given semantic quality; however, the label-based interface was better for finding a matching motion example. Despite this, most participants greatly preferred the map (14 out of 17 participants, 82%), enjoying the ability to quickly scrub through the entire motion set. Although PCA coordinates are not typically intuitive, our participants had no trouble navigating the space of motions due to the fact that changes could immediately be seen on the character. The PCA space simultaneously clusters similar motions together and supports smoothly blending between PCA coordinates. Our user studies demonstrate that novice users easily understand the map and can use it to create new styles from the input motion set.

The browsing interface described in this thesis is very limited, however. For example, overlapping motion points are hard to work with and although we implemented a zoom capability, participants did not use it. Furthermore, we purposefully used minimal labels, which made the map harder to work with. Many participants requested labels in their subjective evaluation of the interface. Labels would help users understand the structure of the space and could furthermore be customizable to ensure they are meaningful to each user. Overall, the map view is good for browsing but lacks precision. Future
work would design a motion picker widget, similar in concept to the color picker in Photoshop, which would address these shortcomings: supporting different layouts which avoid overlap, adding customizable labels, and supporting scrubbing along with precise numerical controls.

Regarding browsing, it is not clear whether displaying motions by similarity aids browsing, especially when motions are in a similar category, such as walking. For example in their 2001 study, Rodden et al. (2001) investigated whether organizing large sets of images by similarity aided designers. They found that although clustering by similarity was generally useful, it also had the drawback of adjacent examples “blending” together and becoming easily lost and overlooked. They speculate that a randomized ordering may be better when the user is not looking for a specific example.

Lastly, motion search research has proposed numerous interfaces for showing motion examples to users, but no work has yet been done to analyze what interfaces are best for performing specific tasks, such as browsing and motion editing. For example, how does the spatial arrangement of the map affect someone’s ability to remember where motions are located? What is the best way to present motions to a user (e.g., one at a time, side by side, or super-imposed into a single image)? Future work should perform side by side comparisons of these interfaces in the context of performing specific tasks. As motion sets continue to grow, there is an increased need for tools that help people search, explore, and customize such sets.

Because style cannot be easily simulated or generated procedurally, methods for working specifically with stylistic motion capture are valuable, but care must be made that such editing does not unintentionally alter the intent of the original performance as viewers can pick up subtle differences in body language. The results from Chapter 3 and 4 help provide metrics and recommendations to ensure that the stylistic content of captured motions is not diluted or changed with editing. Compact and robust models for blending, sampling, and searching – such as the one in Chapter 5 – can also aid working
with stylistic motion capture by reducing the space and computation needed to work with a large set. For example, in Chapter 6, we demonstrate a possible use of the PCA model for helping non-expert users browse our stylistic motion set. These findings can help enhance systems for working with stylistic motion capture for preserving content, creating new content, and leveraging existing content.
Bibliography


perception from dynamic and static body expressions in point-light and full-light displays. *Perception, 33*(6), 717–746. 17, 18, 19, 20, 22


Frank, M. G., & Stennett, J. (2001). The forced-choice paradigm and the perception of facial expressions of emotion. Journal of Personality and Social Psychology, 80(1), 75–85. 21


Tournier, M., Wu, X., Courty, N., Arnaud, E., & Reveret, L. (2009). Motion Compres-


