EXPRESSIVE EYES AND BODIES: TRACKING, ALGORITHMS, EVALUATION

By

YUZHU DONG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2020

© 2020 Yuzhu Dong

This dissertation is dedicated to my parents, for their endless love, support, and encouragement.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Eakta Jain, for her continued support and encouragement. I thank my committee, Dr. Lisa Anthony, Dr. Jorg Peters, and Dr. Daniel Ferris, for their insightful comments and encouragement. I would like to thank all the past and present members of Jain lab and the HCC community whom I had the pleasure of meeting, for their collaboration and support in this and other research projects. Finally, I thank my parents for their unconditional love and support.

TABLE OF CONTENTS

page

ACKNOWLEDGEMENTS 4

LIST OF TABLES 7

LIST OF FIGURES 9

ABSTRACT 10

CHAPTER

1 INTRODUCTION 12

2 BACKGROUND 15

2.1 Domain Knowledge about Child Motion 15
2.1.1 Biomechanics 15
2.1.2 Perception 15
2.1.3 Motor Control 16
2.1.4 Energy Level and Physical Activities 17
2.2 Motion Translation 17
2.2.1 Motion Retargetting 17
2.2.2 Dynamic Scaling 18
2.2.3 Learn from Corpus 19
2.2.4 Deep Learning and Motions 20
2.3 Motion Evaluation 21
2.3.1 Heuristic Rules 21
2.3.2 Numerical Metrics 22

3 EXPRESSIVE BODIES: ALGORITHMS AND PERCEPTUAL EVALUATION 24

3.1 Linear Scaling Algorithm 26
3.2 Generating Stimuli 27
3.3 Non-linear Scaling Algorithm 30
3.3.1 Method 30
3.3.2 Results 33
3.3.3 Summary 35
3.4 Perceptual Evaluation 35
3.4.1 Study 1 Results 37
3.4.2 Study 2 Results 42
3.4.3 Study 3: child2adult 46
3.4.4 Discussion 50
3.4.5 Summary 52

4 EXPRESSIVE EYES: PERCEPTUAL EVALUATION 55

4.1 Experimental Framework 57
4.2 Study 1: Identify the Scared Avatar 60
4.2.1 Description 60
4.2.2 Results 60

4.3 Study 2: Pupil as a Cue . . . 63
4.3.1 Description . . . 64
4.3.2 Results . . . 64
4.4 Study 3: Pupil Diameter Variation as a Cue . . . 67
4.4.1 Description . . . 67
4.4.2 Results . . . 68
4.5 Respondents' Self-Reports . . . 70
4.5.1 Task Difficulty . . . 70
4.5.2 Recognition Accuracy by Exposure Level . . . 71
4.5.3 Self-reported Cues . . . 71
4.6 Discussion . . . 72

5 NEURAL NETWORK FOR CREATING EXPRESSIVE CHILDLIKE MOTIONS . . . 75

5.1 Motion Tracking Children . . . 75
5.2 Motion Stylization for Creating Childlike Motions . . . 78
5.2.1 Data Representation . . . 78
5.2.2 Network Architecture . . . 81
5.2.3 Loss Function . . . 81
5.2.4 Post-processing . . . 84
5.3 Results . . . 85
5.3.1 Results on Existing Dataset . . . 85
5.3.2 Ablation Studies on Existing Dataset . . . 87
5.3.3 Discussion of the Existing Dataset . . . 89
5.3.4 Discussion of our Novel Child Dataset . . . 90
5.3.5 Results on our Novel Child Dataset . . . 92
5.3.6 Discussion and Limitation . . . 98

6 SUMMARY AND CONCLUSIONS . . . 101

REFERENCES . . . 104

BIOGRAPHICAL SKETCH . . . 111

LIST OF TABLES

Tables page

3-1 The scaling factors according to the dynamic scaling laws. 28

3-2 Means and standard deviations of number of child responses for each actor type in Study 1. 41

3-3 Means and standard deviations of number of child responses for each action type in Study 1. 41

3-4 Means and standard deviations of number of child responses, confidence score and naturalness score in Study 2. 46

3-5 Means and standard deviations of number of child responses for each action in Study 2. . 46

3-6 Means and standard deviations of number of child responses for each actor type in Study 3. 49

3-7 Means and standard deviations of number of child responses for each action type in Study 3. 50

4-1 Results of Study 1: identify the scared avatar. 62

4-2 Results of Study 2: pupil as a cue. 66

4-3 Stimuli generation of Study 2: pupil as a cue. 67

4-4 Stimuli generation of Study 3: pupil diameter variation as a cue. 68

4-5 Results of Study 3: pupil diameter variation as a cue. 69

LIST OF FIGURES

Figures page

1-1 Examples from animated movies. 14
3-1 Visualization of the skeleton topology. 29
3-2 Visualization of the local coordinate frame. 30
3-3 Results of the style translation algorithm in stick figures. 34
3-4 Results of the style translation algorithm on rendered characters. 34
3-5 Distribution of the overall correctness for Study 1. 40
3-6 The distribution of the number of child responses in Study 1 40
3-7 Bar charts of the number of child responses for each action type and actor type in Study 1 41
3-8 Bar charts of the number of child responses for each action type and actor type in Study 2 45
3-9 Distribution of the percent correct responses in Study 3. 48
3-10 Boxplot and bar chart of the number of child responses for each actor type in Study 3 49
3-11 Bar charts of the number of child responses for each action type in Study 3 50
3-12 Rendered results on child actor, dynamically scaled adult actor, and adult actor. 54
4-1 Visualization of the pupil diameter and the gaze direction from subject s015 57
4-2 Raw pupil diameter recordings from the three selected eye motions of people watching the scary movie. 60
4-3 The distribution of percent correct responses for Study 1. 63
4-4 Pupil diameter values of Eye Motion ID s015 66
4-5 Results of Study 2: pupil as a cue. 66
4-6 Pupil diameter values of Eye Motion ID s015 used in Study 3. 69
4-7 Results of Study 3: pupil diameter variation as a cue results. 69
4-8 Respondents' Self-Reports. 72
4-9 Comparison across three studies. 72
5-1 Architecture of CycleGAN. 79
5-2 Architecture of the generator. 80
5-3 Architecture of the discriminator. 80
5-4 Results on fast walking action. 85
5-5 Results on normal walking action. 86
5-6 Results on jumping action. 86

5-7 Results on running action. 86
5-8 Ablation study on cycle consistency loss for fast walking action. 87
5-9 Ablation study on transition loss for fast walking action. 88
5-10 Ablation study on temporal coherence loss for fast walking action. 88
5-11 Ablation study on the ordering of the joints for normal walking action. 89
5-12 An example of improvisation during capture. 92
5-13 The 1D and 2D convolutional networks tested in our experiments. 94
5-14 Architecture of the 1D generator. 94
5-15 Architecture of the 1D discriminator. 95
5-18 Timelapse on walking action on our dataset from the front view. 95
5-19 Timelapse on walking action on our dataset from the side view. 95
5-16 Results on “jump-as-high-as-you-can”. 96
5-17 Results on “walk-as-fast-as-you-can”. 97

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

EXPRESSIVE EYES AND BODIES: TRACKING, ALGORITHMS, EVALUATION

By

Yuzhu Dong

December 2020

Chair: Eakta Jain
Major: Human-Centered Computing

Animated characters have gained increasing popularity in various applications. To create expressive human-like characters that emote, it is essential to model the non-verbal behaviors properly. Humans use a wide range of verbal and non-verbal cues to convey emotions and messages. However, modeling non-verbal behaviour is a difficult and time-consuming task. In this dissertation, I investigated the two extremes of non-verbal cues: creating expressive eyes for scared characters and generating expressive body animations for child characters.

In the first part of my dissertation, I studied whether pupil diameter would be a noticeable social cue for animated characters. Pupils are an indicator of emotional arousal and cognitive load for humans. I conducted a series of perception studies on animated characters that examined different aspects of pupillary motions. Our findings showed that adding pupil dynamics helps characters convey the emotion of fear effectively at a negligible cost.

In the second part of my dissertation, I looked into modeling whole body motions for child characters. Child motions are challenging to model because they are very different from adult motions. Child motions are unpredictable and less coordinated. However, modeling child motions is still not well studied. I explored various models ranging from a simple linear model (dynamic scaling procedures) to a complex deep learning model (CycleGAN). I found that CycleGAN architecture is able to capture the characteristics from child motions, though it also requires more training data for learning. Therefore I collected a motion capture(mocap) dataset from children as part of the training data. Capturing children is a challenging process because children are less

collaborative and less patient. To this end, I compiled and summarized a set of guidelines and insights for capturing child motion. I have also released the dataset I collected for future research along similar threads.

CHAPTER 1
INTRODUCTION

There is an increasing demand in the entertainment and education industry to create vivid characters for a variety of applications such as games, movies, and online education. The latest computer graphics breakthroughs have made the appearance of the characters more and more compelling. To truly bring the characters to life, we need to model their behaviors to be human-like. Humans are experts in interpreting non-verbal behaviors. Studies have shown that 55% of emotion is communicated through body language [55]. Body language consists of facial expressions, body posture, gestures, eye movement, touch and the use of space, etc.

Body language can be displayed on different scales. On the large scale of the spectrum, when an angry child throws a tantrum, his/her emotion may be manifested in the aggressive way of running and stomping. On the small scale of the spectrum, pupils would dilate when someone is surprised or scared. If we take two scenes from the Pixar movies *Toy Story 4* and *Up* as examples (Figure 1), the girl sits there with a bent upper body, weak limbs, and rolling eyes. Apparently, she was disappointed or throwing a tantrum. For the boy on the left-hand side, his spreading arms and enlarged eyes show that he is happy and excited. I studied how to model these behaviors mathematically from motion data collected from participants. I also investigated if these non-verbal behaviors are perceivable on characters, and if modeling these motions correctly would add extra fidelity and expressiveness to them.

The impact of my dissertation work is helping animators or digital content creators be more efficient in creating expressive characters. The process of making the motion right can be costly and time-consuming. In the animation industries, sequences that involve complex motions such as Kung-fu or dancing could take up to weeks and months for animators to finish [42]. By using tracking and algorithms, we can free them from part of the laborious works.

From the wide range of non-verbal behaviors, I studied the two ends of the spectrum: subtle non-verbal cues and obvious non-verbal cues. More specifically, my dissertation details my work along with two thrusts: how to create expressive eye motions for animated characters, and how to create expressive body animations for child characters.

In the first thrust, I investigated subtle social cues that are often ignored in the animation

pipeline: creating expressive eyes for animated characters that can bring the emotions out of them. Pupil diameter is an indicator of emotional arousal and cognitive load but often overlooked in the animation pipeline. Animating pupils comes with additional costs, so our study tests if this cue can be detected inside virtual characters and if adding pupillary motions would make them appear more scared to the viewers.

In the second thrust, I focused on creating expressive body animations for child characters using linear and non-linear algorithms. We chose children as the target population because they have become important market targets in video games industries and electronic entertainment. Children have a very distinctive way of performing actions. Adult motions are well studied, and there are abundant mocap databases ready to use for adult characters. Child motions are less available compared to adult motions.

To create the expressive body for child characters, we investigated three models: dynamic scaling model, linear time-invariant model, and generative adversarial networks. From the investigation of the first model, we learned that the dynamic scaling procedure is simple to implement and does not require training data. However, it did not completely convince the viewers that the motions are from children. The linear time-invariant model can be trained on temporally aligned motion pairs. It can successfully stylize input motions that are similar to the training data. However, the model is not generalizable to different motion types. The complex GANs model enables us to capture the nuances of child motions from a variety of motion types without a temporally aligned training set. But such a complex model requires more training data. As part of this investigation, we created a novel dataset from a collection of adult participants and child participants.

The dissertation is organized as follows: In Chapter 2, I review the related works and background literature on domain knowledge of child motions and eye motions, how to create expressive eyes and body motions, as well as how to perceptually evaluate non-verbal behaviors. In Chapter 3, I introduce my work on creating expressive eyes by adding pupillary motions. I also include two studies on expressive body motions (dynamic scaling study and style translation

algorithm). In Chapter 4, I discuss the motion capture(mocap) process for children and train a CycleGAN to learn the characteristics of child motions.

The contributions of this dissertation are listed below:

1. We found that adding pupil dynamics helps characters to convey the emotion of fear. By modeling the right pupil dynamics, content creators and animators can make the characters more expressive.

2. We compiled a set of guidelines for capturing child motions for future researchers who intend to collect more data from children through the mocap process.

3. CycleGAN is capable of extracting the style from different motion types and work with heterogeneous motion inputs. It also requires more training data. We recommend using the CycleGAN architecture to create child motions from adult motions and enrich the child mocap database.

Figure 1-1. Examples from animated movies. A) Russel from movie *Up*. B) Bonnie from movie *Toy story 4*.

CHAPTER 2
BACKGROUND

2.1 Domain Knowledge about Child Motion

In this section, I aim to answer the first research question: Which study has investigated the differences between the adult motions and child motions, and what have they found? We reviewed the previous works from four aspects: perception, biomechanics, motor control, and energy level.

2.1.1 Biomechanics

A child body is not a mini-version of an adult's. The body proportion keeps changing when they grow up. The ratio of the size of the head to the body height continues to decrease. For a newborn, the head is about one-fourth of the body height, whereas for an adult, the head is about one-seventh of the body height [32]. The growth of the upper limbs vs. the growth of the lower limbs also changes. For new-born babies, their upper limbs are longer than their lower limbs. However, for adults, their lower limbs are longer than the upper limbs [57]. Another finding is that up until puberty, a child's limbs grow much faster than the size of the head and the torso. When they reach puberty, the growth of limbs reduces to the level that is similar to that of the torsos. When they reach the adolescent period, the growth rate of the torso will surpass the growth rate of the limbs. The ratio of the torso to the body height decreases from 70% at birth to 57% at 3 years old, to 50% for a 13 years old girl or a 15 years old boy. The mid-point of the body starts at the navel for a newborn, then gradually move down to a point below the navel for two years old. For a teenager, the mid-point is close to the pelvis. Another important metric to look at is the center of gravity (CoG). CoG for children is located higher vertically than adults'. The relative body mass of the upper body for children is greater than that of adults' [32].

2.1.2 Perception

[36] collected adult mocap data and child mocap data from Kinect device. They asked actors to perform 6 actions ranging from relatively static action to energetic actions. From the main experiment, they found that naive viewers can identify if a motion was performed by a child or an adult at levels significantly above chance. They also noticed that the recognition level depends on the actions being performed. For more energetic actions like "Jump high", the perceptual differences between adult motion and child motion are larger than more static actions

such as “Wave” and “Fly like a bird”. They also noticed that viewers attributed more coordinated action such as “Jumping to an adult more than a child. One takeaway message is that when creating energetic or dynamic motions for children, the motions can be made more exaggerated and child-like than creating static motions for them.

2.1.3 Motor Control

New motor skills open the door to opportunities for learning and doing new things. Infants usually stay in a supine position. They can sit upright when they are about six months old. They start to crawl around nine months old and walk independently in about 1-year-old [76]. All these milestones are achieved by the growth of the motor skills [1]. Motor behaviors include almost all kinds of physical motions, from involuntary leg motions of babies to the last breath before death. Every movement depends on generating and controlling physical forces. Managing the forces requires coordinating multiple muscles, mastering biomechanical skills, and more importantly, operating high-level functions such as perception and cognition [1].

Researchers looked into the natural history of immature walking. Before the fourth year, the maturation in the central nervous system and the growth in height lead to the changes in time-distance parameters in gait, such as step length, cadence (steps/min) and walking velocity. Starting from the fourth year, the growth of limbs and height alone explains the majority of changes in those time-distance parameters. Cadence is found dropping rapidly from 2.5 to 3 years old and continues to decrease until seven years old. In other word, children shuffle their legs more rapidly than adults. Stride length is reported increasing sharply to 4 years old. Then, it continues to increase but at a slower rate. Walking velocity increase rapidly until 3.5 to 4 years of age, which means the increasing rate of the step length outpaced the decreasing rate of the cadence. The step factor (step length divided by leg length) continues to increase until four years old and remains constant and adult-like [80, 39]. [14] used a machine learning approach to classify adult walkers and child walkers. They found that relative stride length and stride frequency are shown to differentiate between them.

2.1.4 Energy Level and Physical Activities

It has been established that when performing dynamic, whole-body activities, such as cycling, sprinting, vertical jumps and isokinetic activities, untrained adults are easier to experience fatigue than pre-pubertal children [69, 15, 43]. Another study even found that children are metabolically comparable with well-trained adult endurance athletes [9]. Trost et al. noticed that there is a negative correlation between the age and the level of moderate-to-vigorous physical activities. They also found that boys were more active than girls [85].

This phenomenon has been mainly attributed to the fact that children have a lower peripheral (i.e., muscular) fatigue than untrained adults [70]. Furthermore, children have a higher level oxidative capacity in the muscles and more fatigue-resistant muscle fibers [70, 48]. We also see a strong negative correlation between age and oxygen consumption. Child oxygen consumption (6–12 years old) is higher than adolescents'(13 – 19 years old). The rate of oxygen consumption for adolescents is higher than that of the adults [90]. In addition, prepubertal children show faster oxygen uptake and heart rate recovery following vigorous activities than untrained adults [25]. Taken together, children are able to complete repeated vigorous exercise easier than untrained adults.

2.2 Motion Translation

Because there are a great number of adult motion available, can we translate those motion into child-like motion? Previous works have translated motions to fit characters with different skeleton ratios, emotions and identities. In this section, we review those papers with respect to the approach they adopted.

2.2.1 Motion Retargetting

Mapping a motion from a character to another with different skeletons and maintaining the physical constraints in the same time is challenging [23, 44]. The resulting motion cannot be exactly the same as the original motion. For example, it is impossible to match both the joint angles and the end-effector positions simultaneously for characters with different size or proportion. To preserve the important content of the original motion, we have to change some

aspects of the motion without making the changes noticeable. This adaption is called motion retargetting [23, 44].

Space-time optimization method does not exploit the kinematic or dynamic features of the motion. Instead, it focused on the relationship between the end-effectors and the environment across the entire duration of motion. [23] used space-time solver to maintain specific features or constraints, such as feet touching the ground while walking, or hands touching the object while fetching. If the constraints are violated, they adapted the motion in a way that preserves the quality of the motion. They re-establishes the motion by distributing the changes across the entire duration of the motion. By considering the entire motion sequence, the system avoids introducing high-frequency glitch or snap when enforcing constraints at certain frames. In this way, the frequency characteristics of the original motion signal got maintained. [81] applied space-time solvers to correcting an unbalanced motion into a dynamically balanced one. Because the space-time optimization can take an incredible amount of computation subject to dynamic constraints (i.e., Newton's law), [93] introduced a character simplification technique to make the problem tractable.

2.2.2 Dynamic Scaling

The applications of dynamic scaling laws range from robotics to computer graphics. It allows the reapply of the motion to a geometrically similar character with a different scale [67, 26]. Raibert and Hodgins first implemented these scaling rules to adapt the control systems of biped animals including human to the control system of their scaled versions [67]. [26] adapted the parameters of a control system of an adult running motion to generate a child running. They mentioned that dynamic scaling laws could be applied to both robotic control systems and motion capture data.

The dynamic scaling laws state that for a character scaled uniformly with the proportion maintained. When an actor is L times shorter, we can derive that the height he can jump is L times lower and it takes $1/\sqrt{L}$ times the time needed to finish the jump. The scaling laws ensure that gravity is maintained in the motion.

Hodgin et al. [26] expanded the scaling rules to adapt motion for new characters that are not only in size, but also in mass distribution, limb length, and moment of inertia. They used the dynamic scaling rules for the first step, geometric scaling. Then they further scale the motion to compensate the differences not captured in geometric scaling such as body proportion. The output motion of a scaled child character is very similar to the motion from video footage of a 4-year-old boy.

2.2.3 Learn from Corpus

Data-driven motion translation allows users to create highly realistic and convincing motion. Such an approach has been a great help to animators to produce animation by passing high-level instructions instead of diving into details frame by frame. A variety of work in computer animation have exploited large motion capture databases and state-of-art machine learning models.

Many researchers have developed system that allow the style to be transferred from one motion to another [3, 87, 13, 31, 75, 94, 50]. Such technique is known as style transfer. The system can learn from two large labeled corpus with different styles, and it allows the input data to be used to define the motion style. [31] performed style translation between two motions with the same content but different styles. They used iterative motion warping to compute the temporal correspondences between the two motions. For each degree of freedom, they trained a separate linear time-invariant model [49] to extract the stylistic aspect of motion. Once the system is trained, the translation can be performed on a new motion in real time. However, their method requires a large amount of training data set with minutes of motion clips. [94] also aimed to use a data-driven approach to model the spatial-temporal differences between styles of motion using training examples. Instead of using a linear invariant model proposed by [31], they constructed a series of local mixtures of autoregressive models (MAR) to capture the motion differences between the styles. The MAR models are estimated by searching the nearest example in the training dataset. This local mixture model enables the system to handle motion data that are either unlabeled or heterogeneous. Once the model is estimated, it adapts the input motion with a simple

linear transformation in real-time.

2.2.4 Deep Learning and Motions

Deep learning has taken off and shows very promising results in areas such as autonomic driving, image processing and so on. Researchers also started to use this powerful tool to model human motions. [27] implemented deep learning in motion synthesis. It does not require any manual processing such as labeling and segmentation. After training on the raw data for 20 hours, the system can synthesize motion with a variety of styles based on the terrains or specific tracks. They stylized the motion in the hidden units space to preserve the naturalness and the smoothness of the motion. The weights in the model were learned by minimizing the differences of the Gram matrix in the hidden layer between the original motion and the stylized motion. Reinforcement learning has also shown great success in character control. Characters learn how to respond to user behavior in real-time, perform complicated tasks or infer behavior styles from examples [47, 45, 54]. However, it is always a challenge to choose the right reward function. [45] presented an automated framework to determine the reward function based on sample motions. The controller learns different behavioral style and transfers the extracted style to a motion in a new environment or new task.

[77]. developed a style transfer model that can adjust the style in latent space. They decompose the problem to three components: pose, timing and foot contact. Each aspect was modeled using a fully connected multilayer neural network. [28] learned high-level manifold of human motions using a convolutional autoencoder, which can be used for motion interpolation, clean up, and searching in the mocap database. Mode adaptive neural networks model [52] leveraged the existing trained Phase-Functioned Neural Network and used a few stylistic examples to achieve style translation with transfer learning. However, their method only works for cyclic motions. [19] learned a generative model using a conditional variational autoencoder from large amount of motions in neutral style. Then they learned style information as a prior distribution.

Generative models can also be extended to motion prediction tasks. Recurrent neural

network is a class of network that is designed to process sequential or time-series data. Several studies have applied it in human motion modeling. [63] modeled human motion by using Gated Recurrent Unit(GRU) connected with a forward kinematics layer. The model takes joint angle in quaternion representation as input but penalize the absolute joint position values in the loss function. Researchers have also cooperated adversarial learning in motion modeling. [5] combined adverserial learning to predict motions in a non-deterministic fashion. They adopted Wasserstein generative adversarial networks (WGAN-GP) with sequence-to-sequence model as generator. It can predict multiple future poses given previous poses and a random vector mapped into the hidden state. One of the use cases for adversarial learning is to generate meaningful instances from noise. [89] investigated conditional generator that could synthesize motions with controls of high level features. Their style translation module can separate styles from content and enable style transfer between different motion type. However, their approach can only alter the poses, not the timing of the motions.

2.3 Motion Evaluation

Evaluation is a critical step after motion synthesis. How do we know if the motion translation algorithm is successful? In this section, I reviewed articles that approach motion evaluation problem from two angles: heuristic rules, and numerical metrics.

2.3.1 Heuristic Rules

One motion evaluation method is establishing a set of heuristic rules to guide the motion synthesis process. It is often used to evaluate if the motion is natural or realistic. If a given motion clip violates any of the rules, it could be considered unrealistic or unnatural. The rules could be either summarized from daily experiences or derived from physics. Example of rules could be angular momentum conservation in flight; friction cone should be maintained when floor and foot are in contact to avoid sliding happened, different body parts should not penetrate each other or any surface in the environment, the motion should be smooth in general instead of joints jump here and there. However, these rules are not sufficient enough to define the quality of the human motion. A motion can satisfy all the heuristic rules but still look unnatural.

2.3.2 Numerical Metrics

One intuitive metric compares the differences between the output motion of a given algorithm with an existing reference motion. Although it is not applicable to motion outside the database, it is nevertheless a practical approach to see what is missing in the synthesis motion and how to improve to be closer to the reference motion. The comparison can happen in different scales. For example, the comparing object could be an entire motion sequence, a short clip or a single frame. The metrics applied on whole animation sequences or clips are often used for action classification or database search. This is also the first step to compare a synthesized motion to a motion from a large amount of dataset, which requires identifying and extracting motions that are logically similar to target motions. Also, the algorithm should work with all types of motion length. [58] introduced spatio-temporal invariance into the geometric features of skeletons based on the specific body points to retrieve motions that are logically similar. [61] aimed to find similar sequences and creating clusters from mocap dataset by testing many feature metrics. They found that a near-perfect classification by combining the right features.

Machine learning techniques provide a powerful tool to classify motions as high or poor qualities based on human-labeled data [34, 71]. [71] explored the possibility of evaluating the perceptual realism of animation using a machine learning model. They trained several learning algorithms, including mixture of Gaussians, hidden Markov models, only on positive (natural) examples. The model can be used to asses a variety of motion whether the motion is natural or not. To increase the variations of the motion dataset, [34] introduced a novel limb transplanting approach that best matched the judgments of users. They trained a support vector machine (SVM) on a motion dataset that manually labeled by viewers as “looks human” or “not human”. The SVM serves as a classifier for the good and bad results.

The work from [84] is very similar to our problem. He identified the motion features that are useful for classifying if a motion is coming from a male or a female. Such a model can also be applied to identify if a motion is from a child or an adult. He used principal component analysis to reduce the dimensions of the motion and found the important features for the classifier. Their

model works well with periodical motion such as walking, but it could not be generalized well to a heterogeneous dataset.

CHAPTER 3
EXPRESSIVE BODIES: ALGORITHMS AND PERCEPTUAL EVALUATION

Surveys have shown that in the UK, children from 8 to 11 years old spent 8 hours on video games and 10 hours on TV shows weekly in 2017 [78]. Games and animations serve not only as tools of entertainment and relaxation but also as motivations for learning and exercise [12, 65]. A study has shown that involving characters that are similar to the users could significantly increase their enjoyment and engagement in their experiences [33]. Hence, in order to attract more young users, many animated movies and video games adopted child characters in their stories. Animated movie *Boss Baby* from DreamWorks Animation told the story of a 7-year-old boy and his toddler brother [92]. One of the lead characters in the adventure game *The Walking Dead* by Telltale Games is an 8-year-old girl [91].

To create motions for the characters, one common approach in animation industries is keyframing [93]. Animators pose the characters in every critical frame and animation software will automatically interpolate the frames in between. In this way, they can create very expressive and exaggerated motions to reflect the characters' personalities. This approach also costs many hours of repetitive works. On the other hand, in the gaming and live-action industries, motion capture (mocap) data is widely used for animating avatars. The data is captured from actual actors and then retargeted to the avatars. It saves animators from creating every pose from scratch. Mocap data also has the advantages of maintaining the realism and kinematics of the motions, as well as following physics rules and constraints from outside world [56].

Across the existing mocap databases, adult mocap data are more abundant and comprehensive compared to child mocap data. The most well-known motion capture database such as CMU database and OSU mocap database only contains adult motions. However, the only open-source motion capture database that contains child motion available online is Kinder-gator dataset [2]. The unbalance between the amount of adult mocap data, and child mocap data stems from the fact that it is very difficult to motion capture children and hire child actors. Children may be confused with the instructions, lack of patience, and make it very challenging to collaborate with them [64].

Based on our personal experiences interacting with children or looking back when we were

kids, we can all agree that children move very differently from adults. Researchers have noticed that child motions, and gestures are perceivably different from adults' even when they are completing the same actions [36]. There are several reasons for these differences. First, studies have shown that there are biomechanical differences between children and adults in terms of body mass distribution and body proportion [32]. Researchers have also found that young children are more energetic than adults [59], but less skillful when it comes to body motor control and proficient performance [29, 30].

Given the fact that there is a large amount of adult motion available, is it possible to put adult motion directly on child characters? Because of all the differences, there will be a noticeable inconsistency if we put adult motion capture data directly on child characters without modification. It would be time-saving if the existing adult motions can be translated into child motions. Hence, we introduce motion translation algorithms to tailor the motion to be more childlike. Such a model would be highly useful in the animation industry, enabling adult actors to be employed for child animations. Motion translation opens a door for animators to change the style of the existing motion or adapt the same motion to characters with different skeletons, limb ratios, or emotions [31, 23, 94, 27]. For example, a normal walking can be modified to sideway shuffle, sneaky crouch, happy stroll, or sad walk [31]. This approach could significantly expand the current motion database.

Researchers developed both machine learning-based algorithms and procedural algorithms to learn the mapping between two labeled mocap corpus with different styles. In my prior work, I looked at using dynamic scaling laws to transfer motion from adult to child. These approach does not rely on training examples. We found that the use of dynamic scaling led to an increase in the number of people identifying the motion as belonging to a child compared to the original adult motion, but not as much as original child motion. The accuracy levels indicated that there are still stylistic differences between adult and child motion that are not captured by linear scaling. My next work [17] explores a machine learning-based style translation algorithm. In particular, I implemented [31]'s method which transforms input motion into a new style through a linear

invariant model. This approach generates motions that closer to actual child motions compares to linear scaling. However, it requires training data consists of pairs of mocap data that are temporally aligned. Also, it can achieve mapping between two motion instances instead of two motion distributions. These limitations also hold for other style translation method [94, 27]. Making motions temporally aligned can be a tedious and time-consuming process because the duration of the motion, repetition of the motion vary. Note that our goal is to create realistic childlike motions by changing the style of adult motions. By using data-driven method, we extracted the characteristics of child motions in the mocap dataset. Artistic stylizing or over exaggerating is out of the scope of this dissertation.

3.1 Linear Scaling Algorithm

As the first study of the adult2child motion translation, I investigated if dynamic scaling laws could help us generate child motions from adult motions. The application of dynamic scaling laws has ranged from robotics to computer graphics. Raibert and Hodgins first implemented these scaling rules to adapt the control systems of animal and human motion to geometrically similar scaled versions [67]. Hodgins and Pollard [26] adapted the parameters of a control system of adult running motion to generate child running motion. They mentioned that dynamic scaling laws could be applied to motion capture data. However, the open question is: from the perception perspective is whether the scaled adult actor appears to be running like a real child. We address this open question

The dynamic scaling laws state that for a character scaled uniformly with all the proportion maintained, Table 3-1 gives rules for scaling the motion parameters. When an actor is L times shorter, we can derive that the height he can jump is L times lower and it takes $1/\sqrt{L}$ times the time needed to finish the jump. The scaling laws ensure that gravity is maintained in a jumping action. According to Newton's laws of motion, in a jumping action where the acceleration is due to gravity, $g = d/T^2$, the ratio between the vertical displacement of the jump d and the square of the time during jumping T is constant. Since velocity $v = L/T$, the changes in the displacement and time duration will result in the velocity to be $\sqrt{L}$ times faster.

We assume that adult and child characters are geometrically similar. Because all the actions we investigated were in place motion and they primarily involved up and down movement, we scaled the vertical displacement of the original adult motion along y-axis L times and ignored the lateral movement along x and z-axes. The time duration for completing each action was shortened by $\sqrt{L}$.

3.2 Generating Stimuli

We used motion capture data from [2]. This data was collected by a Microsoft Kinect 1.0 at 30 fps. We used the same subset of the data containing four adult actors 22 to 32 years old and four child actors 5 to 9 years old as Jain et al. [36]. Each actor performed six actions: "Run as Fast as You Can" ("Run Fast"), "Walk in Place" ("Walk"), "Jump as High as You Can" ("Jump High"), "Fly Like a Bird", "Wave Your Hand" ("Wave"), and "Do 5 Jumping Jacks" ("Jumping Jacks"). The first several frames containing initial T-pose and the last few frames of actors idling were removed.

The mocap data was recorded as three-dimensional joint positions in world coordinates ${}^{x}_{w}u^{j}_{i}$, where j denotes the number of joint and i denotes the frame number; the left superscript x represents x-coordinate, and the left subscript w represents the world coordinate frame (Figure 3-2). The mocap data is stored in a matrix M with dimensions T $\times$ 3N, where T is the total number of frames and N is the total number of joints.

$$M = [{}^{w}u^{1}, {}^{w}u^{2}, \ldots, {}^{w}u^{20}] \tag{3-1}$$

where ${}^{w}u^{j}$ is a $T \times 3$ dimension vector. To simplify the computing process, we translated the joint position from the world coordinate frame into a local coordinate frame where the x, y, z-axes were parallel to the world coordinate frame, but the origins were located at the parent joint of each joint. The hip joint adopted the origin of the world coordinate frame as the parent joint. The joint position in local coordinates was calculated by subtracting the world position of the parent joint

Table 3-1. The scaling factors according to the dynamic scaling laws. If the height of an actor is scaled by factor L, its motion attributes will also be scaled according to the table [26].

Quantity	Units	Geom. Scaling	Mass Scaling
basic variables			
length	L	L	–
time	T	$L^{1/2}$	–
Motion variables			
displacement	L	L	–
velocity	LT^{-1}	$L^{1/2}$	–
acceleration	LT^{-2}	1	–
angular displacement	–	1	–
angular velocity	T^{-1}	$L^{-1/2}$	–
angular acceleration	T^{-2}	L^{-1}	–

from the world position of the child joint. This process was repeated from the leaf joint to the root joint for each branch of the skeleton tree.

The scaling factor L was computed as the ratio of the height of the adult actor and the average height of 5 to 9 years old child according to the World Health Organization's (WHO) Growth Reference [60]. The steps of implementing dynamic scaling laws are shown as follow.

1. The first step is spatial scaling. We assume that a dynamically scaled adult is geometrically similar to an adult. Therefore, we scaled each limb uniformly by the factor L.

2. The displacement scaling described below only applies to hip joint. Because all the actions were in-place motion which involved minimal lateral movement along x and z-axis, therefore we only scaled the displacement along y-axis. The displacement was computed with respect to the hip position in the first frame.

$$u_y'^{hip}(i) = u_y^{hip}(i) + L \times (u_y^{hip}(i) - u_y^{hip}(1)) \tag{3-2}$$

3. Then we perform temporal scaling in the motion sequence. The displacement is shortened by a factor of L. The time duration is shortened by $\sqrt{L}$. Since we keep consistent frame rate across all the videos, the total number of frames is reduced $\sqrt{L}$ times. We calculate the

corresponding frame number between the original adult motion i and dynamically scaled motion i',

$$i = 1 + (i' - 1) \times T / (round(\sqrt{L} \times T)) \tag{3-3}$$

4. We find a mapping between the hip position of dynamically scaled motion in frame i' and original adult motion sequence in frame i through linearly interpolating the original hip position between frame $\lfloor i \rfloor$ and frame $\lceil i \rceil$ (Equation 3-4).

$$u_y^{hip}(i') = (\lceil i \rceil - i) \times u_y^{hip}(\lfloor i \rfloor) + (i - \lfloor i \rfloor) \times u_y^{hip}(\lceil i \rceil) \tag{3-4}$$

5. We then translate the joint position in local coordinate back to world coordinate by traversing the skeleton tree in a top-down manner. We calculate the joint position in world coordinate by adding the world position of the parent joint with the local position of the child joint. In order to remove the cues about actor's height, we normalized the stimuli to a canonical height as in [36].

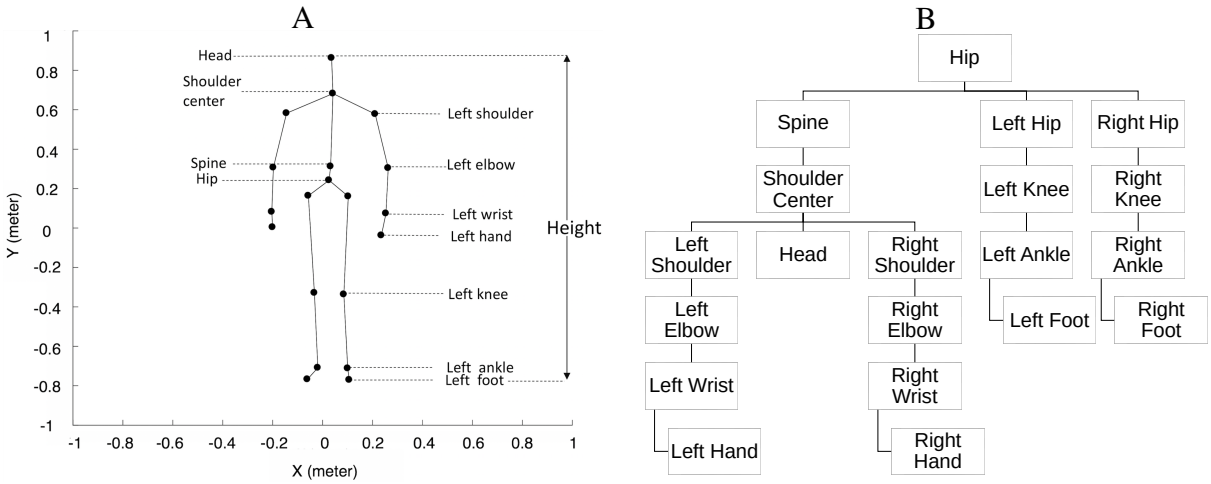


Figure 3-1. Visualization of the skeleton topology. A) sample skeleton shows the name of each joint. B) The tree structure of all the joints with hip joint as root.

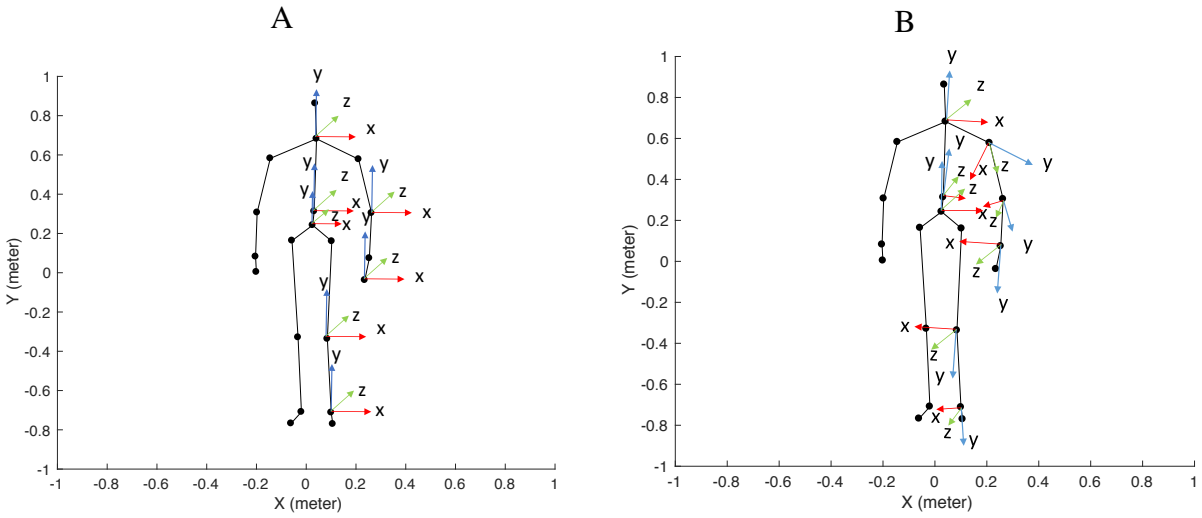


Figure 3-2. Visualization of the local coordinate frame. A) the local coordinate frame which has the origin of each joint in its parent joint and axes parallel to the world coordinate frame. B) the local coordinate frame which has the origin of each joint in its parent joint and *y*-axis parallel to parent limb.

3.3 Non-linear Scaling Algorithm

My first work showed that linearly scaling adult motion in space and time dimensions makes it more likely for viewers to think that such motion came from child actors than adult actors. However, the accuracy levels indicated that there are still stylistic differences between adult and child motion that are not captured by linear scaling. In this work, I explored a non-linear warping based style translation algorithm. In particular, we implement [31]'s method which transforms input motion into a new style through a linear invariant model. Our main contribution is to apply this method towards the new problem of transforming adult motion to appear child-like.

3.3.1 Method

We use the database of [2] for matched pairs of child and adult motion. We first translate the joint position in world coordinates into local coordinates as in 3.1. We denote the frame number as i, the joint number as j, the input adult motion as u, the output child motion as v. Therefore, the joint position along the x-axis in local coordinates is written as ${}^{x}u_i^j$ for adult motion, and the output child motion is written as ${}^{x}v_i^j$. The matrix u is used to describe the entire motion of all the joints in the skeleton, where each frame is a row of the matrix. Each joint is considered separately. The vector $\mathbf{u}$ is concatenated by the position of the jth joint across all the frames,

$\mathbf{u} = [{}^{x}u_1^j, {}^{x}u_2^j, {}^{x}u_i^j \ldots, {}^{x}u_T^j]^T$. The data is normalized to zero mean and unit standard deviation.

Because frame correspondences are the key to successful style translation, iterative motion warping is performed to find correspondences between the two stylistically different motions, child and adult motion. Each iteration has two steps: dynamic time warping followed by space warping. The two steps were iterated until the cost function comes to convergence.

The iterative motion warping procedure minimizes the following cost function:

$$E(\mathbf{a},\mathbf{b},\mathbf{W}) \equiv \|\mathbf{W}(\mathbf{U}\mathbf{a}+\mathbf{b}) - \mathbf{v}\|^2 + \|\mathbf{F}\mathbf{a}\|^2 + \|\mathbf{G}\mathbf{b}\|^2 \tag{3-5}$$

where $\mathbf{U}$ is $diag(\mathbf{u})$, $\mathbf{W}$ is the time warping matrix, $\mathbf{a}$ is the scale vector, and $\mathbf{b}$ is the offset vector. The matrices $\mathbf{F}$ and $\mathbf{G}$ provide weighted finite differences of $\mathbf{a}$ and $\mathbf{b}$ respectively.

Dynamic time warping (DTW) minimizes the Euclidean distance between the joint positions of the input and output motions. For each frame i of the input motion and frame i' of the output motion, the Euclidean distance between the input motion and the output motion is $d(i,i') = \sqrt{(u_i - v_{i'})^2}$, and the overall distance from $d(1,1)$ to $d(T,T')$ is $d = \sum d(i,i')$.

The dynamic time warping process optimize $\mathbf{W}$ independently of a and b. We initialize $\mathbf{a} = \mathbf{1}$, $\mathbf{b} = \mathbf{0}$. We define the Euclidean distance between input motion at time t_u and output motion at time t_v as follow

$$d(t_u,t_v) = \sqrt{(u_{t_u} - v_{t_v})^2} \tag{3-6}$$

DTW find the optimal match between the two signal by finding the path from (1,1) to (T_u, T_v) that could minimize equation 3-7. The acceptance path only allow several move. vertical moves $(t_u,t_v) \rightarrow (t_u+1,t_v)$, horizontal moves: $(t_u,t_v) \rightarrow (t_u,t_v+1)$, diagonal moves: $(t_u,t_v) \rightarrow (t_u+1,t_v+1)$.

$$d = \sum_{t_u,t_v} d(t_u,t_v) \tag{3-7}$$

We then represent the warping path as matricies shown in equation 3-8. We denote r_i as The number of repeated frames at frame i. The input motion after warping can be computed as $\mathbf{W_u u}$

and the output motion after warping can be computed as $\mathbf{W_v v}$.

$$\mathbf{W_{u,v}} = \begin{bmatrix} 1_{r_1} & \dots & \dots & 0 \\ \vdots & 1_{r_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1_{r_T} \end{bmatrix} \tag{3-8}$$

where

$$\mathbf{1}_{r_i} = [1, 1, \dots, 1]^{\mathbf{T}} \ (r_i \ ones) \tag{3-9}$$

$$\mathbf{W_v^{-1}} = \begin{bmatrix} 1 & \dots & \dots & 0 \\ 0_{r_2} & 1 & \dots & 0 \\ 0_{r_2} & \vdots & \ddots & \vdots \\ 0_{r_n} & \dots & \dots & 1 \end{bmatrix} \tag{3-10}$$

where

$$\mathbf{0}_{r_i} = [0, 0, \dots, 0](r_i \ zeroes) \tag{3-11}$$

While [31] only allow repetitions, we modified the time warping step to include deletions as well. This modification was necessary for the type of alignment we needed to see: for example, we noticed that a child's jump had a longer anticipatory phase and a shorter actual jump compared to an adult's jump.

In the space warping stage, we equate the first derivative of cost function E with respect to $\mathbf{a}$ and $\mathbf{b}$ to zero. The parameter $\mathbf{W}$ from the time warping stage is fixed. This leads to the following normal equation:

$$\begin{bmatrix} \mathbf{U^T W^T W U + F^T F} & \mathbf{U^T W^T W U} \\ \mathbf{W^T W U} & \mathbf{W^T W + G^T G} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{U^T W^T v} \\ \mathbf{W^T v} \end{bmatrix} \tag{3-12}$$

We solve this equation via LU (lower-upper) factorization using MATLAB's built in function

linsolve.

In the style translation stage, given a sequence of input frames, we estimate a sequence of output frames as follows:

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u'}_t + \mathbf{e}_t \\ \mathbf{v}_t = \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u'}_t + \mathbf{f}_t \end{cases} \tag{3-13}$$

where $\mathbf{u}' = \mathbf{W}\mathbf{u}$, $\mathbf{x}_t$ is the state of the system at frame t, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ are constant system matrices. $\mathbf{e}_t^j$, $\mathbf{f}_t^j$ are white noise terms. We treat each joint separately and trained different models for different joints. The training data are fed into MATLAB's system identification method, *ssest* [**?**], to estimate parameters in matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$. The model order is set to 40. The state vector $\mathbf{x}_t$ is initialized as zero vector. Once the model parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ have been learned, any input adult motion capture sequence can be style translated using this model.

The result of our algorithm was translated into Euler angles using the method described in 3.1. The result of our algorithm was then converted into a .bvh file where it can be directly imported into Maya.

3.3.2 Results

We show the results of the style translation algorithm on the action "Jump High" from the dataset of 3.1. Child actor 723 and adult actor 734 are used as training data, while adult actor 934 is the new input data. Because these motion sequences are very short (40-50 seconds), we repeated each sequence 28 times after iterative motion warping, before estimating the system matrices.

The result of the motion "Jump High" is shown in Figure 3-3. The skeleton in green is the translated motion. The figure captures the moment when both adult and child are bending to the lowest point in the landing phase of the jump. Compared to adult 934 which has the green skeleton, the result of the algorithm in the red skeleton and the child 723 in the blue skeleton both bend lower. Also, adult 934 does not bend his knees to the same extent. Finally, the motion of child 723 and the result of our algorithm both spend longer in the landing phase because it takes

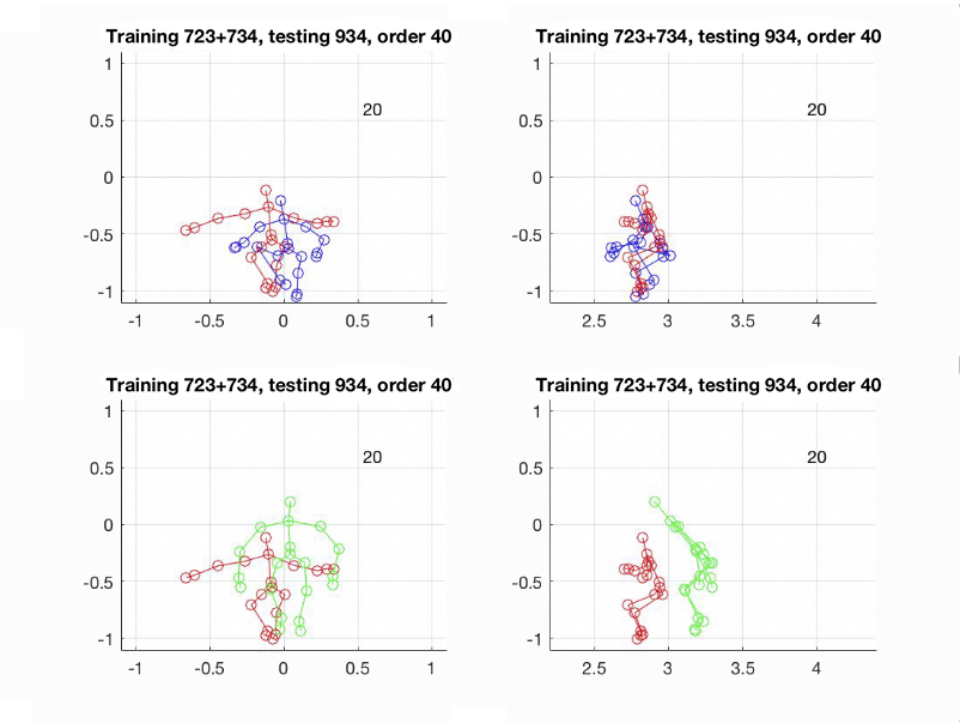


Figure 3-3. Results of the style translation algorithm in stick figures. Child motion (red), adult motion (green), our result (blue).

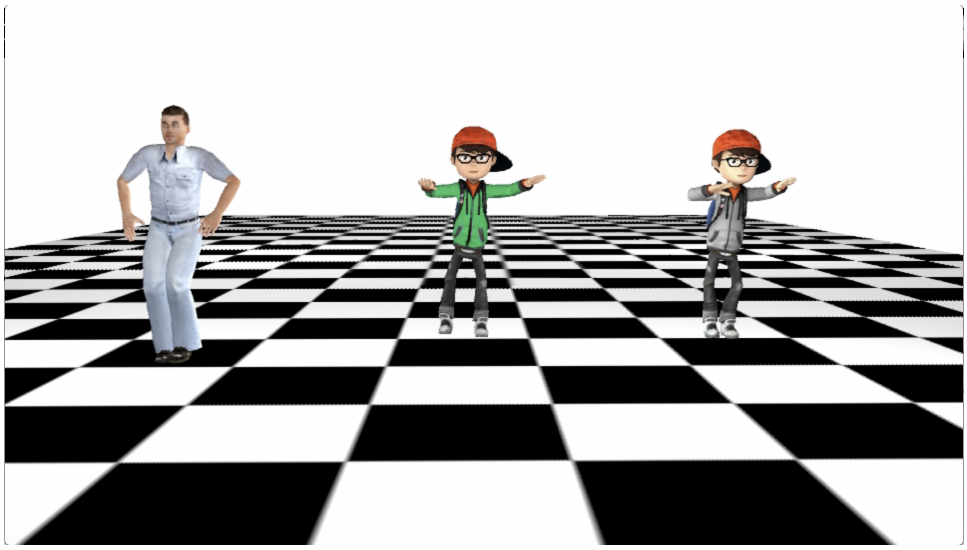


Figure 3-4. Results of the style translation algorithm on rendered characters. Training data for adult motion is from actor 934, and for child motion is from actor 723. Our result motion is in green.

longer to bend to a lower position.

In the rendered result in Figure 3-4, the arm of child 723 (grey jacket) and our result (green jacket) both swing to a larger extent compared to adult 934. Thus, we see that the gesture and the timing of our translated motion is very similar to child motion.

3.3.3 Summary

We generated motion for animated child characters by algorithmically transforming adult motion capture data to child-like motion capture data. We showed that a warping based style translation algorithm could create child-like motion that mimics the original child motion. This method can be used for a variety of applications, such as games, movies, educational content, training and simulation environments. As future work along this direction, we can apply this algorithm to more actions and actors, and to evaluate how close the result is to actual child motion by conducting a perception study.

3.4 Perceptual Evaluation

We evaluated the results from the linear scaling algorithm using perceptual metrics. We first looked at if the motion appears natural and realistic to viewers, and if the motion breaks any kinematic constraints or physics constraints. Second, we evaluate if the translated motions are similar to the reference motions. More specifically, how child-like are translated adult motions look like compared to real child motions. We visualized the motions as stick figures and rendered the snapshots of keyframes. We also rendered the motions on high fidelity characters to further inspect the motions in detail.

We conducted two perception studies to evaluate how effective the dynamic scaling procedure in creating childlike motion from input adult motion. The protocol was approved by our Institutional Review Board (IRB). We adopted within-study designs to test our hypothesis that dynamically scaled adult motion is identified as child motion.

In both studies, each participant watched videos from three actor types (adult, child, dynamically scaled adult). In each actor type, participants watched videos from four actors performing six actions (“Run Fast”, “Walk”, “Jump as High as You Can”, “Jump High Fly Like a

Bird", "Wave" and "Jumping Jacks"). The videos were shown in random order. Participants were asked three questions after watching each video. The first one "Does this motion belong to a Child or an Adult?" offered two options to select from: "Adult", "Child". In Study 2, participants were also asked to indicate their confidence level on a 7-point unipolar Likert scale. This question was not asked in Study 1. Finally, participants were asked "What is the action being performed?" by entering free-form text. The purpose of this question was to ensure that they watched the videos. We generated the videos for dynamically scaled adult motion using the method described in Section 3.2. All the videos were played at 30fps.

In Study 1, participants watched videos from four actors for each actor type. In total, we had 72 videos (3 actor types $\times$ 4 actors $\times$ 6 actions). From a pilot study, we learned that participants experienced fatigue after watching 72 videos. Therefore, we created 4 surveys such that each survey contains 54 videos (54 videos = 6 actions $\times$ 3 child actors + 6 actions $\times$ 3 adult actors + 6 actions $\times$ 3 adult actors, dynamically scaled). We used the same adult actors for adult motion and dynamically scaled adult motion in each survey. Each actor appeared in three surveys and the actor number was counterbalanced across all the subsets of actor type. In total, 24 participants (13 female, age range 21 to 29 years, mean = 23.96, standard deviation (SD) = 2.9) completed the survey either for extra credit in a class or voluntarily.

Because participants watched videos from the same actor performing the same action twice (once as DS video and once as adult video), they may have noticed the similarity between DS motion and adult motion and this could have affected their judgment of whether the motion belongs to a child or an adult. To make sure that each participant saw the motion from an actor only once, we created Study 2. In Study 2, we used different actors for adult motion and DS motion. We created 4 different surveys in total, and the actor ID was counterbalanced with the same actor appeared in 2 surveys. In each survey, we included 12 adult videos from 2 adult actors (the actor IDs in the four surveys are 934 and 970; 734 and 934; 734 and 921; 921 and 970), 12 DS videos from two adult actors whose actor ID were different from the adult actors (the actor IDs in the four surveys are 734 and 921; 921 and 970; 934 and 970; 734 and 934) and 12 child

videos from two child actors in each survey (the actor IDs in the four surveys are 290 and 723; 337 and 644; 644 and 723; 290 and 337). We added one more question in the survey: "How natural do you find this motion is?" to understand if dynamic scaling procedure alters the naturalness of the input adult motion. We expanded the participant pool to the general public and recruited 80 participants on Amazon MTurk (23 female, age range from 21 to 59 years, mean=31.23, SD= 7.52). Each MTurk worker was compensated 3 dollars for completing the survey. According to pilot study, the survey takes about 30 minutes.

3.4.1 Study 1 Results

We first compared our results on adult and child motion data with Jain et al. [36]. They reported the percentage correctness for the question "Does the motion belong to a child or an adult?" and found that viewers can identify child motion from adult motion significantly above chance. We computed the same measurement in our experiment and compared with theirs. Because there was no correct response for dynamically scaled motion, we excluded the dynamically scaled motion for this portion of the analysis. For each participant, we computed the percentage correct responses for each actor type by counting the number of responses that matched the actual actor type and divided it by 18 (the number of videos in each actor type). The overall percentage correctness was the average of the two actor types for each participant. There is one participant whose percent correctness was below 50%. Hence, we excluded that participant from further analysis. The mean overall percentage correctness for the remaining participants is 67.63% (SD = 13.16%) (Figure 3-5).

We then ran Shapiro-Wilk test which showed us that the distribution of percentage correctness for child motion and adult motion were normal ($p = 0.8151 > 0.05$, $p = 0.8455 > 0.05$). Next, we ran one-tailed t-test for adult motion and child motion individually. We found that the percent correctness of both adult motion (mean = 68.36%, SD = 12.47%) and child motion (mean = 66.91%, SD = 14.06%) were significantly above chance(child: t = 5.7664, df = 22, $p < 0.0001$, adult: t = 7.0592, df = 22, $p < 0.0001$). The results were consistent with the findings was reported in Jain et al. [36]

We tested if dynamically scaled adult motion was recognize as belonging to a child. For each participant, we measured the number of responses when participants identified the motion as a child (child responses) for each actor type. The number of child responses is a measure of how child-like the motions are in each actor type. A Shapiro-Wilk test indicates that the number of child responses for each actor type (child, adult, and dynamically scaled (DS)) was normal ($p = 0.8151 > 0.05$, $p = 0.8455 > 0.05$, and $p = 0.1609 > 0.05$). We also computed the mean and standard deviation of the number of child responses for each actor type and action type (Table 3-3).

We ran a two-way repeated measures ANOVA on number of child responses to examine the influence of two independent variables: actor type (Child, Adult, DS) and action type ("Fly Like a Bird", "Jump High", "Jumping Jacks", "Run Fast", "Walk", "Wave"). We found a significant main effect by actor type ($F_{2,384}=57.57$, $p < 0.0001$). The results from Tukey HSD post-hoc test showed there was a significant difference between every two actor types: Adult vs. Child ($p < 0.001$), Child vs.DS ($p < 0.001$) and Adult vs. DS ($p=0.019 < 0.05$) (see Figure 3-10). There were 18 videos in each actor type. The child motion received more child responses (mean=12.04, SD=2.53 or alternately, mean = 66.89%, SD = 14.06%) than adult motion did (mean = 5.70, SD = 2.24 or alternately, mean = 31.67 %, SD = 12.44 %), which confirmed the finding from Jain et al. [36] that viewers are able to tell child motion capture data and adult motion capture data apart. The number of child responses for dynamically scaled motion (mean = 7.35, SD = 2.17, or alternately, mean = 40.83%, SD = 12.06%) is higher than that of original adult motion (mean = 5.70, SD = 2.24 or alternately, mean = 31.67 %, SD = 12.44 %) but less than that of child motion (mean = 12.04, SD = 2.53 or alternately, mean = 66.89%, SD = 14.06%). The dynamically scaled motion was more likely to be identified as a child than adult motion, but it was not a perfect replacement for motion that was recorded from a child (Figure 3-6).

We also found a significant main effect on action type in the two-way ANOVA. The results indicate that action type affects viewers perception of whether the motion belongs to a child ($F_{5,384}=3.21$, $p=0.0075 < 0.01$). A Tukey HSD post-hoc test shows that the following actions

differ significantly: “Jumping Jacks” vs. “Run Fast” ($p = 0.004 < 0.01$), and “Jumping Jacks” vs. “Jump High” ($p = 0.023 < 0.05$). “Jumping Jacks” action is more likely to be perceived as performing by an adult than “Run Fast” and “Jump High”. Jain et al. [36] pointed out that “Jumping Jacks” is a more coordinated action and adults are generally better at coordination than children. Hence, viewers might relate more coordinated action to adults which led to fewer number of child responses for the coordinated action. The number of child responses for “Run Fast” action was dominated by child responses from child actors (over half of child responses belonged to child actors (50.91%)). Because “Run Fast” action is a very rapid and energetic action, it requires balance control of the body. As suggested by Schaefer and his colleagues that adult outperformed children in balance control [74]. In “Jump High” action, half of the child actors did more than one jump. The more jumps a child actor perform, the higher likelihood that this actor is recognized as a child. For example, for child actor 290 who performed 4 jumps, the percent child responses is 73.91%; for child actor 337 who performed 3 jumps, the percent child responses is 60.87% ; for child actor 723 who performed 1 jump, the percent child responses is 52.17%; for actor 644 who performed 1 jump, the percent child responses is 39.13%.

We analyzed the confidence level with respect to actor type and did not find any significant difference among actor type. The results indicated that viewers are equally confident toward videos from different actor type. Hence our results of the number of child responses are comparable within different actor type.

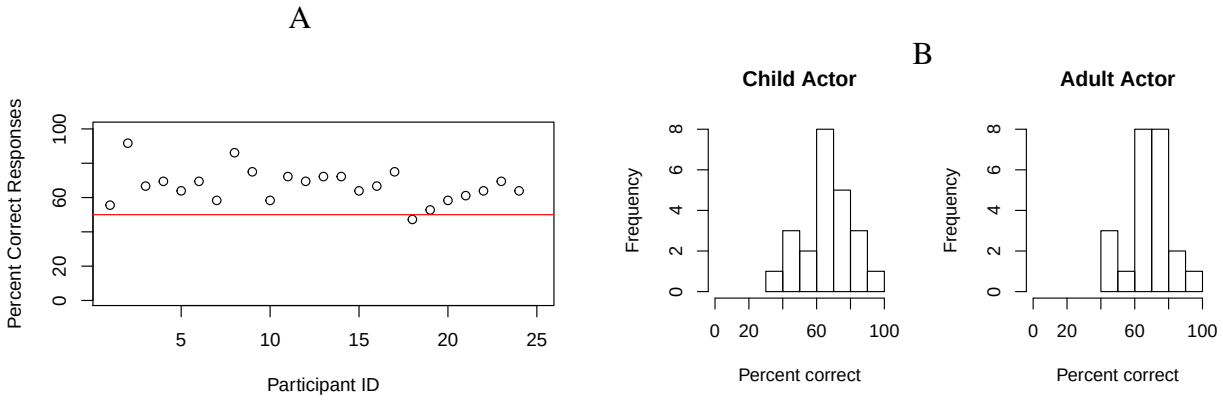


Figure 3-5. Distribution of the overall correctness for Study 1. A) the distribution of overall correctness for each participant. All participants except one have the overall correctness above chance (50%). B) Histogram showing the distributions of the percentage of correctness for child videos and adult videos.

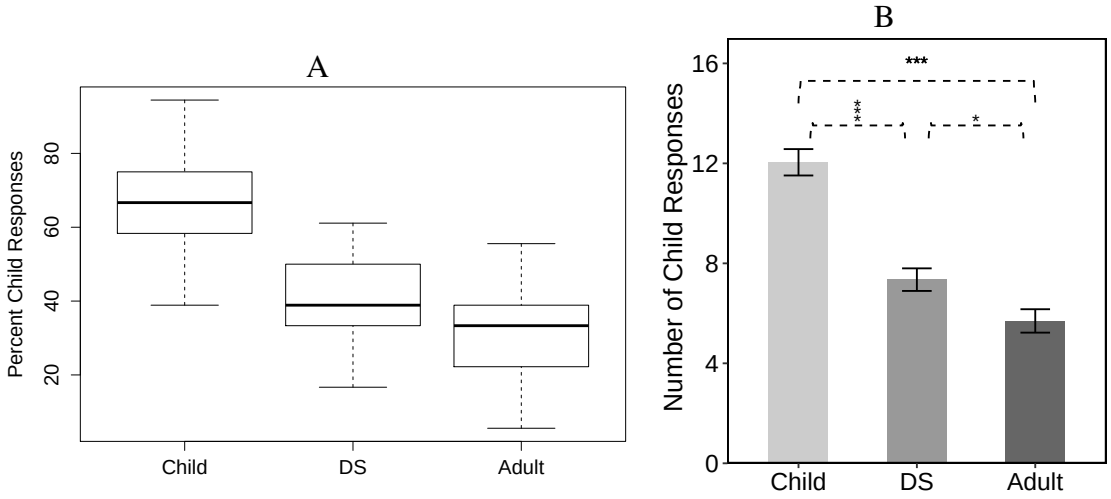


Figure 3-6. The distribution of the number of child responses in Study 1. A) Figure shows the boxplot of the distribution of number of child responses. The bold bar in the middle shows the median value. B) Bar chart contains the mean and standard error of number of child responses for all three actor types. The error bars represent the standard error. * indicates significance at the $p<0.05$ level, ** indicates significance at the $p<0.01$ level whereas *** indicates significance at the $p<0.001$ level.

Table 3-2. Means and standard deviations of number of child responses for each actor type in Study 1. Every participant watched 18 videos of each actor type. The maximal value of number of child response is 18.

Actor Type	Number		Percent	
	Mean	SD	Mean	SD
Child	12.04	2.53	66.89%	14.06%
Adult	5.70	2.24	31.67%	12.44%
DS	7.35	2.17	40.83%	12.06%

Table 3-3. Means and standard deviations of number of child responses for each action in Study 1. Every participant watched 9 videos of each action. The maximal value of number of child response is 9.

Action Type	Number		Percent	
	Mean	SD	Mean	SD
Fly like a bird	4.26	1.76	47.33%	19.56%
Jump high	4.57	1.56	50.78%	17.33%
Jumping jacks	3.22	1.54	35.78%	17.11%
Run fast	4.78	1.54	53.11%	17.11%
Walk	3.96	2.12	44%	23.56%
Wave	4.30	2.10	47.48 %	23.33%

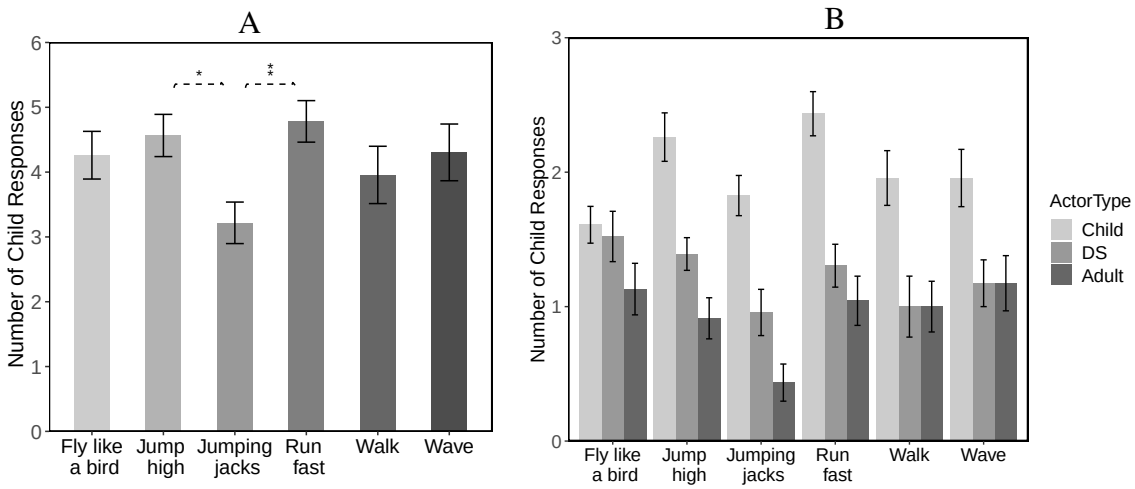


Figure 3-7. The distribution of the number of child responses for each action type and actor type in Study 1. A) Bar chart showing the mean and standard error of child responses for all six actions. The error bars represent the standard error. * indicates significance at the $p<0.05$ level, whereas ** indicates significance at the $p<0.01$ level. B) Bar chart showing the mean and standard error of number of child responses for each action and each actor type. The error bar represents standard error.

3.4.2 Study 2 Results

We computed the overall percentage correctness using the method described in Section 3.4.1. Overall percentage correctness from 30 participants out of 80 participants was below or equal to chance (50%). These results were not surprising considering that MTurk workers may complete tasks without being fully engaged as reported in Downs et al [18]. Therefore, they were excluded in the subsequent analysis. The mean overall percent correctness for the remaining participants is 67.63% (SD = 13.16%).

We then ran a Shapiro-Wilk test for adult videos and child videos. The results showed that the distribution of percentage correctness for adult videos (mean = 69%, SD = 17.82%) was normal ($p = 0.054 > 0.05$). However, the distribution of percentage correctness for child videos (mean = 63%, SD = 19.57%) was marginally normal ($p = 0.003 < 0.05$). We ran a one-tailed t-test for adult videos to check if the percentage correctness is significant above chance. We also a one-tailed Wilcoxon test for child videos for the same purpose. The results showed that participants' percentage correctness for both actor types were significantly above chance ($p < 0.001$ for both types of videos). The results are consistent with the results of Study 1.

We computed the mean and SD for number of child responses for each actor type and found similar results as Study 1. The child actor type received the most number of child responses (mean = 7.56, SD = 2.35 or alternatively, mean = 63.00%, SD = 19.58%) among the three actor types while the adult actor type received the least number of child responses (mean = 3.72, SD = 2.14 or alternatively, mean = 31.00%, SD = 17.83%). The number of child responses for DS actor type (mean = 4.90, SD = 2.78 or alternatively, mean = 40.83%, SD = 23.17%) was in between of the other two actor type. We found a significant main effect by actor type through a two-way repeated ANOVA. The results from Tukey HSD post-hoc test were same as Study 1: the difference between each two actor type was significant (Adult vs. Child ($p < 0.001$), Child vs. DS ($p < 0.001$) and Adult vs. DS ($p = 0.00115 < 0.01$)). This study confirmed that using different actors for adult motion and DS motion did not affect the difference in the number of child responses between different actor types.

Through a two-way ANOVA, we also found a significant main effect of action type ($p<0.001$). We computed the mean and standard deviation of the number of child responses for each action shown in Table 3-5. "Jumping Jacks" (mean = 1.76, SD = 1.42) received least number of child responses among all 6 actions, which was consistent with the result in Study 1. "Fly Like a Bird" received most number of child responses (mean = 3.30, SD = 1.71) among all 6 actions, while in Study 1 "Run Fast" action was recognized as performed by a child most often. Tukey HSD post hoc test found that number of child responses in following action pairs were significantly different: "Jumping jacks" and "Fly Like a Bird"($p <0.001$); "Jumping Jacks" and "Jump High" ($p <0.001$); "Run Fast" and "Jumping Jacks" ($p <0.001$); "Wave" and "Jumping Jacks" ($p <0.001$). In Study 1, we only found significant differences between "Jumping Jacks" and "Jump High" ($p <0.05$) and "Run Fast" and "Jumping Jacks" ($p <0.01$). We went back to the survey responses and found that for action "Fly Like a Bird", the motion was more likely to be attributed to a child in all three actor types. For "Jumping Jacks" action, videos from adult actor type and DS actor type received low child responses and therefore reduced the overall child responses for this action. The same trend can be found in action "Walk" which was a relatively slow and coordinated action. The adult motion and the DS motion were more likely to be attributed to an adult.

Similar as Study 1, We computed the mean and standard deviation for confidence score with respect to actor type. Compared to child motion (mean = 4.96, SD = 1.07) and DS motion (mean = 5.06, SD = 1.35), viewers are slightly more confident about adult motion (mean = 5.12, SD = 1.05). One-way ANOVA test showed that there is no main effect of actor type ($F(2,98)=0.08$). Participants are equally confident about their answer no matter the actor type.

We also asked the viewers to rate the naturalness of each motion on a scale of 1 (very unnatural) to 7 (very natural). We analyzed the naturalness score with respect to actor type. For each actor type, we computed the mean and standard deviation of naturalness score for all the videos from all the participants (Table 3-5). The data indicate that adult motion was considered most natural while child motion was considered least natural. Naturalness score for DS motion

was in between adult motion and child motion but closer to the naturalness score of adult videos. A one-way ANOVA test found that there is a main effect with respect to actor type ($F(2,98)$ ¡ 0.0001). Tukey HSD post hoc test found a significant difference in naturalness score between adult motion and child motion ($p<0.0001$) as well as DS motion and child motion ($p<0.001$). However, there was no significant difference between the naturalness score of DS motion and adult motion. Because the natrualness score is measured in 7-point Likert scale and the value differed in decimal points scale, we can conclude that the motion from different actor types are similarly natural.

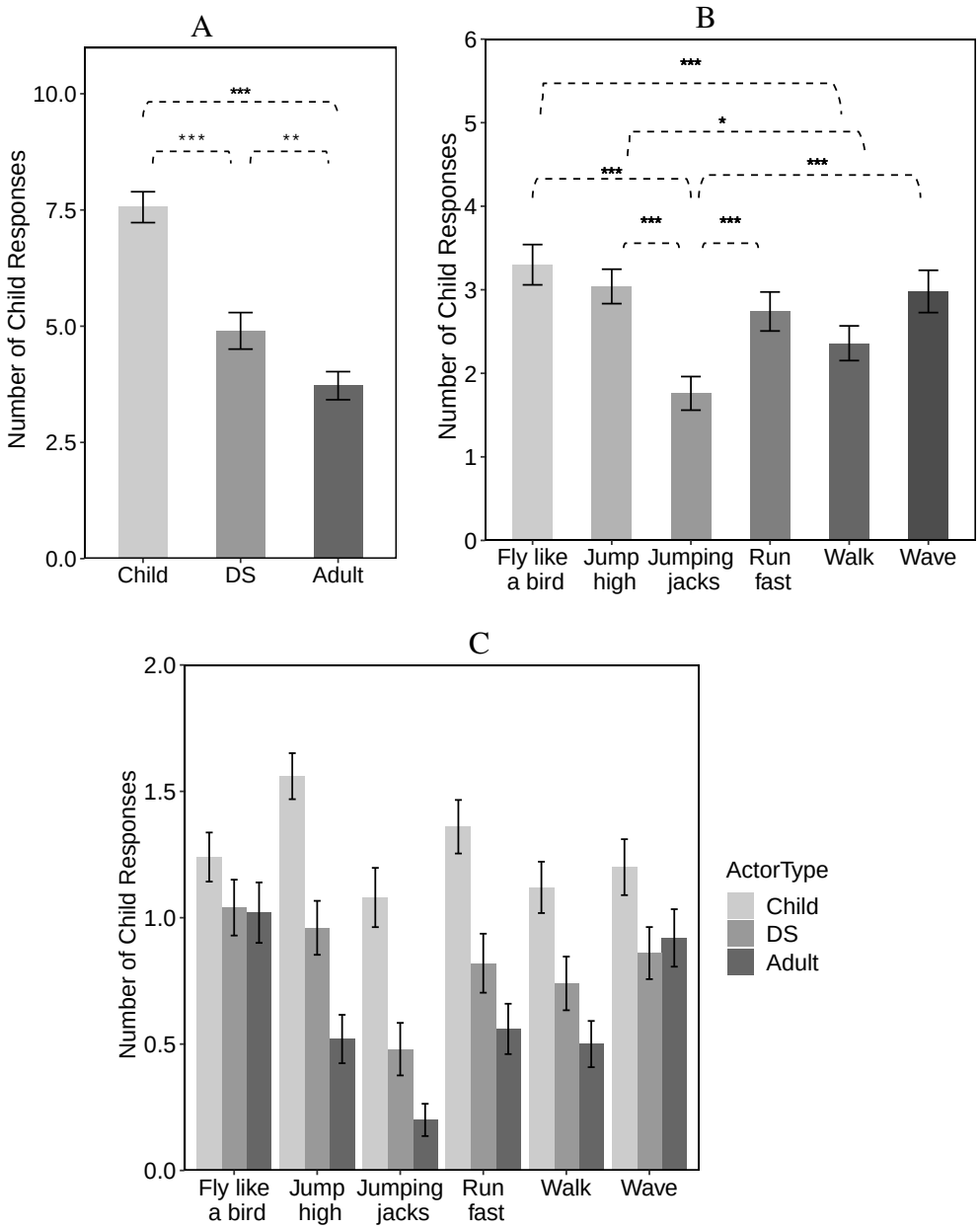


Figure 3-8. Bar charts of the number of child responses for each action type and actor type in Study 2. A) Bar chart contains the mean and standard error of number of child responses for all three actor types. B) Bar chart contains the mean and standard error of number of child responses for all six action types. C) Bar chart contains the mean and standard error of number of child responses break down by actor type and action types. The error bars represent the standard error. * indicates significance at the $p<0.05$ level, ** indicates significance at the $p<0.01$ level whereas *** indicates significance at the $p<0.001$ level.

Table 3-4. Means and standard deviations of number of child responses, confidence score and naturalness for each actor type in Study 2.

Actor Type	Child Responses		Confidence		Naturalness	
	Mean	SD	Mean	SD	Mean	SD
Child	7.56	2.35	4.96	1.07	4.63	1.13
Adult	3.72	2.14	5.12	1.05	5.09	0.96
DS	4.90	2.78	5.06	0.96	4.94	0.99

Table 3-5. Means and standard deviations of number of child responses for each action in Study 2.

Action	Mean	SD
Fly like a bird	3.30	1.71
Jump high	3.04	1.46
Jumping jacks	1.76	1.42
Run fast	2.74	1.65
Walk	2.36	1.47
Wave	2.98	1.79

3.4.3 Study 3: child2adult

During the presentation of my work in MIG 2017 3.1, I received a question: " Have you tried to use dynamic scaling laws to scale a child back to adult?" To answer this question, I did another micro study that scale a child motion to an adult.

Since it is a preliminary work to see if the age regression effect exist when we scale child back to adult, I included a subset of motions that are used in Study 1 and Study 2. The motions include: "Jumping Jacks", "Jump high" and "Run Fast". I used the same 4 adult actors and 4 child actors as Study 1 and Study 2. We noticed that for "Jump high" action, three out of four child actors did more than one jump and which lead to a increasing percentage of child responses. So in this study, I want to find out what if we only keep the first jump, will that reduce the child responses and by how much.

I generated dynamically scaled child motions from original child motions by scale them to an adult of height 1.75 meters. The scaling procedure are the same as Study 1 and Study 2. Then I generated 36 point light display (3 actor type $\times$ 4 actors $\times$ 3 actions) and created a perception study on MTurk which follows the same guideline as Study 1 and Study 2.

We did the same validation procedure. We computed the overall percentage for child videos

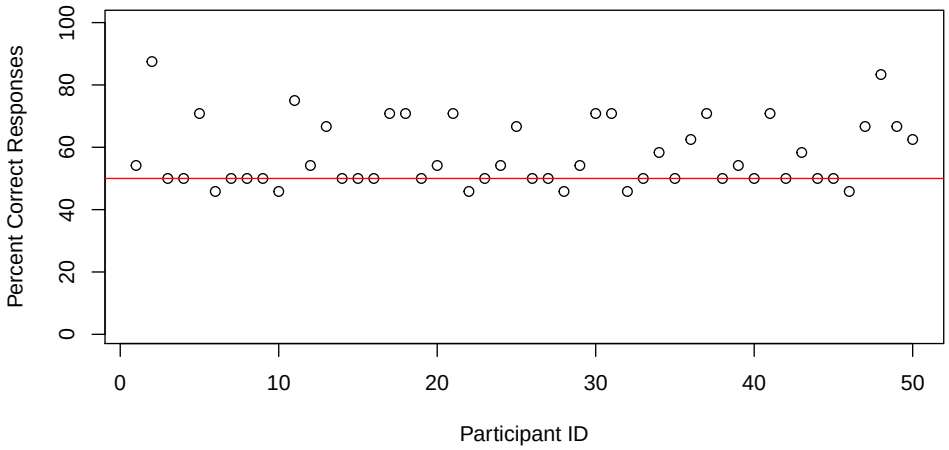
and adult videos. We discarded participants whose overall accuracy are below 50%. The overall accuracy for the rest of the participants is mean = 65.83%, SD = 27.93%.

Shapiro-Wilk test shows that the percentage child responses of child motions belongs to a normal distribution (W = 0.9325, p = 0.10). However the distribution for adult motion is not normal (W = 0.90, p-value = 0.02). For DS videos, the percentage child responses follows a normal distribution (W = 0.96, p = 0.49).

We computed the mean and SD for number of child responses for each actor type(Table 3-6). The child videos received the most number of child responses (mean = 5.84 , SD = 3.37 or alternatively, mean = 48.66% , SD = 28.08%) among the three actor types. Adult videos received the least number of child responses (mean = 2.04 , SD = 1.64 or alternatively, mean = 17% , SD = 13.66%). The number of child responses for DS actor type (mean = 4.68 , SD = 3.05 or alternatively, mean = 39% , SD = 25.41%) was in between of the other two actor type. We found a significant main effect by actor type through a two-way repeated ANOVA. The results from Tukey HSD post-hoc test were same as Study 1: the difference between each two actor type was significant (Adult vs. Child ($p < 0.001$), Child vs. DS ($p < 0.0301$) and Adult vs. DS ($p < 0.001$)). See Figure 3-10.

Through a two-way ANOVA, we also found a significant main effect of action type (p = 0.0014). We computed the mean and standard deviation of the number of child responses for each action shown in Table 3-7. “Jumping Jacks” (mean = 3.16, SD = 1.42) received least number of child responses among all 6 actions, which was consistent with the result in Study 1. “Jump High” received most number of child responses (mean = 4.76, SD = 3.03). Child responses for action ”Run fast” is: mean = 4.64, SD = 2.27 Tukey HSD post hoc test found that number of child responses in following action pairs were significantly different: “Jumping jacks” and “Jump high”(p = 0.0013); “Run Fast” and “Jumping Jacks” (p = 0.003) (Figure 3-11).

A



B

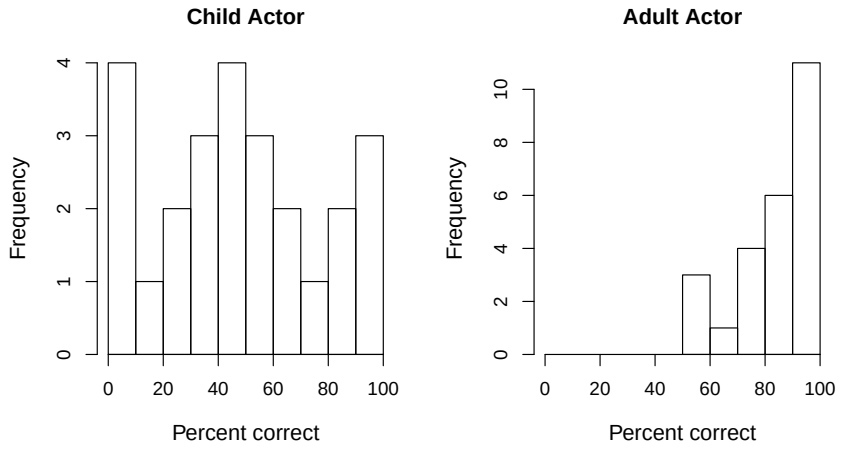


Figure 3-9. Distribution of the percent correct responses in Study 3. A) the distribution of overall correctness for each participant. B) Histogram showing the distributions of the percentage of correctness for child videos and adult videos.

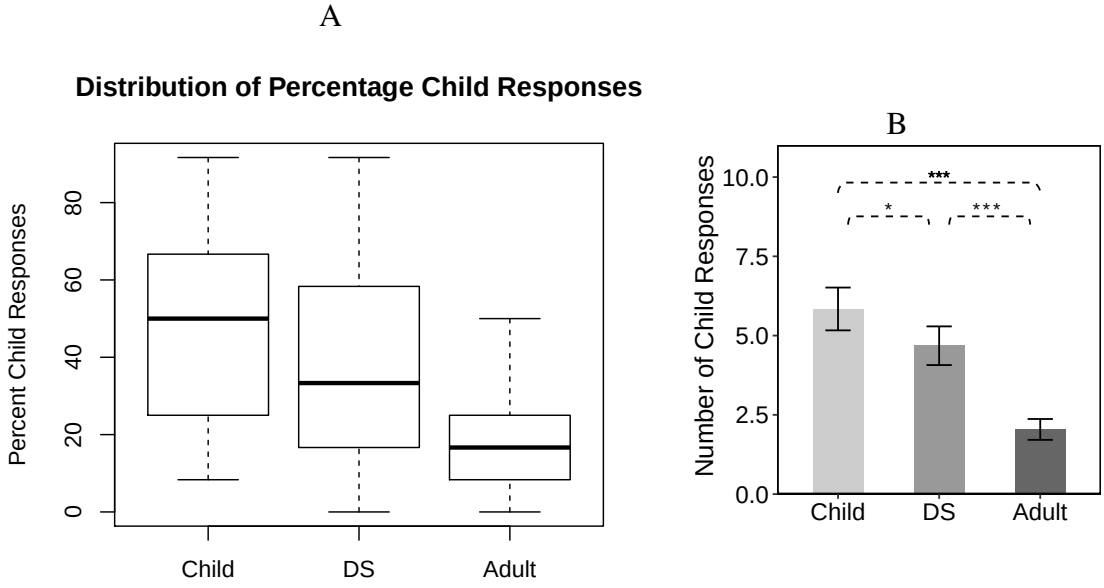


Figure 3-10. Figures of the number of child responses for each action type in Study 3. A) Figure shows the boxplot of the distribution of number of child responses in Study 3. The bold bar in the middle shows the median value. B) Bar chart contains the mean and standard error of number of child responses for all three actor types. The error bars represent the standard error. * indicates significance at the $p<0.05$ level, ** indicates significance at the $p<0.01$ level whereas *** indicates significance at the $p<0.001$ level.

Table 3-6. Means and standard deviations of number of child responses for each actor type. Every participant watched 12 videos of each actor type. The maximal value of number of child response is 12 in Study 3.

Actor Type	Number		Percent	
	Mean	SD	Mean	SD
Child	5.84	3.37	48.66%	28.08%
Adult	2.04	1.64	17%	13.66%
DS	4.68	3.05	39%	25.41%

Table 3-7. Means and standard deviations of number of child responses for each action in Study 3. Every participant watched 12 videos of each action. The maximal value of number of child response is 12.

Action Type	Number		Percent	
	Mean	SD	Mean	SD
Jump high	4.76	3.03	39.66%	25.25%
Jumping jacks	3.16	2.54	26.33%	21.16%
Run fast	4.64	2.78	38.66%	23.16%

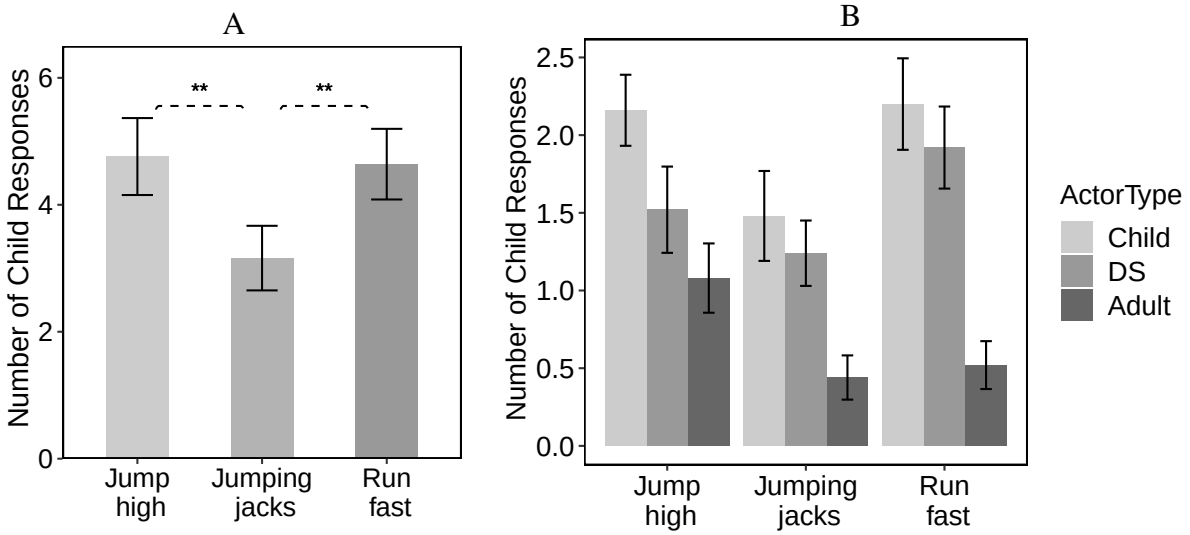


Figure 3-11. Bar charts of the number of child responses for each action type in Study 3. A) Bar chart contains the mean and standard error of number of child responses for three action types. B) Bar chart contains the mean and standard error of number of child responses break down by actor type and action types. The error bars represent the standard error. * indicates significance at the $p<0.05$ level, ** indicates significance at the $p<0.01$ level whereas *** indicates significance at the $p<0.001$ level.

3.4.4 Discussion

We generated dynamically scaled motion from adult motion . Through two perception studies, we examined if the dynamically scaled motion was perceived as child motion compared to original adult motion and original child motion.

In all three studies, we found a significant main effect of the actor type. In Study 1 and Study 2, the dynamically scaled motion was perceived as a child more often than an adult motion. Although viewers were not completely convinced that dynamically scaled motion was performed by a child, it is nevertheless a step in the right direction. For the “Jumping Jacks” action, viewers

are generally more likely to attribute dynamically scaled adult videos as being performed by a child, and less likely to attribute the original adult videos as being performed by a child. However, for the "Walk" and "Wave" actions, viewers are equally likely to attribute dynamically scaled adult videos and the original adult videos as being performed by a child. To better understand how the action being performed affects viewers' perception, we rendered the child, adult, and dynamically scaled adult videos for all 6 actions on wooden androgynous figures (Figure 3-12).

In Study 3, we found that dynamic scaled child motions received significantly less child responses than original child motion, but not as much as than original adult motions. This finding indicate that dynamic scaling laws works in both directions: adult2child and child2adult. Viewers perceived slower motions as from an adult.

We also found a significant main effect of action type. The results indicate that viewers are biased by the action performed no matter the actor type. It is not hard to understand as some motions (such as "fly like a bird") are more easily seen on children than adult and viewers would more likely to attribute these kinds of motion to children regardless of the actors' age.

Although we did not find any interaction effect between actor type and action type in both study, we noticed that the difference in the number of child responses between the actor type vary for each action. For faster and more dynamic motions such as "Jump High", "Jumping Jacks" and "Fly Like a Bird", "Run Fast", viewers are more generally attribute dynamically scaled motion as performed by a child. For slower motions like "Walk" and "Wave", the dynamically scaled motion is more likely to be identified as belonging to an adult. We learned from previous research in physiological gait that children usually move faster than adult [14]. Hence, we assumed that viewers might perceive faster actions as belonging to children. Dynamic scaling procedure took advantage of this motion characteristic from children and modified the adult motion in a way that the time actor complete each action is shortened, and the overall speed increased, which explains why dynamically scaled motion received more child responses than original adult motion. However, for slower motion such as "Walk" and "Wave", although dynamically scaled actor completed the action faster, the visual difference between a dynamically scaled actor and an adult

actor is not that noticeable. The presented method works better for fast dynamic motions than slow motions.

In Study 1, we used the same adult actors in both adult motion and dynamically scaled motion. In other word, participants watched the same actors performing the same actions twice (one in normal speed as adult motion and one in faster speed as dynamically scaled motion). To ensure that the results of Study 1 was not affected by participants recalling the similarity, we conducted Study 2 where we used different set of actors for adult motions and dynamically scaled motions. However, we found similar results in both studies: dynamically scaled motion is more likely to be recognized as child motion than adult motion significantly. This finding implies that when animators implementing dynamically procedure in the animated movies or video games to create child-like motions, they can use the same adult actors that appeared in the scene previously and dynamically scale the same actors. It is not necessary to find different set of actors to create child-like motions because viewers would not find a difference of using the same set of adult actors or not. This will tremendously reduce the overhead of recruiting more actors.

3.4.5 Summary

We investigated the problem of generating child-like motion from an existing adult mocap dataset. We found that by implementing a simple scaling procedure, we could trick viewers into believing the motion comes from a child. We presented the details of the method, implementation and perception study. Results showed that viewers judged dynamically scaled adult motion as belonging to a child significantly more often than the original adult motion, although still less than the original child motion. However, considering the simplicity of implementing dynamic scaling procedure: no training data required and work in real-time, this method is a good preliminary step to create childlike motion. It can be applied in wide range of applications that involves child characters such as gaming, movies, virtual characters in museums and so on.

As a future work along this direction, we could recruit participants from a more diverse background including parents and teachers. Our method did not take into account the difference in child and adult limb ratios and mass distributions. Our algorithm also did not modify other

properties such as force, stiffness, and mass. Future work might investigate more algorithms that take account of these factors.

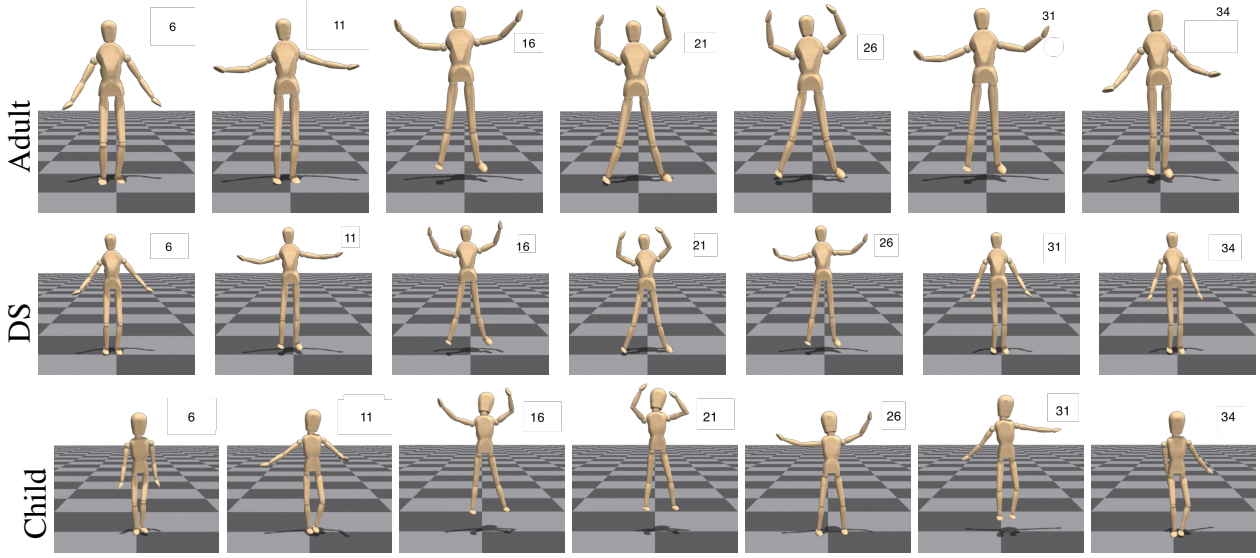


Figure 3-12. Rendered results on child actor 290, dynamically scaled adult actor 921, adult actor 921, performing the “Jumping Jacks” action. Both the child and the dynamically scaled adult complete one jumping jack faster than the adult. In frame 34, the child and DS actors have their arms back down by the side, but the original adult has yet to complete that jump.

CHAPTER 4
EXPRESSIVE EYES: PERCEPTUAL EVALUATION

In this part, we switched gear to focus on eye motions. Eyes are a critical piece for truly compelling virtual avatars, especially avatars intended for close-up communication. To quote, "oculus animi index", in other words, the eye is the window to the soul. Artists carefully create detailed appearance models for their characters' eyes, and spend hours on keyframing to have them move just right [41]. There is much work on algorithms and techniques to create eyes that feel alive [46, 4, 72]. Research has advanced the appearance of eyes to the point of hyper-realism [82, 7, 6] while on the other hand, matching advances have been made on the animation of the eyes [73].

Eyes are typically animated as eyeballs that are rigged to follow a desired target, or to match eye-tracking data from an actor. Previous research has focused on making the dynamics of the eyeball realistic [95], as well as the motions of the eyelids [86, 8]. With advances in eye tracking, VR headsets are being designed with eye trackers built into them, for example, the Tobii Eye Tracking VR and FOVE Eye Tracking VR headset [83, 22]. As a result, the sensors needed to record cues from the eyes can be taken for granted. In particular, it is possible to record not only gaze direction and blinks, but also subtle cues such as pupil dilation. In this work, our first motivation is a pragmatic consideration: though a VR system will have access to this data, should it use the data to drive the social virtual avatar? And if so, which components of this data make a difference?

There is now work on adding subtle pupil dynamics, such as pupil jitter [38] and pupil light reflex [62]. While this past research has performed perceptual studies to show that adding realistic dynamics to the eyeballs, eyelids, and pupils improves the perceived realism and naturalness of a virtual avatar, there has been little investigation into how cues in the eyes impact the perceived internal state of a virtual avatar. We therefore aim to answer the question: do cues as subtle as pupil size impact if an avatar is perceived as scared? Such an investigation is relevant because each cue adds potential complexity to the modeling of a virtual avatar.

The most famous observation about the pupil is that it dilates when you look at someone you love. Less famously but equally importantly, pupil dilation is an indicator of stress and

emotional arousal [11, 10] as well as cognitive load [51, 21]. As a result, the pupil has been used in computational analysis, for example as a physiological index for affective video processing [68, 40]. While the variation in pupil size is a useful feature for computational processing, our motivation for this work is to determine whether it is an intrinsic social cue that humans use to make judgments about the avatar.

Through three studies, we demonstrate the importance of high fidelity pupillary dynamics on how viewers perceive a computer generated avatar. Ours is the first work to consider pupillary dynamics toward the goal of generating eyes that are expressive, not just realistic. This paper makes three contributions: first, we find that naive viewers can identify if an avatar is scared from only the information contained in its eyes, second, we find that naive viewers can make this judgment even if gaze direction and blink rate are not available as a cue, third, we show that correct pupil dynamics are a critical perceptual cue. These findings have implications for digital avatars in social virtual reality, conversational agents, and social robotics.

We make three contributions:

1. We are the first study to investigate how subtle social cues in the eyes communicate the internal emotional state of an avatar, as different from prior art that investigates the avatar's overall realism, naturalness, personality, and trustworthiness.

2. We report a sequence of large scale studies (N≈500) that investigate the role of subtle cues in an avatar's eyes for communicating if this avatar is scared.

3. We find that a cue as subtle as the variation in pupil size is a significant perceptual cue in addition to gaze and blinks. These findings underscore the importance of high fidelity pupillary dynamics for expressive avatars in animated movies, conversational agents, and social virtual reality.

The work on creating realistic eyes and gaze movement for social avatars also has applications for conversational agents and social robots. A recent review by Ruhland and colleagues [73] discusses this literature in detail. Steptoe et al. [79] have shown that high fidelity

visual representation of avatars will increase accuracy with which people identify the avatar's direction of gaze, together with subjective reports of authenticity. Ham et al. [24] show that gaze and gestures are two crucial strategies that help conversational agents become more persuasive during storytelling. Our findings provide motivation for future work towards creating realistic as well as artist-directable models of pupillary dynamics as a way to impact this diverse set of application domains.

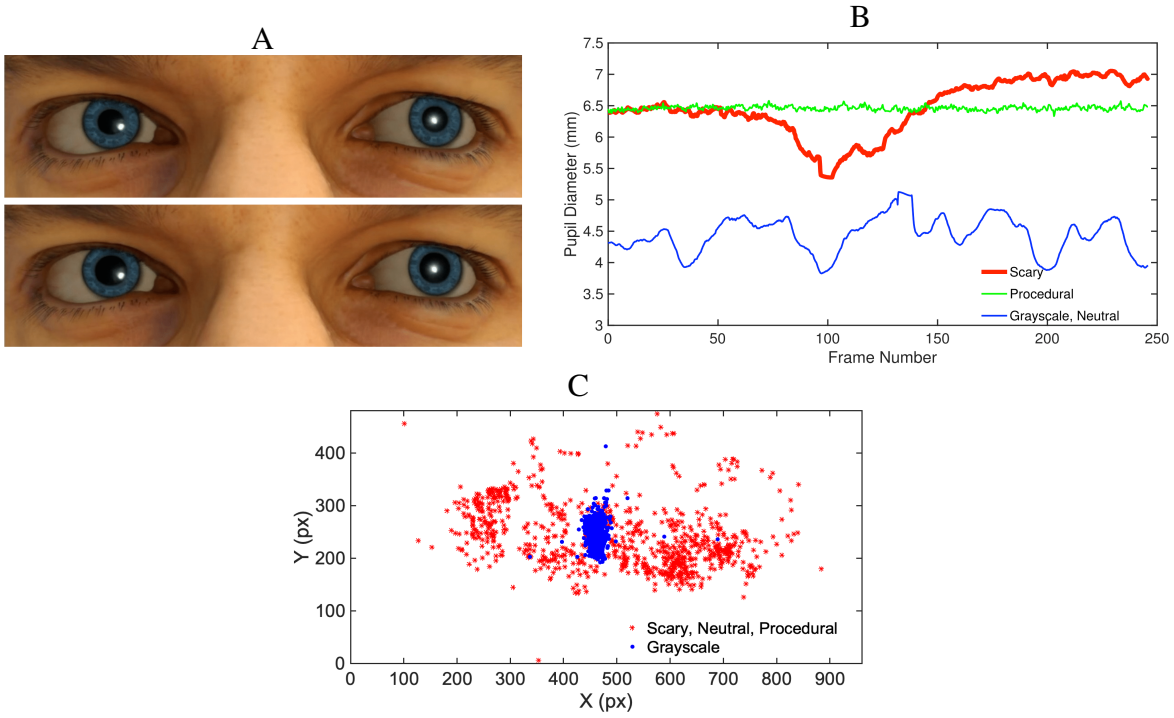


Figure 4-1. We present three studies that that investigate how increasingly subtle social cues in an avatar's eyes impact our judgment of its internal emotional state. The avatar for one of our eye motions (s015) is pictured here for Condition Scary (A top), where gaze, blinks and pupil dynamics were recorded when the person was watching a scary movie, and Condition Procedural (A bottom), where gaze and blinks were the same as Scary, but the pupil diameter was generated procedurally. The numeric values of pupil diameter (B) and gaze positions (C) are graphed for the different experimental conditions in the three studies.

4.1 Experimental Framework

Eye-tracking Data We used the dataset of Raiturkar and colleagues to obtain eye-tracking data for people watching a scary movie. Their dataset contains eye motions from ten people

watching excerpts from movies and YouTube videos[1]. Details of the data collection protocol are available in the associated publication [68]. We selected an excerpt from a zombie movie (*Decay*) that is freely available online under a Creative Commons license. The ten second excerpt we selected contained a zombie attack on the heroine[2]. This attack was marked by Raiturkar as an arousal eliciting event in their publication. We agreed with their assessment of this event being the scariest zombie attack. From the set of available eye motionS, we selected those in which the pupils changed in diameter by more than 1.5mm in the three seconds following the zombie attack (ID: s004, s009 and s015). The selected eye motions were from people who had self-reported an engagement score greater than or equal to 7, when asked to rate their experience on a scale of 1 to 10 (1=least engaging). Finally, we checked that the recorded pupil diameter had less than 40% missing samples.

We obtained additional eye-tracking data from the same people as above. The difference was that this time they were watching a video comprising only grayscale slides without any semantic content. Thus we ensured that in this case there was no reason to expect that they were scared. The grayscale videos were generated as follows: for each person whose eye tracking data was collected, the foveal neighborhood of their gaze position in each frame of the zombie video was identified. The color values of the pixels in this neighborhood were converted into an average grayscale intensity. A corresponding frame was generated for the neutral video such that its grayscale value matched this average intensity. More details are available in the publication by John et al. [37].

We adopted the highly realistic human avatar from Jörg et al. [38]. The avatar's eyeball movement was driven by a point in a 2-D plane through a point constraint to mimic someone watching a screen. We adjusted the plane's size and distance from the character to match the real world eye tracking setup. As our virtual character is not at the scale of a real character, we used its iris size compared to the average iris size of a real person as a reference to scale the pupil sizes

[1]Dataset "Gaze Pupil Diameter III" available at http://jainlab.cise.ufl.edu/publications.html#GAZEPUPIL

[2]The time stamps are 66279ms to 76575ms relative to the longer excerpt available within the dataset of Raiturkar et al.

to match the character.

The captured data (gaze coordinates and pupil size) was manually cleaned by an artist. We extracted the blink occurrence and length from gaps in the eye tracking data. The scene was set up in Maya 2016 and rendered with Solid Angle's Arnold 5 renderer. The lighting was adjusted and the avatar was rendered in a close-up view focusing on the eye area (Figure 4-1). This view displays the eye movement in high detail. We used a resolution of 854×480 with a frame rate of 24 frames/second.

We ran our data collection under a protocol approved by the university review board (IRB). Respondents were redirected to Survey Gizmo to take an online survey. First, we asked demographics questions including age, gender and ethnicity. Then survey respondents were shown the stimuli videos. They were asked to watch each video in its entirety. Because each video was only 10.3 seconds long, we looped it four times and created a stimulus that was 41.2 seconds long to give survey respondents enough time to observe the details. We implemented a timer that constrained the survey respondents to spend at least 35 seconds on each page. The stimuli were presented in randomized order.

Respondents were asked two questions after each stimulus presentation. For the question "One of these people is watching a scary movie. Which one is it?", their response was recorded via a radio button with two options: "Top" or "Bottom". For the question "How difficult did you find this task?", the response was recorded on a 7-point Likert scale with 1(very easy) and 7(very difficult).

At the end, we asked two supplemental questions. The first question was multiple-choice: "How often do you play video games?" Five options were provided: "0 hours/week", "0-7 hours/week", "7-14 hours/week", "14-21 hours/week", "Above 21 hours/week". The second question was "How often do you watch animated movies or TV shows?" The options were the same as the first supplemental question. Finally, survey respondents were asked a free-form text question: "What cue(s) did you use to say whether the person is watching a scary movie?" All the questions in the survey were mandatory.

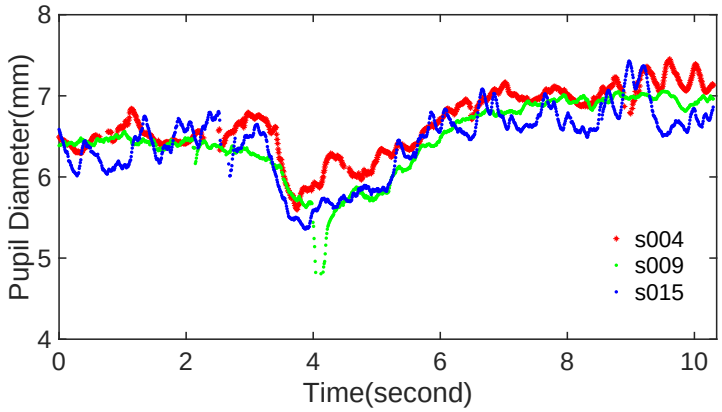


Figure 4-2. Raw pupil diameter recordings from the three selected eye motions of people watching the scary movie.

4.2 Study 1: Identify the Scared Avatar

In Study 1, we investigated if an avatar with the eye motions of people watching scary movies would be perceivably different from the same avatar with eye motions from people watching only light intensity changes. Our motivation was to establish if respondents can identify the scared avatar.

4.2.1 Description

The independent variable in this study was the video that was used to elicit the eye motions: a scary zombie video (Condition Scary) versus a video composed of only grayscale slides that changed in intensity value (Condition Grayscale). The gaze directions, blinks, and pupil diameters came from the corresponding eye-tracking data.

Stimuli videos generated from the same person were shown synchronized and aligned vertically. To control for order effects, we created stimuli videos with Condition Scary on top and Condition Grayscale at the bottom, as well as Condition Grayscale on top and Condition Scary at the bottom. In total, we had six different stimuli videos (3 eye motions $\times$ 2 orders). A screenshot is shown in Figure 4-1.

4.2.2 Results

We consented 104 survey respondents from Amazon Mechanical Turk. 1 survey respondent entered the wrong survey code. 4 survey respondents entered the code without taking the survey.

We received 99 complete survey responses. Respondents were compensated 0.8 dollars when they completed the survey. We excluded those responses where the IP addresses were repeated to have a total of 88 unique respondents. Our survey respondents were fairly well distributed by age (age range 19-72, μ = 31.89, σ =9.34), gender (32 female) and ethnicity (25 Asian, 49 White, 6 Black or African American, 2 Hispanic or Latino, 5 American Indian or Alaska Native, 1 Others).

We created two additional stimuli videos from different eye motions of a person (s010). These were presented such that the top and the bottom videos are identical. Each pair was presented twice, resulting in four dummy stimuli sets. Respondents were asked to respond to two easy questions for each stimuli set, one question from the Top/Bottom pair, and one question from the glasses/eye color pair. The questions were: "The person on the bottom is watching a scary movie. Click "Bottom".", "The person on top is watching a scary movie. Click "Top".", "What's the eye color of this person?" and "Is this person wearing glasses?". We only kept the data from a survey respondent if they got $>= 6$ questions correct. As a result, 76 respondents were kept for the remainder of the analysis.

For the remaining 76 survey respondents, we computed the number of correct responses as follows: we coded each row with the actual answer; if it matched the respondent's answer, then the number of correct responses was incremented. We computed the percentage of correct responses for all six stimuli videos for each respondent. Figure 4-3 (A) illustrates the respondents' of overall percentage of correct responses. The dotted line marks chance performance (50%). Figure 4-3 (B) illustrates the same distribution as a histogram. Table 4-1 shows the percent correct broken down by Eye Motion ID.

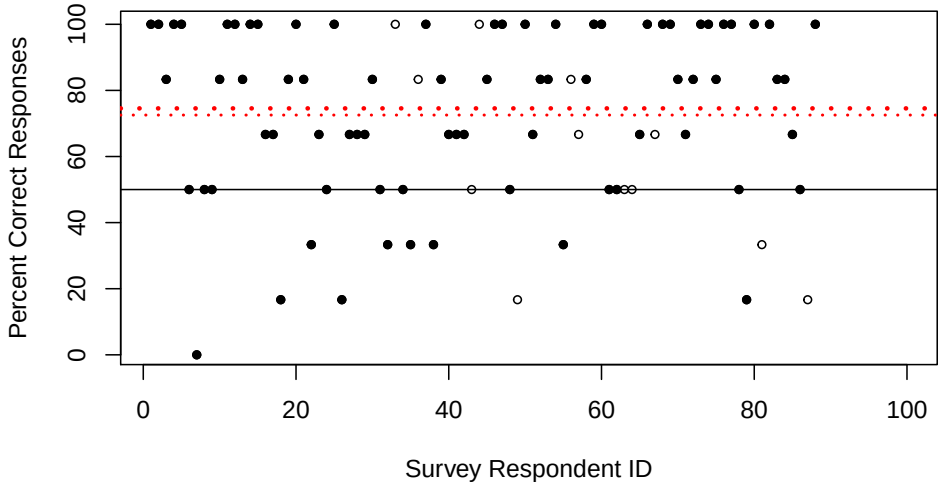
We used the statistical analysis package R to further analyze the data. The distributions of percentage correct responses were not normal (Shapiro-Wilk, W=0.86, $p < 0.001$), as seen in Figure 4-3 (Right). A one-sample Wilcoxon signed-rank test reported that the percent of correct responses (μ = 74.56%, σ = 25.88%) is significantly above chance (V = 1953.5, $p < 0.001$). These results demonstrate that survey respondents are able to distinguish the avatar that was watching a scary movie from the avatar that was watching a baseline neutral video.

We examined if respondents' accuracy in identifying the scared avatars was biased in favor of any individual person whose eye tracking data was used to generate the stimuli videos. We computed the mean and standard deviation of the percentage correctness by each Eye Motion ID in Table 4-1. There was no significant difference between the percentages of correct responses for each Eye Motion ID. A non-parametric Friedman test of repeated measures was conducted (Friedman χ^2 value of 1.11, $p = 0.57, p > 0.05$). As a result, we retained all three eye motions when we generated stimuli videos for subsequent experiments.

Table 4-1. Results of Study 1: identify the scared avatar. Means and standard deviation of percent correct responses broken down by Eye Motion ID.

ID	Percent Correct	
	μ	σ
s004	76.31%	32.12%
s009	73.68%	35.09%
s015	73.68%	35.09%
Overall	74.56%	25.88%

A



B

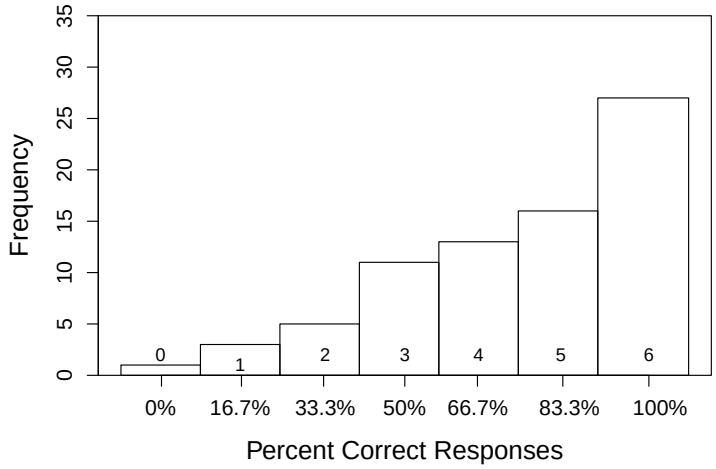


Figure 4-3. Study 1 results. A) Scatter plot of percent correct responses for all survey respondents. Grayed circles met inclusion criteria. The solid line marks chance performance. The thin dotted line shows the average percent correct responses for all respondents and the thick dotted line shows the average percent correct responses for the respondents met our inclusion criteria. B) Histogram of Percent Correct Responses for consistent survey respondents. The number of correct responses is shown inside the bars of the histogram.

4.3 Study 2: Pupil as a Cue

The human pupil changes diameter as part of the autonomic nervous system response to feeling scared or stressed. The pupil diameter also changes in response to changing scene

brightness. If the pupil was simply changing size in response to changes in brightness, would survey respondents still think the avatar was scared? In Study 2, we investigated this question by controlling for gaze and blink data while manipulating the source of the pupil diameter changes.

4.3.1 Description

The independent variable in this study was the source of the pupil diameter values: pupil diameter recorded while people were watching a scary movie (Condition Scary) versus pupil diameter recorded while people were watching a video composed of only grayscale slides that changed in intensity value (Condition Neutral). The gaze directions and blink rates were controlled. These values came from the recordings where the people were watching a scary movie for both conditions. Note that Condition Scary in this study is identical to Condition Scary in Study 1.

Our goal was not to test which cue (gaze direction, blink, pupil) dominates, rather, the goal was to test whether having the correct pupil diameter makes a difference to how often the avatar is judged as being scared. If the correct pupil diameter values make a difference to respondents' judgments, then the message to content creators would be clear: for truly expressive eyes, you need to get the pupil diameter correct.

Avatars were rendered as before. We showed survey respondents the Condition Scary video and Condition Neutral video in a vertical layout. The videos were synchronized to play simultaneously. To account for order effects, each respondent also saw each video pair again, this time with the top and bottom reversed. The questions asked were the same as the previous study. The study was conducted on Amazon Mechanical Turk under an IRB approved protocol. Our dependent variable was the response to "One of these people is watching a scary movie. Which one is it?" We calculated the metric Percent Correct Responses as before.

4.3.2 Results

In total, we paid 500 survey respondents \$0.8 for their time. After eliminating survey respondents with repeated IP addresses, there were 500 distinct respondents left (223 female, age range 18-80 years, μ = 36.04 years, σ = 11.89 years, 97 Asian, 308 White, 38 Hispanic or Latino,

36 Black or African American, 9 American Indian or Alaska Native, 12 Others).

We used the same validation procedure as in Study 1. 466 survey respondents were kept for the remaining analysis.

For the remaining 466 survey respondents, we calculated the metric Percent Correct Responses. The distribution of Percent Correct Responses by survey respondent was not normal (Shapiro-Wilk, W = 0.91, $p < 0.001$). The histogram in Figure 4-5 shows the distribution. As shown in the last row of Table 4-2, on average Condition Scary was selected as 'the person watching the scary movie' in 58.7% of the presentations ($\sigma = 32.6\%$), whereas Condition Neutral was picked in 41.28% of the presentations ($\sigma = 32.6\%$). A one-tailed Wilcoxon signed-rank test found that Percent Correct Responses for Condition Scary (μ = 58.7%, σ = 32.6%), averaged across the 466 respondents is significantly above chance (V = 50209, $p < 0.001$) Thus, we concluded that even if an avatar's gaze and blinks correspond to eye-tracking data while watching a scary movie, it matters where the pupil diameter values drawn from.

We checked if respondents' judgments of whether Condition Scary was the 'person watching a scary movie' depended on the person whose data was used to generate that avatar. For each respondent and each Eye Motion ID, we aggregated the number of times Condition Scary was selected as the response. Table 4-2 shows the mean and standard deviation of percent correct Responses for each condition broken down by Eye Motion ID. We found a main effect of Eye Motion ID (Friedman χ^2 = 135.03, $p < 0.001$). The mean values indicated that for s004 Condition Neutral was selected as the "person watching a scary movie" slightly more often than Condition Scary. If we refer to the average pupil diameters in Table 4-3, we see that the pupil was on average much larger in Condition Scary relative to Condition Neutral for both s009 and s015. In contrast, the average pupil diameter for s004 for the two conditions was more similar. This suggests that respondents could be basing their judgment on the average pupil size over the course of stimulus video. This factor was subsequently controlled in the next study.

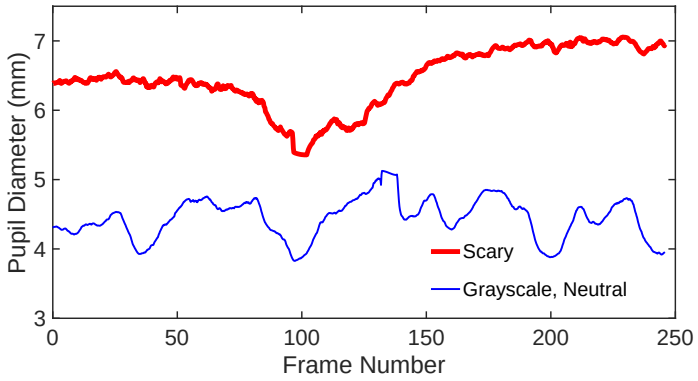


Figure 4-4. Pupil diameter values of Eye Motion ID s015 used in Study 2: pupil as a cue.

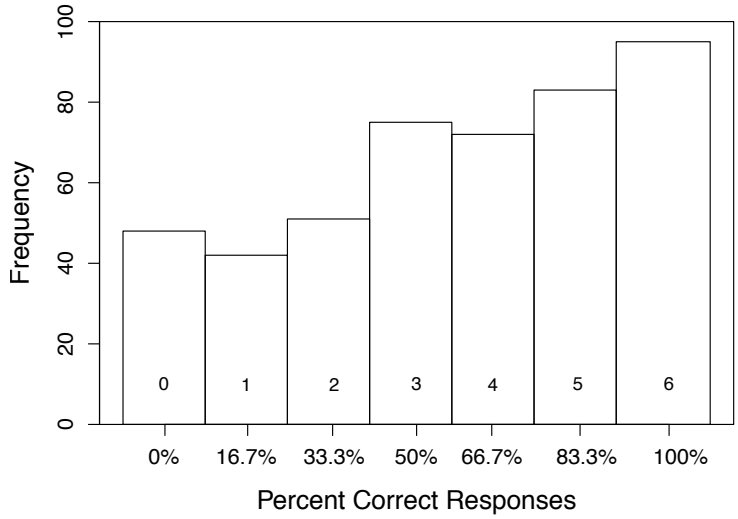


Figure 4-5. Results of Study 2: pupil as a cue. Histogram showing percent correct responses. The number of correct responses is marked inside each bar.

Table 4-2. Results of Study 2: pupil as a cue. Mean and standard deviation of percent correct responses broken down by Eye Motion ID.

ID	Percent Correct	
	μ	σ
s004	45.28%	40.68%
s009	67.70%	39.53%
s015	63.20%	40.73%
Overall	58.72%	32.63%

Table 4-3. Stimuli generation of Study 2: pupil as a cue. Mean and standard deviation of pupil diameter values for each Eye Motion ID.

ID	Condition							
	Scary				Neutral			
	μ	σ	Max	Min	μ	σ	Max	Min
s004	6.66	0.40	7.45	5.6	5.41	0.34	6.12	3.75
s009	6.44	0.43	7.05	4.81	4.23	0.40	5.12	3.83
s015	6.47	0.44	7.43	5.46	4.43	0.28	5.13	3.83

4.4 Study 3: Pupil Diameter Variation as a Cue

In Study 3, we investigated whether a pupil with the same average pupil diameter but a procedurally generated jitter would be as effective in conveying that the avatar is scared as replaying the originally recorded pupil diameter. In other words, if we control for the average size of the pupil, can respondents pick up on the pupil diameter variation as a cue.

4.4.1 Description

The independent variable in this study was the source of the variation in the pupil diameter values: pupil diameter recorded while people were watching a scary movie (Condition Scary) versus pupil diameter generated procedurally (Condition Procedural). The average pupil diameter was kept the same in both conditions (Table 4-4). The gaze directions and blink rates in both conditions were from eye tracking data when people were watching the scary movie.

To create the procedural pupil, we generated pink noise by following the approach by Duchowski et al. [20]. We used MATLAB's Colored Noise function with an alpha value of 1 to generate the noise. We added an offset to the noise to make the mean pupil diameter the same as Condition Scary. We scaled the noise to result in an average amplitude of 0.0023, a value found in Jörg et al. [38] for the same character with a nine point motion as that type of motion is most similar to our stimuli. The resulting values are illustrated in Figure 4-6.

The gaze direction and the blink data of the rendering were from the person watching the scary movie. We showed survey respondents two videos at a time in vertical layout from the same Eye Motion ID, and the videos played simultaneously. To minimize the order effect, we also created a corresponding video for each stimulus with the same avatars but reversed vertical layout.

4.4.2 Results

Study 3 was conducted separate from Study 2. In total, we paid 515 survey respondents \$0.8 to complete the study. After eliminating survey respondents with repeated ID addresses, there were 484 distinct respondents left (203 female, age range 18-70 years, μ = 35.11 years, σ = 11.09 years, 128 Asian, 282 White, 27 Hispanic or Latino, 30 Black or African American, 12 American Indian or Alaska Native, 5 Others).

We used the same validation procedure as in Study 1 and 2. 445 respondents were kept for the remainder of the analysis.

For each of the 445 respondents, we computed the metric Percent Correct Responses. If respondents identified the scared avatar as the one whose pupils were a replay of the original recording then this response was marked as a correct response. If they selected the avatar whose pupils were generated procedurally, it was marked incorrect. The distribution of Percent Correct Responses (Figure 4-7) was not normal (W = 0.93, p<0.001 for both types of the videos). A one-tailed Wilcoxon signed-rank test showed that the mean value of Percent Correct Responses was significantly above chance (μ = 59.77%, σ = 24.93%, W = 36775, $p < .001$).

Similar to Study 1 and Study 2, we examined if respondents' accuracy in identifying the scared avatars was dominated by any individual whose data was used to generate the avatars. The breakdown of Percent Correct Responses by Eye Motion ID is shown in Table 4-5. We did not find a significant main effect of Eye Motion ID (Friedman χ^2 = 3.61, $p = 0.16 > 0.05$).

Table 4-4. Stimuli generation of Study 3: pupil diameter variation as a cue. Mean and standard deviation of pupil diameter for each Eye Motion ID.

ID	Condition							
	Scary				Procedural			
	μ	σ	Max	Min	μ	σ	Max	Min
s004	6.66	0.40	7.45	5.98	6.65	0.03	6.76	6.54
s009	6.43	0.43	7.05	4.81	6.42	0.03	6.53	6.32
s015	6.47	0.44	7.43	5.46	6.46	0.03	6.58	6.33

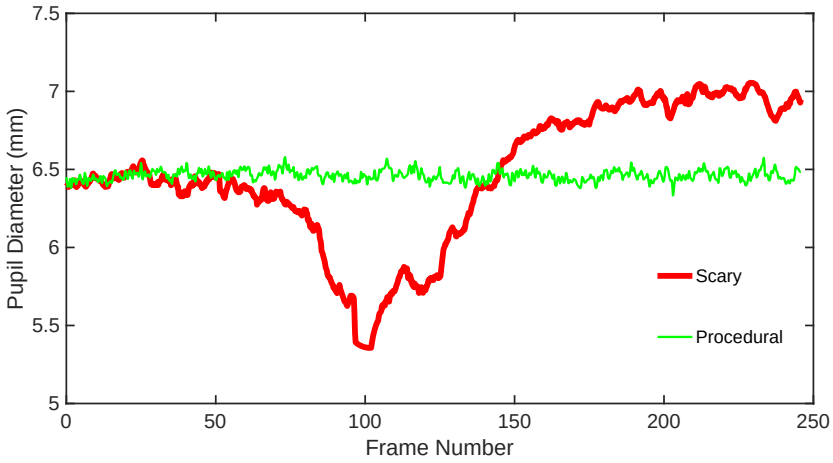


Figure 4-6. Pupil diameter values of Eye Motion ID s015 used in Study 3.

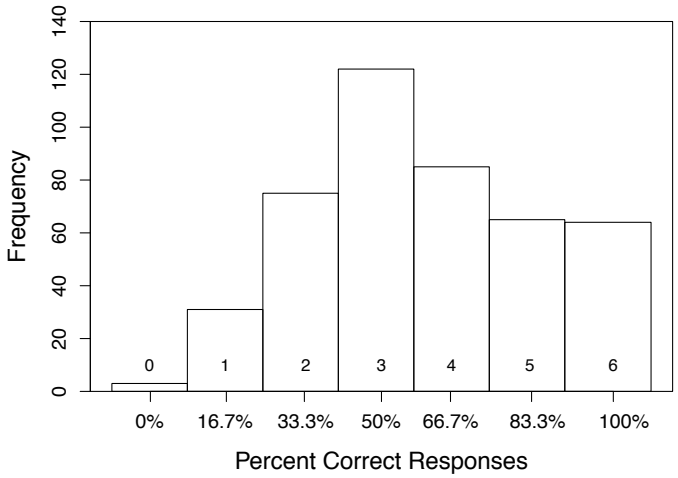


Figure 4-7. Results of Study 3: pupil diameter variation as a cue results. Histogram of percent correct responses. The numbers inside the bars are the number of correct responses.

Table 4-5. Results of Study 3: pupil diameter variation as a cue. Means and standard deviations of percentage correct responses broken down by Eye Motion ID.

ID	Percent Correct	
	μ	σ
s004	58.88%	34.14%
s009	62.13%	35.34%
s015	58.31%	34.69%
Overall	59.77%	24.93%

4.5 Respondents' Self-Reports

After each stimulus presentation, respondents were asked to self-report how difficult they found the task. At the end of the survey, respondents were additionally asked how often they play video games and how often they watched animated movies and shows. Finally respondents were asked to write out in free-form text the cues they thought they used to answer whether the avatar was watching a scary movie. In this section, we analyze the self-reports for each study.

4.5.1 Task Difficulty

We examined survey respondents' self reports on the difficulty of the task (1=very easy, 7=very difficult). Figure 4-8 (a) shows that, generally speaking, survey respondents found the tasks neither too difficult nor too easy. The mean values trend upward from as we move from Study 1 ($\mu = 3.81, \sigma = 1.51$) to Study 2 ($\mu = 4.19, \sigma = 1.49$) and then to Study 3 ($\mu = 4.79$, $\sigma = 1.60$). This is consistent with the design of the experiments. In the first study, respondents have a number of cues to help them identify which avatar is scared, including gaze patterns, pupil diameter changes, and blink rate. In the second study, respondents only have the pupil as a cue because differences in gaze patterns and blink rate are controlled for, in effect making the task harder. In the third study, the average size of the pupil is additionally controlled for, thus leaving only the change in pupil size as a cue, which makes the task even harder.

We ran individual Friedman rank sum tests for each study. In Study 1 and Study 3, there was no significant effect of Eye Motion ID (Study 1: Friedman $\chi^2 = 1.11$, $p > 0.05$, Study 3: Friedman $\chi^2 = 3.61$, $p > 0.05$). In Study 2, there was a significant effect of Eye Motion ID (Friedman $\chi^2 = 135.03$, $p < 0.0001$). Survey respondents reported that they found it more difficult to identify s004 ($\mu = 4.49$, $\sigma = 1.55$) than s009 ($\mu = 4.04$, $\sigma = 1.69$) and s015($\mu = 4.03$, $\sigma = 1.68$). This is consistent with the trend in Percent Correct Responses, and is explained by Table 4-3–the average pupil diameters for the two conditions for s004 were closer to each other than for the other two eye motions. This motivated the design for Study 3, where the average pupil diameters were made equal for both conditions (see Table 4-4 for numerical values).

4.5.2 Recognition Accuracy by Exposure Level

Exposure level on of survey respondents were comparable in our three studies. The majority of respondents spend 0-7 hours/week on video games as well as animated movies and TV shows (between 54.6% - 82.5%).

We analyzed Percentage Correct Responses with respect to survey respondents' exposure to computer generated content to understand how this exposure might impact their judgment. The breakdown of the average percentage of correct responses by exposure level is shown in Figure 4-8 (b), (c).

4.5.3 Self-reported Cues

Survey respondents gave a free-form answer to the question "What cue(s) did you use to say whether the person is watching a scary movie?" We coded the responses as referring to one of three cues: gaze, pupil, blink. For example, if the response was "their eyes darting quickly" then this response was coded as 'gaze'. Example responses coded as 'pupil' include "pupils size increasing" and "I assume the pupil tells it all", while responses coded as 'blink' include "how much they are blinking". A response such as "Eye movements and eye pupil size change" was coded as both 'gaze' and 'pupil'. We then looked at a breakdown of Percent Correct Responses by the self-reported cues (Figure 4-8 (d)). In this figure, we do not consider blinks as they appear infrequently compared to gaze and pupil. In both Study 2 and Study 3, gaze was identical for the two conditions and could not have been a distinguishing cue.We interpreted these findings as indications of how unreliable self-reports can be. In order to make the correct selections, survey respondents had to have been paying attention to the pupil even though they believed that they were using gaze as a cue.

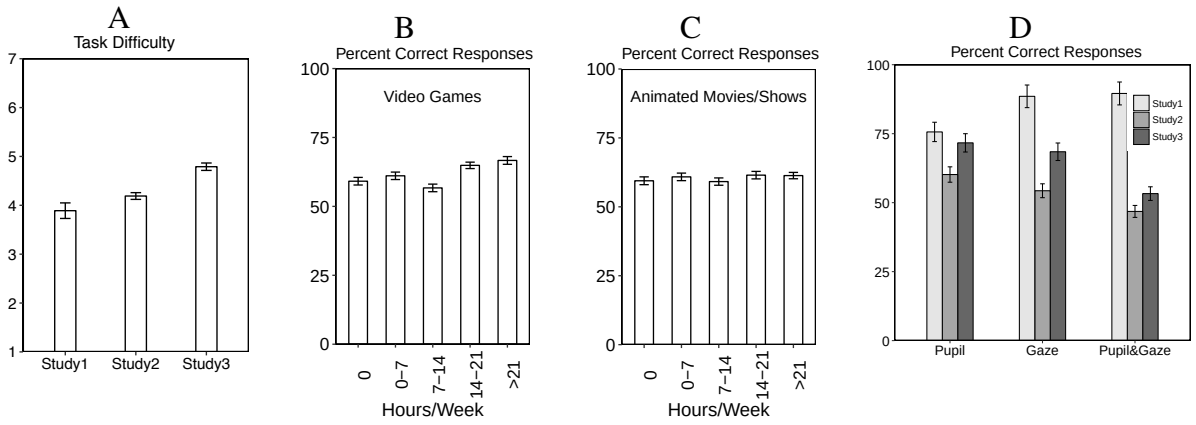


Figure 4-8. Respondents' Self-Reports. A) Means and standard errors of difficulty score for Study 1, Study 2 and Study3. B) Means and standard errors of Percent Correct Responses for all the survey respondents in the three studies broken down by exposure to video games. C) Means and standard errors of Percent Correct Responses for all the survey respondents in the three studies broken down by exposure to animated movies/shows. D) Means and standard errors of Percent Correct Responses for survey respondents in Study 1, Study2 and Study3 broken down by their self-reported cues.

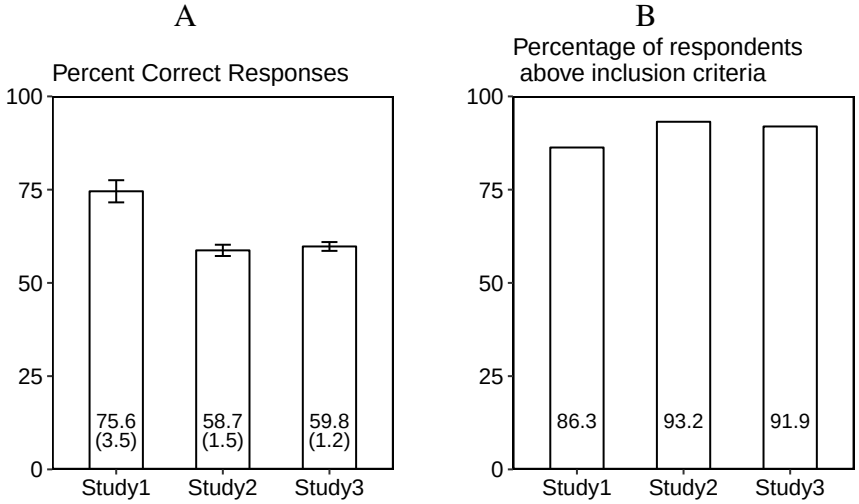


Figure 4-9. Comparison across three studies. A) Means and standard errors of Percent Correct Responses for Study 1, Study 2 and Study3. B) Percentage of respondents above consistency criteria for Study 1, Study 2 and Study3.

4.6 Discussion

In Study 1, we created two types of stimuli: in the first stimuli, the avatar's eye movements including pupil dilation, gaze and blinks matched the eye motions from people watching scary movies while in the second type of stimuli, the avatar's eye movements match the people watching light intensity changes. In Study 2 and Study 3, we presented experiments to evaluate if

survey respondents could identify the avatar watching a scary movie when the avatar's pupil was the only factor changed (gaze direction and blinks were controlled). When we consider the three studies collectively (Figure 4-9), we see that the metric Percent Correct Responses is highest for Study 1, and lower for Study 2 and Study 3 (though significantly above chance in all three cases). This trend is consistent with the increasing difficulty of the task.

In all our studies, the avatar was rendered in close up: our intention with rendering at this scale was that if survey respondents could not perceive pupil diameter changes even at this scale, they would definitely not do so in mid-range shots. Our finding that respondents are sensitive to pupillary dynamics thus holds for close up views, such as zooming in to a character's face, a type of shot quite common for emotional scene in games and movies. Future work could investigate different scales and different camera angles.

Our study design asked respondents to perform a side-by-side comparison to identify which of the two avatars was scared. This forced choice paradigm is provided by signal detection theory as a solution to measure the detection rate of a stimulus independent of respondents' individual internal criteria. The advantage of this paradigm is that the measured detection rates are less variable than single stimulus paradigms, and are independent of internal motivations such as social desirability or the amount of money respondents are being paid. Though users in actual application scenarios will likely not encounter digital avatars side-by-side, our study design is a theoretically solid method to establish if a signal (in this case, pupil dynamics) is detectable.

We considered 'scared' as the emotion to test in part because it is possible to elicit this emotion with the use of movies, and in part because the expression of this emotional state is more universal and less culture specific than anger and disgust, for example. It would be interesting to consider different emotions in future studies. It would also be interesting to expand these studies to include more eye motions from more people, including avatars of females, avatars that represent darker eyes and darker skin color, and a broader range of age groups. This work could inspire models of pupil diameter dynamics for different emotions, personalities, and age groups. For virtual avatars, it may be interesting to consider models of exaggeration and stylization within

pupillary dynamics to match a more stylized render style. With such models it may even become feasible to directly generate eyes from an audio recording or a script.

A recent paper on image stylization served a very similar purpose as ours but in a different domain [98]. Their work successfully translated realistic pictures to oil painting style while preserving the content of the pictures in the absence of paired examples. Their algorithm leveraged Generative adversarial networks (GANs), an unsupervised learning technique inspired by game theory. GANs achieve success in multiple area, such as image generation [16, 66], image style transfer [98, 35], representation learning [53], text, lyrics synthesis [96] and so on. GANs consists of two networks compete with each other. The adversarial loss shapes the generated samples to be indistinguishable from the real samples in the training data. In CycleGAN, given two types of training data from two distribution, even without the paired samples, the network can convert results from one distribution to visually match the other distribution of the training data. However, this concept has not been used in motion translation before. In my dissertation work, I use an adapted CycleGAN network to solve the problem of adult2child motion translation. This model can find the mapping between adult motions and child motions from temporally unaligned training samples.

Previous works applied GANs to motion capture data but for a different purpose. [5] used GANs to predict future frames based on previous frames. They introduced a non-deterministic motion prediction model learning a probability density function that could generate several versions of the future frames from the same input frames. However, motion translation using GANs has not been explored yet. It is an important topic when it comes to creating childlike motion because studies found larger inter-subject and intra-subject variability in child motion data than adult motion data. Learning the mapping of two distributions as a whole than learning the mapping from two instances can maintain the variability inside each distribution.

5.1 Motion Tracking Children

There are existing motion capture databases collected from children [2]. However, it consists of only motions performed in place. This could be unnatural for locomotions such as walking and running. We collected a motion capture database that focused on dynamic motions. We used Vicon optical motion capture system with ten 120 Hz cameras which have better

tracking accuracy and low noise level compared to Kinect 1.0. All the child participants are from 5-10 years old. All the adult participants are 18 years old and above. The database includes a variety of dynamic motions from 10 children and 10 adults. The motion capture study is approved by the university review board (IRB). Each subject signed an informed consent form before participating in the study. For child participants, we also had their parents consent to perform the study. We captured the following actions: “Throw a ball with left arm”, “Walk”, “Walk as fast as you can”, “Hop Scotch”, “Throw a ball”, “Run and Jump”, “Sneaky Walk”, “Walk, step over obstacle”, “Punch”, “Happy Walk”, “Kick”, “Jog”, “Run as fast as you can”, “Idle”, “Jump with one leg”, “Skip”, “Jump”, “Jump as high as you can in place”, “Broad Jump Forward”, “5 Jumping jacks”. We captured 2-4 repetitions for each action type. My work on dynamic scaling algorithm (chapter 4.1) has shown that dynamic motions have a higher recognition rate than static motions. Therefore the actions we chose contains a range of dynamic motions, ranging from low intensity movement such as punch to high intensity movement like jump with one feet. Past research also shows that children have different coordination level from adults because they are still in the developing phase. We want to train the networks to learn modeling the coordination of the body. Most of the motions we collected requires different parts of the body to work together, which facilitate the learning of the neural networks.

We named all the motions with regard to the motion types, subject type(adult or child), and subject ID. The data can be downloaded in fbx, bvh, and csv formats. We used the front waist marker set, also known as the standard-53 marker set. Skeletons were generated using the default skeleton template using Vicon solving. Each skeleton has 33 joints. There are minutes of motions in the database in total.

The motion capture suits are designed to be tight so the markers will remain still on the body while the subjects move. However, this could also create a challenge for children at their young age to put the suits on their own. To make sure children have enough privacy help, we used another closed room with tables and chairs for the parents or guardians to go inside and help the children to get dressed. For older adults (8 years old or older), children were advised to go to the

private room to get dressed by themselves.

For adult participants, most participants did not want to take any break and complete the study in one run. Some participants need one break throughout the study. Adults are more disciplined and showed more perseverance in the process. For child participants they get tired and bored more easily. Only one child participant completed the study without taking any break (except taking some water). He enjoyed the process insisted to resume the study as soon as possible. Most child participants took two or three breaks during the study. When participants started panting we would pause the study and ask them to take a break, because we have to make sure they are not over strained. The motions they performed when they are exhausted can also be unnatural. For example, the motions would be slower and softer than they normally would be. We also gave participants at least one minute to rest in between motions.

Another challenge with motion capturing children is that they are easily distracted. Because of the constraints with the lab space, the computer running Vicon Shogun software is co-located with the motion capture space. During capturing, some child participants turned around and focused on the monitor to see their embodies virtual avatar inside the Shogun software. To avoid this distraction, one way is to turn the monitor around or move it away so the participants can not see it. The second approach is to have experiment staff standing in front of the participants and remind them to always look ahead.

Children sometimes are less willing to communicate their needs to our experiment staff. During the study, we asked the child participants every 15-20mins to see if they need any water, a break, or if there is anything make them uncomfortable. We noticed that most children would not bring their needs up unless the study staff asked them, so it is also necessary to closely monitoring them. Once we noticed the participants panting or sweating heavily, we pause the study and suggested they take a break and resume the study when they are ready. When performing the same action, children usually have a larger range of motions than adults. Their motions tend to be more dramatic than their adult counterpart. The markers are easily moved around. For child participants, we usually have to replace markers once or twice during the study. For adult

participants, we usually do not need to make any adjustment. One of the study members would pay attention to the marker position and make sure they all stay in place during the study. If a marker falls off, we would pause the study immediately and place the marker back to the original spot. Subject recalibration is needed after that. Before capturing a dataset, motion selection is an important design choice. Younger children may have a difficult time interpreting the actions. Actions that require much interpretation is not suitable for children to perform. We chose actions that children already incorporate in their daily life such as hopscotch, skipping, etc. Actions that are less likely to encounter in day to day life such as fly like a bird are not included in the dataset. In addition, to capture motions as training data to our system, we also captured combo motions as part of the testing dataset, such as run then jump.

Vicon provides motion capture suits tailored for children. They provided velcro straps to put on the shoes. The shoe markers were stuck to the straps. However, the participants were asked to perform many locomotions such as walking and running, the straps often fall off during the study. To avoid this problem, we made customized mocap shoes. We used duct tapes to fix the markers to child shoes of different sizes. Participants were asked to try out different sizes to find out the best fit.

The study duration is about 1.5 hours for each child participant. It is also important to keep the children motivated in the study. At the beginning of the study, We tried to excite the participant by relating the motion capture technology to the popular animated movies or video games. We told them their motions will be used to animate their favorite character. During the study, every time they complete an action, we tell them how great they did. Even when they made some mistakes or deviate from the instruction, we always stay positive and patiently correct them.

5.2 Motion Stylization for Creating Childlike Motions

5.2.1 Data Representation

We represented the motions using joint angles. All joint angles are converted into quaternion format except the root joint. Quaternion representation is free of Gimbal lock and it supports interpolation. The motions are captured in 120 fps. The motion sequences were divided

into short motion words with a window size of 24 frames. There were 8 overlapping frames between the adjacent motion words for blending purposes. The remaining frames that were not long enough to be a motion word were discarded. We paired the child and adult motion words that have the same motion type in the training. We calculated the overall max and min values of each Degree of Freedom (DOF) across all the motions and normalized individual DOF to the range of [-1,1] using the overall max and min values. During the postprocessing step, we denormalized the values back to its original range.

Because the subjects may face different directions during the study, we translated the motions into local root space. The global position along the x-axis and z-axis of the root(hip) in each pose is set to zero. We also computed the direction of the front-facing vector(local z-axis) of the hip joint. The hip joint will be rotated along the global y-axis by the angle between the local z-axis and the global z-axis. After the rotation, all the poses in the database faces the same direction.

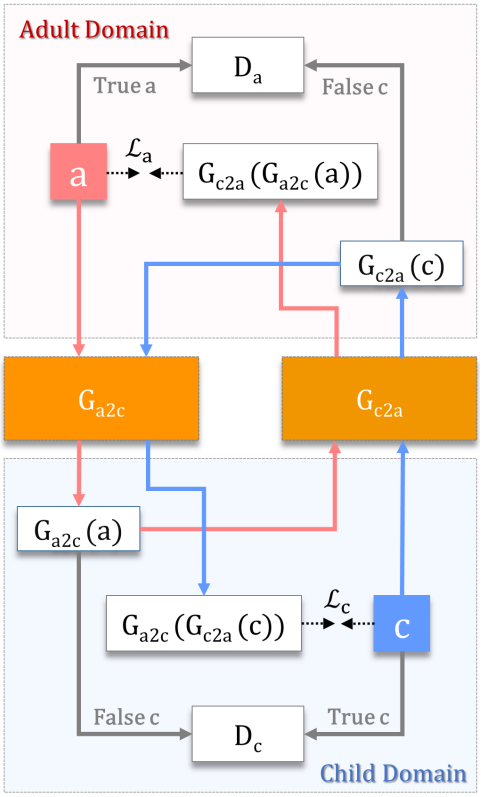


Figure 5-1. Architecture of CycleGAN.

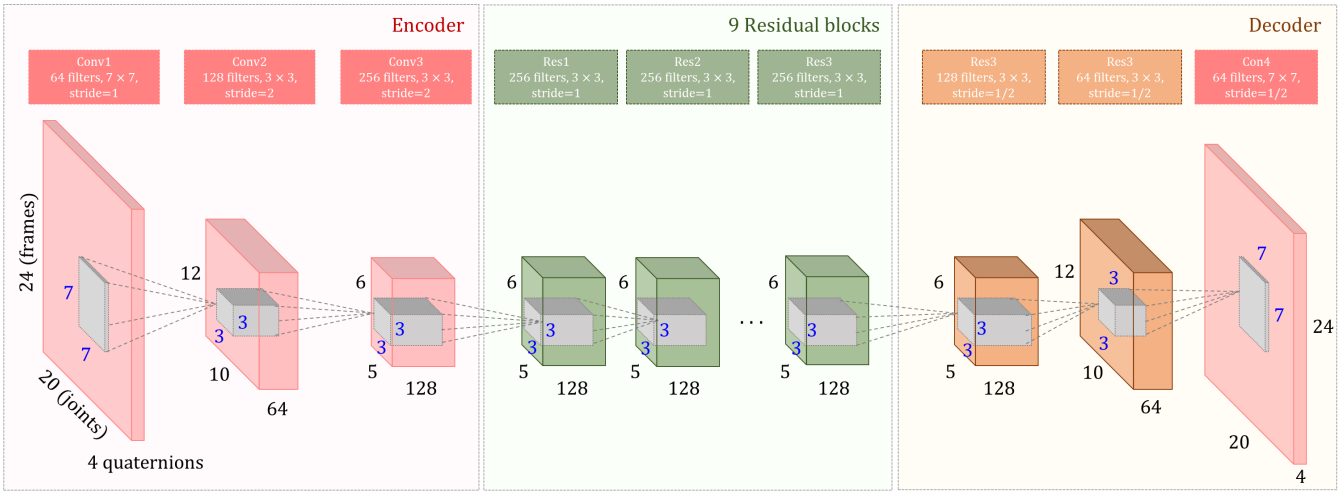


Figure 5-2. Architecture of the generator.

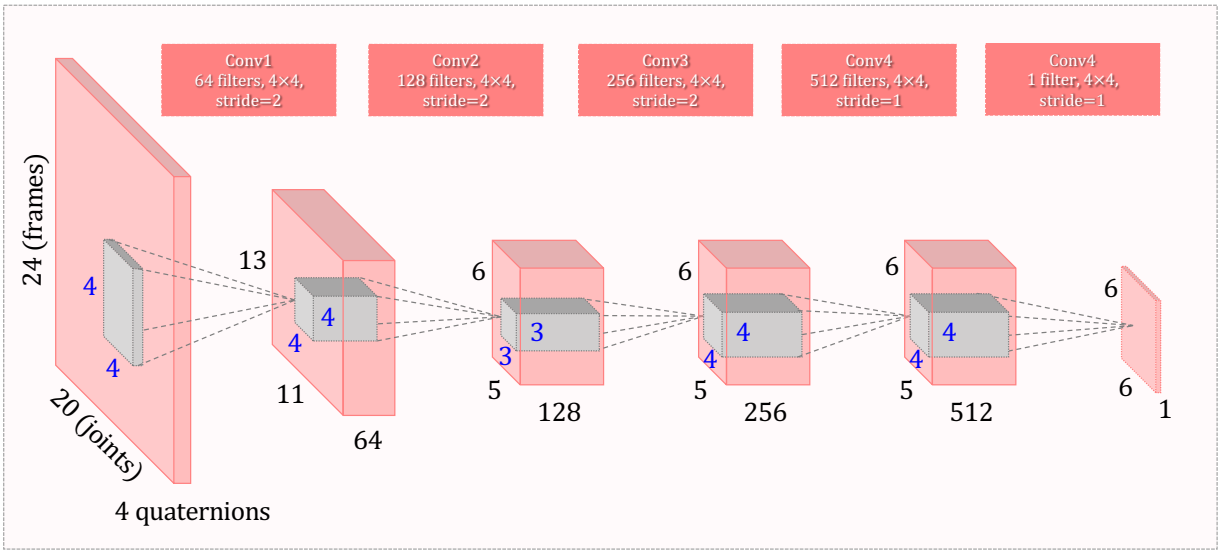


Figure 5-3. Architecture of the discriminator

5.2.2 Network Architecture

We adopted the CycleGAN architecture to learn the mapping between adult motions to childlike motions. The network was trained on one motion word pair at a time. The original CycleGAN paper learned using the unpaired training data from two domains such as a dataset of horse images and a dataset of zebra images. However, in our case, the inter-variability of the two datasets is larger than the intra-variability. In other words, the differences in motion type and the timing in the same style are more distinct than the motion style itself between the two styles. To simplify the problem, we matched the training data with the same motion types. However, this step is trivial because unlike image data, all of the motion capture data available online are labeled with motion type and style type.

The network architecture consists of two GANs. One GAN converts the motion from the first domain to the second domain. The other GAN converts the motion from the second domain back to the first domain (Figure 5-1). The two GANs are symmetrical. Each GAN consist of a generator and a discriminator. The two generators and the two discriminators share the identical architecture. However, the weights in the two networks are different because they translate the motions in two different directions.

For the generator (Figure 5-2), we used a stride-1 convolution layer with filter size 7 followed by two stride-2 convolution layers with filter size 3. The output of the third convolution layer feeds into nine residual blocks. Each residual block consists of two convolution layers with filter size 3 with a skip layer connection. The decoder of the generator consists of two stride-1/2 deconvolution layers and one stride-1 convolution layer of stride-1 with filter size 7. Same as CycleGAN, the discriminator is a four-layer convolutional network with 64, 128, 256, 512 filters in each layer. All the layers have a filter size 4. The first three layers have a stride length of 2 while the fourth and fifth layers have a stride length of 1.

5.2.3 Loss Function

We applied two adversarial training: one for adult2child translation and the other one for child2adult translation. According to [98], having the two GANs forming a cycle enables training

without paired data and prevents modes collapse. We denote adult motion as a and child motions as c. We have two generators: G_{c2a} maps adult motions to fake child motions while G_{a2c} maps the child motions back to adult motions. We also have two discriminators: D_a that distinguish original adult motions a from the fake adult motions$G_{c2a}(c)$. The other discriminator D_c differentiate original child motions c from the fake child motions$G_{a2c}(a)$. The goal of the GANs is to generate instances similar to the target distribution conditional to the input. We also introduced additional loss terms: temporal coherence loss and transition loss. The temporal coherence loss ensures smooth motions and prevents sudden jerk happens. Because we processed motion words, there is no guarantee of a smooth transition between adjacent motion words. We added transition loss to penalize different overlapping frames.

5.2.3.1 Adversarial loss

We adopted the GAN adversarial loss and the loss function for the generator and discriminator is shown below. Equation 5-1 is the loss of mapping the adult domain to the child domain. Equation 5-2 is the loss for the opposite direction. Adversarial learning is a min-max game. The generator tries to minimize the likelihood that the generated instances are recognized as fake while the discriminator maximizes the likelihood of successfully discriminate the generated instances as fake.

$$L_{\mathbf{D_c}} = 0.5 * \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}[\mathbf{D_c}(\mathbf{c}) - 1] + 0.5 * \mathbb{E}_{\mathbf{a}\sim p(\mathbf{a})}[\mathbf{D_c}(\mathbf{G_{a2c}}(\mathbf{a}))] \tag{5-1}$$

$$L_{\mathbf{D_a}} = 0.5 * \mathbb{E}_{\mathbf{a}\sim p(\mathbf{a})}[\mathbf{D_a}(\mathbf{a}) - 1] + 0.5 * \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}[\mathbf{D_a}(\mathbf{G_{c2a}}(\mathbf{c}))] \tag{5-2}$$

$$L_{\mathbf{G_{c2a}}} = 0.5 * \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}[\mathbf{D_a}(\mathbf{G_{c2a}}(\mathbf{c})) - 1] \tag{5-3}$$

$$L_{\mathbf{G_{a2c}}} = 0.5 * \mathbb{E}_{\mathbf{a}\sim p(\mathbf{a})}[\mathbf{D_c}(\mathbf{G_{a2c}}(\mathbf{a})) - 1] \tag{5-4}$$

5.2.3.2 Cycle consistency loss

We adopted the cycle consistency loss from [98] to increases the visual quality of the output motion. Because our dataset contains different motion type, having cycle consistency loss helps

to preserve the content of the input motion It is defined as the differences between $G_{a2c}(G_{c2a}(c))$ and c for child domain (equation 5-5) and the differences between $G_{c2a}(G_{a2c}(a))$ and a for adult domain (equation 5-6).

$$L_{cycle}(c) = G_{a2c}(G_{c2a}(c)) - c \tag{5-5}$$

$$L_{cycle}(a) = G_{c2a}(G_{a2c}(a)) - a \tag{5-6}$$

5.2.3.3 Temporal coherence loss

Temporal coherence loss was introduced to increase the smoothness and stability of the output motion. We computed the first derivative of the output motion or in other words, the delta between two consecutive poses for all the degree of freedom across all the frames in one motion word. The temporal coherence loss is shown as Equation 5-7 and Equation 5-8.

$$L_{coherence}(a) = \sum_{t} \sum_{DOF} ||G_{a2c}(a(t)) - G_{a2c}(a(t-1))|| \tag{5-7}$$

$$L_{coherence}(c) = \sum_{t} \sum_{DOF} ||G_{c2a}(c(t)) - G_{c2a}(c(t-1))|| \tag{5-8}$$

5.2.3.4 Transition loss

During the pre-processing step, we sliced the motion sequence into shorter motion words with overlapping frames. In the post-processing step, the output motion words need to be blended and stitched back to the original length. We used the average of the overlapping frames in the adjacent motion words during blending. The differences in the overlapping frames of output motion can cause undesired artifacts. To reduce the artifact, we added a transition loss term shown in Equation 5-10. We denote overlapping frame number as $t_{overlap}$, the motion word index as i.

$$y = G_{c2a}(c) \tag{5-9}$$

$$L_{transition}(c) = \sum_{t} \sum_{DOF} ||y_i(t_{overlap} : end) - y_{i+1}(0 : t_{overlap})|| \tag{5-10}$$

5.2.3.5 Overall loss function

The overall loss function (Equation 5-11) can be computed as the sum of all the terms described above. λ denotes the weight of each loss term. The weights of each term are decided by ablation studies in the Results Section.

$$\begin{aligned} L(G_{a2c}, G_{a2c}, D_a, D_c) = {} & L_{GAN}(G_{a2c}, D_c, a, c) + L_{GAN}(G_{c2a}, D_a, a, c) + \lambda_{cycle} \times L_{cycle}(c) \\ & + \lambda_{cycle} \times L_{cycle}(a) + \lambda coherence \times L_{coherence}(a) + \lambda_{coherence} \times L_{coherence}(c) \\ & + \lambda_{transition} \times L_{transition}(c) + \lambda transition \times L_{transition}(a) \end{aligned} \tag{5-11}$$

We train the two generators to minimize the overall loss while train the two discriminators to maximize the adversarial loss.

$$L(C) = \arg \min_{G_{a2c}, G_{a2c}} \min_{D_a, D_c} \mathbb{L}(D_a, D_c, G_{a2c}, G_{c2a}) \tag{5-12}$$

5.2.4 Post-processing

In the pre-processing step, we scaled the values to the range of [-1,1]. In the post-processing step, we first used the saved min and max values for the childlike motions and scaled each DOF in the output motions back to the normal range. We also divided the motion sequences into small motion words. Then we stitch the output motion words of the fake childlike motions $G_{a2c}(a)$ back to the original length. The motion words that belonged to the same sequence are stitched together

in order. We blended the adjacent motion words by taking the average of the overlapping frames.

5.3 Results

We implemented and trained the model on Google Colab Pro with P100 or T4 graphics card. The model was written in TensorFlow. We trained the motions for 200 epochs and the training takes 7 hours. The trained model is 8.67MB.

In section 5.3.1, I discussed the results tested on the existing dataset([94]). In section 5.3.4, I showcased the results tested on our existing dataset.

5.3.1 Results on Existing Dataset

The results discussed below are on sequences taken from the testing dataset. General observations We noticed that the algorithm works well for cyclic motions but there are artifacts with non-cyclic motions. The results are shown in Figure 5-4, 5-5, 5-16, 5-7, . Child motions are shown in subpart A, fake child motions are shown in subpart B whereas the adult motions are shown subpart C. We can see that both the timing and the poses for the fake childlike motions are more similar to childlike motion than adult motions. The spine of the fake childlike motion is more straightened up compared to the old motions.

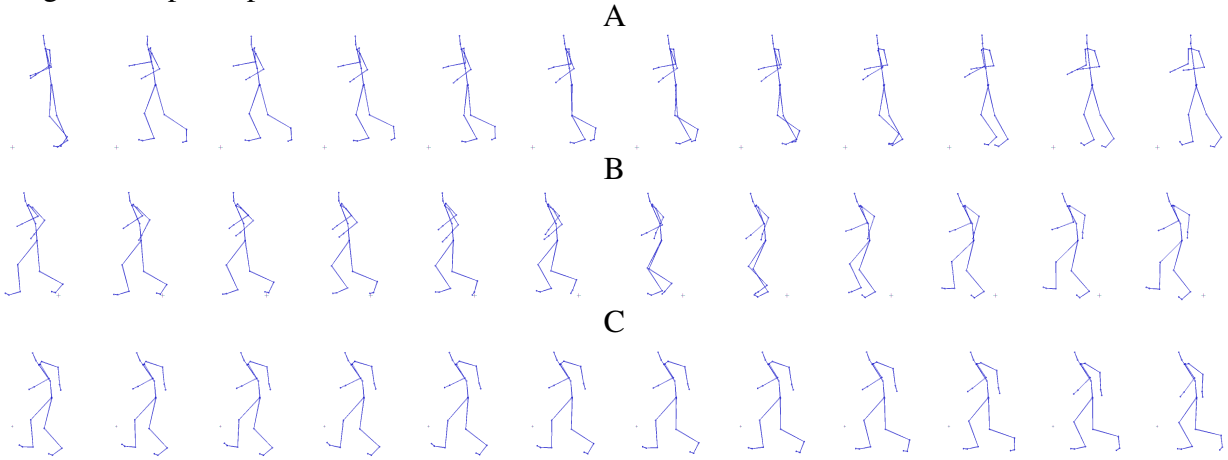


Figure 5-4. Results on fast walking action. A) input childlike motion. B) output fake childlike motion. C) input old motion.

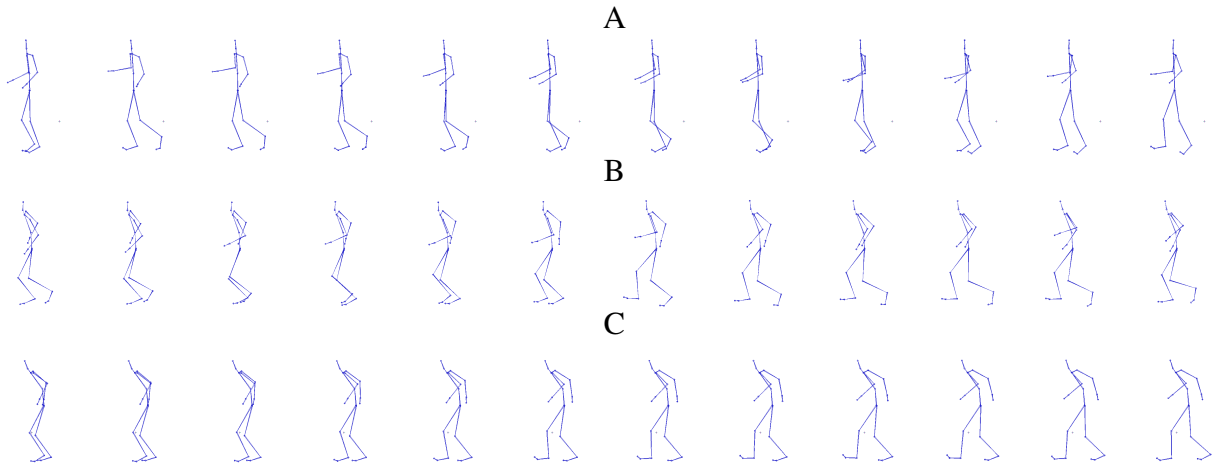


Figure 5-5. Results on normal walking action. A) input childlike motion. B) output fake childlike motion. C) input old motion.

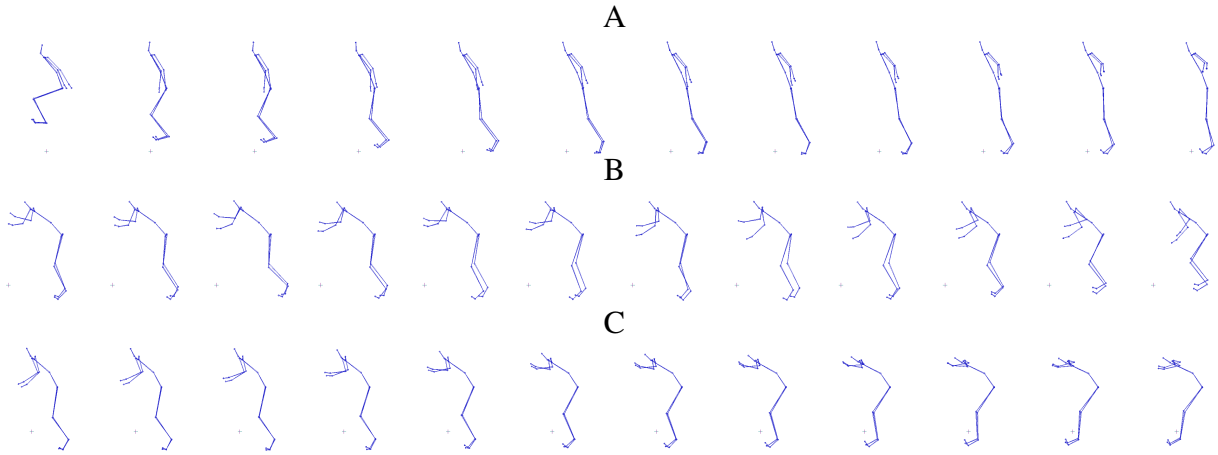


Figure 5-6. Results on jumping action. A) input childlike motion. B) output fake childlike motion. C) input old motion.

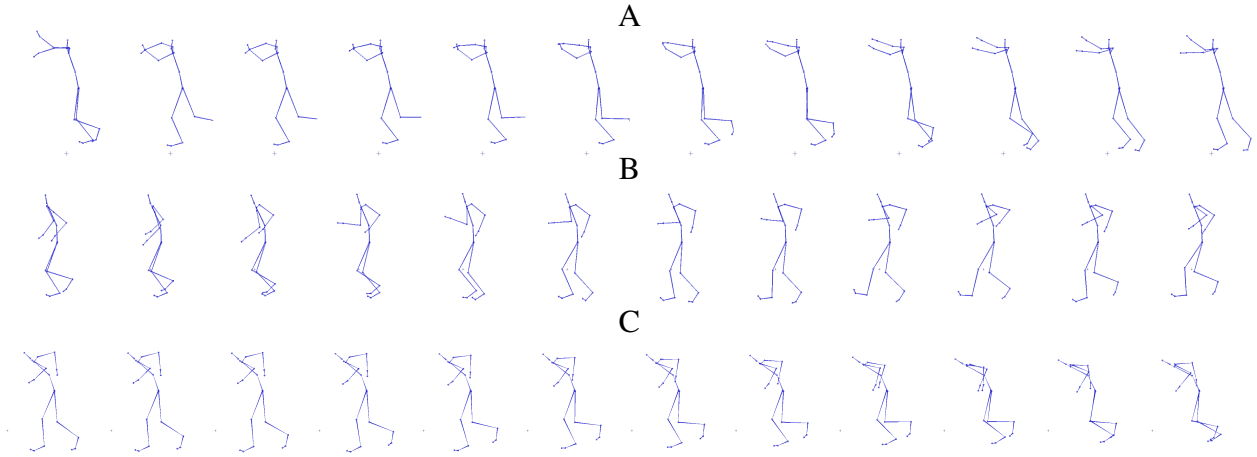


Figure 5-7. Results on running action. A) input childlike motion. B) output fake childlike motion. C) input old motion.

5.3.2 Ablation Studies on Existing Dataset

We conducted ablation study for each term in the the overall loss function $L(G_{a2c}, G_{a2c}, D_a, D_c)$. For each term, we removed that term from the loss function and re-trained the entire network for 200 epochs. In each case, we used this re-trained network and generated an output motion from the same input motion: a fast walk sequence in the style "old". As many research paper pointed out [5], using loss function is not a reasonable metric as GAN loss does not reflect the quality of the motion. Therefore, we visualized the output motion as key frames.

5.3.2.1 Effect of Cycle consistency loss

We first removed the two cycle consistency losses $L_{cycle}(c)$ and $L_{cycle}(a)$ from the loss function. we observed the output motions degraded and become unnatural. We showed the visualization of the output in Figure 5-8. We can see in the output motion without the cycle consistency loss, some poses are not expected, whereas in the output with cycle consistency loss, the motion is smoother and natural.

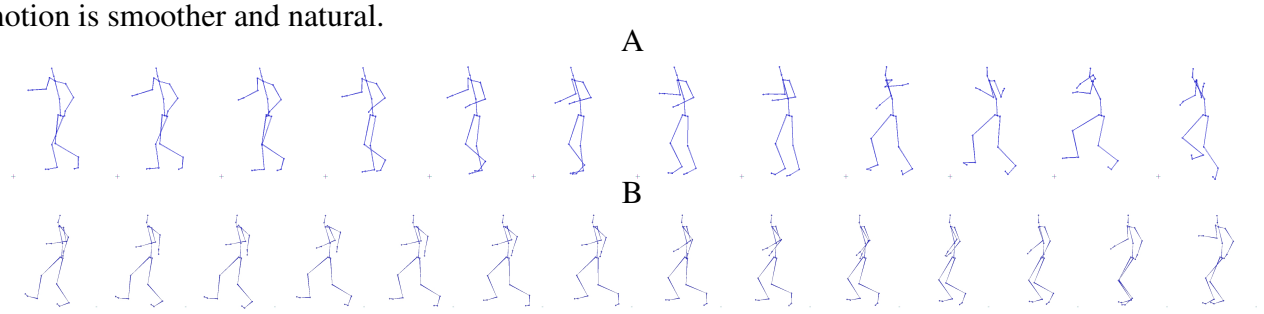


Figure 5-8. Ablation study on cycle consistency loss for fast walking action. A) output fake childlike motion without cycle consistency loss. B) output fake childlike motion with cycle consistency loss.

5.3.2.2 Effect of Transition Loss

We examined the effect of transition loss. The purpose of adding the transition loss is to ensure a smooth transition between motion words. The transition loss penalizes the differences in the overlapping frames of adjacent motion words. We found from the ablation study that without transition loss, there is nothing to stop the output motion word from having incomplete cycles. This can result in the character starting a new step without finishing its previous step. From Figure 5-9, we see the character put one leg forward for a new step before it completes the

previous step cycle. The arms swing in the motion without transition loss displays jumpiness. However, we do not see these artifacts when we train the network with transition loss.

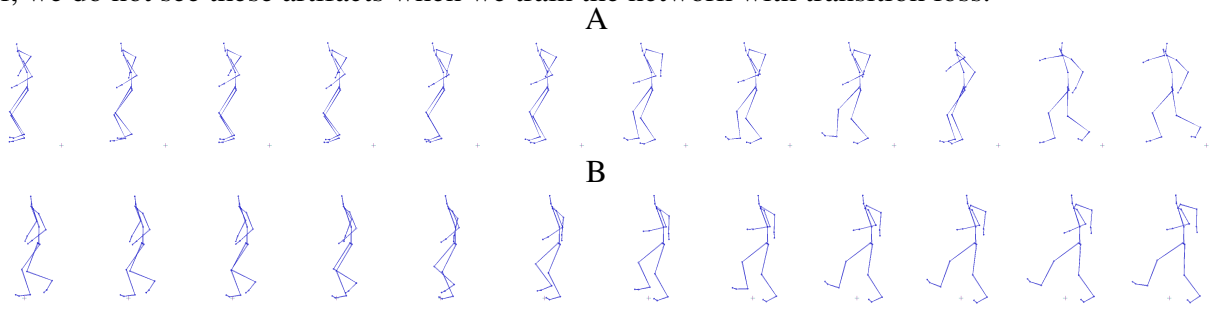


Figure 5-9. Ablation study on transition loss for fast walking action. A) output fake childlike motion without transition loss. B) output fake childlike motion with transition loss.

5.3.2.3 Effect of Temporal Coherence Loss

We trained the network without the temporal coherence loss term and found that the quality of the motion degraded. The output motions become jerky and unstable compared to the results with the temporal coherence loss. We can see in Figure 5-10, the arms from the output motion (A) are not moving in a coherent trajectory. This effect is better seen in the supplementary video.

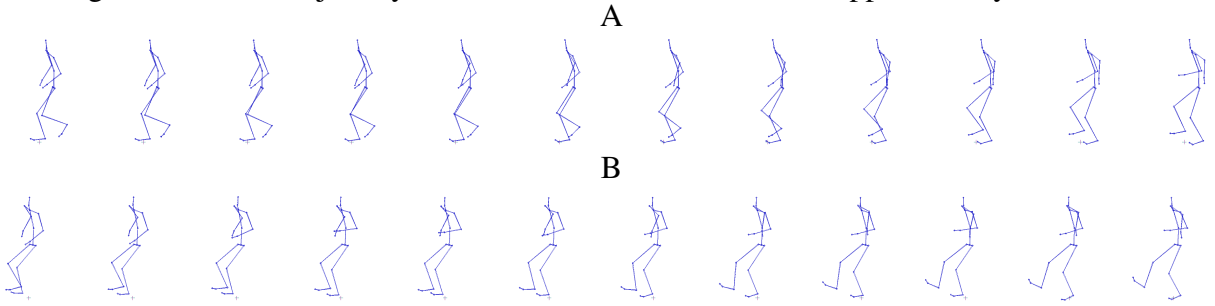


Figure 5-10. Ablation study on temporal coherence loss for fast walking action. A) output fake childlike motion without temporal coherence loss. B) output fake childlike motion with temporal coherence loss.

5.3.2.4 Effect of Joint Ordering: Breadth-First vs. Depth First Ordering of the Skeletal Tree

Because convolution computation convolves the n by n (n is the kernel size) pixels in the motion word, meaning that the adjacent DOF along with the adjacent frames will be computed together. The ordering of the input decides what joints will be convolved.

We experimented with two different types of ordering for the joints: Breadth-First order and Depth First order. The human skeleton can be organized as a tree structure. For depth-first

ordering, we start from the hip joint and go through each branch top to bottom. For breadth-first ordering, we start from the root and go through all the joints in the same depth before moving to the next depth.

We see minor differences between the two orderings. The results of breadth-first ordering and depth-first ordering for “fast walk” are shown in Figure 5-11. We saw that the results of depth-first ordering seem to have a larger stride compared to the breadth-first ordering. We also tried a normal walk sequence. In this case, it appeared that the results from breadth-first ordering seem more natural.

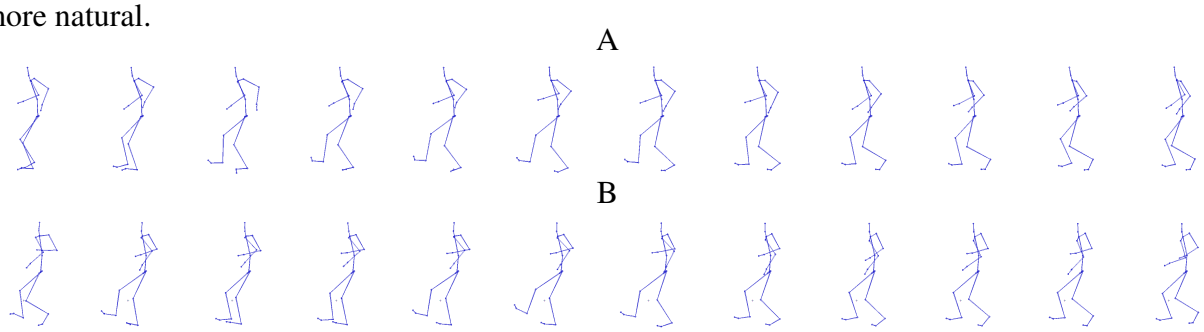


Figure 5-11. Ablation study on the ordering of the joints for normal walking action. A) output fake childlike motion with joints arranged using breadth-first ordering. B) output fake childlike motion with joints arranged using depth-first ordering.

5.3.3 Discussion of the Existing Dataset

We created a system using CycleGAN to translate adult motions to child-like motions without temporally aligned motion pairs. We generated visualization of the motions as snapshots on stick figures. We used the style ”old” and ”childlike” motions from Xia et al. dataset [94] as an alternative for adult and childlike motions. In the dataset, the old motions generally have a crouched spine and very slow speed. For the original childlike motions, the actors performed in a faster cadence and straightened upper body. We noticed that output childlike motions have a faster pace than the input adult motion in walking motions. In addition to the timing change, the output childlike motions have a straightened up spine. These qualities make the old motions look more similar to the childlike motions.

One artifact from the motion is that the output motions can be jittery. We noticed the same artifacts in the original CycleGAN paper. The translated image appeared noisy. We suspect this is

intrinsic to the training process. The temporal loss term helped mitigate this issue and make the results much smoother but it did not solve the problem entirely. Another way to address this artifact is by applying a low-pass filter. We experimented on that and the output motions start to appear much more natural. The comparison will be showcased in the supplementary video during my defense.

CycleGAN trains two generators: one for adult2child and the other one for child2adult translation. The process is supposed to be symmetric. However, the quality of the results is not symmetrical. Fake childlike motions look more natural than the fake adult motions. This is related to how we created the index file. We matched the childlike motions to adult motions. The adult motions are fed into the system in order whereas the order of the child motions are not guaranteed due to they have different length. As future work, we can investigate how to ensure the quality of the translation in both studies.

We conducted 4 ablation studies to see the effect of each term in the loss function as well as the format of the input. We found that each term in the loss function is crucial to the system, The cycle loss ensures the motion to be natural and plausible; transition loss ensures the transition between the motion words to be smooth; temporal coherence loss filter out the jitters and noises in the motions. However, we did not find any impact on the ordering of the input. This finding indicates that the network is robust enough to adapt to the different ordering of the input.

Another limitation of this work is that the foot contact. The output motions have foot sliding issues. CycleGANs learns to generate motions that are similar to the training set, but the networks do not impose any physics constraints. As future work, we will annotate the foot contact in the original training data. We can either train a neural network to learn the foot contact, or we can use inverse kinematics to fix the foot contact in the output motions.

5.3.4 Discussion of our Novel Child Dataset

Our novel child dataset is very different from the one released by [94]. There are three broad categories of differences that directly impact the training of the neural network parameters:

1. Improvisation/Not following instruction exactly. Xia et al.'s dataset captured adult acting

like a child whereas our dataset had actual children, who interpret instructions quite differently, and are prone to improvisation when they are having fun. For example, we noticed in the child subject child004, (s)he did multiple dabs while performing walking motion (Figure 5-12). We observed the improvisation in one of the child participant out of eight child participants. In another fast walking motion from child 005, participant swing their arms to a very large extent (Figure 5-18,5-19). However, for adult participants, they followed the instructions as is and did not perform any extra motions or improvise. We see the same trend in Xia et al's dataset. All the motions match the motion types and there are no redundant motions.

2. Extra movements before and after the prompted action. All the motions from our dataset started with a T-pose. We cut off the t-pose at the beginning using an algorithmic approach. The angles of shoulder joints were used to determine the threshold when the T-pose ends. However, after the motion complete, some child participants would swing their arms around or cross their legs. These extra motions were captured in our dataset. Another child participant got distracted and kept staring at the monitor during walking. We saw the extra motions in five out of eight child participants during the capture. In the future, we can use an automatic method or a manual method to cut off the redundant frames and only keep the relevant motions.

3. Age/biomechanics/skeleton of motion sequences. In the training set, we have 3 adults (adult001, adult002, adult 003) and 3 child participants(child001, child004, child005). The number of participants was not reported in Xia et al.'s paper. However, all the motions share the same skeleton template. Either they retargeted all the motions to the same template or they captured only one adult. In contrast, our dataset consists of motions from actual child participants who are 5-10 years old. Five-year-olds are much smaller than ten-year-olds and have appreciably different skeletal proportions. They also have different cognitive abilities and we found that the ten-year-olds were more likely to keep closer to what we expected

when we gave them a certain prompt.

4. Capture setup. In the dataset of [94], subjects only move in one direction and there is no turning. In our dataset, we asked our subjects to turn when they reached the end of the capture space and do a loop in order to record long continuous sequences. As a result, another difference between the two datasets is that for locomotion, i.e, walk, fast walk, and jog, our dataset included 180-degree turns. During the turns, we noticed the participants would cross their legs to complete the turn. We think that this is the source of the artifact that the fake child motions have crossed legs even when going straight. One way to fix this artifact as future work would be cutting off the turns and only leave locomotions that move straight and train the network again with cropped motions.

5.3.5 Results on our Novel Child Dataset

We trained the generators and discriminators in the following order: $\mathbf{G_{a2c}}, \mathbf{D_c}, \mathbf{G_{c2a}}, \mathbf{D_a}$, and we set $\mathscr{L}_{coherence} = 30$ and $\mathscr{L}_{cycle} = 15$. We used the Adam solver for gradient descent, while the batch size is 1 to incorporate the transition loss. The network was trained 180 epochs with a learning rate of 0.0002, and the training takes approximately 7 hours. After 100 epochs, the learning rate is decreased at a rate of 1% per epoch. The size of the trained model is 8.67MB. The training data includes 90% of motions from three mocap actors. Our test data includes the remaining 10% of data from the training set plus motions from two different actors. The purpose is to test the generalizability of the model, and how it performs on motions that are similar to the training data as well as motions that are different from the training data.

We applied the *adult2child* architecture on our motion dataset, using two different architectures, as shown in Figure 5-13 : joint-wise convolution and overall convolution. For

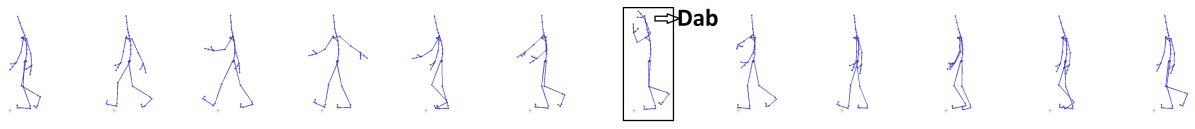


Figure 5-12. An example of improvisation during capture. A child participant did one dab during the walking motion.

joint-wise convolution, we computed separate kernels for each joint. The kernels convolves the motion words along the time axis and the quaternion axis. For overall convolution. the kernels traversed each motion word along the time axis and the joint axis. We have observed that the overall convolution networks failed to create natural and childlike motions. This happens since the same 2D kernels were applied to all the joints, thus the network failed to extract the style from the entire skeleton. On the other hand, the joint-wise convolution networks operates on joint basis. In this case, different kernels were applied on each joint separately, with different set of weights; allowing the network to learn the transfer function on per joint makes style modeling much easier.

In this work, we introduce the *adult2child* framework to translate adult motions to child-like motions without assuming the availability of temporally aligned pairs of motion sequences in our training database. Anecdotally, we have noticed that our output results exhibit some childlike characteristics: faster pace and swinging arms. Figures 5-16 and 5-17 demonstrate "jump-as-high-as-you-can"and "walk-as-fast-as-you-can" motions that have been translated from an adult's motion. We show frame snapshots taken every 5 frames. Subfigure A shows the input adult motion used to generate the fake child motions shown below, subfigure B shows results using our *adult2child* method, whereas the subfigure C shows a reference child motion. It can be observed that both the timing and the poses for our output childlike motions are more similar to the movements of a child than adults. In particular, the spine of the fake childlike motion generated using our method is more exaggerated, while the movement of the hands rise-up appears to be carefree and playful, compared to the tight and structured movement of the adult. Our method takes into consideration the timing; the output childlike motion starts at the same pose as the input adult motion, but towards the end, the output motion is about 10 frames (one snapshot) faster than the input adult motion.

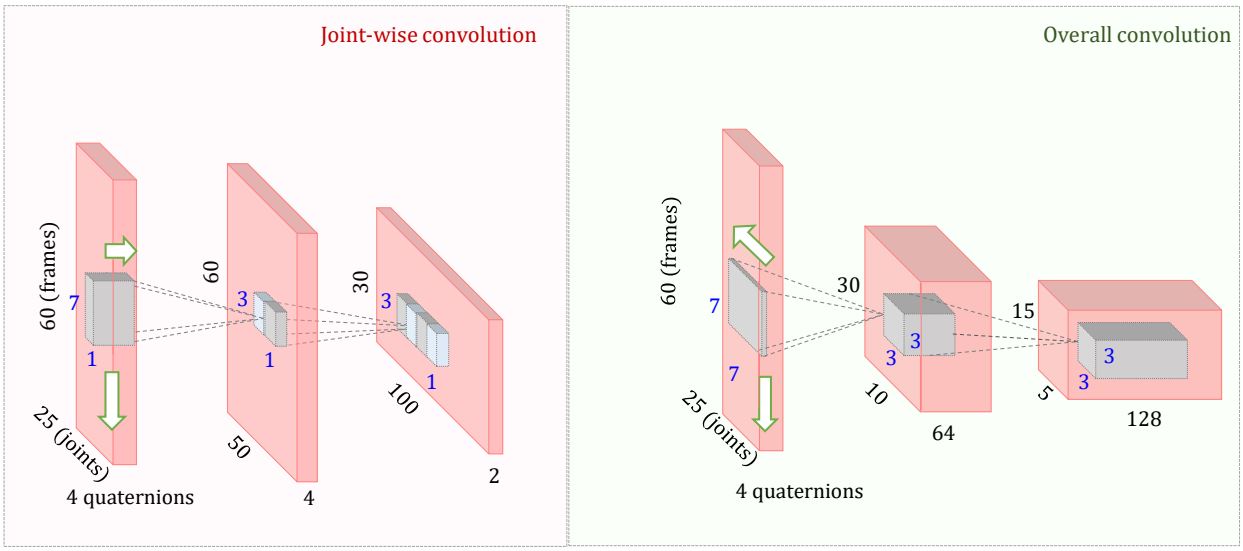


Figure 5-13. The 1D and 2D convolutional networks tested in our experiments. The joint-wise convolution convolves along the temporal axis(height) and the quaternion axis(depth). The overall convolution convolves along the temporal axis and the joint axis(width).

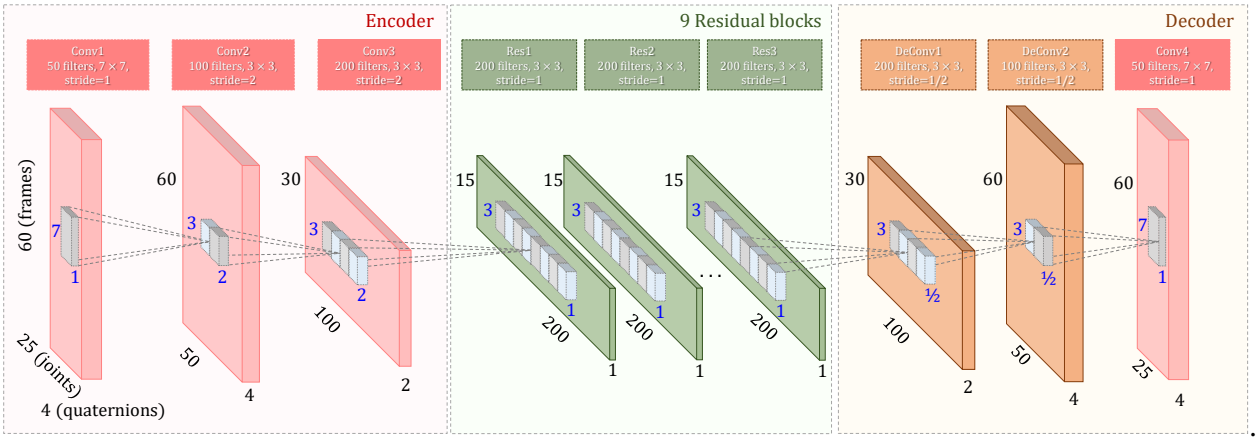


Figure 5-14. Architecture of the 1D generator.

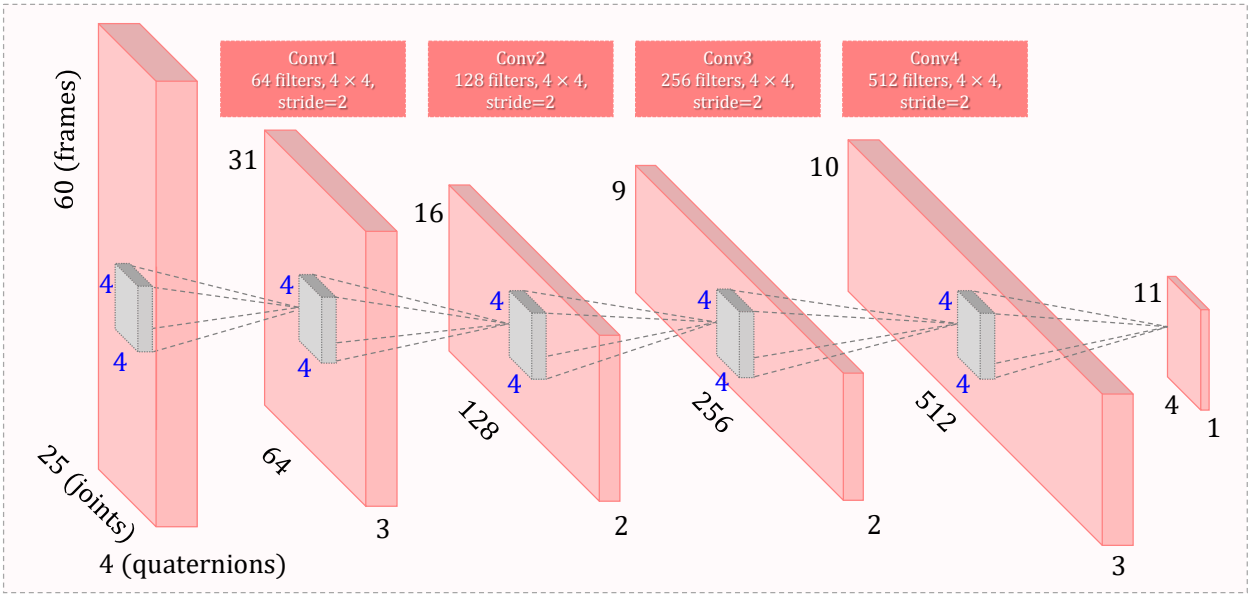


Figure 5-15. Architecture of the 1D discriminator.

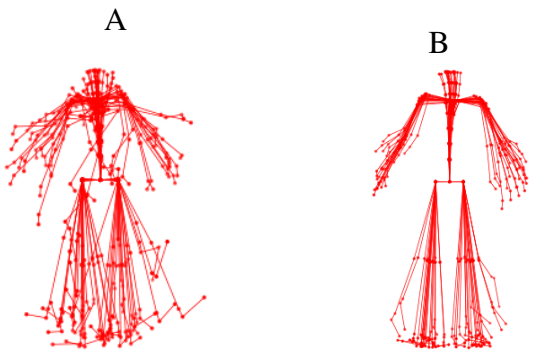


Figure 5-18. Timelapse on walking action on our dataset from the front view. A) child motion. B) input adult motion.

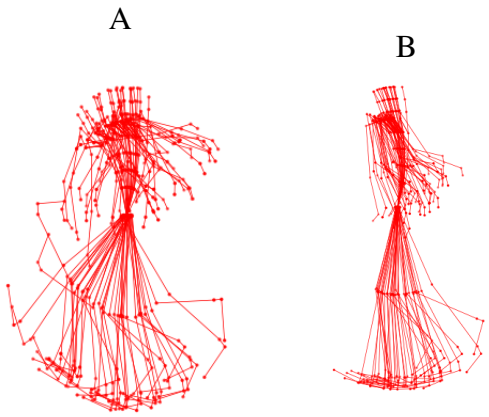


Figure 5-19. Timelapse on walking action on our dataset from the side view. A) child motion. B) adult motion.

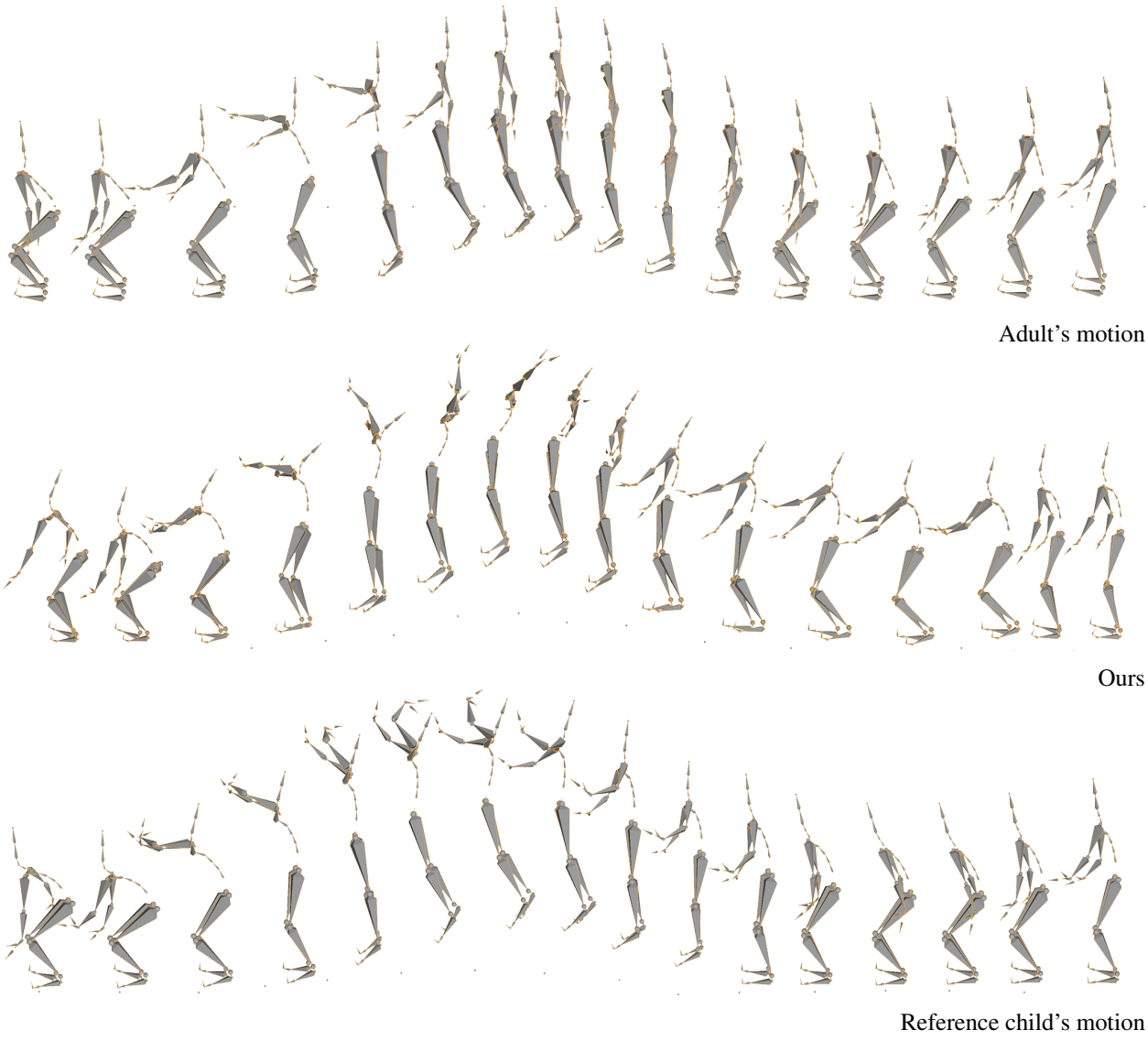


Figure 5-16. Results on "jump-as-high-as-you-can". It can be seen that the spine of the translated childlike motion using our method is more exaggerated than input adult motion, to look more similar to the reference child motion, while the hands are risen up in a more playful way.

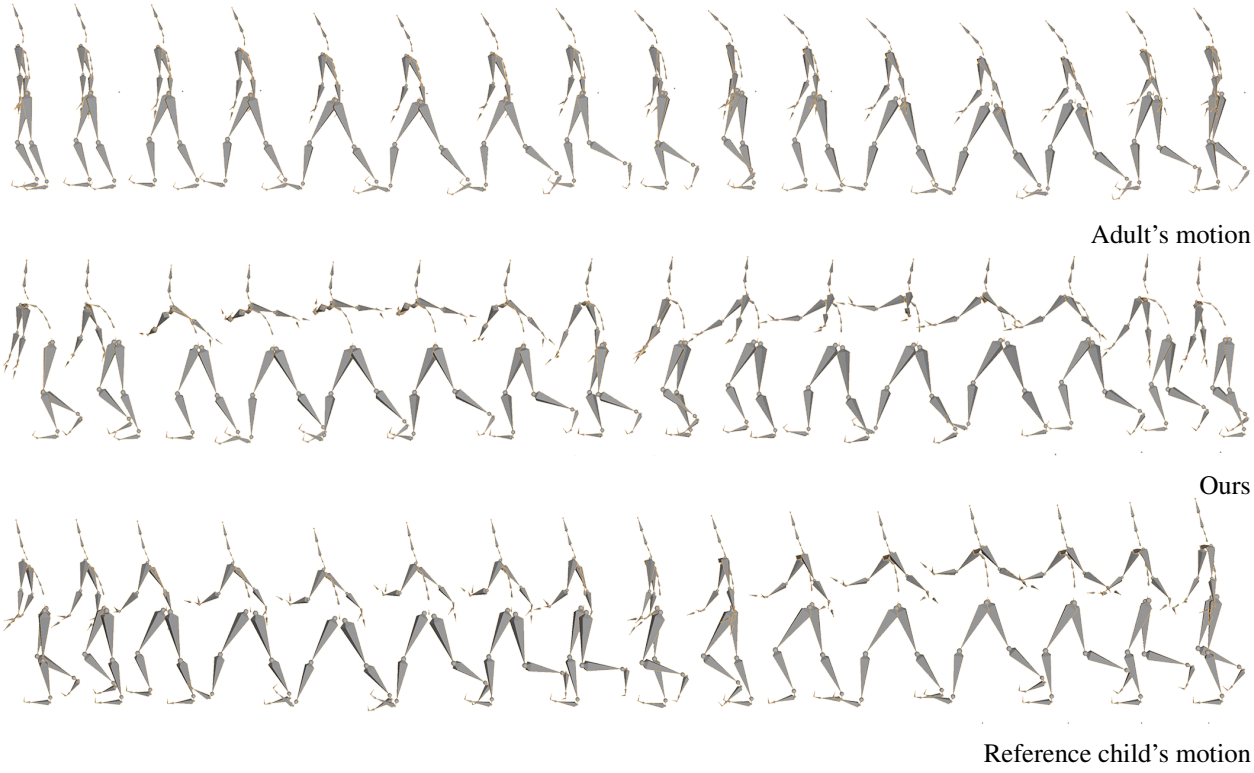


Figure 5-17. Results on "walk-as-fast-as-you-can". It can be observed that the output childlike motion using our method has the arms move in a carefree, uncoordinated, and playful mode during the walking cycle to mimic the child walking, whereas this is not seen in the original adult motion.

5.3.6 Discussion and Limitation

1. Working with heterogeneous motion types. Adapted CycleGAN can be used to learn human motions and extract style component from the motions. The output motions appeared more childlike compared to the original adult motions. It is also capable of learning different motion types within one network. This feature enables style translation for heterogeneous motions without first segmenting the motions. Another advantage compared to the traditional style translation methods that use time-warping [77, 31], is that our style translation algorithm can work on unaligned data and learn the temporal mapping from adult motions to child motions. Dynamic time warping is known to be a time-consuming process. Skipping this process can save us computation time, which is crucial for real-time applications such as games and animations.

   From the output childlike motion, we found that the network can translate style from one motion to a different motion type. For example, when we tested our architecture on [94]'s dataset, the actor had his/her arms swinging in the air in the running motion inside training data to act childlike. However, childlike walking and fast walk motions in the training data did not show this pattern. The output childlike walking motion picked up this behavior and also lifted the arms in the air. This indicates that the cycleGAN architecture can learn the characteristics from an entire dataset and translate what it learned from one action to another.

2. The size of the training set. For any machine learning model, the amount of data is key to the training. The adversarial loss can be very unstable. Abundant data can prevent the explosion of the loss function. Another issue with machine learning is that the model can overfit the training data when the training set is too small. If we include more data, the model will be tuned to deal with various cases and therefore prevent fine-tuning to a few examples. Including more data can improve the generalizability of the model. Our training data consists of three child actors and three adult actors performing ten actions. Our bottleneck is training time. We trained the model on Google Colab pro, and the training

time is capped to 24 hours. Training with more data could be a good step if we could get more computation resources as future work. Another limitation is that the network cannot do extrapolation, meaning it does not perform well for motions that are not in the training set. This limitation is inherited from neural networks. Neural networks are good at learning the underlying rules from the training data. However, for unseen examples, the networks can produce problematic and unnatural motions.

3. Architecture design. We trained our model with both 1D convolution and 2D convolution neural networks. From both results, we found that for a very diverse dataset like the child dataset we collected, 2D convolution is not capable of learning the style of the motions. For 1D convolution, each joint convolves with a separate filter, and the filter moves only along the time direction. For 2D convolution, although we trained multiple filters, each filter goes through all the joints and time frames. Because each joint has different behavior and a lot of the time, motions are not coordinated or balanced. We found 1D convolution capture the style better than the 2D convolution (Figure 5-14,5-15).

   Inside CycleGAN, we have two GANs to translate adult motions to a child motions and child motions back to adult motions. The child2adult translation didn't work as well as we expected. The fake adult motions are jumpy and disconnected. This could also be the reason the fake child motions are jittery for some motions. Another future work is to improve fake adult motion and see if it will compensate for the fake child motion.

4. Neural networks vs. Other approaches. Traditional style translation approaches have certain limitations when we want to achieve adult2child style translation. Dynamic scaling procedure accounts for the scaling of characters' size while maintaining the physics laws. In chapter 3 I found that dynamically scaled adult motions are perceived as more child-like but still not quite as much as the actual child motions. Learning from corpus leverages the existing motion dataset and learns the style from the training data. I adopted style translation algorithm [31] to transform adult motion capture data to look child-like. The

output motion is very similar to original child motion. However, given the child dataset is very limited, this approach only apply to the six motions in their training data. Another common approach to alter the style is through signal processing [88, 75, 97]. It achieves very promising results using simple computation in the time domain or frequency domain. The algorithms can modify the amplitude or frequency of the motion curve directly. However, algorithms that use Fourier transformation fit well for periodical motion but could not generalize well for non-periodical motions. Previous works also explored using procedural approach. Researchers integrate knowledge from personal experiences or literature to model the style in continuous space. It is ideal for creating motion for characters in all ages. But it can be mathematically challenging and labor consuming to generate realistic and compelling motion.

The neural network architecture are excel at learning patterns and features from data. By showing the networks tons of examples, they can adjust the weights towards creating examples that are similar to what they are exposed to. After training on the raw data for 20 hours, the system can synthesize childlike motions with a variety of motion types from adult motions. GANs are a type of generative architectures that can learn the mapping from two distributions instead of two examples, like the style translation algorithm. As we point out in section 5.3.4, child motions have a wide spectrum and large randomness. Even when they are asked to perform the same motion, they would improvise and perform extra motions. However, there are also commonality inside child motions such as exaggerations for poses and faster pace. GANs can learn features by observing adult motions and child motions we collected. Second, cycleGAN achieved success in style transfer without paired training data. finding temporally aligned adult and child mocap data is difficult and sometimes impossible. Traditional time warping technique can slow down the whole pipeline and therefore not suitable for real-time applications.

CHAPTER 6
SUMMARY AND CONCLUSIONS

Human bodies and human eyes are what we see every day in our life either from observing ourselves or looking at others. We are most familiar with how the bodies and eyes move and we are all experts on telling if a motion is natural. From my work, we realized that viewers can detect non-verbal cues as small as pupil dilation. In short, modeling human motions is a challenging, and important task. Avatar-based applications such as animations and games are big part of people's childhood, and even adulthood. Surveys have shown that in the UK, children from 8 to 11 years old spent 8 hours on video games and 10 hours on TV shows weekly in 2017 [78]. Games and animations serve not only as tools of entertainment and relaxation but also as motivations for learning and exercise [12, 65]. A study has shown that involving characters that are similar to the users could significantly increase their enjoyment and engagement in their experiences [33]. Hence, in order to attract more young users, many animated movies and video games adopted child characters in their stories.

In this dissertation, data-driven approaches were explored to create expressive eyes and bodies for virtual characters in distinct styles. we also developed mathematical models to generate motions with less laborious work. The work of this dissertation is dedicated to helping people in computer graphics related fields to create expressive avatars without going through the tedious process. My dissertation work is developed around two threads: in the first thread, we studied eye motions under scared emotion. In the second thread, the whole body motion from children and adults.

In the first thread that focuses on eyes, we animated characters' pupil diameters to match participants' pupil size when they were watching scary movies. From the perceptual studies, we learned that viewers can distinguish which character is scared based only on pupil dynamics. Modeling pupil dynamics right is a key for characters to convey emotions. To animate close-up shot of faces, animators can tweak the pupil diameter based on human physiology such as light reflex and emotional arousal. For characters who are scared or excited, their pupils can be enlarged. For VR content creators, they can consider using the actual pupil size from users collected by eye trackers. These findings open up several future work directions. First, it would

be interesting to see how this effect would interact with other facial animations that are commonly seen in characters such as raising eyebrows and enlarged eyes. Next for VR applications, researchers can test out how viewers would perceive this cues from avatars in VR environment.

In the second thread of my dissertation for creating whole body animation, we collected a dataset from adult subjects and child subjects to learn the patterns of child motions and how to show the patterns on adult motions. By leveraging the state-of-art neural network architecture, the network can model the complex characteristics of child motions including the randomness and coordination by observing the data. This study shows that a neural network architecture designed for image processing and computer visions can be applied to human motions with adaptations such as customized loss functions or types of kernels.

Human motion has a more compact representation than images. For example, the motion words we used has a size of 64 by 24 by 4. For images, it usually have sizes larger or equal to 256 by 256 by 3. Each data entry in the human motion is a degree of freedom bounded by kinematics constraints. For example, our limbs cannot bend backward. Deviations in one data point can cause the character look like having a broken arms and legs. For images, semantics meaning emerges when viewers connect thousands of pixels together. Imprecise values for one or two pixels would not break the entire image.

Generative adversarial networks are really good at generating instances that are similar to the training data, however, the result can be noisy. We introduced novel loss terms: temporal coherence loss and transition loss to ensure the motions are smooth and natural. We also adopted 1D convolution instead of 2D to learn the patterns in each join independently. GANs can be very unstable during training, causing loss functions to diverge. The solution for this is to train the model with more data. Compared to images, motion capture data are usually more difficult to acquire. Hence we collected our own training data. One of my contributions in my dissertation research is to collect this dataset from adults and children. The dataset as well as the code and the model will be released by the end of September 2020 and will benefit the animation and research community, especially for animators or content creators who want to model child motions and

researchers who are learning how child motions developed.

This work opens floor to additional research in this direction. For example, one limitation for our work is foot contacts and physical constraints. Physics-based neural network that can enforce such constraints would be very interesting to explore. Another future work could be developing a deep learning model for not only children and adults but people in all age groups such as teenagers and elderly people.

To conclude, the two threads in my dissertation research are interwoven with respect to methods and impact. I have developed and evaluated algorithms to generate expressive motion for human bodies and eyes. I have thus studied non-verbal expressiveness in human motion at two scales: whole body, consisting of the simultaneous movement of the entire human skeleton, and just the eyes, where I focused on subtle movement of just one component of the face. My evaluation methods have involved perceptual studies conducted in local institute, and over a crowd sourcing platform, again two scales of conducting human subjects experiments. Taken together, my dissertation work advances computer animation research by integrating algorithms development with perceptual evaluation so as to focus the technical contributions in those areas which have the greatest perceptual impact.

REFERENCES

[1] Karen E Adolph and John M Franchak, *The development of motor behavior*, Wiley Interdisciplinary Reviews: Cognitive Science **8** (2017), no. 1-2, e1430.

[2] Aishat Aloba, Gianne Flores, Julia Woodward, Alex Shaw, Amanda Castonguay, Isabella Cuba, Yuzhu Dong, Eakta Jain, and Lisa Anthony, *Kinder-gator: The UF kinect database of child and adult motion*, Eurographics, 2018, p. [To Appear].

[3] Kenji Amaya, Armin Bruderlin, and Tom Calvert, *Emotion from motion*, Graphics interface, vol. 96, Toronto, Canada, 1996, pp. 222–229.

[4] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher, *A head-eye coordination model for animating gaze shifts of virtual characters*, Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, ACM, 2012, p. 4.

[5] Emad Barsoum, John Kender, and Zicheng Liu, *Hp-gan: Probabilistic 3d human motion prediction via gan*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1418–1427.

[6] Pascal Bérard, Derek Bradley, Markus Gross, and Thabo Beeler, *Lightweight eye capture using a parametric model*, ACM Transactions on Graphics (TOG) **35** (2016), no. 4, 117.

[7] Pascal Bérard, Derek Bradley, Maurizio Nitti, Thabo Beeler, and Markus H Gross, *High-quality capture of eyes.*, ACM Trans. Graph. **33** (2014), no. 6, 223–1.

[8] Amit Bermano, Thabo Beeler, Yeara Kozlov, Derek Bradley, Bernd Bickel, and Markus Gross, *Detailed spatio-temporal reconstruction of eyelids*, ACM Transactions on Graphics (TOG) **34** (2015), no. 4, 44.

[9] Anthony Birat, Pierre Bourdier, Enzo Piponnier, Anthony John Blazevich, Hugo Maciejewski, Pascale Duché, and Sébastien Ratel, *Metabolic and fatigue profiles are comparable between prepubertal children and well-trained adult endurance athletes*, Frontiers in physiology **9** (2018), 387.

[10] Margaret M Bradley, *Natural selective attention: Orienting and emotion*, Psychophysiology **46** (2009), no. 1, 1–11.

[11] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang, *The pupil as a measure of emotional arousal and autonomic activation*, Psychophysiology **45** (2008), no. 4, 602–607.

[12] Thomas F Braeunig, *Body-mounted video game exercise device*, May 15 1990, US Patent 4,925,189.

[13] Matthew Brand and Aaron Hertzmann, *Style machines*, Proceedings of the 27th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., 2000, pp. 183–192.

[14] James W Davis, *Visual categorization of children and adult walking styles*, International Conference on Audio-and Video-based Biometric Person Authentication, Springer, 2001, pp. 295–300.

[15] Mark BA De Ste Croix, Martine A Deighan, Sebastien Ratel, and Neil Armstrong, *Age-and sex-associated differences in isokinetic knee muscle endurance between young children and adults*, Applied Physiology, Nutrition, and Metabolism **34** (2009), no. 4, 725–731.

[16] Emily L Denton, Soumith Chintala, Rob Fergus, et al., *Deep generative image models using a laplacian pyramid of adversarial networks*, Advances in neural information processing systems, 2015, pp. 1486–1494.

[17] Yuzhu Dong, Aishat Aloba, Lisa Anthony, and Eakta Jain, *Style translation to create child-like motion*, Eurographics, 2018.

[18] Julie S Downs, Mandy B Holbrook, Steve Sheng, and Lorrie Faith Cranor, *Are your participants gaming the system?: screening mechanical turk workers*, Proceedings of the SIGCHI conference on human factors in computing systems, ACM, 2010, pp. 2399–2402.

[19] Han Du, Erik Herrmann, Janis Sprenger, Noshaba Cheema, somayeh hosseini, Klaus Fischer, and Philipp Slusallek, *Stylistic Locomotion Modeling with Conditional Variational Autoencoder*, Eurographics 2019 - Short Papers (Paolo Cignoni and Eder Miguel, eds.), The Eurographics Association, 2019.

[20] Andrew Duchowski, Sophie Jörg, Aubrey Lawson, Takumi Bolte, Lech Świrski, and Krzysztof Krejtz, *Eye movement synthesis with 1/f pink noise*, Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games, ACM, 2015, pp. 47–56.

[21] Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos, *The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation*, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (New York, NY, USA), CHI '18, ACM, 2018, pp. 282:1–282:13.

[22] FOVE, *Home - fove eye tracking virtual reality headset*, 2018.

[23] Michael Gleicher, *Retargetting motion to new characters*, Proceedings of the 25th annual conference on Computer graphics and interactive techniques, ACM, 1998, pp. 33–42.

[24] Jaap Ham, Raymond H Cuijpers, and John-John Cabibihan, *Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures*, International Journal of Social Robotics **7** (2015), no. 4, 479–487.

[25] HELGE Hebestreit, KL Mimura, and ODED Bar-Or, *Recovery of muscle power after high-intensity short-term exercise: comparing boys and men*, Journal of Applied Physiology **74** (1993), no. 6, 2875–2880.

[26] Jessica K Hodgins and Nancy S Pollard, *Adapting simulated behaviors for new characters*, Proceedings of the 24th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., 1997, pp. 153–162.

[27] Daniel Holden, Jun Saito, and Taku Komura, *A deep learning framework for character motion synthesis and editing*, ACM Transactions on Graphics (TOG) **35** (2016), no. 4, 138.

[28] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce, *Learning motion manifolds with convolutional autoencoders*, SIGGRAPH Asia 2015 Technical Briefs, 2015, pp. 1–4.

[29] TOMOKI Horita, Kiyokazu Kitamura, and Nobuhiro Kohno, *Body configuration and joint moment analysis during standing long jump in 6-yr-old children and adult males.*, Medicine and science in sports and exercise **23** (1991), no. 9, 1068–1072.

[30] Marijana Hraski, Željko Hraski, and Ivan Prskalo, *Comparison of standing long jump technique performed by subjects from different age groups*, Baltic Journal of Sport and Health Sciences **98** (2015), no. 3, 2.

[31] Eugene Hsu, Kari Pulli, and Jovan Popović, *Style translation for human motion*, ACM Transactions on Graphics (TOG), vol. 24, ACM, 2005, pp. 1082–1089.

[32] Donald F Huelke, *An overview of anatomical considerations of infants and children in the adult world of automobile safety design*, Annual Proceedings/Association for the Advancement of Automotive Medicine, vol. 42, Association for the Advancement of Automotive Medicine, 1998, p. 93.

[33] Shang Hwa Hsu, Feng-Liang Lee, and Muh-Cherng Wu, *Designing action games for appealing to buyers*, CyberPsychology & Behavior **8** (2005), no. 6, 585–591.

[34] Leslie Ikemoto and David A Forsyth, *Enriching a motion collection by transplanting limbs*, Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation, Eurographics Association, 2004, pp. 99–108.

[35] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, *Image-to-image translation with conditional adversarial networks*, arxiv (2016).

[36] Eakta Jain, Lisa Anthony, Aishat Aloba, Amanda Castonguay, Isabella Cuba, Alex Shaw, and Julia Woodward, *Is the motion of a child perceivably different from the motion of an adult?*, ACM Transactions on Applied Perception (TAP) **13** (2016), no. 4, 22.

[37] Brendan John, Pallavi Raiturkar, Arunava Banerjee, and Eakta Jain, *An evaluation of pupillary light response models for 2d screens and virtual reality hmds*, Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST) (To Appear), 2018.

[38] Sophie Jörg, Andrew Duchowski, Krzysztof Krejtz, and Anna Niedzielska, *Perceptual adjustment of eyeball and pupil diameter jitter amplitudes for virtual characters*, ACM Transactions on Applied Perception **15** (2018).

[39] James Oat Judge, Sylvia Õunpuu, and Roy B Davis III, *Effects of age on the biomechanics and physiology of gait*, Clinics in geriatric medicine **12** (1996), no. 4, 659–678.

[40] Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Chua Tat-Seng, *Affective video summarization and story board generation using pupillary dilation and eye gaze*, 2011 IEEE International Symposium on Multimedia (ISM), IEEE, 2011, pp. 319–326.

[41] Darryn King, *Building a better digital eye*, Jul 2017.

[42] Gosia Kucharska, *How long does it take to produce an animation?*, 2017.

[43] Savvas Lazaridis, Dimitrios A Patikas, Eleni Bassa, Themistoklis Tsatalas, Konstantinos Hatzikotoulas, Christos Ftikas, and Christos Kotzamanidis, *The acute effects of an intense stretch-shortening cycle fatigue protocol on the neuromechanical parameters of lower limbs in men and prepubescent boys*, Journal of sports sciences **36** (2018), no. 2, 131–139.

[44] Jehee Lee and Sung Yong Shin, *A hierarchical approach to interactive motion editing for human-like figures*, Proceedings of the 26th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., 1999, pp. 39–48.

[45] Seong Jae Lee and Zoran Popović, *Learning behavior styles with inverse reinforcement learning*, ACM Transactions on Graphics (TOG), vol. 29, ACM, 2010, p. 122.

[46] Sooha Park Lee, Jeremy B Badler, and Norman I Badler, *Eyes alive*, ACM Transactions on Graphics (TOG), vol. 21, ACM, 2002, pp. 637–644.

[47] Yongjoon Lee, Seong Jae Lee, and Zoran Popović, *Compact character controllers*, ACM Transactions on Graphics (TOG), vol. 28, ACM, 2009, p. 169.

[48] Jan Lexell, Michael Sjöström, Ann-Sofie Nordlund, and Charles C Taylor, *Growth and development of human muscle: a quantitative morphological study of whole vastus lateralis from childhood to adult age*, Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine **15** (1992), no. 3, 404–409.

[49] Lennart Ljung, *System identification: theory for the user*, Prentice-hall, 1987.

[50] Wanli Ma, Shihong Xia, Jessica K Hodgins, Xiao Yang, Chunpeng Li, and Zhaoqi Wang, *Modeling style and variation in human motion*, Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Eurographics Association, 2010, pp. 21–30.

[51] Sandra P Marshall, *The index of cognitive activity: Measuring cognitive workload*, Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on, IEEE, 2002, pp. 7–7.

[52] Ian Mason, Sebastian Starke, He Zhang, Hakan Bilen, and Taku Komura, *Few-shot learning of homogeneous human locomotion styles*, Computer Graphics Forum, vol. 37, Wiley Online Library, 2018, pp. 143–153.

[53] Michaël Mathieu, Junbo Jake Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun, *Disentangling factors of variation in deep representations using adversarial training*, CoRR **abs/1611.03383** (2016).

[54] James McCann and Nancy Pollard, *Responsive characters from motion fragments*, ACM Transactions on Graphics (TOG), vol. 26, ACM, 2007, p. 6.

[55] Albert Mehrabian, *Nonverbal communication*, Transaction Publishers, 1972.

[56] Alberto Menache, *Understanding motion capture for computer animation and video games*, Morgan kaufmann, 2000.

[57] DA Minimas, *A critical evaluation of the lund and browder chart*, WOUNDS UK **3** (2007), no. 3, 58.

[58] Meinard Müller, Tido Röder, and Michael Clausen, *Efficient content-based retrieval of motion capture data*, ACM Transactions on Graphics (ToG), vol. 24, ACM, 2005, pp. 677–685.

[59] Philip R Nader, Robert H Bradley, Renate M Houts, Susan L McRitchie, and Marion O'Brien, *Moderate-to-vigorous physical activity from ages 9 to 15 years*, Jama **300** (2008), no. 3, 295–305.

[60] Mercedes de Onis, Adelheid W Onyango, Elaine Borghi, Amani Siyam, Chizuru Nishida, and Jonathan Siekmann, *Development of a who growth reference for school-aged children and adolescents*, Bulletin of the World health Organization **85** (2007), no. 9, 660–667.

[61] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins, *Fmdistance: A fast and effective distance function for motion capture data.*, Eurographics (Short Papers), 2008, pp. 83–86.

[62] Vitor F Pamplona, Manuel M Oliveira, and Gladimir VG Baranoski, *Photorealistic models for pupil light reflex and iridal pattern deformation*, ACM Transactions on Graphics (TOG) **28** (2009), no. 4, 106.

[63] Dario Pavllo, David Grangier, and Michael Auli, *Quaternet: A quaternion-based recurrent model for human motion*, arXiv preprint arXiv:1805.06485 (2018).

[64] Jean Piaget, *The grasp of consciousness (psychology revivals): Action and concept in the young child*, Psychology Press, 2015.

[65] Marc Prensky, *Digital game-based learning*, (2001).

[66] Alec Radford, Luke Metz, and Soumith Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, arXiv preprint arXiv:1511.06434 (2015).

[67] Marc H Raibert and Jessica K Hodgins, *Animation of dynamic legged locomotion*, ACM SIGGRAPH Computer Graphics, vol. 25, ACM, 1991, pp. 349–358.

[68] Pallavi Raiturkar, Andrea Kleinsmith, Andreas Keil, Arunava Banerjee, and Eakta Jain, *Decoupling light reflex from pupillary dilation to measure emotional arousal in videos.*, Symposium on Applied Perception, 2016, pp. 89–96.

[69] S Ratel, P Duche, A Hennegrave, E Van Praagh, and M Bedu, *Acid-base balance during repeated cycling sprints in boys and men*, Journal of Applied Physiology **92** (2002), no. 2, 479–485.

[70] Sébastien Ratel, Anne Tonson, Yann Le Fur, Patrick Cozzone, and David Bendahan, *Comparative analysis of skeletal muscle oxidative capacity in children and adults: a 31p-mrs study*, Applied Physiology, Nutrition, and Metabolism **33** (2008), no. 4, 720–727.

[71] Liu Ren, Alton Patrick, Alexei A Efros, Jessica K Hodgins, and James M Rehg, *A data-driven approach to quantifying natural human motion*, ACM Transactions on Graphics (TOG), vol. 24, ACM, 2005, pp. 1090–1097.

[72] Kerstin Ruhland, Sean Andrist, Jeremy Badler, Christopher Peters, Norman Badler, Michael Gleicher, Bilge Mutlu, and Rachel Mcdonnell, *Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems*, Eurographics state-of-the-art report, 2014, pp. 69–91.

[73] Kerstin Ruhland, Christopher E Peters, Sean Andrist, Jeremy B Badler, Norman I Badler, Michael Gleicher, Bilge Mutlu, and Rachel McDonnell, *A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception*, Computer Graphics Forum, vol. 34, Wiley Online Library, 2015, pp. 299–326.

[74] Sabine Schaefer, Ralf Th Krampe, Ulman Lindenberger, and Paul B Baltes, *Age differences between children and young adults in the dynamics of dual-task prioritization: Body (balance) versus mind (memory).*, Developmental Psychology **44** (2008), no. 3, 747.

[75] Ari Shapiro, Yong Cao, and Petros Faloutsos, *Style components*, Proceedings of Graphics Interface 2006, Canadian Information Processing Society, 2006, pp. 33–39.

[76] Mary Dorothy Sheridan, *The developmental progress of infants and young children*, vol. 102, HM Stationery Office, 1975.

[77] Harrison Jesse Smith, Chen Cao, Michael Neff, and Yingying Wang, *Efficient neural networks for real-time motion style transfer*, Proceedings of the ACM on Computer Graphics and Interactive Techniques **2** (2019), no. 2, 1–17.

[78] Statista, *Hours children spend gaming weekly in the united kingdom (uk) from 2013 to 2017, by age group (in hours).*, 2018, [Online; Retrieved April 21, 2018].

[79] William Steptoe and Anthony Steed, *High-fidelity avatar eye-representation*, Virtual Reality Conference, 2008. VR'08. IEEE, IEEE, 2008, pp. 111–114.

[80] David H Sutherland, RICHARD Olshen, L Cooper, and S LY Woo, *The development of mature gait*, J Bone Joint Surg Am **62** (1980), no. 3, 336–353.

[81] Seyoon Tak and Hyeong-Seok Ko, *A physically-based motion retargeting filter*, ACM Transactions on Graphics (TOG) **24** (2005), no. 1, 98–117.

[82] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, *Facevr: Real-time facial reenactment and eye gaze control in virtual reality*, CoRR **abs/1610.03151** (2016).

[83] Tobii, *Tobii eye tracking: the next natural step in vr*, 2018.

[84] Nikolaus F Troje, *Decomposing biological motion: A framework for analysis and synthesis of human gait patterns*, Journal of vision **2** (2002), no. 5, 2–2.

[85] Stewart G Trost, Russell R Pate, James F Sallis, Patty S Freedson, Wendell C Taylor, Marsha Dowda, and John Sirard, *Age and gender differences in objectively measured physical activity in youth*, Medicine & Science in Sports & Exercise **34** (2002), no. 2, 350–355.

[86] Laura C Trutoiu, Elizabeth J Carter, Iain Matthews, and Jessica K Hodgins, *Modeling and animating eye blinks*, ACM Transactions on Applied Perception (TAP) **8** (2011), no. 3, 17.

[87] Munetoshi Unuma, Ken Anjyo, and Ryozo Takeuchi, *Fourier principles for emotion-based human figure animation*, Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, ACM, 1995, pp. 91–96.

[88] Jue Wang, Steven M Drucker, Maneesh Agrawala, and Michael F Cohen, *The cartoon animation filter*, ACM Transactions on Graphics (TOG) **25** (2006), no. 3, 1169–1173.

[89] Qi Wang, Thierry Artières, Mickael Chen, and Ludovic Denoyer, *Adversarial learning for modeling human motion*, The Visual Computer (2018), 1–20.

[90] Robert L Waters, Helen J Hislop, Lynn Thomas, and Joyce Campbell, *Energy cost of walking in normal children and teenagers*, Developmental Medicine & Child Neurology **25** (1983), no. 2, 184–188.

[91] Wikipedia, *Clementine (the walking dead) — wikipedia, the free encyclopedia*, 2017, [Online; accessed 4-July-2017].

[92] Wikipedia contributors, *The boss baby — Wikipedia*, 2018.

[93] Andrew Witkin and Zoran Popovic, *Motion warping*, Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, ACM, 1995, pp. 105–108.

[94] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins, *Realtime style transfer for unlabeled heterogeneous human motion*, ACM Transactions on Graphics (TOG) **34** (2015), no. 4, 119.

[95] Sang Hoon Yeo, Martin Lesmana, Debanga R Neog, and Dinesh K Pai, *Eyecatch: simulating visuomotor coordination for object interception*, ACM Transactions on Graphics (TOG) **31** (2012), no. 4, 42.

[96] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu, *Seqgan: Sequence generative adversarial nets with policy gradient*, Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[97] M Ersin Yumer and Niloy J Mitra, *Spectral style transfer for human motion between independent actions*, ACM Transactions on Graphics (TOG) **35** (2016), no. 4, 137.

[98] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

BIOGRAPHICAL SKETCH

Yuzhu Dong received her B.S. degree in space science from Nanjing University in 2015 and the M.S. degree in computer engineering from University of Florida in 2017. She received her Ph.D. in human-centered computing from the University of Florida in 2020. Her research interests include deep learning, computer graphics and character animation.