LEARNING-BASED 3D HUMAN MOTION CAPTURE AND ANIMATION SYNTHESIS

by IKHSANUL HABIBIE

Dissertation zur Erlangung des Grades des
Doktors der Ingenieurwissenschaften (Dr.-Ing.)

der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

Saarbrücken, 2023

Date of Colloquium:	May 16, 2023
Dean of the Faculty:	Prof. Dr. Jürgen Steimle
Chair of the Committee:	Prof. Dr. Philipp Slusallek
Reviewers:	Prof. Dr. Christian Theobalt
	Prof. Dr. Michael Neff
	Prof. Dr. Antonio Krüger
	Prof. Dr. Gerard Pons-Moll
Academic Assistant:	Dr. Mohamed Elgharib

ABSTRACT

Realistic virtual human avatar is a crucial element in a wide range of applications, from 3D animated movies to emerging AR/VR technologies. However, producing a believable 3D motion for such avatars is widely known to be a challenging task. A traditional 3D human motion generation pipeline consists of several stages, each requiring expensive equipment and skilled human labor to perform, limiting its usage beyond the entertainment industry despite its massive potential benefits.

This thesis attempts to explore some alternative solutions to reduce the complexity of the traditional 3D animation pipeline. To this end, it presents several novel ways to perform 3D human motion capture, synthesis, and control. Specifically, it focuses on using learning-based methods to bypass the critical bottlenecks of the classical animation approach. First, a new 3D pose estimation method from in-the-wild monocular images is proposed, eliminating the need for a multi-camera setup in the traditional motion capture system. Second, it explores several data-driven designs to achieve a believable 3D human motion synthesis and control that can potentially reduce the need for manual animation. In particular, the problem of speech-driven 3D gesture synthesis is chosen as the case study due to its uniquely ambiguous nature. The improved motion generation quality is achieved by introducing a novel adversarial objective that rates the difference between real and synthetic data. A novel motion generation strategy is also introduced by combining a classical database search algorithm with a powerful deep learning method, resulting in a greater motion control variation than the purely predictive counterparts.

Furthermore, this thesis also contributes a new way of collecting a large-scale 3D motion dataset through the use of learning-based monocular estimations methods. This result demonstrates the promising capability of learning-based monocular approaches and shows the prospect of combining these learning-based modules into an integrated 3D animation framework.

The presented learning-based solutions open the possibility of democratizing the traditional 3D animation system that can be enabled

using low-cost equipment, e.g., a single RGB camera. Finally, this thesis also discusses the potential further integration of these learning-based approaches to enhance 3D animation technology.

ZUSAMMENFASSUNG

Realistische virtuelle menschliche Avatare sind ein entscheidendes Element in einer Vielzahl von Anwendungen, von 3D-Animationsfilmen bis hin zu neuen AR/VR-Technologien. Die Erzeugung glaubwürdiger Bewegungen solcher Avatare in drei Dimensionen ist bekanntermaßen eine herausfordernde Aufgabe. Traditionelle Pipelines zur Erzeugung menschlicher 3D-Bewegungen bestehen aus mehreren Stufen, die jede für sich genommen teure Ausrüstung und den Einsatz von Expertenwissen erfordern und daher trotz ihrer enormen potenziellen Vorteile abseits der Unterhaltungsindustrie nur eingeschränkt verwendbar sind.

Diese Arbeit untersucht verschiedene Alternativen um die Komplexität der traditionellen 3D-Animations-Pipeline zu reduzieren. Zu diesem Zweck stellt sie mehrere neuartige Möglichkeiten zur Erfassung, Synthese und Steuerung humanoider 3D-Bewegungen vor. Sie konzentriert sich auf die Verwendung lernbasierter Methoden, um kritische Teile des klassischen Animationsansatzes zu überbrücken: Zunächst wird eine neue 3D-Pose-Estimation-Methode für monokulare Bilder vorgeschlagen, um die Notwendigkeit mehrerer Kameras im traditionellen Motion-Capture-Ansatz zu beseitigen. Des Weiteren untersucht die Arbeit mehrere datengetriebene Ansätze zur Synthese und Steuerung glaubwürdiger humanoider 3D-Bewegungen, die möglicherweise den Bedarf an manueller Animation reduzieren können. Als Fallstudie wird, aufgrund seiner einzigartig mehrdeutigen Natur, das Problem der sprachgetriebenen 3D-Gesten-Synthese untersucht.

Die Verbesserungen in der Qualität der erzeugten Bewegungen wird durch eine neuartige Kostenfunktion erreicht, die den Unterschied zwischen realen und synthetischen Daten bewertet. Außerdem wird eine neue Strategie zur Bewegungssynthese beschrieben, die eine klassische Datenbanksuche mit einer leistungsstarken Deep-Learning-Methode kombiniert, was zu einer größeren Variation der Bewegungssteuerung führt, als rein lernbasierte Verfahren sie bieten.

Ein weiterer Beitrag dieser Dissertation besteht in einer neuen Methode zum Aufbau eines großen Datensatzes dreidimensionaler Bewegungen, auf Grundlage lernbasierter monokularer Pose-Estimation-Methoden. Dies demonstriert die vielversprechenden Möglichkeiten

lernbasierter monokularer Methoden und lässt die Aussicht erkennen, diese lernbasierten Module zu einem integrierten 3D-Animations-Framework zu kombinieren.

Die in dieser Arbeit vorgestellten lernbasierten Lösungen eröffnen die Möglichkeit, das traditionelle 3D-Animationssystem auch mit kostengünstiger Ausrüstung, wie z.B. einer einzelnen RGB-Kamera verwendbar zu machen. Abschließend diskutiert diese Arbeit auch die mögliche weitere Integration dieser lernbasierten Ansätze zur Verbesserung der 3D-Animationstechnologie.

ACKNOWLEDGMENTS

This thesis would not be possible without the support of many people.

First of all, I would like to thank my advisor Christian Theobalt for his guidance throughout my study. Christian is a great academic mentor who not only taught me important research-related lessons but also showed full support and encouragement whenever I faced non-academic challenges during the completion of this thesis. He is an inspiring leader who has done a fantastic job creating an outstanding research environment within his group. I am very grateful for the opportunity to have a first-hand learning experience from him.

I would like to thank my supervisors and postdocs, Gerard Pons-Moll, Michael Neff, Justus Thies, Weipeng Xu, Mohamed Elgharib, Kripasindhu Sarkar, and Diogo Luvizon for their tremendous help and advice in my projects. I also thank my collaborators, Yuxiao Zhou, Balamurugan Thambiraja, Simbarashe Nyatsanga, and Ahsan Abdullah. I really appreciate the meaningful discussions we have as well as some of the eventful moments we encountered during the submissions. Moreover, I also would like to thank the proofreaders of the thesis, Jiayi Wang, Rishabh Dabral, Marc Habermann, Mallikarjun BR, Diogo Luvizon, Soshi Shimada, Gereon Fox, and Mohamed Elgharib for their feedback.

I want to thank my friends and colleagues who have helped and shaped me in various ways during my time at MPI Informatics. I am grateful to Dushyant Mehta for his mentorship during my early years as a Ph.D. student. I thank Mohamed Elgharib for the encouraging discussions we had over some of the best Arabian cuisines in town – and for his regular visits to check on my well-being. I thank Oleksandr Sotnychenko for his ever-reliable help in all of my recording sessions. I also want to thank Soshi Shimada, Marc Habermann, Mallikarjun BR, Neng Qian, Rishabh Dabral, and Victor Rudnev for inviting me to the much-needed non-research related activities in Saarbruecken. Furthermore, I would like to say a big thank you to the inspiring and brilliant members of the GVV group, D4, and D6 departments for creating a vibrant and supportive atmosphere that I am proud to be a

part of. It is certainly one of the most wonderful environments that I will fondly remember.

A special mention goes to my officemates Marc Habermann, Lingjie Liu, Gereon Fox, and Heming Zhu for the memorable day-to-day interactions and the funny stories we shared over the years.

I also thank Onkel Lucas, Tante Tine, and Onkel Hansi for offering me a wonderful place to stay in Saarbruecken and for providing me with all the necessary preparations before (and after) coming to Germany.

Finally, I would like to thank my family for their eternal support and love.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 MOTIVATION

Producing realistic 3D human motion plays a vital role in many traditional computer graphics applications, such as the generation of virtual characters in movies and video games. Representing motion in 3D is also key to simulating the lifelike behavior of humans in a virtual environment, which is essential to creating an immersive experience for interactive augmented and virtual systems.

The standard techniques for generating 3D human animation are generally classified into two main categories: direct motion capture and motion synthesis. Motion capture is operated by tracking and recording a real human performer's body joint positions using a motion capture (mocap) system into a digital 3D representation. Once the captured motion data is cleaned, it can then be used directly to animate a virtual character. In contrast to direct capture, synthesis approaches generate a new 3D animation sequence either through a manual animation process or by using a particular synthesis algorithm. A typical motion synthesis procedure is conducted by combining and editing a collection of pre-recorded motion capture data into a new 3D animation. Alternatively, a synthesis algorithm can be designed to associate the motion information with a set of intuitive signals, e.g., directional locomotion information, driven by an analog controller. By manipulating these signals, motion control can then be performed to direct how the motion should be generated, thus, enabling a real-time motion synthesis. However, despite numerous available techniques to choose from, 3D motion generation is still widely known to be a challenging task to perform. Many existing approaches have considerable limitations and bottlenecks that demand expensive tools and skilled labor to overcome.

Animating 3D human motion has been generally reserved for a handful of dedicated practitioners in the movie or video game industries. The motion capture system commonly used today is typically performed in a studio setting using a specialized multi-view camera

system to track the position of several optical markers attached to the actor's body. This setup has several limitations: it requires costly equipment, often uses a tight body suit that limits both performers' movement flexibility and appearance, and cannot be used in outdoor or heavily occluded environments. Similarly, motion synthesis approaches also come with their own set of limitations. Manually animating a collection of motion capture data is a tedious task requiring hours of human labor to complete.

Motion control algorithms used in 3D video games are also not trivial to design and implement, as they rely on a large-scale motion state machine from where the final motion output is queried. The data of such a state machine is usually stored in a database consisting of tens of thousands motion captured data performing specific labeled actions. Constructing this state machine involves manually arranging tens of thousands of different motion sequences into hundreds of distinct action hierarchies along with the complex logic to select the correct action given the current state and control input.

Therefore, it is natural to examine a new set of techniques that can simplify and reduce the bottlenecks that hinder the current 3D generation approaches. This effort has the potential to democratize 3D animation technology in ways that will not only benefit the existing 3D animation pipeline but also has the potential to enable novel applications beyond the graphics-related industry.

Several recent attempts have demonstrated the benefits of simplifying the 3D animation process. Motion capture can now be alternatively performed using depth-based sensor cameras (Baak et al., 2011; Ganapathi et al., 2012; Girshick et al., 2011; Moon et al., 2018; Shotton et al., 2011; Ye et al., 2016) or markerless multi-view capture systems (Elhayek et al., 2016; Rhodin et al., 2015; Starck and Hilton, 2003; Stoll et al., 2011b) (see also Figure 1.1a). Depth-based camera systems can produce reasonably accurate motion capture results without the need for multiple sensors, and they have shown to be successful in commercial applications such as gaming consoles. Recent progress in markerless multi-view mocap technologies have also shown promising results, enabling motion capture in outdoor environments without the trouble of wearing a tight body suit. Lately, this technology has been used to guide soccer game officials in conducting critical decision-making by considering the 3D body positions of the players in the field (see also Figure 1.1b). Similarly, a lot of progress has also been made to

(a) Markerless Mocap (The Captury)

(b) Offside Detection (FIFA)

Figure 1.1: Markerless multi-view mocap system demonstrates several benefits of relaxing the constraints of the current 3D animation technology. Left: a commercial system offered by The Captury allows for a production-level motion tracking quality without the need for a specific body suit. Right: The recently introduced semi-automated markerless multi-view offside detection technology used by FIFA exhibits the potential of what a less constrained mocap system can achieve beyond the graphics industry.

reduce the bottlenecks in the task of synthesizing 3D motion. As an alternative to building an animation database, novel data-driven motion synthesis algorithms such as deep neural networks promise a new way to simplify 3D animation design by removing the need to create a complex motion state machine (Fragkiadaki et al., 2015; Habibie et al., 2017; Holden et al., 2017, 2016; Taylor and Hinton, 2009). For instance, the Motion Matching algorithm (Büttner and Clavet., 2015) used in the AAA-level video game The Last of Us 2 demonstrated considerably more expressive motion control results than the previous state-of-the-art approaches. While these alternatives are not without their limitations, it is clear that such efforts have elevated the impact of 3D animation technology compared to the existing methods.

Inspired by the recent progress, this thesis contributes to further expediting and simplifying several bottlenecks in the space of 3D motion generation. By leveraging predictive models that can learn from data, the proposed approaches presented in this thesis can be used together to automate the task of capturing, synthesizing, and controlling human motion using only low-cost commodity equipment. In particular, this thesis proposes a method for improving monocular 3D pose capture through a weak supervision strategy, two approaches to generating and controlling 3D gestures from speech, and a method

to perform 3D human motion synthesis from a text description. These are briefly explained in the following.

1.2 MONOCULAR 3D HUMAN POSE ESTIMATION

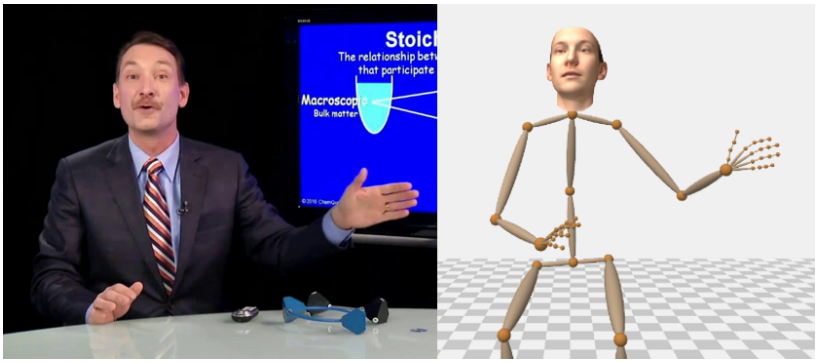
To design a democratized and easy-to-use motion capture technology, we ideally want a system that can be deployed and applied with an affordable and ubiquitous commercial device. With this in mind, this thesis considers the task of accurately estimating 3D human body pose from a single RGB image. Unfortunately, monocular human motion capture is an underconstrained problem that is known to be hard to solve. Without multi-view information or dedicated body markers, the capture process must deal with ambiguity, occlusion, and arbitrary foreground and background appearance variations. A popular way to solve this problem is by building a predictive capture model, which incorporates a strong prior to accurately infer the most likely 3D pose of the person in the given input image. Constructing such a prior involves extracting meaningful statistical information from a large collection of data.

Following their success in other vision-related tasks, recent data-driven approaches for monocular 3D pose estimation in the form of deep learning have shown significant improvement over the existing ideas, such as template-based body model fitting or graph-based priors (Andriluka et al., 2009; Guan et al., 2009; Sigal et al., 2007; Stoll et al., 2011b). These new methods usually employ a deep convolutional neural network, which has more parameters than the number of sample points, allowing it to learn a strong prior from the patterns found in the dataset. Most of these methods are trained in a supervised manner and assume that the training data consists of a pair of RGB images and its corresponding 3D body joint annotations. However, due to the inherent limitation of the contemporary commercial 3D motion capture setup, the 3D annotations are challenging to collect. The 3D labeled datasets used for this task are recorded in a studio setting with uniform background and show only a handful of unique subjects. The domain gap between the studio-specific training data and the endless potential appearance variations of the test data can lead to a significant performance drop when the model is tested on images containing outdoor scenes, which the model has never observed during training.

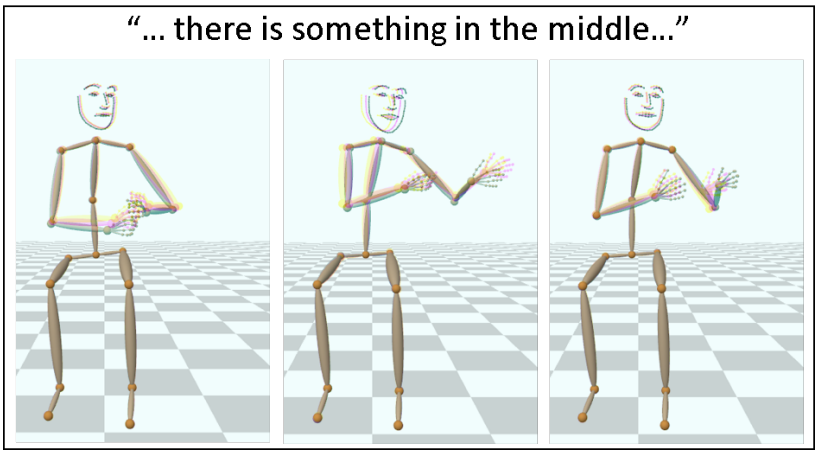
On the other hand, annotating the 2D pose of a human from a general RGB image taken in-the-wild is an easier task than its 3D counterparts as it can be performed manually. Learning-based 2D pose estimation models also do not suffer from the domain gap problem, as the images can be manually labeled regardless of the environment. Based on this observation, this thesis presents a weakly supervised 3D pose estimation approach that can learn 3D pose priors from 2D labels. This allows the model to be trained on the 2D-only labeled image when the 3D label is unavailable, such as the in-the-wild images. The key idea of presented in Chapter 3 of this thesis is a novel network architecture that can learn to update its 3D pose representation from 2D pose information by leveraging several weak supervision approaches. This results in a state-of-the-art model that generalizes better on outdoor scenes, even when it has never been explicitly trained on any 3D labeled in-the-wild images.

1.3 MOTION SYNTHESIS AND CONTROL FROM MULTI-MODAL SIGNALS

Recently, deep learning approaches have also shown superior performance in 3D motion synthesis and control, despite being traditionally known as a challenging task to solve due to the high-dimensional nature of the problem. Several works have demonstrated the capability of deep neural networks to synthesize high-quality 3D human locomotion from a given directional control information (Henter et al., 2020; Holden et al., 2020, 2017; Lee et al., 2014). This is a significant step towards automating the design process of creating 3D human motion animation. However, there are scenarios where directional information is not the most suitable signal to guide motion synthesis. As an example, it is not reasonable to use an analog controller to drive the body gesture of a talking character in virtual telepresence or Role-playing Games (RPG). Designing a system that can perform such a mapping will naturally improve the quality of the virtual interaction between users. The availability of such tools could allow animators to easily create new motion without having to capture excessive amounts of data. Therefore, this thesis examines novel techniques for generating human motion using non-directional control input, notably from audio signals.



(a) Stage 1: data capture



(b) Stage 2: speech-driven 3D motion synthesis

Figure 1.2: An illustration of the proposed speech-driven gesture synthesis approach by Habibie et al. (2021a). In the first stage (a), a large scale 3D annotations of hand and body poses, as well as facial expressions are collected from in-the-wild videos using monocular estimation methods. Then, in the second stage (b), a neural network learns how to map the speech input into 3D gestures of the hand, body, and facial expression using the captured training data. This work demonstrates the first step toward designing a unified capture and animation pipeline using learning-based approaches.

One particularly well-known multi-modal human motion synthesis topic is the task of predicting body gestures from the speech audio of a talking person. The existence of such correspondence has been supported by a number of psycho-linguistic studies, suggesting a strong correlation between speech and the body gesture that accompanies it. From the animation perspective, this problem can be modeled as a mapping from the input speech, which can be treated as the control signal, to the 3D gesture of the human body as output. However, the speech-to-gesture correlation is known to be ambiguous, as there exist multiple correct gestures that can satisfy the same speech input. Modeling this correspondence naively using a standard supervised learning

approach typically leads to a dampened gesture result, as such models do not take the stochastic nature of the gesture into consideration. This causes the model to predict an averaged-out gesture output, making the results appear unnatural. Another major hurdle in training a predictive speech-to-gesture model is the prospect of collecting hours of 3D gesture annotations, which is tedious and exhausting to perform, especially using the traditional marker-based motion capture system that limits the naturalness of the gesture. Chapter 4 of this thesis proposes a new adversarial training procedure to resolve the averaging problem commonly found in the task of gesture synthesis. We overcome the challenge of capturing hours of speech-gesture annotations by using 3D monocular capture approaches to annotate the 3D body, hand, and facial expression from in-the-wild monocular videos (see also Figure 1.2).

Furthermore, Chapter 5 of this thesis also proposes a further enhanced version of the previously described method that combines the adversarial learning-based approach with database matching, thus enabling the model to generate multiple plausible gestures from a single speech input. The idea is to pre-select a number of plausible motion candidates that satisfy a certain similarity score with respect to the audio and the previously generated motion. These pre-selected gestures are then combined with the original speech input to produce the final 3D gesture using an adversarial network. Interestingly, the proposed database search algorithm can also be conditioned by constraining the search space using a certain quantifiable aspect of the gesture like the position or speed of the hand. This feature allows the animator far greater flexibility in designing the motion towards a particular style or emotion, such as fast, subdued, or angry gestures.

1.4 STRUCTURE

This thesis consists of six main chapters. Each of chapter can be summarized as follows:

- Chapter 1 explored the motivation of this research topic and provides an overview of the content of the thesis.
- Chapter 2 introduces several technical background concepts required for later chapters.

- Chapter 3 (published as Habibie et al. (2019)) presents an approach to improve the performance of learning-based 3D monocular human pose estimation methods on in-the-wild images. The key contribution of the proposed method is the introduction of an explicit 2D pose representation in the latent space and a 3D-to-2D projection layer. This allows the neural network to be jointly trained on 2D-only labeled data, which is far easier to collect than the studio-captured 3D pose annotations.

- Chapter 4 (published as Habibie et al. (2021a)) proposes a learning-based method that can simultaneously drive the face and the prosody-correlated 3D upper body and hand of a 3D virtual human avatar from the corresponding speech input. This task is challenging to solve using a naive learning-based approach, as the multi-modal nature of the speech gesture can result in an averaged motion generation that appears unnatural. This work introduces an adversarial learning strategy to resolve this issue by classifying whether the predicted motion can be distinguished from the real gesture distribution or not. Furthermore, this work also provides a large scale 3D upper body, hand, and face annotations from more than 30 hours of in-the-wild videos of talking people, which is expensive to conduct in a controlled setting.

- Chapter 5 (published as Habibie et al. (2022)) introduces a novel framework for a speech-driven 3D upper body motion synthesis that can incorporate different levels of gesture style control. This includes high-level control signals such as hand height and velocity or even lower-level control, such as frame-based gesture matching. The proposed approach combines the flexibility of a database search algorithm that allows for convenient constraint-based look-up with the effectiveness of the adversarial learning approach, thus, further enhancing the quality of the previously retrieved gesture sequences.

- Chapter 6 concludes this thesis through a brief summary of insights that have been gathered from the previous chapters and provides an outlook on the potential future ideas.

1.5 PUBLICATIONS

All methods presented in this thesis can be found in the following publications:

- Ikhsanul Habibie et al. (2019). "In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ikhsanul Habibie et al. (2021). "Learning Speech-driven 3D Conversational Gestures from Video". In: *ACM International Conference on Intelligent Virtual Agents (IVA)*.
- Ikhsanul Habibie et al. (2022). "A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech". In: *SIGGRAPH '22 Conference Proceedings*.

Additionally, contributions were also made in the following publications. However, these papers are not considered a part of this thesis.

- Yuxiao Zhou et al. (2020). "Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuxiao Zhou et al. (2021). "Monocular Real-time Full Body Capture with Inter-part Correlations". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

2

PREREQUISITES

This chapter introduces several background concepts that will support the description of the methods proposed in this thesis. First, relevant 3D data representation regarding the human body, hands, and face is introduced. Second, it also discusses an approach to extract useful features of human speech, which will be used as a control signal for the motion generation models introduced in Chapter 4 and Chapter 5.

2.1 3D REPRESENTATION OF HUMAN BODY, HAND, AND FACE

Due to the complex structure of human anatomy, modeling a digitally accurate representation of its parts and interactions is a computationally challenging task. For this reason, specific data structures are often required to capture essential information regarding the visual 3D model of the human body to avoid high computational penalty. This thesis uses two separate models to simplify the 3D representation of the human body.

The 3D body and hands of the virtual characters are modeled using a *skeleton representation*, which describes the 3D articulation of human bodies using the 3D position of several key body joints. These joints are represented by 3D vertices, with the bones connecting them represented as hypothetical lines with fixed distances (see Figure 2.1). This skeleton

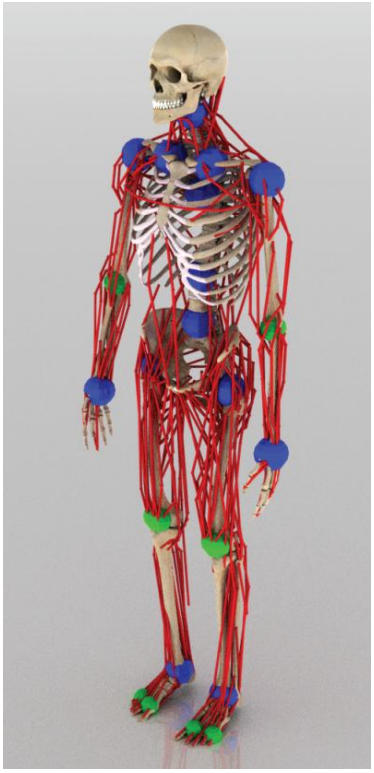


Figure 2.1: A muscoskeletal model of a human body. This representation can be simplified by only using its kinematic structure, where only relevants joints (shown as blue and green spheres) and their connections are considered (Figure taken from Lee et al. (2019b)).

representation can easily be extended by adding more complexities depending on the application, such as by introducing the muscles to model the dynamical constraints.

To model the face, this thesis uses a *3D morphable model* (3DMM) representation (Blanz and Vetter, 1999) that parameterizes a 3D face mesh using semantic attributes such as shape, expression, albedo, lighting, and head rotation. The 3DMM is used to represent the facial expression of a virtual talking human introduced in Chapter 4. Essential information such as lip movement and facial expression can be extracted from the mesh to obtain key 3D landmarks that represent the change of expression.

2.1.1 *Skeleton Model of Body and Hands*

Psychological studies on simplified body representation have shown that humans can perceive and recognize natural human movement behavior using only a few key points of the body (Johansson, 1973). This study suggests that a collection of body joints is a sufficient data points to represent human pose and motion. In this data structure, the body pose at frame t can be represented using a vector that stores the global x, y, and z positions of the N number of key body joints.

For the task of capturing or synthesizing the animation of the 3D body and hands, the goal is to predict the joint location at every frame. Key point locations of the body often serve as a more effective and accurate way to represent the target prediction for such models. Consequently, this thesis mainly uses 3D joint positions to represent output prediction for 3D monocular pose estimation task (Chapter 3) and motion synthesis from speech (Chapter 4 and Chapter 5).

Alternatively, the temporal evolution of the skeleton can also be modeled based on its kinematic properties. To model changes between different body poses, every joint j in the skeleton $\mathcal{S}$ is associated with a transformation $T_j \in SE(3)$ that can be used to map its own local coordinate to the local coordinate of its parent $par(j)$. Every joint is assumed to be at the origin of its own local coordinate. This rigid transformation can be described in homogeneous coordinate using

$$J_j^p = T_j \cdot \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^\mathsf{T}, \tag{2.1}$$

$$J_j^p = \left[\begin{array}{ccc|c} & & & t_x \\ & \mathbf{R} & & t_y \\ & & & t_z \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \tag{2.2}$$

where $\mathbf{R} \in SO(3)$ is the rotational component of the transformation, and the vector $\mathbf{t} = [t_x, t_y, t_z]^\mathsf{T}$ is the offset which represent the bone distance from a joint j to its immediate parent. Hence, a particular 3D pose of the skeleton in the global coordinate can be performed by recursively applying a series of local transformations along the kinematic chain. In general, the global position of a joint i can be found by iteratively applying the local transformation J_j^p from the root node:

$$J_i^g = \Big(\prod_{parents(j)} T_j\Big) \cdot T_i \cdot \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^\mathsf{T}, \tag{2.3}$$

where $parents(j)$ denotes the list of parental nodes of joint j from the root node.

This kinematic representation is also commonly found in commercial 3D animation systems, and it can be directly extended with a more detailed surface or volumetric representation of the body. The motion of a mesh-based 3D virtual character is typically controlled by the kinematic skeleton representation, which is achieved by rigging the model to the skeleton. Consequently, it is often necessary to transform a set of 3D human keypoints, e.g. from a 3D monocular prediction of human pose, into the kinematic representation. The parameters can be recovered by applying inverse kinematic algorithms (Aristidou et al., 2018). This is achieved by estimating the optimal rigid transformation parameters T_j with respect to the given global 3D joint positions.

2.1.2 *3D Morphable Model of Human Faces*

Compared to the sparse 3D keypoints for kinematic representation, a denser 3D model is typically required to capture the shape and expression of a human face accurately. One popular approach to represent this information is by using a 3D morphable model (3DMM). A 3DMM

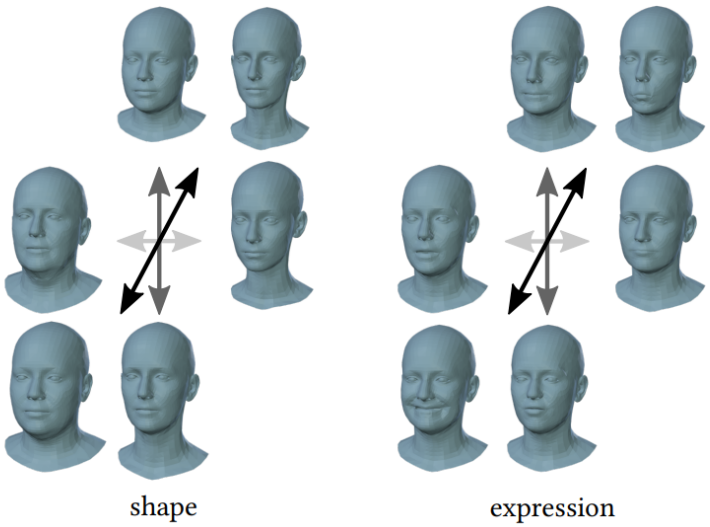


Figure 2.2: The 3D morphable model of a face can be used to generate a certain face based on a specific shape, expression, as well as skin reflectance by varying the coefficient values of each attributes. (Figure taken from Li et al., 2017)

of a human face (Egger et al., 2020) captures both the geometry and appearance of the face by parameterizing a 3D mesh consisting of N number of vertices. With this parameterization, a 3DMM can be used to either represent or generate a new instance of a face based on some semantically meaningful attributes, e.g., face identity or expression. In this way, a 3DMM offers an efficient way to describe a specific deformation of the data, offering a significant parameter reduction compared to using the 3D vertex positions directly. As a result, the 3DMM is often used for various tasks related to 3D face reconstruction, including learning-based 3D geometry prediction. This thesis makes use of this capability by employing a state-of-the-art monocular dense face reconstruction method to obtain the 3D annotations in terms of 3DMM parameters for the speech-gesture training data introduced in Chapter 4. Similarly, the task of predicting 3D facial expression from speech can be defined as predicting the expression parameters of a 3DMM.

The parameters of a face 3DMM are usually learned from a collection of real-world 3D scans of human faces involving a number of subject performing various expressions. By carefully curating the samples of this captured data, the variation of the face can be modeled using several semantically meaningful criteria, including shape geometry, expression, and skin reflectance (see also Figure 2.2). This

process is commonly performed by decomposing the data using dimensionality reduction techniques such as Principal Component Analysis (PCA). The semantically disentangled attributes are modeled by the eigenvectors obtained during the PCA decomposition.

The semantic attributes of the geometry of a 3DMM are represented as matrices of the face shape model $\boldsymbol{E}_s \in \mathbb{R}^{3N \times N_s}$ and face expression model $\boldsymbol{E}_e \in \mathbb{R}^{3N \times N_e}$ where N_s and N_e denote number of eigenvectors which correspond to the N_s and N_e highest eigenvalues in their respective PCA decomposition. A specific instance of a 3D face can then be expressed as a linear combination between the template 3D face and the semantic attributes. By defining the weight of each facial attributes in terms of the face shape coefficients $\boldsymbol{\alpha}_s \in \mathbb{R}^{N_s}$ and expression coefficients $\boldsymbol{\alpha}_e \in \mathbb{R}^{N_e}$, an instance of a 3D face mesh geometry $\boldsymbol{V} \in \mathbb{R}^{3N}$ can be expressed using

$$\boldsymbol{V} = \boldsymbol{T}_s + \boldsymbol{E}_s\boldsymbol{\alpha}_s + \boldsymbol{E}_e\boldsymbol{\alpha}_e, \tag{2.4}$$

where $\boldsymbol{T}_s \in \mathbb{R}^{3N}$ denotes the vertex position of the average face mesh template.

2.2 HUMAN SPEECH FEATURE EXTRACTION

To properly control human motion from speech signals using a learning-based approach, meaningful features need to be extracted from the original waveform representation of speech. In speech recognition, one popular choice of speech feature representation is the Mel Frequency Cepstrum (MFC) (Davis and Mermelstein, 1980; Mermelstein, 1976). These features are designed to capture the distinct phonemes generated by the human vocal tract. In particular, human speech characteristic is embedded in the envelope of the log power spectrum of the original audio signal. The MFC features are formulated to extract this information effectively.

The commonly used algorithm to extract MFC features can be described by the following steps:

1. *Windowing*. The speech signal is first divided into multiple chunks of short audio frames with some overlap. The window size of each frame typically falls between 25-40 ms with an overlapping step size of 10 ms between them. These operations ensure that the

window size is sufficient to perform spectral analysis, while at the same time also short enough to center and isolate individual phonemes.

2. *Compute Power Spectrum Periodogram*. In the next stage, the frame of the signal $x(n)$ is transformed into the frequency domain by applying the Discrete Fourier Transform (DFT):

$$X(k) = \sum_{n=0}^{N-1} h(n) \cdot x(n) \cdot \exp\left\{\frac{-i2\pi nk}{N}\right\}, \tag{2.5}$$

where n, N, and k respectively denote the signal sample index, the number of signal frames, and the frequency bin index. A window function $h(n)$ such as the Hanning window is often used in this operation to counter the side effects of performing a Fourier Transform on a short-time signal.

For the purpose of creating a meaningful speech feature, we are interested in analyzing the power distribution of the speech in terms of its frequencies. To this end, the periodogram of the power spectrum is computed according to the following:

$$P(k) = \frac{1}{N}|X(k)|^2. \tag{2.6}$$

3. *Mel Spectrum*. Human ears perceive audio frequencies in a non-linear way, with a stronger sensitivity in the lower frequency range. The Mel-scale is introduced to adjust the physical frequency of the data f to match the sensitivity of human ears, defined by

$$f_{Mel} = 2595 \log_{10}(1 + \frac{f}{700}). \tag{2.7}$$

The next objective is to compute the Mel spectrum $s(m)$, which is obtained by applying m number of band-pass filters known as the Mel-filter bank to the power spectrum of the audio frame $P(k)$. The filter bank is constructed by converting a collection of evenly-spaced triangular filters in the frequency domain to the Mel scale according to the formula introduced in Equation 2.7. As shown in Figure 2.3, the Mel-filter bank usually consists of $m = 40$ different band-pass filters.

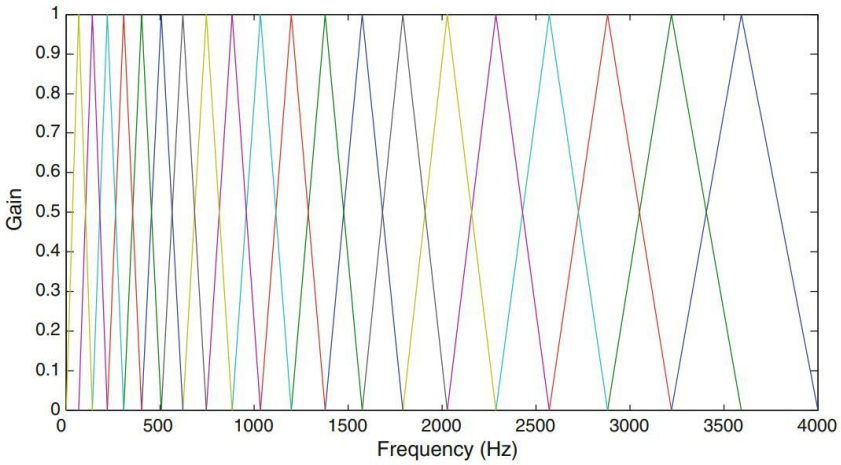


Figure 2.3: A visualization of the Mel-filter bank (Figure taken from Rao and Manjunath, 2017).

4. *Inverse Discrete Fourier Transform (DFT)*. In the last step, an Inverse DFT is applied to the log of the Mel spectrum $log_{10}(s(m))$. The purpose of this step is to decorrelate the signal. Since the signal is real-valued, a Discrete Cosine Transform (DCT) is commonly used to implement the Inverse DFT. The DCT transforms the spectral coefficient of the signal into the cepstral domain. The results of this transformation are commonly referred to as the Mel-frequency Cepstral Coefficients (MFCC). Since the DCT can also be seen as a compression step, we can choose the first few coefficients of the resulting cepstrum as the final feature of the audio signal. In this thesis, the first 13 MFCC coefficients are used as the input feature for the gesture synthesis models proposed in Chapter 4 and Chapter 5.

3

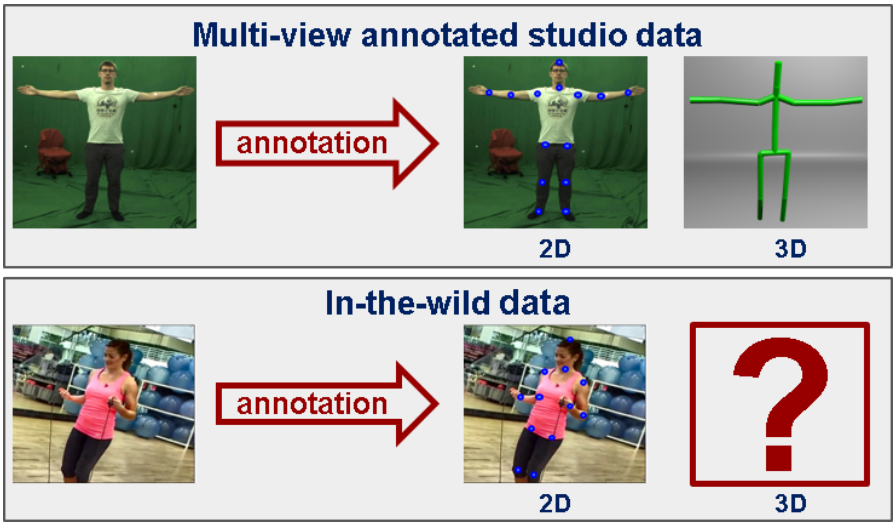
IN-THE-WILD 3D MONOCULAR POSE ESTIMATION

This chapter introduces a new deep learning-based monocular 3D human pose estimation method that shows high accuracy and generalizes better to in-the-wild scenes (published as Habibie et al., 2019). Unlike many prior learning-based approaches that primarily rely on studio-captured 3D pose annotations, the proposed network architecture enables novel ways to weakly supervise a 3D pose estimator using 2D pose labels when the 3D pose information is unavailable. This architecture allows the network to be trained on large corpora of 2D pose annotated images that are easier to obtain, achieving state-of-the-art accuracy on challenging in-the-wild data.

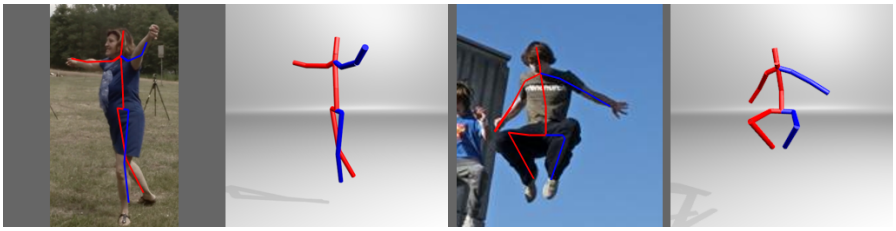
3.1 INTRODUCTION

Human motion capture has a wide range of applications in computer animation and also other areas such as biomechanics, medicine, and human-computer interaction. However, the standard 3D human motion capture systems typically require marker suits and/or multiple cameras recording in a controlled setting which are expensive and complicated to set up, and are impractical outside of the lab or studio environments. Methods that infer 3D pose only from monocular images overcome many such limitations and make 3D pose estimation more widely applicable. However, due to the under-constrained nature of monocular 3D pose estimation, achieving accurate 3D prediction is still a challenging task.

Recent progress of Convolutional Neural Networks (CNN) (Krizhevsky et al., 2012, He et al., 2016) has enabled promising learning-based methods for 3D human pose estimation from a single color image. Training such methods typically requires a large amount of RGB images annotated with reference 3D poses from either marker-based or markerless multi-camera motion capture systems (Elhayek et al., 2015; Joo et al., 2018; Rhodin et al., 2016; Stoll et al., 2011a), synthetic data (Chen et al., 2016), or IMU-based systems (Huang et al., 2018; Marcard et al.,



(a) 2D and 3D labeled multi-view studio vs. 2D labeled in-the-wild data



(b) Prediction results of the proposed method

Figure 3.1: (a) This chapter aims to improve the quality of learning-based 3D monocular pose estimation approaches by leveraging the 2D annotations of the in-the-wild-images. (b) Example 3D pose prediction results of the weakly supervised method for general scenes proposed in this chapter.

2018, 2017). Owing to this complex reference data capturing, diversity in real-world appearance or pose is hard to achieve in training data, which limits the generalization of trained networks on in-the-wild scenes.

Previous work has leveraged features learned on in-the-wild annotated 2D pose data to improve in-the-wild generalization. Some methods (Mehta et al., 2017a,b) proposed to finetune this learned representation on 3D pose prediction using 3D pose datasets captured in a studio. Others (Zhou et al., 2017) use this learned representation as initialization to jointly predict 2D key points and depth information. For images where the 3D annotations are available, both 2D keypoints and depth are supervised, with supervision coming from geometric constraints otherwise. In this way, networks carry features useful for in-the-wild 2D in order to achieve better 3D pose estimation in out-of-studio settings.

Using a strong pre-existing pose prior, like a parametric body model, can also help a network to predict more accurate 3D poses if labeled 3D training data is scarce (Kanazawa et al., 2018; Yang et al., 2018).

Since 3D pose labels on general scene images are hard to obtain while larger annotated 2D training corpora exist, several deep learning based methods resort to using 2D pose as the target prediction, followed by an additional 3D pose lifting step (Bogo et al., 2016; Chen and Ramanan, 2017; Insafutdinov et al., 2016; Martinez et al., 2017b; Tomè et al., 2017; Wang et al., 2018; Yasin et al., 2016). Using such, Martinez et al. (2017b) showed that 2D pose data alone is enough to train a network that achieves promising 3D pose estimation accuracy. However, solely predicting 3D from 2D pose is an inherently ambiguous task, and in these approaches, important 3D pose cues from the image are neglected.

This chapter introduces a new convolutional neural network architecture for 3D pose estimation that achieves state-of-the-art accuracy on challenging in-the-wild data. It has two main innovations that enable us to effectively train the network using both, more scarcely available image data with 3D annotation and more easy to generate image data with only 2D annotation.

The first innovation is inspired by 2D-to-3D pose lifting (Martinez et al., 2017b), but maintains the network's capability to utilize 3D cues in images explicitly. To this end, explicit 2D keypoint features are encoded as joint heatmaps in some channels of the convolutional latent space, leaving the rest of the features to contain "depth" information about the human pose. Separating the 2D and depth, and supervising 2D with additional in-the-wild data, which has been the primary driver of accurate 2D pose estimation methods (Cao et al., 2017; Newell et al., 2016; Wei et al., 2016), allows the network to consequently predict 3D pose more reliably even under a significant shift of the input appearance between the training and testing time. These 2D pose features can be trained jointly with depth features on data with 3D annotations or trained independently on data with 2D annotations, while in both cases improving overall network performance.

The second innovation is a supervised approach that reduces 3D-to-2D ambiguity when training on data with 2D annotations only. To this end, this chapter introduces a neural network that learns how to estimate the location of 2D body joints by using the 3D human pose predicted from the earlier network layers as latent features. More specif-

ically, the proposed network learns to predict the weak perspective camera parameters of the given monocular image input that projects the predicted 3D pose to the 2D space. During training, this projection loss can be used to update the information on 3D joint positions regardless of whether the training image has 3D labels or only 2D labels.

The proposed approach achieves a state-of-the-art accuracy of 70.4% 3D PCK on the MPI-INF-3DHP benchmark with challenging outdoor scenes, even when trained only using images with 3D pose labels from the H3.6M (Ionescu et al., 2014) studio dataset. When jointly training on larger corpora of in-studio images with 3D labels and in-the-wild data with 2D labels, we achieve 91.3% 3D PCK on MPI-INF-3DHP, which outperforms all previous methods.

3.2 RELATED WORK

Human pose estimation is an actively studied area in computer vision. This section focuses discussion on recent learning-based approaches that are relevant to the proposed method.

3.2.1 *3D pose from 2D keypoint detection*

Due to the robustness of some recent CNN-based 2D pose detection methods (Cao et al., 2017; Newell et al., 2016; Tompson et al., 2015; Toshev and Szegedy, 2014; Wei et al., 2016), many 3D pose estimation methods reformulate the task as a combination of 2D keypoints prediction and body depth regression. Mehta et al. (2017b) combine 2D heatmap prediction with 3D location maps to estimate the position of each joint in the 3D space. Zhou et al. (2017) propose a weak supervision training scheme using a stacked hourglass network by Newell et al. (2016) on both in-the-wild 2D data and studio data with 3D labels. The network is trained to predict 2D pose on both studio and outdoor datasets and, at the same time, also learns to predict depth information from the 3D labeled data. Yang et al. (2018) also use similar weak supervision, but they extend this idea by introducing an adversarial network that learns how to differentiate between ground truth and a predicted pose generated by the 3D pose prediction network. Another similar line of work is proposed by Dabral et al. (2018),

which improves this approach further by using body symmetry constraints and a separate temporal prediction network to achieve better 3D prediction stability across sequential frames. To take full advantage of the detection-based method, Pavlakos et al. (2017) proposed using a volumetric representation as an extension of the 2D joint heatmaps in the 3D space. However, this formulation is computationally expensive to perform even after using the coarse-to-fine strategy proposed to mitigate this issue.

3.2.2 *Direct 3D pose prediction*

Instead of using the combination of 2D and depth prediction, several works regress 3D body keypoints directly. Tekin et al. (2016) enhance a direct 3D prediction network by learning human body structure using a pose autoencoder.

Mehta et al. (2017a) use multiple intermediate supervision tasks, such as predicting the output at various network levels and predicting 2D heatmaps as an additional objective. They use two-step training approach to improve generalization. The network is first trained to learn 2D joint heatmaps and then refined on the task of directly predicting 3D joint location maps from 3D annotated studio data. Instead of directly predicting the keypoints, Zhou et al. (2016) regresses the joint angles on a kinematic body model, assuming that the bone length of the subject is known. Sun et al. (2017) uses a geometry-aware formulation that also predicts bone length and vector orientation instead of only regressing 3D keypoint locations.

Rhodin et al. (2018b) proposed a multi-view consistent prediction approach during training to refine neural network's monocular pose prediction on general scenes, but it requires synchronized multi-camera footage to train. Multi-view settings can also be used to perform unsupervised or semi-supervised learning on human pose estimation by training the network to learn a geometry-aware latent space that can generate novel views on different cameras (Rhodin et al., 2018a).

3.2.3 *3D lifting without depth information*

Some methods compute 3D pose by estimating the depth from the detected 2D keypoints only. Tomè et al. (2017) performs a sequence

of 3D lifting and reprojection to improve prediction quality iteratively. Chen and Ramanan (2017) find the closest 3D pose from a library of human poses that best matches the detected 2D pose. Martinez et al. (2017b) use a fully connected neural network with a residual connection can achieve accurate 3D pose estimation performance using 2D ground truth or a very accurate 2D keypoint detection as input. Regardless, these approaches cannot overcome the principled ambiguity that there are many possible 3D body poses that can be correctly projected into the corresponding 2D pose. To reduce this ambiguity of 3D lifting from 2D estimates, Pavlakos et al. (2018) use ordinal depth annotation between joint pairs, which is a special case of posebits introduced by Pons-Moll et al. (2014).

3.2.4 *Estimating 3D pose using 2D projection information*

Bogo et al. (2016) fit the 2D keypoints projection of the parametric SMPL body model (Loper et al., 2015) to 2D predictions from a separate method using an optimization approach. Brau and Jiang (2016) demonstrated that 2D projection, body pose prior, and body part length information could be used as the training loss objectives for 3D pose prediction. The method proposed in this chapter extends the idea of Brau and Jiang (2016) by introducing additional 3D supervision and paired training on the in-the-wild dataset. Kanazawa et al. (2018) showed that pose and shape parameters of the SMPL body model from monocular images could be learned using a neural network. While their method uses a 2D projection loss of the body model as the main objective, their method also requires an adversarial regularizer against parametric body models. This method can be further improved by using additional labels of 3D pose and SMPL parameters if available. Omran et al. (2018) proposed another deep learning approach to infer the parameters of the SMPL body model and analyzed performance when varying the input representation (silhouettes, 2D keypoints, part segmentations) and the proportion of 2D and 3D data.

The above review shows that many methods tackle generalizability on the in-the-wild images using either transfer learning from the 2D pose task or by decoupling the 3D pose estimation into separate 2D keypoint detection and depth regression problems. For methods that decouple the 3D representation [Dabral et al., 2018; Yang et al.,

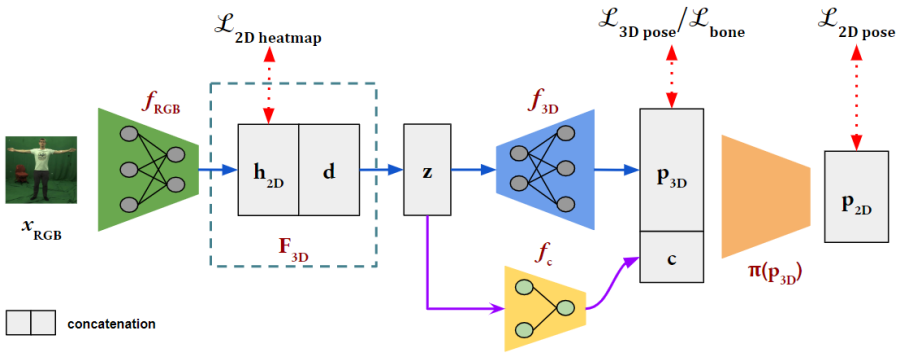


Figure 3.2: The overview of the proposed architecture. It uses a CNN f_{RGB} to learn 3D pose features represented as 2D heatmap locations $\mathbf{h}_{2D}$ and additional 3D pose cues $\mathbf{d}$ in the latent space. Both information are used to predict a root centered 3D pose $\mathbf{p}_{3D}$ and viewpoint parameters $\mathbf{c}$ using networks f_{3D} and f_c, respectively. Finally, $\mathbf{p}_{3D}$ and $\mathbf{c}$ are concatenated to learn 2D keypoint information $\mathbf{h}_{2D}$, allowing the network to update 3D pose information even if 3D labels are not available.

2018; Zhou et al., 2017], depth information is predicted if 3D labels are available. Otherwise, some weak supervision constraints (e.g., a parametric body model) are used for regularization. This chapter proposes a new architecture that combines explicit encoding of separate 2D and 3D depth features in hidden space instead of operating on vectorized 2D predictions as in previous lifting schemes. The trained projection network further stabilizes overall 3D prediction accuracy.

3.3 APPROACH

The proposed method estimates the root (pelvis) relative 3D locations of K human body joints $\mathbf{P} = \{\mathbf{J}_1, \ldots, \mathbf{J}_K\}$ in the camera reference frame from a monocular RGB image. The method also assumes that a crop around the subject is available.

A baseline strategy for the approach can be described as follows: Given a training set consisting of pairs of RGB images and their corresponding 3D pose labels $\mathcal{D} = \{(\mathbf{I}_n, \mathbf{P}_n^{GT})\}_{n=1}^{N}$, train a convolution-based neural network $f_{RGB}(\mathbf{I}_n, \boldsymbol{\theta})$ that could predict a vectorized representation of 3D joint locations. Network parameters $\boldsymbol{\theta}$ could be trained by minimizing the difference $\mathcal{L}_{3D}$ between pose prediction and ground truth

$$\mathcal{L}_{3Dpose} = \frac{1}{N}\sum_{n=1}^{N} \parallel f_{RGB}(\mathbf{I}_n, \boldsymbol{\theta}) - \mathbf{P}_n^{GT} \parallel_2^2 . \tag{3.1}$$

By training on currently available image data sets with 3D pose annotation, such direct supervision approach can already enable the network to achieve reasonable performance on studio test images. However, such a baseline method is still constrained in its ability to generalize to in-the-wild scenes due to the limited amount of available real-world images with ground truth 3D poses.

Therefore, several strategies need to be introduced to augment such a 3D pose network such that it performs better in in-the-wild scenes. The proposed augmented network can be trained on both images with 3D labels and in-the-wild images with only 2D labels. First, using an explicit 2D pose representation in the feature space of the CNN combined with 2D pre-training can significantly boost the quality of the prediction. Second, additional supervision is proposed by using a trained projection sub-network that learns weak perspective camera information for projecting 3D pose estimates to the 2D image space. The overview of the proposed network is shown in Figure 3.2.

3.3.1 *Explicit 2D feature representation for 3D pose prediction*

Martinez et al. (2017b) showed that a simple neural network is capable of directly regressing 3D human pose with good accuracy by using only vectorized 2D poses as input. This indicates that a neural network is able to estimate the structure of a natural 3D human pose from corresponding 2D information to some extent. However, such a lifting scheme can only remedy to some extent the fundamental ambiguity that multiple 3D poses can look the same in 2D. Pavlakos et al. (2018) showed that additional weak ordinal depth supervision could partially resolve the ambiguity of the problem.

Based on this observation, it can be argued that a 2D-to-3D lifting approach can also be applied to 2D heatmap input instead of the vectorized 2D pose representation. To test this hypothesis, this chapter proposes a pose estimator CNN where the convolutional features are designed to encode 2D pose heatmap information explicitly. The idea behind this decision is to explicitly decouple 2D pose information from other learned features in the convolutional latent space. The network can use the rest of the feature maps to capture additional image information related to 3D human pose, such as 3D depth. In this way, the network is guided to learn 3D pose features that are

more reliable due to the robust 2D pose prediction and are easier to interpret. Furthermore, by using a 2D training loss on this component, the network is allowed to learn valuable features from images when 3D pose labels are not available.

To this end, a convolutional feature map $\mathbf{F}_{3D} = [\mathbf{h}_{2D}, \mathbf{d}]$ is introduced after the extractor network f_{RGB}. This feature map consists of 64 output channels with a spatial dimension of 16×16. The first 14 channels are used to capture the 2D pose information. This region is optimized during training by minimizing the loss compared to the 2D ground truth heatmap in the least square sense. The rest of the feature channels $\mathbf{d}$ are not directly constrained by any explicit loss and will be supervised through the 3D pose, 2D projection, as well as additional pose constraint losses explained later.

To infer 3D pose from $\mathbf{F}_{3D}$, the explicit 2D heatmaps $\mathbf{h}_{2D}$ are first combined with the additional features $\mathbf{d}$ learned by the convolutional encoder by using a simple fully connected layer into a latent vector $\mathbf{z} \in \mathbb{R}^{1024}$. Then, a fully connected network with residual connections f_{3D} is used to learn the vectorized 3D pose representation $\mathbf{p}_{3D}$. The network f_{3D} is designed to be similar to the lifting architecture in Martinez et al., 2017b. More specifically, a series of four fully connected layers and ReLU activations are used, each with the width of 1024. A residual connection is also incorporated to connect $\mathbf{z}$ with the output of the second layer of f_{3D}.

Several earlier works reported that detection-based approaches using a heatmap or volumetric representation tend to achieve better performance on both 2D and 3D pose estimation tasks than approaches regressing vectorized predictions. However, additional structure-aware supervision can lift the performance of vectorized prediction to a competitive level (Sun et al., 2017). Since the proposed method also performs vectorized 3D pose prediction, the 3D training loss $\mathcal{L}_{3Dpose}$ (equation 3.3) is complemented with a bone supervision loss $\mathcal{L}_{bone}$. For 3D training data, $\mathcal{L}_{bone}$ measures the similarity of the vector between a joint $\mathbf{J}_k$ to its corresponding parent in the kinematic chain to ground truth. For 2D data, it measures the difference of scalar bone lengths to ground truth.

3.3.2 *Predicting 2D projection from 3D pose*

To further improve the method's ability to utilize 2D pose data for training 3D pose prediction, a sub-network is trained to project the predicted 3D pose to the image space. The camera network f_c predicts the principal coordinate (c_x, c_y) and the focal length (α_x, α_y) parameters of a weak perspective camera model from the given input image. By using the features extracted from the latent representation $\mathbf{z}$, a multi-layer perceptron is used to infer the camera parameters $\mathbf{c} \in \mathbb{R}^4$. During training, a 2D loss $\mathcal{L}_{2Dpose}$ measures the L2 distance between ground truth 2D pose and 2D projection $\mathbf{p}_{2D}$ of the predicted 3D pose:

$$\mathbf{p}_{2Dproj.} = \begin{bmatrix} \pi_x(\mathbf{p}_{3D}) \\ \pi_y(\mathbf{p}_{3D}) \end{bmatrix} = \begin{bmatrix} \alpha_x \mathbf{p}_{3D}(x) + c_x \\ \alpha_y \mathbf{p}_{3D}(y) + c_y \end{bmatrix} \tag{3.2}$$

The projection formulation allows the network to learn partial information about the 3D pose even when only 2D pose annotations are available. However, no constraints can guarantee the correctness of the predicted depth information. To regularize 3D pose prediction when training on 2D data, the additional bone loss $\mathcal{L}_{bone}$ is used to enforce bone length similarity to ground truth for additional supervision. The bone length is selected by randomly picking one of the training subjects as the ground truth for every training instance.

3.3.3 *Network design*

An adapted ResNet-50 (He et al., 2016) is used as the basis of the backbone subnetwork f_{RGB} (Figure 5.4) that extracts pose features from 2D images. This offers a good trade-off between prediction accuracy and inference time, allowing the network to be optionally used in real-time applications. The original ResNet-50 architecture is used up to level *Res4f*, and we train level *Res5a* from scratch without striding while also reducing its number of output channels to 1024. This extractor network is then followed by the 3D pose regressor network described in 3.3.1.

The studio datasets with 3D labels and the outdoor data sets with 2D labels tend to have slightly different image statistics due to contrast differences and foreground-background augmentations on the 3D data

sets. To further mitigate this residual domain gap beyond what the new network architecture can already do by its design, a pre-training approach is employed, similar to several earlier 3D pose prediction methods, e.g. Mehta et al., 2017a. To this end, the ResNet-50 network is pre-trained on ImageNet features to perform 2D heatmap prediction only. Here, intermediate 2D pose supervision is used on the first 14 channels of the *res4d* and *res5a* feature maps. The same intermediate supervision is also used later when finetuning the complete network on both 2D and 3D pose data. After pre-training, the final training of the full network on both outdoor images with 2D annotations and studio images with 3D annotations results in learned features that generalize well to in-the-wild scenes and yield high accuracy in 3D pose estimation.

The algorithm can be modified to handle input images of arbitrary framing around the human because the subnetwork f_{RGB} is convolutional. For example, a tight bounding box cropping can be performed around the detected 2D keypoints before passing the rescaled image into the subsequent sub-network.

3.3.4 *Loss functions*

Given an input image $\mathbf{I} \in \mathbb{R}^{w \times h \times 3}$, the extractor network f_{RGB} will predict the features $\mathbf{F}_{3D}$ which consist of the explicit 2D pose features $\mathbf{h}_{2D}$ and additional pose cues $\mathbf{d}$ as feature maps. The predicted 2D pose features are defined as 2D per-joint heatmaps Tompson et al., 2014

$$\mathbf{h}_{2D} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K), \qquad \mathbf{m}_k \in \mathbb{R}^{\frac{w}{s} \times \frac{h}{s}},$$

where $s = 16$ is the heatmap down-sampling factor. Similarly, the ground truth heatmaps are defined as

$$\mathbf{h}^{GT} = (\mathbf{m}_1^{GT}, \mathbf{m}_2^{GT}, \ldots, \mathbf{m}_K^{GT}), \qquad \mathbf{m}_k^{GT} \in \mathbb{R}^{\frac{w}{s} \times \frac{h}{s}}.$$

To train the 2D pose features, the difference between the predicted 2D joint heatmaps and the ground truth maps is minimized using an L2 loss

$$\mathcal{L}_{2Dheatmap} = \sum_{k=1}^{K} b_k \parallel \mathbf{m}_k - \mathbf{m}_k^{GT} \parallel_2^2, \tag{3.3}$$

where $b_k \in \{0, 1\}$ is a binary mask to ensure that the objective is not evaluated if the annotation of a particular joint is not available.

The latent features $\mathbf{F}_{3D}$ are then used to predict the 3D pose $\mathbf{P}_{3D} \in \mathbb{R}^{21\times3}$ by the sub-network f_{3D}. Given a 3D pose annotation $\mathbf{P}^{GT} \in \mathbb{R}^{21\times3}$, the 3D joint position loss is calculated as follows

$$\mathcal{L}_{3Dpose} = \| \mathbf{P}_{3D} - \mathbf{P}^{GT} \|_2^2 . \tag{3.4}$$

Given that $Parent(\mathbf{J}_k)$ is the position of the parent of a joint $\mathbf{J}_k \in \mathbb{R}^3$ in the kinematic chain, when the 3D joint position ground truth $\mathbf{J}_k^{GT} \in \mathbb{R}^3$ is available, the bone during training is defined as

$$\mathcal{L}_{bone} = \sum_{k=1}^{K} \left\| \Big(Parent(\mathbf{J}_k) - \mathbf{J}_k \Big) - \Big(Parent(\mathbf{J}_k^{GT}) - \mathbf{J}_k^{GT} \Big) \right\|_2^2 . \tag{3.5}$$

On the other hand, if we train on data for which only 2D joint annotations, but no 3D annotations are available, then we instead only compare the bone length magnitude between the predicted joint with a bone length $\mathbf{J}_k^S$ randomly selected from a training annotation

$$\mathcal{L}_{bone} = \sum_{k=1}^{K} \left\| \| Parent(\mathbf{J}_k) - \mathbf{J}_k \|_2^2 - \| Parent(\mathbf{J}_k^S) - \mathbf{J}_k^S \|_2^2 \right\|_2^2 . \tag{3.6}$$

Finally, given a predicted 2D pose from the projection layer $\mathbf{p}_{2D}$ and its corresponding ground truth 2D joint coordinates in the image space $\mathbf{p}_{2D}^{GT}$, the projection loss is defined as

$$\mathcal{L}_{2Dpose} = \| \mathbf{p}_{2D} - \mathbf{p}_{2D}^{GT} \|_2^2 . \tag{3.7}$$

The final training loss can be expressed as

$$\begin{aligned} \mathcal{L}_{all} = {} & \lambda_{2Dheatmap}\mathcal{L}_{2Dheatmap} + \lambda_{3Dpose}\mathcal{L}_{3Dpose} \\ & + \lambda_{bone}\mathcal{L}_{bone} + \lambda_{2Dpose}\mathcal{L}_{2Dpose}, \end{aligned} \tag{3.8}$$

where $\lambda_{3Dpose} = 10$, $\lambda_{2Dheatmap} = 0.1$, $\lambda_{2Dpose} = 10$. $\lambda_{bone} = 10$ if the bone direction is considered (i.e. 3D pose annotations are given) and $\lambda_{bone} = 100$ if only the bone length scalar is estimated (i.e. only 2D annotations are given).

Table 3.1: 3D PCK (higher is better) on the MPI-INF-3DHP dataset after training with MPI-INF-3DHP and H3.6M 3D training sets, and MPII and LSP 2D training sets. The proposed method outperform all other methods that use a similar combined 2D and 3D training on this benchmark with both indoor and in-the-wild scenes. This holds true for all evaluation protocols (*unscaled* (US), *glob. scaled* (GS), *Procrustes* (PA)).

Method	**PCK GS**	**PCK No GS**	**PCK Outdoor**	**PCK All**	**AUC All**	**MPJPE All**
Mehta et al. (2017a)	84.6	72.4	69.7	76.5	-	-
Mehta et al. (2017b)	-	-	-	76.6	40.4	124.7
Dabral et al. (2018)	-	-	-	76.7	39.1	103.8
Proposed (US)	87.8	80.2	73.8	81.5	44.5	90.7
Proposed (GS)	88.0	80.5	74.8	82.0	44.7	91.0
Proposed (PA)	94.9	92.4	84.0	91.3	57.5	65.4

3.4 RESULTS AND EVALUATION

This section discussed the datasets, training strategy, as well as quantitative and qualitative comparisons of the proposed method against the prior arts.

The H3.6M data set (Ionescu et al., 2014) is used to compare general 3D pose estimation accuracy on in-studio data. Furthermore, the proposed method is compared against previous methods on the more general MPI-INF-3DHP benchmark set (Mehta et al. (2017a)). The latter features more diverse motions and scenes, including indoor scenes with green screen background (**GS**), as well as more in-the-wild scenes with general backgrounds, both indoors (**No GS**) and outdoors (**Outdoor**). An ablation analysis shows the significance of the individual components in the proposed approach.

3.4.1 *Datasets and evaluation metrics*

As training data with ground truth 3D pose, the proposed method uses a combination of the H3.6M training set, as well as both background augmented and unaugmented MPI-INF-3DHP training sets, which consist of 350k training images in total. As in-the-wild training images with only 2D pose annotation, the method is trained using the MPII (Andriluka et al., 2014) and LSP (Johnson and Everingham, 2011;

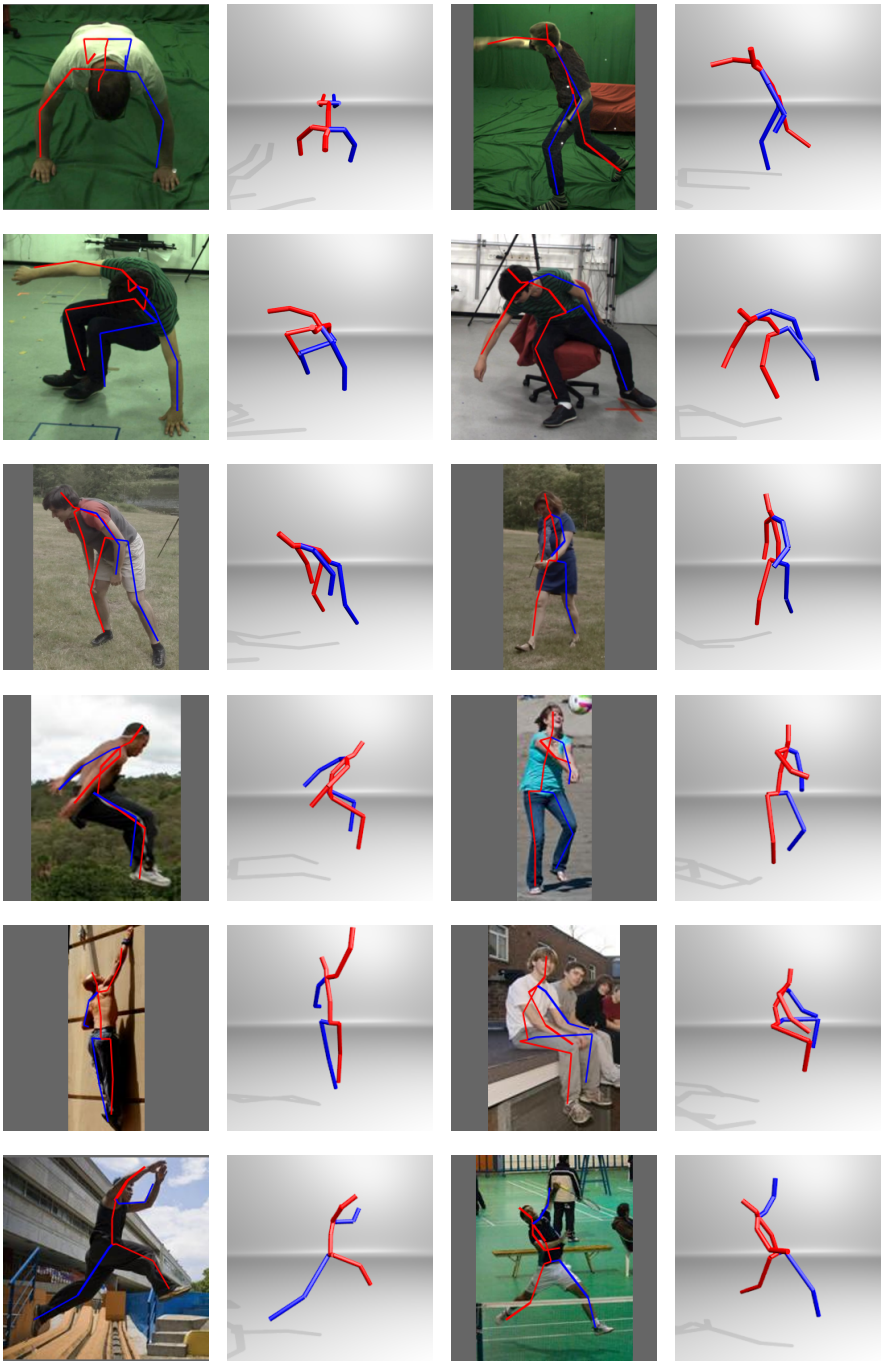


Figure 3.3: Qualitative examples from the MPI-INF-3DHP test set (row 1, 2, and 3) and LSP (row 4, 5, and 6).

Figure 3.4: Qualitative examples of applying the proposed method to the MPI-INF-3DHP test set.

Figure 3.5: Additional qualitative examples of applying our method on the LSP test set.

Johnson and Everingham, 2010) datasets which are augmented by randomly cropping, translating and rotating the images.

At test time, the proposed approach is compared against other previous methods on both standard H3.6M and MPI-INF-3DHP test data to show both general 3D pose prediction accuracy as well as state-of-the-art generalization on outdoor scenes. The qualitative result of the proposed algorithm on in-the-wild images is visualized in Figure 3.3, Figure 3.4, and Figure 3.5.

The quantitative performance is evaluated by comparing the Mean Per Joint Position Error (MPJPE), the Percentage of Correct 3D Keypoints (3D PCK) under a 150 mm radius from the reference joint location (Mehta et al., 2017a), as well as the Area Under Curve (AUC) metric which corresponds to the thresholds of the 3D PCK. Since evaluation protocols in previous work are not uniform, the quantitative analysis is evaluated under the three most commonly used protocols: (i) 3D joint predictions are neither scaled nor aligned to ground truth (*unscaled*), (ii) 3D joint predictions are globally scaled with ground truth scale before evaluation (*glob. scaled*), and (iii) 3D joint predictions are aligned to ground truth with full Procrustes alignment (*Procrustes*). The evaluation conducted in this chapter follows standard practice in monocular 3D pose estimation by cropping a tight bounding box in test images using 2D ground truth information. Since cropping essentially performs a virtual rotation from the original camera, perspective correction (Mehta et al., 2017a) is used to re-align the pose to the correct view.

3.4.2 *Training procedure*

As outlined earlier, the proposed network is trained in two stages. First, the feature extractor network is pre-trained on the 2D heatmap regression task on both MPII (Andriluka et al., 2014) and LSP (Johnson and Everingham, 2011; Johnson and Everingham, 2010) datasets. At this stage, the network is trained for 186k iterations with a minibatch size of 21. The initial learning rate is 0.05, which is decayed exponentially.

After pre-training, the learned weights are used to initialize the weights of the whole 3D pose prediction network. The entire network is then trained on both the 3D labeled studio data as well as the in-the-wild data with only 2D annotations. Image data with 3D and 2D

annotations are both fed into the network with a minibatch size of 10 to train for 240k iterations. For this second stage, the training again starts by using a learning rate of 0.05 with decay over 60k iterations. Adadelta with a momentum of 0.9 is used as the optimization algorithm in both training stages.

Using learning rate discrepancy on the pre-trained layers is empirically found to preserve in-the-wild features, as suggested by Mehta et al. (2017a), and is necessary to achieve good generalization if 3D training data is very limited or more biased. The experiments indicate that a learning rate discrepancy with a factor of 100 when training using H3.6M data as the only source of 3D pose labels yields the best result when tested on the MPI-INF-3DHP dataset. On the other hand, the best performance is achieved without using any such discrepancies when training on both H3.6M and the augmented data of MPI-INF-3DHP as the source of 3D labels. This suggests that foreground and background augmentation of the 3D data can further close the domain gap between the indoor and outdoor scenes.

Table 3.2: Mean Per Joint Position Error (MPJPE) on H3.6M when trained on H3.6M (the proposed method is *glob. scaled* for evaluation). (*) indicates methods that also use 2D labeled datasets during training or pre-training.

	Direct.	Discuss	Eat	Greet	Phone
Mehta et al. (2017a)*	59.7	69.7	60.6	68.8	76.4
Mehta et al. (2017b)*	62.6	78.1	63.4	72.5	88.3
Pavlakos et al. (2017)	67.4	72.0	66.7	69.1	72.0
Martinez et al. (2017b)*	51.8	56.2	58.1	59.0	69.5
Zhou et al. (2017)*	54.8	60.7	58.2	71.4	62.0
Yang et al. (2018)*	51.5	58.9	50.4	57.0	62.1
Sun et al. (2017)*	52.8	54.8	54.2	54.3	61.8
Kanazawa et al., 2018*	-	-	-	-	-
Luvizon et al., 2018*	49.2	51.6	47.6	50.5	51.8
Dabral et al., 2018*	46.9	53.8	47.0	52.8	56.9
Proposed* (H80K)	57.1	69.6	61.6	66.0	73.4
Proposed* (5 fps)	54.0	65.1	58.5	62.9	67.9
	Pose	Purch.	Sit	SitD	Smoke
Mehta et al. (2017a)*	59.1	75.0	96.2	122.9	70.8
Mehta et al. (2017b)*	63.1	74.8	106.6	138.7	78.8

Pavlakos et al. (2017)	65.0	68.3	83.7	96.5	71.7	
Martinez et al. (2017b)*	55.2	58.1	74.0	94.6	62.3	
Zhou et al. (2017)*	53.8	55.9	75.2	111.6	64.1	
Yang et al. (2018)*	49.8	52.7	69.2	85.2	57.4	
Sun et al. (2017)*	53.1	53.6	71.7	86.7	61.5	
Kanazawa et al., 2018*	-	-	-	-	-	
Luvizon et al., 2018*	48.5	51.7	61.5	70.9	53.7	
Dabral et al., 2018*	45.2	48.2	68.0	94.0	55.7	
Proposed* (H80K)	57.1	70.9	89.8	109.2	68.6	
Proposed* (5 fps)	54.0	60.6	82.7	98.2	63.3	

	Photo	Wait	Walk	WalkD	WalkP	Avg.
Mehta et al. (2017a)*	85.4	68.5	54.4	82.0	59.8	74.1
Mehta et al. (2017b)*	93.8	73.9	55.8	82.0	59.6	80.5
Pavlakos et al. (2017)	77.0	65.8	59.1	74.9	63.2	71.9
Martinez et al. (2017b)*	78.4	59.1	49.5	65.1	52.4	62.9
Zhou et al. (2017)*	65.5	66.1	63.2	51.4	55.3	64.9
Yang et al. (2018)*	65.4	58.4	60.1	43.6	47.7	58.6
Sun et al. (2017)*	67.2	53.4	47.1	61.6	53.4	59.1
Kanazawa et al. (2018)*	-	-	-	-	-	88.0
Luvizon et al. (2018)*	60.3	48.9	44.4	57.9	48.9	53.2
Dabral et al. (2018)*	63.6	51.6	40.3	55.4	44.3	55.5
Proposed* (H80K)	81.3	65.8	54.3	78.4	58.2	71.1
Proposed* (5 fps)	75.0	61.2	50.0	66.9	56.5	65.7

3.4.3 *Quantitative comparison*

Table 3.1 compares the proposed method on the MPI-INF-3DHP benchmark against the closest competing approaches that can be trained on both, images with 2D and 3D annotations. All methods were trained using both H3.6M and augmented and unaugmented MPI-INF-3DHP 3D datasets, and the LSP and MPII 2D datasets. Unless stated otherwise, the training procedure used the H80K samples of the H3.6M dataset which consists of around 41K training samples before augmentation. The proposed algorithm achieves by far the highest accuracy (across all evaluation protocols), yielding 82.0% 3D PCK, 44.7% AUC and 91.0

Table 3.3: Mean Per Joint Position Error (MPJPE) on H3.6M when trained on H3.6M. (*) indicates methods that also use 2D labeled datasets during training or pre-training. (*Procrustes* for evaluation).

	Direct.	Discuss	Eat	Greet	Phone	
Sun et al. (2017)*	42.1	44.3	45.0	45.4	51.5	
Kanazawa et al. (2018)*	-	-	-	-	-	
Dabral et al. (2018)*	32.8	36.8	42.5	38.5	42.4	
Omran et al. (2018)	-	-	-	-	-	
Proposed* (H80K)	46.1	51.3	46.8	51.0	55.9	
Proposed* (5 fps)	43.7	46.9	45.4	48.0	50.2	
	Pose	Purch.	Sit	SitD	Smoke	
Sun et al. (2017)*	43.2	41.3	59.3	73.3	51.0	
Kanazawa et al. (2018)*	-	-	-	-	-	
Dabral et al. (2018)*	35.4	34.3	53.6	66.2	46.5	
Omran et al. (2018)	-	-	-	-	-	
Proposed* (H80K)	43.9	48.8	65.8	81.6	52.2	
Proposed* (5 fps)	40.6	41.6	60.7	75.6	48.8	
	Photo	Wait	Walk	WalkD	WalkP	Avg.
Sun et al. (2017)*	53.0	44.0	38.3	48.0	44.8	48.3
Kanazawa et al. (2018)*	-	-	-	-	-	56.8
Dabral et al. (2018)*	49.0	34.1	30.0	42.3	39.7	42.2
Omran et al. (2018)	-	-	-	-	-	59.9
Proposed* (H80K)	59.7	51.1	40.8	54.8	45.2	53.4
Proposed* (5 fps)	54.9	46.8	36.9	47.5	43.9	49.2

mm MPJPE overall (using *glob. scaled* for evaluation). It also achieves the state-of-the-art result specifically on the outdoor scenes with 74.8% 3D PCK. Further, the average 3D PCK of 91.3% is the highest ever reported by all algorithms that evaluated on the MPI-INF-3DHP, irrespective of what training data they used.

Table 3.4 further shows the comparison of the proposed approach to other methods on MPI-INF-3DHP, when all methods are trained using only H3.6M as the source of 3D pose labels. The proposed method achieves the highest accuracy in terms of 3D PCK and AUC on the basis of all three evaluation protocols.

Finally, the proposed method is compared by only using the H80K samples of H3.6M as the 3D pose dataset and testing on every 64*th* frame of the S9 and S11 subjects in H3.6M, see Table 3.2 (here we use

Table 3.4: Comparison on MPI-INF-3DHP after training only on H3.6M dataset. The proposed approach outperforms all other competing approaches in all metrics and testing protocols.

Method	**PCK**	**AUC**	**MPJPE**
Mehta et al. (2017a)	64.7	31.7	-
Yang et al. (2018)	69.0	32.0	-
Zhou et al. (2017)	69.2	32.5	-
Proposed (*unscaled*)	69.6	35.5	127.0
Proposed (*glob. scaled*)	70.4	36.0	129.1
Proposed (*Procrustes*)	82.9	45.4	92.0

glob. scaled following Zhou et al., 2017, Yang et al., 2018, Dabral et al., 2018) and Table 3.3 (*Procrustes*). On this test set which is heavily biased to in-studio data of a single background, the proposed method geared for in-the-wild generalization cannot beat the best-performing methods. However, it still achieves competitive accuracy. When the number of training data is increased by sampling from H3.6M at 5 frames per second, the proposed method achieved a better MPJPE of 65.7 mm while maintaining competitive result when tested on MPI-INF-3DHP with 71.2% 3D PCK and 36.3% AUC. When using *Procrustes* during comparison, a state-of-the-art accuracy of 53.4 mm average MPJPE is achieved when trained using H80K samples and 49.2 mm average MPJPE when trained using H3.6M data sampled at 5 fps. Notably, here the proposed approach also outperforms other methods that use some pose projection operation related to the proposed architecture and regularization with a statistical body model, namely Kanazawa et al., 2018 and Omran et al., 2018.

3.4.4 *Activity-wise result*

The activity-wise performance of the proposed method tested on MPI-INF-3DHP is shown in Table 3.5. The proposed method achieves a very high 3D PCK of more than 80% on almost all categories, except for the on-the-floor activities (60.7%), which are in general also challenging for other methods.

Table 3.5: Activity-wise 3D PCK of the proposed method on the MPI-INF-3DHP test set. The proposed method achieved more than 80% 3D PCK in most actions except for the challenging on-the-floor examples (60.7% 3D PCK).

Action	**PCK**					**AUC**
	Head	**Neck**	**Shou**	**Elbow**	**Total**	
Standing/Walking	93.2	100.0	99.6	89.8	89.7	51.2
Exercising	91.3	98.2	98.2	87.6	85.6	47.2
Sitting	81.7	92.8	91.8	76.7	80.0	43.7
Reaching/Crouching	76.6	91.1	91.3	83.3	84.6	47.6
On The Floor	62.8	83.9	78.9	54.7	60.7	28.5
Sports	90.0	99.2	98.7	84.9	87.0	49.3
Miscellaneous	80.8	96.8	95.3	71.3	80.4	43.4
All	82.3	94.9	93.7	78.0	81.5	44.7
Action	**PCK**					**AUC**
	Wrist	**Hip**	**Knee**	**Ankle**	**Total**	
Standing/Walking	74.3	100.0	90.0	77.3	89.7	51.2
Exercising	75.6	100.0	77.6	65.5	85.6	47.2
Sitting	65.1	99.8	75.8	63.9	80.0	43.7
Reaching/Crouching	78.0	98.7	84.2	73.2	84.6	47.6
On The Floor	40.9	94.6	53.9	28.6	60.7	28.5
Sports	67.8	100.0	90.6	72.4	87.0	49.3
Miscellaneous	53.8	100.0	86.5	66.9	80.4	43.4
All	64.5	99.3	81.2	65.5	81.5	44.7

3.4.5 *Ablation study*

An ablation study is conducted to measure the effectiveness of the proposed contributions (Table 3.6). A direct 3D pose regression method with 2D pose pre-training without the explicit 2D pose loss in the feature space and without the 2D-from-3D projection loss is used as a baseline. The baseline is trained on 3D data only and uses both joint position and bone losses as training objective. All of the comparison results were trained on the H80K samples of H3.6M and then performed the evaluation tests on the MPI-INF-3DHP dataset.

The baseline reaches 62.3% 3D PCK. Using the explicit 2D pose in the latent feature space allows us to use the outdoor data during

Table 3.6: Ablation study on MPI-INF-3DHP test data (split into scene subcategories: in-studio with green screen (GS), and more in-the-wild scenes indoors (No GS) and outdoors (Outdoor)). Only H3.6M data with ground truth 3D labels were used for training. 3D predictions are globally scaled.

Method	**PCK**	**AUC**
Baseline (direct 3D prediction + bone loss)	62.3	30.3
+ 2D latent loss + outdoor data	66.4	33.0
+ 3D-to-2D projection loss + outdoor data	69.5	35.3
+ 2D latent loss + outdoor data + 3D-to-2D projection loss	70.4	36.0

the training. This addition improves the performance by 3.1% against the baseline. Similarly, adding the 3D-to-2D projection loss improves the performance of the method even without the explicit 2D pose in latent feature space. Using both the proposed components advances the result to the state-of-the-art result with 70.4% 3D PCK.

3.4.6 *Qualitative results*

Example prediction results is visualized on MPI-INF-3DHP and LSP test images in Figure 3.3. The proposed method performs consistently well on studio, general indoor and in-the-wild images.

Several failure cases are shown in Figure 3.6. The proposed method can fail on challenging poses that are heavily (self-) occluded, on poses seen from unusual camera angles, or poses that are from what was seen in the training set. Such failure cases are common to many monocular 3D pose estimation approaches.

3.5 ADDITIONAL COMPARISONS ON MPI-INF-3DHP

At some point in the past, the authors of MPI-INF-3DHP released a correction to the ground truth annotations of a subset of two their six test sequences. For all tests, this chapter reported the corrected data.

Their very first version of the test set contained small errors on the in-studio sequences with general, i.e. no green screen, background (test subject 3 and 4, meaning sequences labelled as **No GS**). On these sequences, before correction, the annotations were temporally mis-

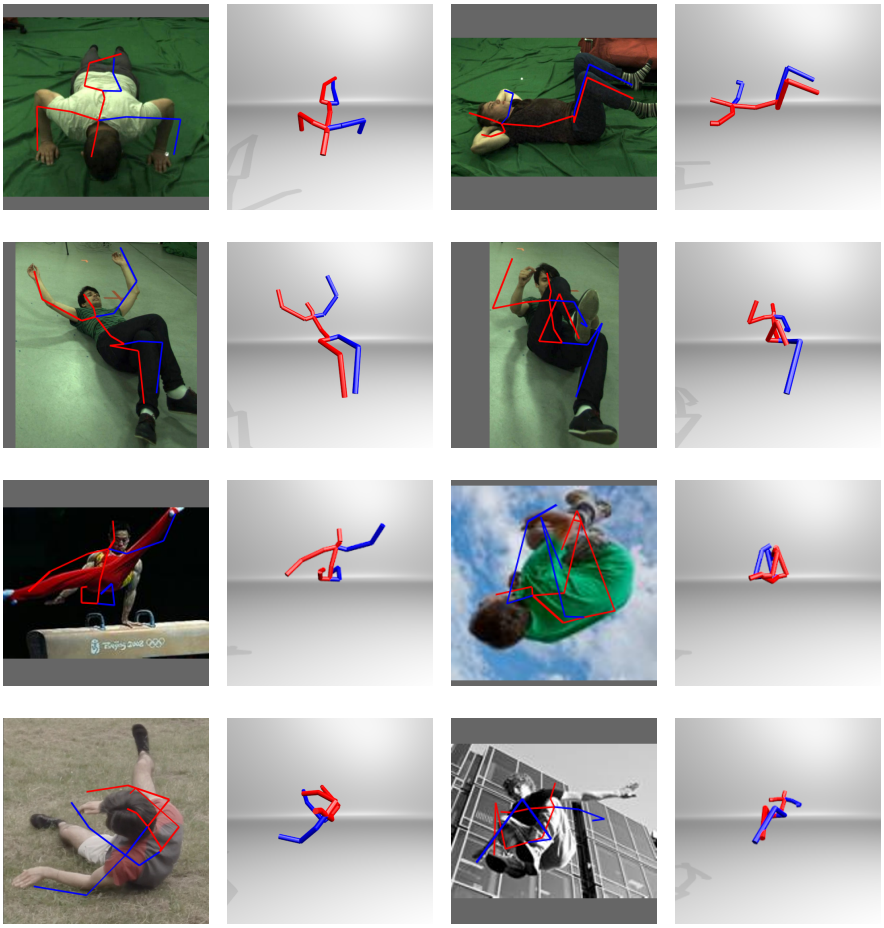


Figure 3.6: Examples of prediction failures by the proposed method.

aligned by one or two frames. It is difficult to say what previous paper we compared against may have unknowingly used the uncorrected subset of sequences.

For the tests to be as transparent and fair as possible, we therefore also provide a comparison on the subset of 4 out of 6 MPI-INF-3DHP test sequences (**GS** and **Outdoors**) that were always correct.

Table 3.7 shows the comparison for methods trained on both H3.6M and MPI-INF-3DHP using the mentioned subset for testing. We include methods that in their original papers reported the respective results on the subsets of test sequences. All evaluations of the proposed method are performed with the corrected annotations. The proposed method is also state-of-the-art when tested on this subset of sequences.

Table 3.7: Comparison on the subset of MPI-INF-3DHP test sequences that was not corrected at some point by the authors of MPI-INF-3DHP (GS and Outdoors). All here refers to the average on this subset of sequences. Unless stated otherwise, all H3.6M training data mentioned in this table use H80K samples.

Method	**3D training data**	**PCK GS**	**PCK Outdoor**	**PCK All**	**AUC All**	**MPJPE All**
Mehta et al. (2017a)	H3.6M + 3DHP	84.6	69.7	78.8	-	-
Mehta et al. (2017a)	H3.6M	70.8	58.5	66.0	-	-
Zhou et al. (2017)	H3.6M	71.1	72.7	71.7	-	-
Ours (*unscaled*)	H3.6M + 3DHP	87.8	73.8	82.3	45.3	91.4
Ours (*unscaled*)	H3.6M	74.6	64.0	70.5	36.3	128.7
Ours (*glob. scaled*)	H3.6M (5 fps)	75.4	66.9	72.1	37.2	125.5
Ours (*glob. scaled*)	H3.6M + 3DHP	88.0	74.8	82.9	45.6	91.8
Ours (*glob. scaled*)	H3.6M	75.2	65.3	71.4	36.9	131.4
Ours (*glob. scaled*)	H3.6M (5 fps)	75.8	67.9	72.8	37.8	128.6
Ours (*Procrustes*)	H3.6M + 3DHP	94.9	84.0	90.7	58.0	66.1
Ours (*Procrustes*)	H3.6M	85.9	78.8	83.2	46.6	91.1
Ours (*Procrustes*)	H3.6M (5 fps)	86.2	78.0	83.0	47.5	89.6

3.6 CONCLUSION

While the proposed method can still fail to produce accurate estimates in some scenarios, such as heavily occluded scenes or highly uncommon pose articulation, the experiments show that the method can achieve state-of-the-art 3D monocular pose estimation results that appear consistent without the need for temporal smoothing, especially if the given test images contain poses that have been observed in the training data. Based on this observation, it is natural to consider the feasibility of deploying such monocular approaches as a motion capture system on in-the-wild images. Leveraging a learning-based monocular motion capture approach for downstream 3D animation tasks also demonstrates the possibility of synthesizing 3D human motion in a democratized manner. The following two chapters of this thesis attempt to explore this plausibility by making use of 3D monocular annotations to train speech-driven 3D motion synthesis and control models. This setting is also particularly suitable to benchmark the monocular capture system, as the in-the-wild target videos may contain various

challenging scenarios in 3D pose estimation, such as partial occlusions as well as appearance and articulation variations.

4

LEARNING SPEECH-DRIVEN 3D CONVERSATIONAL GESTURES FROM VIDEO

The previous chapter discusses a method that performs 3D pose estimation of a human body from monocular images. The ability to estimate 3D human body joints can be utilized to directly drive the motion of a character in the virtual world. Alternatively, the captured 3D data can be used to train downstream 3D animation models that can generate new human motion from a certain control input. This offers a major advantage compared to classical approaches, as it allows us to collect a much larger amount of 3D data from monocular videos, avoiding the need to comply with the multi-view studio capture constraints. In particular, this chapter proposes an approach to synthesize the synchronous 3D conversational body and hand gestures, as well as 3D face and head animations, of a virtual character from speech input (published as Habibie et al., 2021a). Leveraging the capability of the monocular tracking approaches similar to the one presented in Chapter 3, this chapter introduces a new corpus that contains more than 33 hours of annotated data from in-the-wild videos of talking people to train the model. To this end, several state-of-the-art monocular approaches for 3D body and hand pose estimation as well as 3D face performance capture are applied to the video corpus to obtain 3D annotations. In particular, we use a multi-person 3D pose estimation approach of Mehta et al. (2020) to capture the 3D pose of the body, as this method is found to be robust to partially visible subjects in the scene commonly found in the target speech gesture videos. In this way, orders of magnitude more training data can be used than previous algorithms that resort to complex in-studio motion capture solutions and thereby train more expressive synthesis algorithms.

In addition to the new way of capturing speech gestures, this chapter also proposes a new algorithm that uses a CNN architecture to leverage the inherent correlation between facial expressions and hand gestures. This task is traditionally challenging, as synthesizing conversational body gestures is a multi-modal problem where several gesture variations can plausibly accompany the same input speech. To this

end, a Generative Adversarial Network (GAN) based model is trained to measure the plausibility of the generated sequences of 3D body motion when paired with the input audio features.

4.1 INTRODUCTION

Animating the motion of virtual human characters is a crucial task in many computer graphics pipeline. Traditionally, their generation requires a combination of complex motion capture recordings and tedious work by animation experts to generate plausible appearance and movement. The particular challenges include the animation of the conversational body gestures of a talking avatar, as well as the facial expressions accompanying the audio in conveying the emotion and mannerisms of the speaker. Both are traditionally achieved by manually specified key-frame animation. Automated tools for animating body gestures and facial expressions directly from speech would tremendously ease the effort required and allow non-experts to author higher-quality character animations. Further, such tools would enable users to drive real-time embodied conversational avatars of themselves populating shared virtual spaces and animate them with on-the-fly body gestures and facial expressions in tune with speech. In psycho-linguistics studies, it has been shown that user interfaces showing avatars with plausible body gestures, facial expressions, and speech are perceived as more believable and trustworthy (Van Mulken et al., 1998). Studies have shown that non-verbal behavior is essential for conveying information (Goldin-Meadow, 1999), providing a view into the speaker's internal state, and both speech and body gestures are tightly correlated, arising from the same internal process (Kendon, 2004; McNeill, 2000).

Prior work on speech-driven virtual characters has been limited either to the generation of co-verbal body gestures through heuristic rule-based (Marsella et al., 2013) or learning-based (Ferstl and McDonnell, 2018; Levine et al., 2010, 2009) approaches, or the generation of facial expressions (Karras et al., 2017) and head movements (Sadoughi et al., 2017) in tune with speech. Many learning-based approaches use motion and gesture training data captured in a studio with complex motion capture systems (Alexanderson et al., 2020a; Ferstl and McDonnell, 2018; Ferstl et al., 2019; Lee et al., 2019a; Levine et al., 2010, 2009;

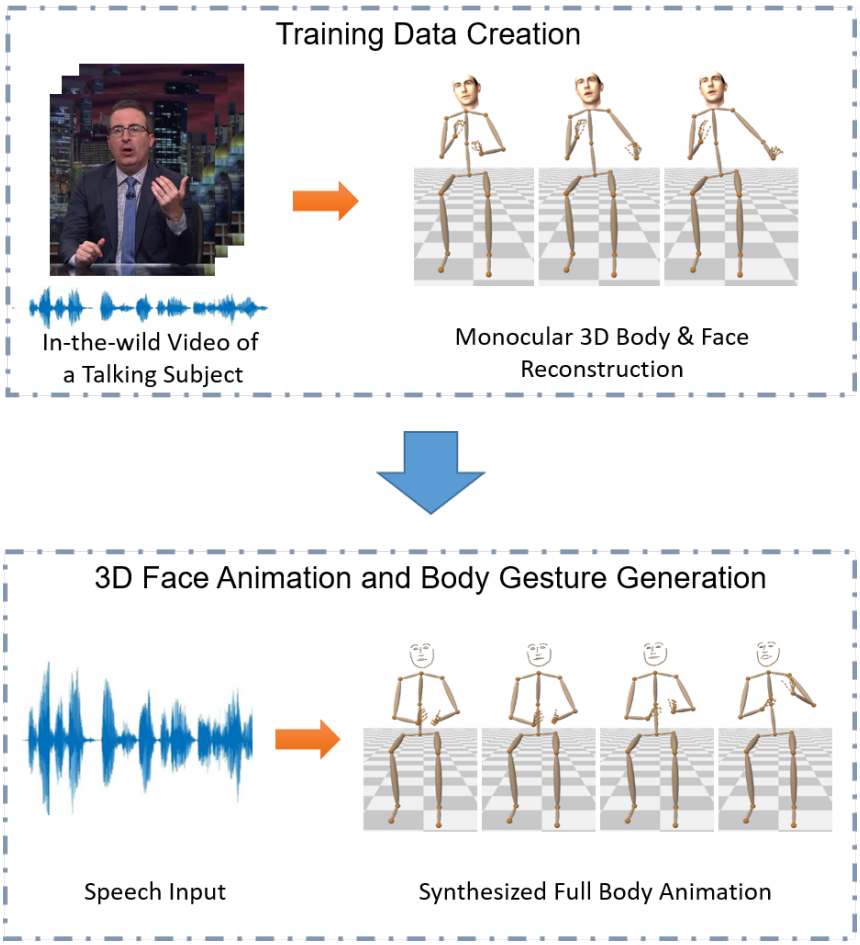


Figure 4.1: This chapter proposes the first approach to jointly synthesize the synchronous 3D conversational body gestures and 3D face animations of a virtual character from speech input. It is trained using the 3D facial expression, body, and hand pose annotation for a large corpus of in-the-wild video of talking people introduced in this work.

Takeuchi et al., 2017b). In this way, it is hard to record large corpora of data reflecting gesture variation across subjects, or subject-specific idiosyncrasies revealed only in long term observation.

This chapter introduces the first approach to jointly generate synchronized conversational 3D gestures of the arms, torso, and hands, as well as a simple but expressive 3D face and head movement of an animated character from speech. It is based on the following contributions:

1. A new set of 3D training data annotations from more than 33 hours of in-the-wild videos of talking subjects, which was used for learning a purely 2D gesturing model, without face expression synthesis, before (Ginosar et al., 2019a). The ground truth annotations are created by applying monocular in-the-wild 3D body pose reconstruction (Mehta et al., 2020), 3D hand pose

reconstruction (Zhou et al., 2020), and monocular dense 3D face reconstruction (Garrido et al., 2016) on these videos.

2. A novel CNN architecture that synthesizes face, body, and hand gestures from speech input. It has a common encoder for body and hands gesture as well as facial expression, which learns the inherent correlation between them and three decoder heads to jointly generate realistic motion sequences for body, hands, and face. In addition to facial expressions and head poses in tune with audio, it synthesizes plausible conversational gestures, such as beat gestures that humans use to emphasize spoken words, and gestures that reflect mood and personal conversational style. Note that, the goal is not to generate gestures relating to semantic speech content, or carrying specific language meaning, like in sign language.
3. Synthesis of body gestures is a multi-modal problem; several gestures could accompany the same utterance. To prevent convergence to the mean pose in training and ensure expressive gesture synthesis, the prior 2D work of Ginosar et al. (2019a) used adversarial training (Goodfellow et al., 2014). This proposed work improves upon this idea by not only designing a discriminator that can measure whether the synthesized body and hand gestures look natural, but also the plausibility of the synthesized gestures when paired with the ground truth audio features. Figure 4.1 summarizes the proposed contribution.

4.2 RELATED WORK

Human gestures and facial expressions are known to be highly correlated to speech and often convey meaningful information. This section discusses various techniques that have been proposed to learn the correlation between gesture and speech, especially data-driven approaches which are relevant to the method proposed in this chapter. Like many other data-driven models (Ferstl et al., 2019; Ginosar et al., 2019b), the main goal is to model the generation of beat gestures, which are the repetitive motion used to emphasize certain parts of the speech (McNeill, 2000). However, the proposed formulation also allows the model to generate a specific type of gesture at a particular

time by leveraging additional information to produce body motion beyond beat gestures.

4.2.1 *Speech-Driven Body Gestures and Head Motion*

The prior art can be grouped into rule-based and data-driven methods. The seminal works by Cassell et al. (1994) and Cassell (2000) show that automatic body gesture and facial expression generation of a virtual character can be synchronized with the audio by using a set of manually defined rules. Other works incorporate linguistic analysis (Cassell et al., 2004) into an extendable rule-based framework. Marsella et al. (2013) develop a rule-based system to generate body gesture (and facial expression) by analyzing the text input and audio content. However, such methods heavily rely on the study of language-specific rules and cannot easily handle non-phoneme sounds.

To overcome these problems, data-driven approaches, which do not rely on expert knowledge in the linguistic domain, have attracted increasing attention. Neff et al. (2008) propose a method to create a person-specific gesture performance using manually annotated video corpora, given the spoken text and performer's gesture profile. The gesture script is then used to animate a virtual avatar. Levine et al. (2009) use a complex motion capture setup to capture 45 minutes of training data and trained a Hidden Markov Model to select the most probable body gesture clip based on the speech prosody in real time. Levine et al. (2010) map the audio signal into a latent kinematic feature space using a variant of Hidden Conditional Random Fields (CRF). The learned model is then used to select a gesture sequence via reinforcement learning approach. Mariooryad and Busso (2012) use a combination of Dynamic Bayesian Networks (DBN) to synthesize head pose and eyebrow motion from speech. Sadoughi et al. (2014) extend this approach by modeling discourse functions as additional constraints of the DBNs. Sadoughi et al. (2017) use a learning-based approach that can leverage text-to-speech (TTS) system to synthesize head motion and propose a method that can solve the mismatch between real and synthetic speech during training. Chiu and Marsella (2011) train a Conditional Restricted Boltzmann Machine (CRBM) to directly synthesize sequences of body poses from speech. Chiu and Marsella propose using Gaussian Process Latent Variable Models

(GPLVM) to learn a low dimensional embedding to select the most probable body gestures from a given speech input (Chiu and Marsella, 2014).

In recent years, deep learning has demonstrated its superiority in automatically learning discriminative features from big data. Bidirectional LSTM is used by Takeuchi et al. (2017a), Hasegawa et al. (2018), and Ferstl and McDonnell (2018) to synthesize body gestures from speech. Similarly, Haag and Shimodaira (2016) use LSTM to synthesize head motion from speech. Kucherenko et al. (2019) propose a denoising autoencoder to learn lower dimensional representation of body motion and then combines it with an audio encoder to perform audio-to-gesture synthesis at test time. Lee et al. (2019a) contribute a large scale motion capture dataset of synchronized body-finger motion and audio, and propose a method to predict finger motion based on both audio and arm position as input. Ferstl et al. (2019) use a multi-objective adversarial model and make use of a classifier that is trained to predict the gesture phase of the motion to improve gesture synthesis quality.

Recent works also try to incorporate text-based semantic information to improve generation quality of body gestures from speech (Kucherenko et al., 2020; Yoon et al., 2020). Alexanderson et al. (2020a) propose a normalizing flow-based generative model that can synthesize multiple plausible 3D body gesture from the same speech input and also allows some degrees of control to the synthesis. Ahuja et al. (2020) show that a single learning-based mixture model can be trained to perform gesture style transfer between multiple speakers. In contrast, the focus of the method presented in this chapter is to find the best solution of predicting all relevant body modalities from audio using a single framework, which is a challenging problem even when trained in a person-specific manner.

Deep learning approaches typically require a large scale training corpus of audio and 3D motion pairs, which is usually captured with complex and expensive in-studio motion capture systems. To tackle this problem, Ginosar et al. (2019a) propose a learning-based speech-driven generation of 2D upper body and hand gesture model from a large scale in-the-wild video collection. With this solution, they are able to build an order of magnitude larger corpus from community video. Similarly, Yoon et al. (2019) train a speech-to-gesture method using ground truth 2D poses extracted from TED Talk videos via OpenPose

(Cao et al., 2017). Their model employs a Bidirectional LSTM to map audio input into a sequence of 2D human body pose. In contrast to existing methods, the method proposed in this chapter synthesizes not only the 3D upper body and hand gestures, but also head rotation and facial expression of the speaker.

4.2.2 *Speech-Driven Facial Expressions*

Current techniques can be classified into: 1) face model-based (Cha et al., 2018; Cudeiro et al., 2019; Liu et al., 2015; Pham et al., 2018; Taylor et al., 2017; Tzirakis et al., 2019) and 2) model-free based. Model-based approaches parameterize expressions in terms of blendshapes and estimate these parameters from the audio input. Model-free based approaches, however, directly map the audio into 3D vertices of a face mesh (Karras et al., 2017) or 2D point positions of the mouth Suwajanakorn et al., 2017. In Karras et al. (2017), an LSTM is used to learn this mapping, and in Suwajanakorn et al. (2017), final photorelistic results are generated. Cudeiro et al. (2019) use DeepSpeech voice recognition (Hannun et al., 2014) to produce an intermediate representation of the audio signal. This is then regressed into the parameters of the FLAME face model (Li et al., 2017). Taylor et al. (2017) use an off-the-shelf speech recognition method to map the audio into phoneme transcripts. A network is trained to translate the phonemes into the parameters of a reference face model. Tzirakis et al. (2019) use a Deep Canonical Attentional Warping (DCAW) to translate the audio into expression blendshapes. Pham et al. (2018) directly map the audio to the blendshape parameters even though their results suffer from strong jitter. While current audio-driven facial expression techniques produce interesting results, most of them show results on voice data recorded in controlled studios with minimal background noise (Cha et al., 2018; Cudeiro et al., 2019; Karras et al., 2017; Liu et al., 2015; Pham et al., 2018; Taylor et al., 2017). Cudeiro et al. (2019) show interesting results in handling different noise levels.

Nevertheless, there is currently no audio-driven technique that estimates high quality facial expressions in-the-wild, as well as estimates the head motion and body conversational gestures. The proposed method uses the face model as the first category. In contrast to other methods, the proposed approach applies a simple but effective ap-

Figure 4.2: The 3D annotations are created from monocular in-the-wild videos using the monocular dense 3D facial reconstruction (dense face mesh visualized) approaches of Garrido et al. (2015), 3D hand pose estimation of Zhou et al. (2020), and 3D body pose estimation approach of Mehta et al. (2020).

proach to jointly learn the 3D head and face animation with body gestures, by directly regressing the facial parameters captured from a large corpus of community video.

4.3 DATASET CREATION

4.3.1 *Creating 3D Annotations from Video*

A major bottleneck for previous speech-driven animation synthesis work is the generation of sufficient training data. Many methods resort to complex in-studio capture of face and full-body motion with multi-camera motion capture systems. Therefore, this chapter proposes the first approach to extract automatic annotations of 3D face animation parameters, 3D head pose, 3D hands, and 3D upper body gestures from a large corpus of community videos with audio. In this way, much larger training corpora spanning over several long temporal windows and diverse subjects can be created more easily.

In particular, the approach uses the dataset of Ginosar et al. (2019a), which features 144 hours of in-the-wild video of 10 subjects (e.g., talk show hosts) talking into the camera in both standing and sitting poses. From these videos, Ginosar et al. (2019a) extracted 2D keypoints of the arms and hands, as well as 2D sparse face landmarks. They used a subset of these annotations to train a network synthesizing only 2D

arm and finger motion from speech. While showing the potential of speech-driven animation, their approach does not synthesize 3D body motion; does not synthesize 3D motions of the torso, such as leaning, which is an element of personal speaker style; and does not predict 3D head pose and detailed face animation parameters. To train a method jointly synthesizing the latter more complete 3D animation parameters in tune with input speech, the approach introduced in this chapter annotates the dataset with state-of-the-art 3D face performance capture and monocular 3D body and hand pose estimation algorithms, see Figure 4.2 and Figure 4.3.

For monocular 3D face performance capture, the optimization-based tracker of Garrido et al. (2015) is used to predict parameters of a parametric face model, specifically: 64 expression blend shape coefficients, 80 PCA coefficients of identity geometry, 80 PCA coefficients of face albedo, 27 incident illumination parameters, and 6 coefficients for 3D head rotation and position. The face tracker expects tightly cropped face bounding boxes as input. The face tracker from Saragih et al. (2011) is used for bounding box extraction and temporally filter the bounding box locations; the experimental results found this to be more stable than using the default 2D face landmarks in Ginosar's dataset for bounding box tracking. To train the algorithm, we use the face expression coefficients $\boldsymbol{\theta}_{Face} \in \mathbb{R}^{64}$ and head rotation coefficients $\mathbf{R} \in SO(3)$ are used.

For 3D body capture, the approach needs to be robust to body self-occlusions, occlusions by other people, occlusions of the body by a desk (sitting poses by talk show hosts), or occlusions by camera framing not showing the full body, even in standing poses. Therefore, the XNect (Mehta et al., 2020) monocular 3D pose estimation approach is used as it is designed to handle these cases. Specifically, in each video frame, the 3D body keypoint predictions from XNect's network output are extracted to obtain the 13 upper body joints (2 for head, 3 for each arm, 1 for neck, 1 for spine, 3 for hip/pelvis). This results in a 39-dimensional representation $\mathcal{K} \in \mathbb{R}^{39}$ for the body pose. The head rotation $\mathbf{R}$ predicted by the face tracker is grouped together with the 3D body keypoints $\mathcal{K}$ in a 42-dimensional vector $\boldsymbol{\theta}_{Body} \in \mathbb{R}^{42}$.

To perform hand tracking, the state-of-the-art monocular 3D hand pose estimation method of Zhou et al. (2020) is employed. To ensure good prediction results, we first tightly crop the hand images using the 2D hand keypoint annotations provided by Ginosar et al. (2019a)

before feeding it to the 3D hand pose predictor. Since the hands can be occluded or out of view, we also employ an off-the-shelf cubic interpolation method to fill in potentially missing 3D hand poses information. This results in 21 joints prediction for each hands, which we group into a 126-dimensional vector $\boldsymbol{\theta}_{Hand} \in \mathbb{R}^{126}$.

To improve the robustness of our data, we exclude the data if the prediction confidence of the face landmarks or hand keypoints within a certain number of frames falls below a stipulated threshold. This is obtained by reinterpreting the maximum value of the 2D joint heatmap prediction of the body parts produced by the tracker as a confidence measure. We also remove 4 out of 10 subjects provided by Ginosar et al. (2019a) due to the low resolution of the videos which lead to poor quality 3D face reconstruction results. Our final 3D dataset consists of more than 33 hours of videos from 6 subjects. We use the same training, validation, and test split as the original 2D dataset, which comprise around 80%, 10%, and 10% of the total data respectively, even after accounting for the excluded data.

The 3D body and hand pose prediction as well as the head rotation results are temporally smoothed using a Gaussian filter with a standard deviation of $\sigma = 1.5$ to improve the visual quality of the output. The same filter is also applied to the ground truth sequences in the video results.

4.3.2 *Annotation Quality*

The training data is created by annotating in-the-wild videos using 3D monocular tracking approaches. Naturally, the quality of our pseudo ground truth is less accurate compared to using a standard multi-view motion capture or performance capture system. Unlike Ginosar et al. (2019a) which evaluate their automatic 2D labels against human annotations, it is not possible for us to quantitatively measure the quality of our 3D annotations. However, due to the controlled setting of the recordings, most of the tracked videos consist of commonly observed poses often found in the training set of the monocular tracking methods we employ. Through manual inspection of the tracking results, we found that the prediction is quite reliable for our task.

To filter out low-quality data, the average confidence of the 2D keypoint predictions over a sequence is used based on a certain threshold.

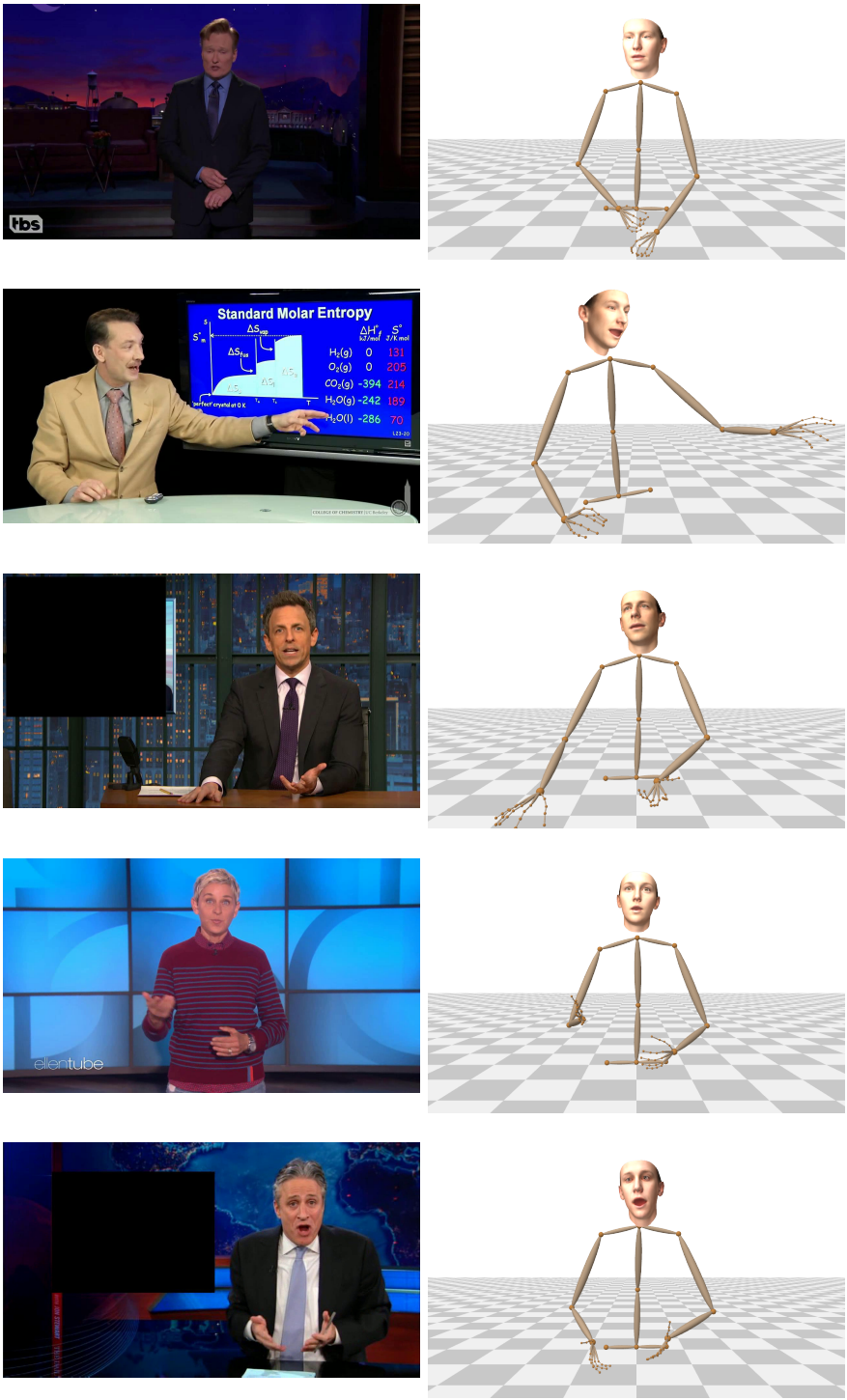


Figure 4.3: Additional subject-specific examples of the 3D face, body, and hand annotations in the proposed training data obtained from monocular estimation approaches.

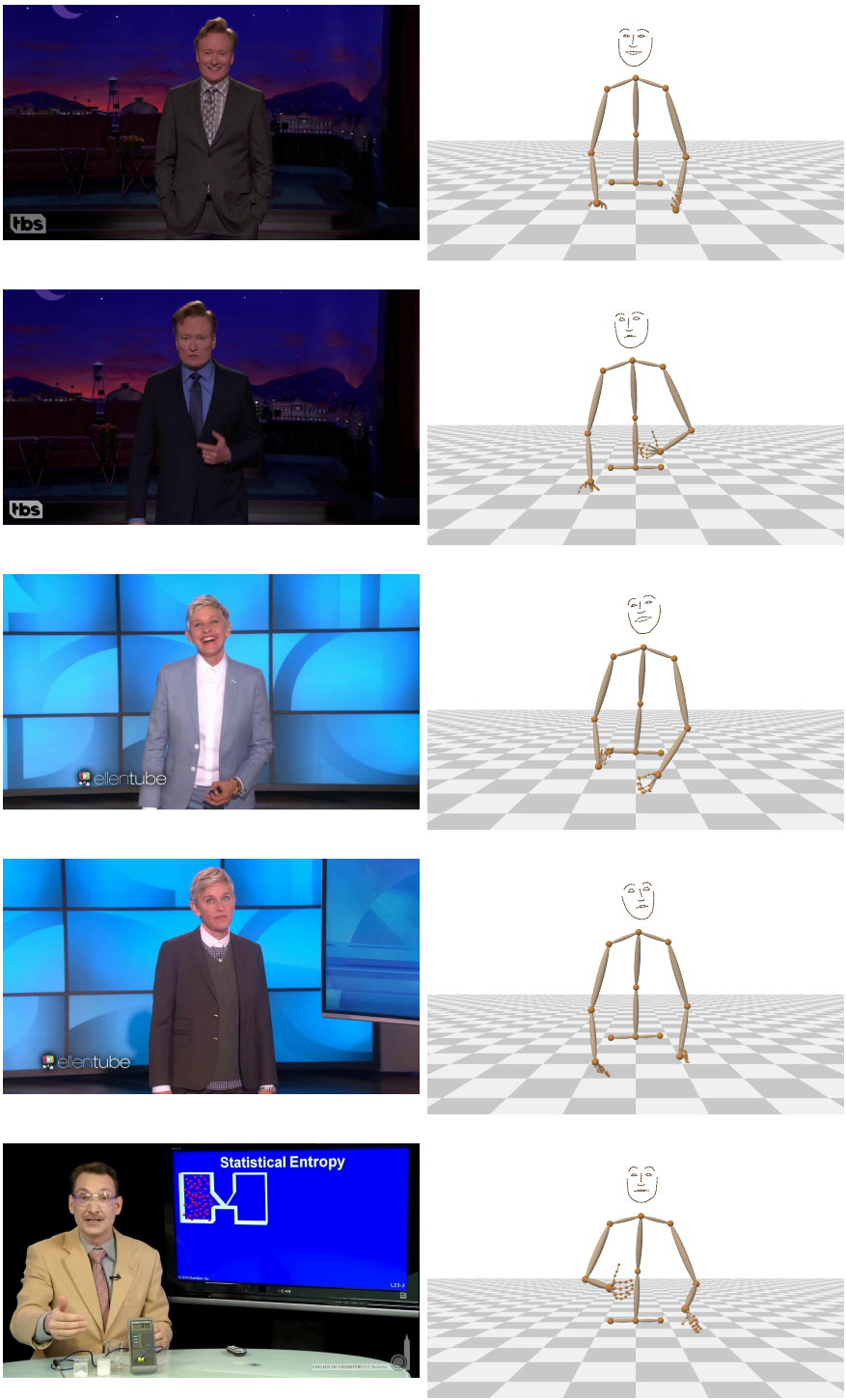


Figure 4.4: Occlusion scenarios are commonly observed in the video corpus, and they happen frequently for standing subjects. To alleviate this issue, a confidence-based filter is applied to remove these occluded frames. To ensure sufficient training data can be collected for each subject, the filtering threshold is designed to tolerate hand occlusion cases if they occur over a short time period.

However, the hands of the subject are also often out of view, especially in the videos where the subjects are standing, e.g. Ellen and Conan. To prevent losing a significant amount of data, our confidence threshold is designed to allow some of the occluded hand cases to be included in the dataset if they occur over a short time window. Some of these hand occlusion examples are shown in Figure 4.4.

Please refer to the original paper of the 3D body tracker XNect (Mehta et al., 2020), the hand tracker (Zhou et al., 2020), and the face tracker (Garrido et al., 2015) for the detailed quantitative performance of their respective methods on several benchmark datasets.

4.3.2.1 *2D-to-3D Lifting vs. Image-to-Body Pose*

The footage from Ginosar et al. is annotated with 2D body keypoint locations. One way to achieve 3D annotations is by running state-of-the-art 2D-to-3D lifting methods such as Martinez et al. (2017a) and Pavllo et al. (2019) on the provided 2D keypoints. However, multiple subjects in the dataset are either seated behind a desk or have parts of the upper body outside the image frame. The occluded pelvis (by the desk) causes the already ambiguous 2D-to-3D lifting approaches to have no information for correctly predicting the torso leaning, which is crucial for conveying conversational gestures. Approaches such as XNect (Mehta et al., 2020) are designed to be robust against partial occlusion and utilize image cues to predict the potentially missing torso.

4.3.3 *Audio features pre-preprocessing*

Similar to Suwajanakorn et al. (2017), we compute the Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980; Mermelstein, 1976) of each input video frame after normalizing the audio using FFMPEG (FFmpeg Developers, 2016; Robitza, 2019). This work uses CMU Sphinx (Lamere et al., 2003) for computing the coefficients, and use 13 MFC coefficients and an additional feature to account for the log mean energy of the input. These, together with their temporal first derivatives, yield a 28-dimensional vector $\mathcal{F}_{MFC} \in \mathbb{R}^{28}$ representing the speech input at each time step. MFCC encodes the characteristics of how human speech is perceived, which make it useful for a wide range of applications such as speech recognition. Encoding the characteristics

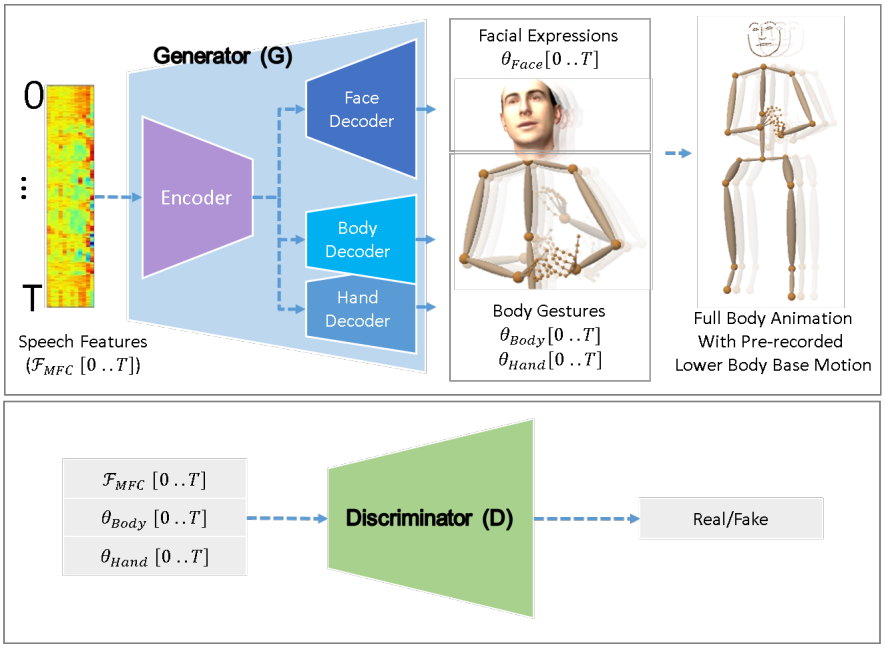
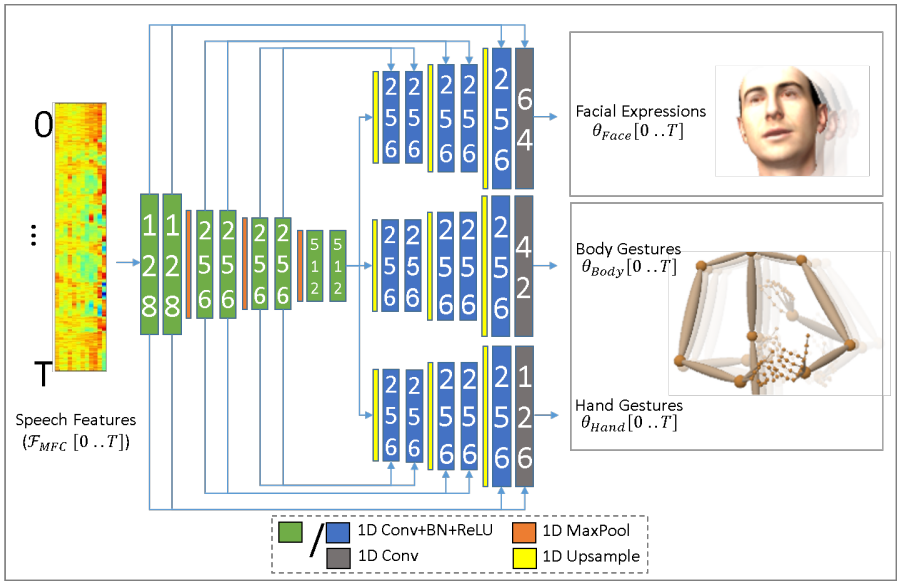


Figure 4.5: The proposed approach produces a temporal sequence of 3D facial expression parameters, head orientation, and 3D keypoints of the upper body and hands given a speech signal as input. An adversarial loss is employed in which the discriminator network tries to distinguish whether the input audio and body pose features are real or generated by the generator network.

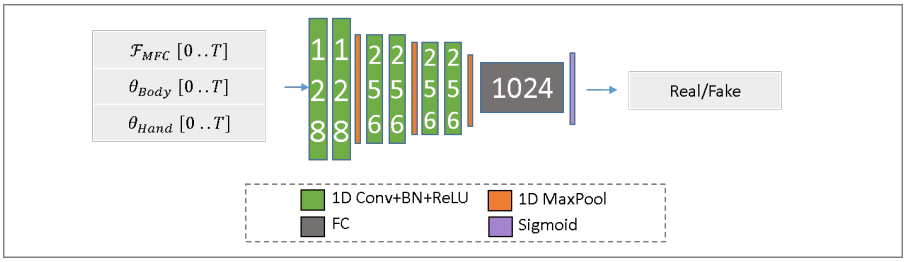
of speech perception make MFC coefficients a good representation for predicting facial expressions because modulation of face shapes is a part of the speech production process. For predicting body gestures, the change of MFCC features over the sequence carries the rhythm information needed to produce beat gestures.

4.4 APPROACH

The proposed approach produces a temporal sequence of 3D facial expression parameters, head orientation, 3D body, and 3D hand pose keypoints given a speech signal as input. Temporal variations in these aforementioned parameters contain gestural information. As described in section 4.3.3, the speech input is pre-processed to yield MFC based feature frames $\mathcal{F}_{MFC}[t] \in \mathbb{R}^{28}$ for each discrete time step t. The facial expression parameters at each time step are indicated as $\boldsymbol{\theta}_{Face}[t] \in \mathbb{R}^{64}$, 3D keypoints for both hands as $\boldsymbol{\theta}_{Hand}[t] \in \mathbb{R}^{126}$, and the head orientation and 3D body keypoints are represented together as $\boldsymbol{\theta}_{Body}[t] \in \mathbb{R}^{42}$. The temporal sequences are sampled at $15Hz$.



(a) Generator architecture.



(b) Discriminator architecture.

Figure 4.6: Details of the proposed network architecture. The design of the generator has a common encoder, but separate decoders for facial expression parameters, hands, and body pose (including head pose). The numbers in the blocks represent the number of feature channels output by the block. The discriminator uses the body and hand parameters which we concatenate with the audio features as input to the network.

4.4.1 *Network Architecture*

Similar to other adversarial learning-based approaches, the proposed model consists of 2 main neural networks, referred as the generator network G and discriminator network D. The proposed design follows prior human motion and gesture synthesis approaches (Ginosar et al., 2019a; Holden et al., 2016, 2015) by using 1D convolutional networks to model the temporal relationship of the audio and body features across different frames.

A 1D convolutional Encoder-Decoder architecture is employed for the generator network G to map the input audio feature sequence $\mathcal{F}_{MFC}[0:T]$ to 3D face expression parameter sequence $\hat{\boldsymbol{\theta}}_{Face}[0:T]$,

3D body pose parameter sequence $\hat{\boldsymbol{\theta}}_{Body}[0:T]$, and 3D hand pose parameter sequence $\hat{\boldsymbol{\theta}}_{Hand}[0:T]$ which is also trained in a supervised manner. Here, $\hat{\boldsymbol{\theta}}_{Face}[0:T]$, $\hat{\boldsymbol{\theta}}_{Body}[0:T]$, and $\hat{\boldsymbol{\theta}}_{Hand}[0:T]$ refer to the predicted outputs for the 3D face, body, and hands, respectively.

The proposed 1D convolutional architecture for the generator G is adapted from a reference implementation (Usuyama, 2018) of the U-Net (Ronneberger et al., 2015) architecture originally proposed for 2D image segmentation. The architecture utilizes a single encoder, comprised of 8 1D [Conv-BN-ReLU] blocks with a kernel size of 3, and is interleaved with MaxPool after every second block except the last. The last block is followed by an upsampling layer (nearest neighbor). Each face, body, and hand sequences utilize a separate decoder to learn body-part-specific motion characteristics. The decoders are symmetric with the encoder and comprised of 7 1D [Conv-BN-ReLU] blocks and a final 1D convolution layer, interleaved with upsampling layers after every second block. The decoders, being symmetric with the encoder, utilize skip connectivity from the corresponding layers in the encoder. The discriminator network is designed to predict whether its input audio and pose features are real or not. This network is comprised of 6 1D [Conv-BN-ReLU] blocks with a kernel size of 3, and is interleaved with MaxPool after every second block. Afterwards, it is followed by a linear layer and a sigmoid activation layer.

A schema of the architecture is shown in Figure 5.2. Figure 5.4 shows a diagrammatic representation of the network architecture.

4.4.2 *Training Details*

The proposed networks are trained on subject specific training sets in order to capture the particular gesture characteristics of the subject.

Each mini-batch for training comprises of a random sampling of such 64-frame sub-sequences extracted from all training sequences. Adam optimization algorithm (Kingma and Ba, 2014) is used for training, with a learning rate of $5e-4$, a mini-batch size of 25, and trained until 300,000 iterations per subject. Since the generator network is fully convolutional, the network can handle input speech features of arbitrary duration during test time.

The proposed generator network G is supervised with the following loss terms:

$$\mathcal{L}_{Reg} = w_1 * \mathcal{L}_{Face} + w_2 * \mathcal{L}_{Body} + w_3 * \mathcal{L}_{Hand}. \tag{4.1}$$

$\mathcal{L}_{Face}$ is the L2 error of facial expression parameters

$$\mathcal{L}_{Face} = \sum_{t=0}^{T-1} \|\boldsymbol{\theta}_{Face}[t] - \hat{\boldsymbol{\theta}}_{Face}[t]\|_2.$$

$\mathcal{L}_{Body}$ is the L1 error of 3D body keypoint locations and head orientation, and $\mathcal{L}_{Hand}$ is the L1 error of 3D hand keypoint locations

$$\mathcal{L}_{Body} = \sum_{t=0}^{T-1} \|\boldsymbol{\theta}_{Body}[t] - \hat{\boldsymbol{\theta}}_{Body}[t]\|_1,$$

$$\mathcal{L}_{Hand} = \sum_{t=0}^{T-1} \|\boldsymbol{\theta}_{Hand}[t] - \hat{\boldsymbol{\theta}}_{Hand}[t]\|_1.$$

The hyperparameters are defined as $w_1 = 0.37$, $w_2 = 600$, and $w_3 = 840$ to ensure that each term is equally weighted during training.

In practice, experiments show that only employing L1 or L2 error for body keypoints results in less expressive gestures, as has also been pointed out in prior work on 2D body gesture synthesis of Ginosar et al. (2019a). Inspired by the adversarial training approach of Ginosar et al. (2019a), the proposed experiments show that incorporating an adversarial loss using a discriminator network D which is trained to judge whether an input pose is real or fakely generated by the generator G, can lead to more expressive gestures that are also in-sync with the speech input. When trained together with the generator network in a minimax game scenario, it will push the generator to produce a higher quality 3D body and hand pose synthesis in order to fool the discriminator. The proposed method follows similar approach to the work of Ferstl et al. (2019) by using not only the pose but also the audio features as input to the discriminator. This way, the discriminator is not only tasked to measure if the input gesture looks real, but it also needs to determine if the gesture is in-sync with the input audio features or not. Since the multi-modality of the body gestures mainly occurs for the body and hands, the facial expression parameters are excluded from the adversarial loss formulation:

$$\mathcal{L}_{Adv}(G, D) = \mathbb{E}_{\mathcal{F}_{MFC}}[\log(1 - D(\mathcal{F}_{MFC}, G^*(\mathcal{F}_{MFC})))] + \mathbb{E}_{\mathcal{F}_{MFC}, \boldsymbol{\theta}_{Body}, \boldsymbol{\theta}_{Hand}}[\log D(\mathcal{F}_{MFC}, \boldsymbol{\theta}_{Body}, \boldsymbol{\theta}_{Hand})] \quad (4.2)$$

where G^* indicates that only use the predicted $\boldsymbol{\theta}_{Body}$ and $\boldsymbol{\theta}_{Hand}$ outputs of the original generator network G.

Combined with the direct supervision loss, the overall loss is

$$\mathcal{L} = \mathcal{L}_{Reg} + w \cdot \min_D \max_G \mathcal{L}_{Adv}(G, D) \quad (4.3)$$

where w is set to be 5.

4.5 RESULTS

The proposed approach addresses essential aspects of animating virtual humans: synthesizing facial expressions, body, and hand gestures in tune with speech. For visualization of the results, as well as for the user study, to allow observers to focus on the face and body motion, we render an abstract 3D character that showcases all the important skeletal and facial elements without the risk of falling in the uncanny valley, following similar approaches in prior work (Ginosar et al., 2019a; Levine et al., 2009). Since the proposed approach only predicts upper body motion, we fuse it with a pre-recorded base motion of the lower body in both sitting and standing scenarios.

Since the synthesis of conversational gestures is a multi-modal problem, direct comparison with the tracked annotations would not be meaningful for all aspects of the synthesized results, particularly for evaluating the realism of the synthesized gestures. The proposed method is evaluated through extensive user studies to judge the quality and the plausibility of our results, and it is compared against various baselines. Further, the prediction of the facial expressions is measured by comparing the 3D lip keypoints extracted from selected vertices of the predicted 3D face model against the automatically generated ground truth lip keypoints obtained from the source image. A qualitative example of the synthesis result is shown in Figure 4.7.

Table 4.1: User study result measuring both the naturalness and synchronization between the synthesized face+body+hand gesture and speech.

Method	Naturalness	Sync.
Ground truth	$\mathbf{4.29 \pm 0.86}$	$\mathbf{4.39 \pm 0.77}$
Direct regression	3.54 ± 1.11	3.78 ± 1.08
LSTM (adopted from Shlizerman et al. (2018))	3.15 ± 1.03	3.21 ± 1.11
Adv. loss on velocity (adopted from Ginosar et al. (2019a))	3.03 ± 0.98	3.38 ± 0.95
Adv. loss on audio+3D pose (proposed)	$\mathbf{4.05 \pm 0.85}$	$\mathbf{4.00 \pm 0.91}$

4.5.1 *Baseline Comparisons*

The proposed approach is evaluated against other methods that perform body gesture prediction, which use audio features as input. Other baseline methods are trained using the same MFCC features described in 4.3.3. The first baseline is the direct regression 1D CNN model of the proposed network architecture without using adversarial loss. Next, the method is compared against a Recurrent Neural Network (RNN)-based Long Short-term Memory (LSTM) architecture by Shlizerman et al. (2018), which is originally designed to temporally predict 2D hand and finger poses. Since the original method is not designed to handle multi-modal data, we train three LSTM models for face, body, and hand gesture are trained separately on the newly proposed 3D data.

An adaptation of Ginosar et al. (2019a) is trained using the proposed model with the adversarial loss to distinguish between the real and fake synthesis of the gesture in the velocity space similar to their proposed approach and use this version as the baseline comparison. The proposed method is also compared against the work of Alexanderson et al. (2020a) by retraining their method on our in-the-wild 3D data. Their model is originally trained on clean mocap data of 3D body pose without face or hand annotations. The model of Alexanderson et al. (2020a) is found to be sensitive to the hyperparameters used. Because of this, it is only trained on the body and hand data to simplify the problem. An optimal set of hyperparameters that can produce the best results in terms of naturalness and synchronization are manually searched based on the recommendation of the original authors of the work. Following their instruction, multiple experiments were

conducted by varying the number of units H between 512, 700, and 800 and the number of flow-steps K from 8 up to 16. The MoGlow-based model is found to produce the best results when using the number of units $H = 800$ and the number of steps $K = 10$.

4.5.2 *Gesture Synthesis User Study Evaluation*

Two separate user studies are conducted for the qualitative evaluation of the proposed method. The first user study compares methods that synthesize the 3D face, body, and hand gestures from audio. In particular, the participants were shown 3 out of 6 randomly selected video sequences (12 seconds/sequence) synthesized by the proposed method, along with the baselines, and the ground truth (tracked) annotations. This study involved 67 participants. Each user was asked to judge the naturalness and the synchronization between the audio and the generated 3D face and body gestures on a scale of 1 to 5, with 5 being the most plausible and 1 being the least plausible. As shown in Table 4.1, the ground truth sequences are perceived as both the most natural and in-tune with the input speech compared to other synthesized gesture videos, which is rated at 4.29 ± 0.86 and 4.39 ± 0.77, respectively. Compared to other baseline methods, the participants agree that the results produced by the proposed method look more natural and in tune with the speech audio with the score of 4.05 ± 0.85 in terms of naturalness and 4.00 ± 0.91 in terms of synchronization with the speech.

Table 4.2: User study result measuring both the naturalness and synchronization between the synthesized body+hand gesture and speech. The users were specifically asked to ignore the quality of the facial expression.

Method	Naturalness	Synchron.
MoGlow (Alexanderson et al. (2020a))	2.88 ± 1.02	3.11 ± 1.13
Proposed	$\mathbf{4.01 \pm 0.82}$	$\mathbf{3.93 \pm 0.92}$

A second user study is conducted to evaluate the synthesis of the 3D body and hand gestures and compare the proposed method with the MoGlow-based model of Alexanderson et al. (2020a). For this study, the participants were specifically asked to ignore the quality of the facial expressions in the video. To ensure a fair comparison, all videos presented in this study were synthesized by using the 3D

facial expression predicted by the proposed method. Similar to the first study, each of the 45 participants was asked to rate the quality of the gestures from 3 out of 6 possible videos for each method on a scale between 1 to 5. As shown in Table 4.2, the proposed method is rated as both more natural and in-sync with the audio.

Table 4.3: Quantitative comparison to baseline methods for lip motion prediction error against the ground truth (in mm). (*) indicates that the method is adopted and retrained on the proposed 3D speech-to-gesture dataset.

Method	Oliver	Meyers	Ellen
Direct regression	**0.29**	**0.35**	0.29
LSTM* (Shlizerman et al. (2018))	0.3	0.36	0.30
Adv. loss* (Ginosar et al. (2019a))	**0.29**	**0.35**	0.3
Random	0.49	0.57	0.47
Proposed	**0.29**	**0.35**	**0.28**
Method	Kubinec	Stewart	O'Brien
Direct regression	**0.28**	**0.39**	**0.37**
LSTM* (Shlizerman et al. (2018))	0.32	0.41	0.39
Adv. loss* (Ginosar et al. (2019a))	0.29	0.39	0.38
Random	0.43	0.57	0.52
Proposed	**0.28**	**0.39**	**0.37**

4.5.3 *Facial Expression Evaluation*

Table 4.3 compares the 3D lip keypoints of the generated face vertices corresponding to the facial expressions predicted by various approaches against the image-based face tracker's 3D lip keypoints in a neutral head pose. The comparison was performed on the whole test set, which consists of 578 sequences (12 seconds/sequence) across all subjects. As a sanity check baseline, the evaluation also computes the difference between the optimization-based tracked annotations of one sequence to the optimization-based annotations on a *different* sequence *chosen randomly*. The evaluation shows that the proposed method achieves similar or slightly better performance against other proposed baselines. This result also demonstrates that the proposed

unified whole-body architecture is suitable for simultaneous face expression synthesis at decent quality and better than simultaneous face synthesis with other body gesture synthesis architectures. Note here that the work proposed in this chapter is not claiming that the proposed design advances the state-of-the-art in face-only expression synthesis. This is outside the scope of this chapter and left for future work.

4.6 DISCUSSION

While mouth expressions are strongly correlated with speech, the rest of the intended generation targets, such as body gestures, do not have a one-to-one mapping. Coupled with the noisy nature of the proposed monocular data, as observed in the experiments, this multi-modal nature of the problem makes both designing and analyzing a stable expressive model challenging. It is also observed that a lower value of $L1$ or $L2$ loss on the validation set does not always guarantee to produce a qualitatively better gesture synthesis, which further shows the importance of the adversarial loss. The data is also inherently noisy due to the use of 3D monocular trackers, which may lead to jittery 3D motion that can affect the performance of the proposed model and comparison baselines. However, the effect of the noise is observed to be minimal, as it can be suppressed by applying temporal filters to the prediction output.

Based on the observed results, it can be argued that the discriminator network can potentially be used as a plausibility metric to rate the quality of a gesture synthesis from speech, similar to how the inception score is used Salimans et al. (2016), if trained with enough gesture and noise variations. One way to validate this idea is to train the model to classify whether its audio-gesture pair input is in-sync or off-sync. The ground truth audio-gesture pairs can be used directly as in-sync (positive) samples, while off-sync (negative) samples can be prepared by pairing the audio sequence with a different gesture from a random pair.

When the discriminator network is trained in this setup, it can reliably classify unseen test pairs of subject "Oliver" with high accuracy of 87.4%, see Table 4.4. Unfortunately, since the classifier is trained only on ground-truth motion sequences, it is not yet possible to extend

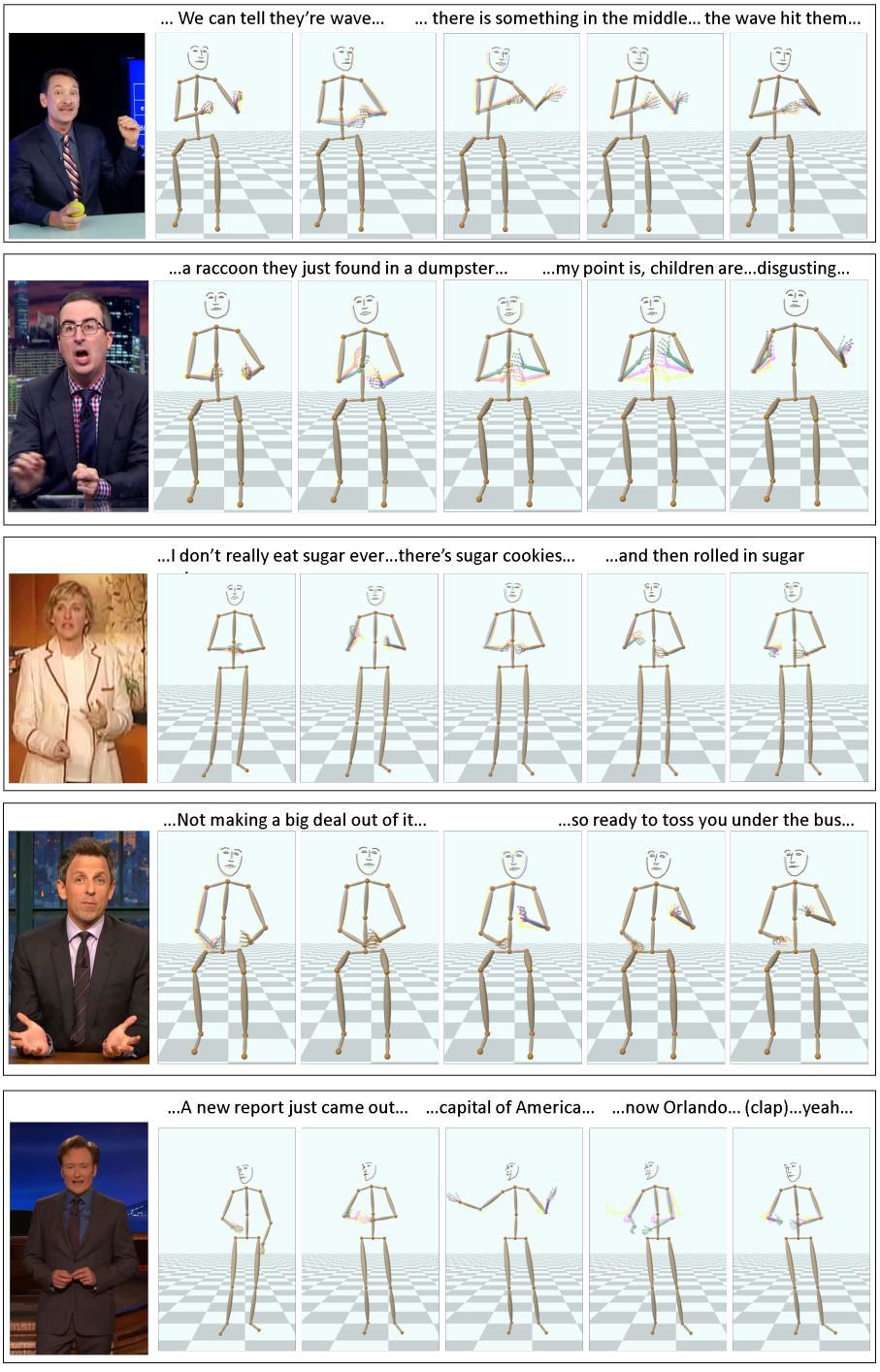


Figure 4.7: Several qualitative result examples of the proposed approach. As demonstrated in the user study (Table 4.1), the method can generate a plausible gesture of 3D body and hands, as well as 3D facial expression from speech input across multiple speakers when trained in a subject-specific manner. Motion visualization is based on the Harris shutter effect.

this model as a quantitative metric for gesture synthesis methods. When this classifier is tested on the baseline models, it produces inconsistent results. For example, it rates the proposed model to be more plausible than the ground truth sequences, which contradicts the result of the user study. A specific dataset containing different gesture noise characteristics may be required if one wants to extend this classifier into a more general gesture plausibility metric.

Table 4.4: Quantitative result of the discriminator when trained as an audio-to-body sync/off-sync pair classifier on Oliver test sequences (higher is better).

Seq. length	In-sync pair	Off-sync pair	Combined
64 frames	82.7%	92.1%	87.4%
32 frames	81.4%	83.1%	82.2%
16 frames	74.1%	80.8%	77.5%

4.7 CONCLUSION

This chapter proposes the first approach for full 3D face, body, and hand gesture prediction from speech to automatically drive a virtual character or an embodied conversational agent. Inspired by the promising monocular human pose estimation result presented in Chapter 3, this work leverages monocular 3D face reconstruction and body pose reconstruction approaches on in-the-wild footage of talking subjects to acquire training data for the proposed learning-based approach, generating 3D face, body, and hand pose annotations for approximately 33 hours of footage. The key insight on incorporating an adversarial penalty not only on the 3D pose but also its combination with the audio input allows the model to successfully generate expressive body gestures that are in-sync with the speech.

The method presented in this chapter demonstrates the ability of a deep neural network model to resolve the problem of regressing to the mean pose commonly found in speech-driven synthesis tasks. The following chapter extends this idea by allowing user-guided gesture modification from various control signals, which is a desirable feature in a 3D animation pipeline.

5

CONTROLLABLE SPEECH-DRIVEN 3D GESTURE SYNTHESIS USING DATABASE MATCHING

Chapter 4 presents a deep learning-based approach that shows promising results for synthesizing 3D human gestures from speech input. However, that approach offers limited freedom to incorporate additional user control. Furthermore, training an audio-to-gesture mapping in a supervised manner often does not capture the multi-modal nature of the data, mainly because the same audio input can produce different gesture outputs. To address these problems, this chapter presents an approach for generating controllable 3D gestures that combine the advantage of database matching and deep generative modeling (published as Habibie et al., 2022).

The method proposed in this chapter predicts 3D body motion by sequentially searching for the most plausible audio-gesture clips from a database using a k-Nearest Neighbors (k-NN) algorithm that considers the similarity between the input audio and the previous body pose information. To further improve the synthesis quality, this chapter proposes a conditional Generative Adversarial Network (cGAN) model to provide a data-driven refinement to the k-NN result by comparing its plausibility against the ground truth audio-gesture pairs. The novel approach enables direct and more varied control manipulation that is not possible with prior learning-based counterparts, such as user-guided manipulation that alter the position or velocity of the generated gesture to reflect a particular speaking style or emotional state. Experiments show that the proposed approach outperforms recent models on control-based synthesis tasks using high-level signals such as motion statistics, e.g., hand speed or wrist height placement. Moreover, this strategy also enables flexible and effective user control for lower-level signals, such as direct frame-based gesture replacement based on specific matching criteria, thus opening the possibility for semantic control based on specific keywords observed in the given input speech.

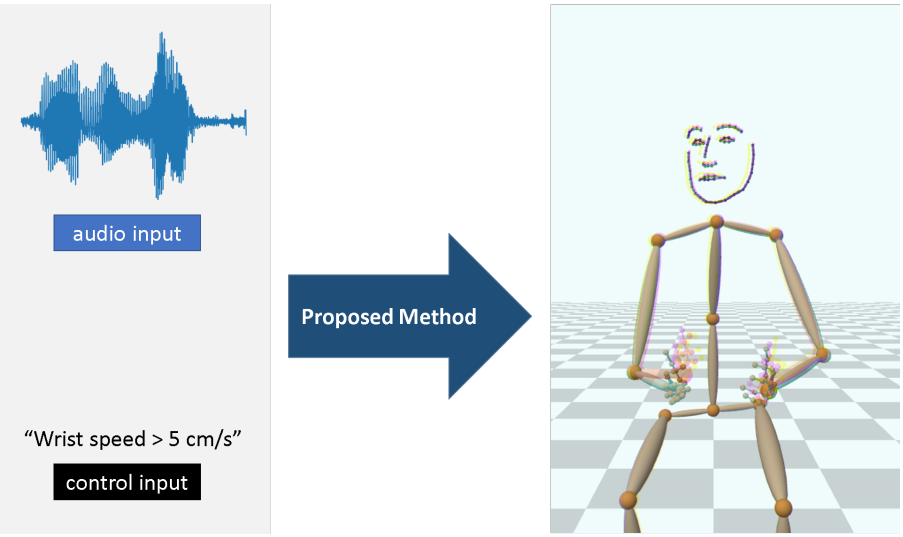


Figure 5.1: The approach proposed in this chapter allows for controllable speech-to-gesture synthesis by combining a novel database matching algorithm and a conditional adversarial network.

5.1 INTRODUCTION

Creating human-like 3D avatars is important in order to provide immersive experiences in virtual worlds. Advances in 3D vision and graphics can now synthesize human-like virtual characters that emulate various aspects of human anatomy, thus potentially simplifying the production of personalized avatars. Designing an easy and accessible way to control such avatars can improve social interactivity between users and such avatars in shared virtual environments.

An appealing approach for developing intuitive character control is to synthesize character gestures from input speech. However, developing an algorithm for speech-to-gesture synthesis is known to be a challenging task (Alexanderson et al., 2020b; Ferstl et al., 2019; Ginosar et al., 2019b; Habibie et al., 2021b). This is partly due to the nature of the audio-to-gesture relationship, where many different gesture sequences may be appropriate for a given speech input. Hence, training a regression-based model in a supervised manner can lead to unnatural "averaged" gesture results as the consequence of regressing multiple outcomes of a single input signal. While recent methods (Ferstl et al., 2019; Ginosar et al., 2019b; Habibie et al., 2021b) use adversarial learning to mitigate the problem of "averaged" synthesis, they provide very limited options for controlling the output, and hence they predict only one particular motion sequence for every speech input. Since human

gesture is known to be related to the personality and internal state of the speaker (Smith and Neff, 2017), the ability to control body motion based on a specific input signal, such as their emotional state can significantly improve the usability of the method. Recently, generative models were employed to introduce probabilistic synthesis to allow some degree of high-level gesture control (Alexanderson et al., 2020b). However, such methods typically need high-quality training data to work well, are slow to train, and require separate pre-trained models to produce different types of control, thus limiting their usability when multiple aspects of the control signal should be varied.

This chapter proposes a new approach for controllable 3D body gesture generation from speech inspired by the popular Motion Matching algorithm commonly found in locomotion synthesis (Büttner and Clavet., 2015) (see Figure 5.2). At the core of the method is a novel k-Nearest Neighbor-based (k-NN) algorithm that selects short clips from a database and is specifically designed to leverage the similarity both in the audio and in the motion space to ensure a continuous, natural, and synchronous gesture output. The quality of the initial synthesis is further improved by passing the k-NN output into a conditional Generative Adversarial Network (cGAN). The cGAN is conditioned on both the audio and the motion synthesized by the k-NN as input and is tasked with producing a new motion that looks similar to the real ground truth motion in the database. This allows the network to perform correction on any less plausible motion generated by the k-NN, especially around the transition boundary between segments. The proposed gesture synthesis formulation can be naturally extended to select motion based on control information by only considering motion candidates that match the control criteria. Unlike the most related approach of Alexanderson et al. (2020b), the synthesis output can be controlled at any particular time and with various types of conditioning without the need to re-train the model for every given type of control signal. Furthermore, the control criteria extend beyond high-level signals, such as motion statistics, to include direct per-frame manipulation, which can be exploited for various exciting applications such as semantic-level control.

Experiments show that the proposed model outperforms related control-based audio-driven synthesis (Alexanderson et al., 2020b), both in terms of naturalness and audio synchronization. Furthermore, even in the absence of control, the proposed technique achieved better

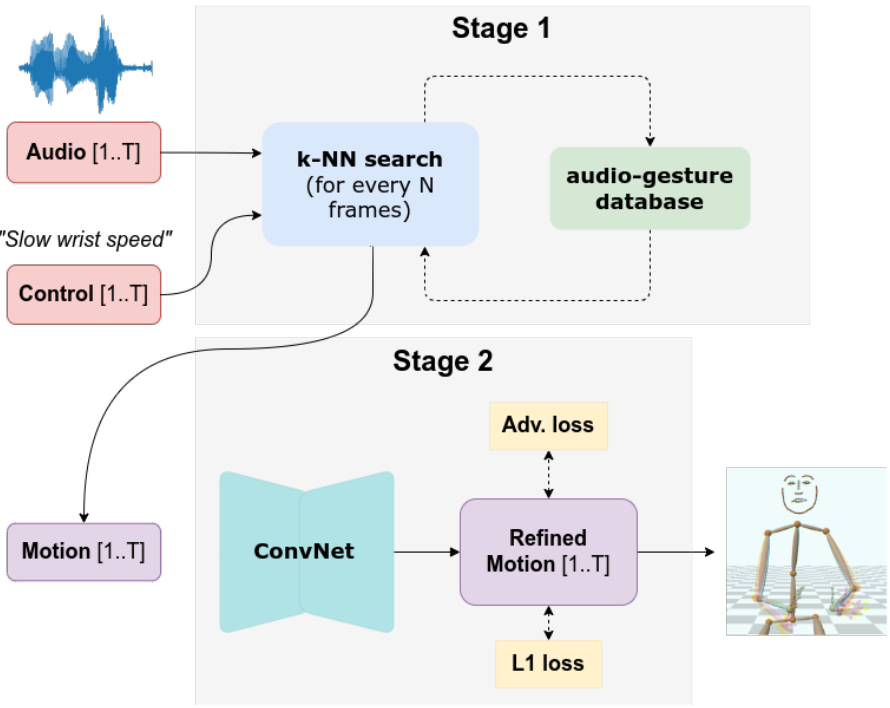


Figure 5.2: The proposed pipeline consists of two main stages. In Stage 1, the method first employ a k-Nearest Neighbor search to find the most plausible sequence considering the audio and previous pose similarity in the database. At any given time step, additional information can be provided to incorporate further control over of the synthesis output. The 3D gesture generated through Stage 1 is then passed to a conditional GAN trained to produce a refined gesture sequence by comparing the output against real audio-gesture sequences.

synthesis quality than the previous state-of-the-art approach (Habibie et al., 2021b). In summary, the contributions proposed in this chapter are three-fold:

1. A novel Motion Matching-based algorithm for gesture synthesis from speech,
2. A deep generative modeling approach to resynchronize and enhance the synthesis quality from the k-NN by leveraging the whole training data that cannot be fully exploited using database features alone,
3. A more interpretable design that enables a greater set of control signals than other previous learning-based approaches, thus facilitating a more comprehensive range of potential applications.

5.2 RELATED WORK

Several relevant related works on the problem of speech-driven gesture synthesis have been discussed in Chapter 4. This section additionally surveys prior arts for 3D human motion control since it is one of the main objectives of the method introduced in this chapter.

Graph-based algorithms were amongst the most popular choice for classical data-driven character animation and control (Arikan and Forsyth, 2002; Kovar et al., 2002; Lee et al., 2002; Safonova and Hodgins, 2007). For example, a motion graph can be constructed to model connections and transitions between motion clips in a dataset, and motion synthesis can be achieved by traversing the graph. However, graph-based approaches do not scale well with data and at times, can be imprecise and unresponsive to the control input. Instead of directly using the actual pose representation in the search space of the graph-based approaches, Lee et al. (2014) propose *motion fields* to generalize the motion representation into a higher dimensional vector space, enabling more responsive synthesis. To achieve control, a reinforcement learning (RL) model is trained to find the best action that can satisfy the user's input. Büttner and Clavet. (2015) propose Motion Matching to further simplify this process and approximate the RL algorithm by casting it as a k-Nearest Neighbor search. Motion generation and control are performed by selecting the most suitable clip from the database that best matches the previous pose and the user's desired trajectory, see also Figure 5.3. Holden et al. (2016) propose one of the earliest deep learning based framework to achieve 3D human motion synthesis and control. Habibie et al. (2017) explore the possibility of using a deep generative model for 3D motion control based on a combination of a variational autoencoder (Kingma and Welling, 2014) and an LSTM network. Holden et al. (2017) propose a real-time character control approach using a phase-functioned neural network. Zhang et al. (2018) extend this approach to achieve high-quality synthesis on locomotion data where the walking phase is unstructured. Starke et al. (2019) propose a further extension of this idea to enable scene-aware motion control.

Holden et al. (2020) propose a fully learning-based approach to Motion Matching for locomotion by emulating the database look-up process with a neural network regressor. This is possible for locomotion

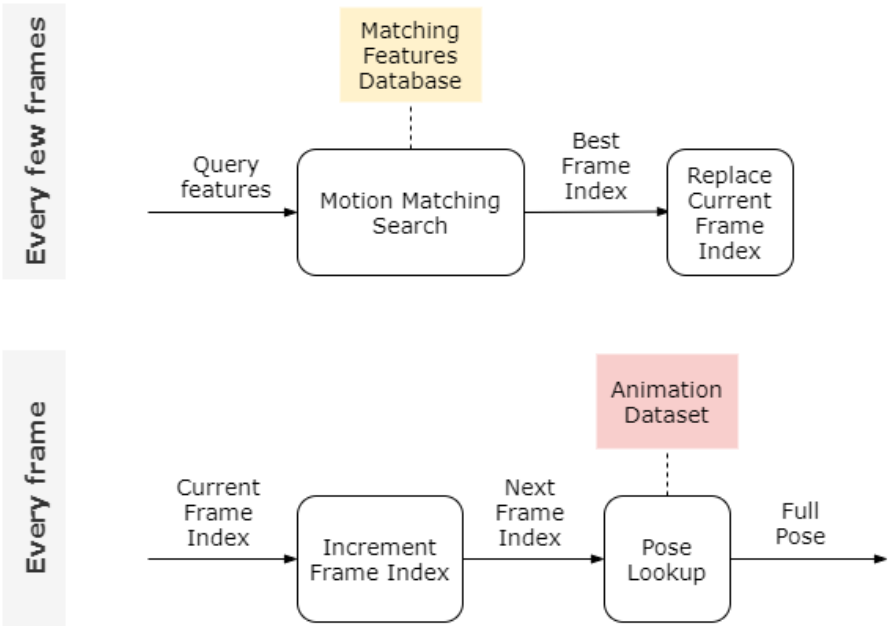


Figure 5.3: A schema of the Motion Matching algorithm. Figure taken from Ubisoft (2020).

control since there is a precise mapping between motion trajectory and the corresponding 3D joint position of the character. Unfortunately, this approach does not work well for speech-to-gesture synthesis. This is because there are multiple plausible 3D gestures that can be mapped from a single speech input, making it very difficult to replace database search using a standard regression algorithm. In contrast, this chapter used a classical nearest neighbor-based approach to perform synthesis. Since the k-NN approach makes direct use of the input data, the proposed algorithm can be extended to allow various types of gesture control by restricting access to subsets of this data. A learning-based model is further used to refine and re-synchronize the outcome of the k-NN.

5.3 APPROACH

The proposed system consists of two main components. This section first describes the design of the proposed nearest neighbor (k-NN) algorithm for gesture synthesis and control. Then, it also describes the design choices to improve the k-NN result through the use of a cGAN to transform the gesture into a more natural and synchronized motion.

5.3.1 *Nearest Neighbor-based Gesture Synthesis*

The proposed k-NN is inspired by the Motion Matching algorithm, which has become a popular method of choice for locomotion synthesis in the gaming industry due to its flexibility and good visual quality (Büttner and Clavet., 2015; Holden et al., 2020). Direct selection over the database using k-NN naturally avoids the problem of regressing to the mean pose while also providing more flexible control options. Multi-modal synthesis can be generated by either selecting different k-values or choosing different pose initializations.

5.3.1.1 *Input/Output Parameters.*

The proposed k-NN algorithm takes as input a sequence of audio features $\mathbf{F} = [\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{f}_{T-1}]$, one frame of previous initial pose features $\mathbf{p}_{-1}$, and optionally a sequence of control masks $\mathbf{C} = [\mathbf{c}_0, \mathbf{c}_1, ..., \mathbf{c}_{T-1}]$, where T is the number of frames in the sequence. The output is a sequence of 3D body poses $\mathbf{G} = [\mathbf{g}_0, \mathbf{g}_1, ..., \mathbf{g}_{T-1}]$. Each audio feature frame $\mathbf{f}_t$ and pose feature frame $\mathbf{p}_t$ encode information about the relevant future frames. The audio feature consists of the first 13 coefficients of the Mel-frequency cepstral coefficients (MFCC) as well as the audio log mean energy. The pose feature is derived from the 3D locations of the wrists, elbows, index finger roots, and little finger roots of both hands.

To find the output sequence, the input features need to be matched with the sequences in the *Matching Database*. The database is constructed from a collection of ground truth audio and 3D body motion pairs. It consists of M sequences of training audio features $\mathcal{F} = [\tilde{\mathbf{F}}^0, ..., \tilde{\mathbf{F}}^{M-1}]$, training pose features $\mathcal{P} = [\tilde{\mathbf{P}}^0, ..., \tilde{\mathbf{P}}^{M-1}]$, as well as the corresponding gesture sequence $\mathcal{G} = [\tilde{\mathbf{G}}^0, ..., \tilde{\mathbf{G}}^{M-1}]$ where $\tilde{\mathbf{F}}^m = [\tilde{\mathbf{f}}^m_0, ..., \tilde{\mathbf{f}}^m_{T_{match}-1}]$, $\tilde{\mathbf{P}}^m = [\tilde{\mathbf{p}}^m_0, ..., \tilde{\mathbf{p}}^m_{T_{match}-1}]$, and $\tilde{\mathbf{G}}^m = [\tilde{\mathbf{g}}^m_0, ..., \tilde{\mathbf{g}}^m_{T_{match}-1}]$. The sequences in the database are prepared by segmenting the original videos into $T_{match} = 64$ frame chunks.

5.3.1.2 *Details on Constructing the Audio and Pose Features.*

Each audio feature at frame t stores the relevant audio information at time t, $t + 2$, ..., $t + 14$ (at 15 fps). For the audio, we use the first 13 coefficients of the Mel-frequency cepstral coefficients (MFCC) as well as the log mean energy of the input audio $\mathbf{m}_t \in \mathbb{R}^{14}$. We stack

these together into a 1-D vector $\{\mathbf{m}_t, \mathbf{m}_{t+2}, ..., \mathbf{m}_{t+14}\} = \mathbf{f}_t \in \mathbb{R}^{112}$ (112 = 14 features $\times$ 8 frames). Similarly, every pose feature frame at time t stores information about the 3D pose coordinate of both left and right wrists, elbows, index finger root, and little finger root at time t, $t+2$, $t+4$, and $t+6$. We combine the 3D positions of all 8 joints into a single vector $\mathbf{p}_t \in \mathbb{R}^{96}$ (96 = 2 hands $\times$ 4 joints $\times$ 3 dims $\times$ 4 frames). This is equivalent to storing the hand trajectory within the next 0.5 seconds. The output pose at frame t contains the 3D coordinate of 13 body joints (pelvis-relative), and 21 joints for each hand, which we combine into a 1-D vector $\mathbf{g}_t \in \mathbb{R}^{165}$ (165 = 55 joints $\times$ 3 dims).

5.3.1.3 *Proposed Search Algorithm.*

To find the optimal output gesture sequence $\mathbf{G}$ from the database, the algorithm considers both the similarity with respect to the current test audio features $\mathbf{f}_t$, as well as the previously searched pose features $\mathbf{p}_{t-1}$ for every N frame interval. Please note that each feature frame contains information of future frames. In the first iteration, the previous pose feature $\mathbf{p}_{-1}$ is initialized by either randomly sampling a frame from the database or set to be the mean pose. Weighting the importance of audio and pose terms is a challenging task since their quantities cannot be directly compared. The proposed algorithm resolves this issue by aggregating the similarity rank of the candidates in both audio and pose space. For every iteration, a gesture sequence candidate is picked if the sum of its audio and pose similarity rank is the lowest compared to other candidates.

To speed up search computation, the best candidate from each training sequence $(\tilde{\mathbf{F}}^m, \tilde{\mathbf{P}}^m)$ in the database is pre-selected based on either the pose similarity ("pose pre-selected") or audio similarity ("audio pre-selected") before scoring them based on both pose and audio similarity scores. This section focuses on the description of the "pose pre-selected" k-NN version of the algorithm, which is also used as input to the later stage, even though experiments also find the result of "audio pre-selected" compelling.

The gesture selection is performed at a regular interval of $N = 8$ frames. During each iteration, given the current frame t, M pose sequence candidates are first pre-selected from the database by comparing the Euclidean distance to the previous pose feature $\mathbf{p}_{t-1}$ at frame $t-1$, denoted as $\{\hat{\mathbf{p}}^0_{0:(N-1)}, \hat{\mathbf{p}}^1_{0:(N-1)}, ..., \hat{\mathbf{p}}^{M-1}_{0:(N-1)}\}$. The similarity

of the audio features is also measured by comparing the current test audio feature $\mathbf{f}_t$ against the corresponding audio feature frame candidates from the database $\{\hat{\mathbf{f}}^0_{0:(N-1)}, \hat{\mathbf{f}}^1_{0:(N-1)}, ..., \hat{\mathbf{f}}^{M-1}_{0:(N-1)}\}$ using a cosine distance metric.

To aggregate both metrics, two separate rankings are created based on audio match quality and pose match quality using their similarity scores. Afterward, both the pose similarity and audio similarity ranks for every candidate are combined by adding their respective ranks in both lists. This combined rank list $R_{combined}$ is then used as the new metric to select the best gesture sequence. A gesture output candidate $\mathbf{g}^*_{0:(N-1)}$ is selected as the best output gesture $\mathbf{g}_{t:(t+N-1)}$ for the current frame t if its corresponding audio and pose features result in the lowest rank in $R_{combined}$.

Algorithm 1 summarizes the proposed matching-based approach.

5.3.2 *Search Database*

The *Matching Database* consists of 9624 unique gesture $\tilde{\mathbf{G}}$, audio $\tilde{\mathbf{F}}$, and pose $\tilde{\mathbf{P}}$ feature sequences, each of which is $T_{match} = 64$ frames, as is commonly used for this dataset. We use the non-overlapping "Oliver" sequences of the in-the-wild speech-to-gesture data originally prepared by Ginosar et al. (2019b) and 3D tracked by Habibie et al. (2021b) as the search database. This data consists of more than 11 hours of audio-gesture pairs of 3D face, body, and hand poses tracked using state-of-the-art monocular trackers, and contain various range of conversational 3D upper body and hand gestures. Table 5.1 describes the detail of the search database used in the experiments.

5.3.2.1 *Enabling Gesture Control with k-NN search.*

Since the algorithm performs an explicit comparison between features in the database, this process can be naturally extended to enforce high to mid-level control over the synthesis, allowing a more direct and interpretable way to achieve the desired behavior. For example, simulating gesture generation that follows a particular motion statistic can be achieved by simply labeling parts of the training data that satisfy the criteria. For example, to produce a sequence of body gestures where the left hand is always higher than the specified threshold $\mathbf{r}$, the search can be restricted to only consider frames where hand heights are higher

Algorithm 1: audio-to-gesture k-NN search

Data:
list of audio feat. sequence $\mathcal{F} = [\tilde{\mathbf{F}}^0, \tilde{\mathbf{F}}^1, ..., \tilde{\mathbf{F}}^{M-1}]$,
list of pose feat. sequence $\mathcal{P} = [\tilde{\mathbf{P}}^0, \tilde{\mathbf{P}}^1, ..., \tilde{\mathbf{P}}^{M-1}]$,
list of gesture sequence $\mathcal{G} = [\tilde{\mathbf{G}}^0, \tilde{\mathbf{G}}^1, ..., \tilde{\mathbf{G}}^{M-1}]$,
$\tilde{\mathbf{F}} = [\tilde{\mathbf{f}}_0, \tilde{\mathbf{f}}_1, ..., \tilde{\mathbf{f}}_{T_{match}-1}]$, $\tilde{\mathbf{P}} = [\tilde{\mathbf{p}}_0, \tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_{T_{match}-1}]$,
$\tilde{\mathbf{G}} = [\tilde{\mathbf{g}}_0, \tilde{\mathbf{g}}_1, ..., \tilde{\mathbf{g}}_{T_{match}-1}]$,
$\tilde{\mathbf{f}} \in \mathbb{R}^{112}$, $\tilde{\mathbf{p}} \in \mathbb{R}^{96}$, $\tilde{\mathbf{g}} \in \mathbb{R}^{165}$
Input : $k \in \mathbb{Z}$, the desired k-best neighbors,
audio feat. sequence $\mathbf{F} = [\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{f}_{T-1}]$,
control $\mathbf{C} = [\mathbf{c}_0, \mathbf{c}_1, ..., \mathbf{c}_{T-1}]$,
initial pose feat. $\mathbf{p}_{(-1)}$,
$\mathbf{f} \in \mathbb{R}^{112}$, $\mathbf{p} \in \mathbb{R}^{96}$, $\mathbf{c} \in \{0, 1\}$
Output: gesture sequence $\mathbf{G} = [\mathbf{g}_0, \mathbf{g}_1, ..., \mathbf{g}_{T-1}]$,
$\mathbf{g} \in \mathbb{R}^{165}$
t = 0;
initialize $\mathbf{G} = []$, $\mathbf{P} = [\mathbf{p}_{(-1)}]$;
while $t < T$ **do**
  $\hat{\mathbf{P}} = []$, $\hat{\mathbf{F}} = []$, $\hat{\mathbf{G}} = []$;
  for $m \leftarrow 0$ **to** $M - 1$ **do**
    $r = 0$;
    $pdist = \infty$;
    for $s \leftarrow 1$ **to** $T_{match} - 1$ **do**
      if $\mathbf{c}_s == 1$ **then**
        if $d(\hat{\mathbf{p}}_s^m \in \tilde{\mathbf{P}}^m, \mathbf{p}_{t-1}) < pdist$ **then**
          $pdist = d(\hat{\mathbf{p}}_s^m \in \tilde{\mathbf{P}}^m, \mathbf{p}_{t-1})$;
          $r = s$;
        end
      end
    end
    $append(\hat{\mathbf{P}}, \hat{\mathbf{p}}^m_{r:(r+N-1)})$;
    $append(\hat{\mathbf{F}}, \hat{\mathbf{f}}^m_{r:(r+N-1)})$;
    $append(\hat{\mathbf{G}}, \hat{\mathbf{g}}^m_{r:(r+N-1)})$;
  end
  $\hat{\mathbf{P}} = \{\hat{\mathbf{p}}^0_{0:(N-1)}, \hat{\mathbf{p}}^1_{0:(N-1)}, ..., \hat{\mathbf{p}}^{M-1}_{0:(N-1)}\}$;
  $\hat{\mathbf{F}} = \{\hat{\mathbf{f}}^0_{0:(N-1)}, \hat{\mathbf{f}}^1_{0:(N-1)}, ..., \hat{\mathbf{f}}^{M-1}_{0:(N-1)}\}$;
  $\hat{\mathbf{G}} = \{\hat{\mathbf{g}}^0_{0:(N-1)}, \hat{\mathbf{g}}^1_{0:(N-1)}, ..., \hat{\mathbf{g}}^{M-1}_{0:(N-1)}\}$;
  $R_{audio} = relrank[d(\hat{\mathbf{f}}^0_0, \mathbf{f}_t), d(\hat{\mathbf{f}}^1_0, \mathbf{f}_t), ...]$;
  $R_{pose} = relrank[d(\hat{\mathbf{p}}^0_0, \mathbf{p}_{t-1}), d(\hat{\mathbf{p}}^1_0, \mathbf{p}_{t-1}), ...]$;
  $R_{combined} = R_{audio} + R_{pose}$ (elem. wise);
  sort $R_{combined}$, sort its indices into $I_{combined}$;
  $i = I_{combined}[k]$;
  $append(\mathbf{G}, \hat{\mathbf{g}}^i_{(0:(N-1))})$;
  $append(\mathbf{P}, \hat{\mathbf{p}}^i_{(0:(N-1))})$;
  $t = t + N$;
end

Table 5.1: Summary of the search database for the "Oliver" sequences. The data is recorded from an "in-the-wild" setting, and thus contain various types of speech gestures unseen in other studio-captured datasets.

Total duration	11.4 hours
Total unique videos	105 videos
Total unique clips	9624 clips
Duration per clip	64 frames @ 15 fps

than $\mathbf{r}$. Formally, this controlled synthesis can be performed by using a binary control mask matrix $\mathbf{c} \in \{0,1\}^{M \times T_{match}}$ which is constructed by checking if the gesture at a particular frame t is eligible according to the control signal or not. This allows us to limit the search space to the desired data effectively. In practice, only the first and the last frame of each search window ($N = 8$ frames long) are considered to allow a broader range of possible options. Unconstrained gesture synthesis can be seen as a special case where the value of $\mathbf{c}$ is always one at every frame. Compared to the controlled synthesis performed by MoGlow (Alexanderson et al., 2020b), the proposed control design is more flexible as it can be used to mix different control criteria at either the same or different frames seamlessly without requiring hours of re-training the neural network for each given criteria. A range of masks C can easily be calculated for various control features of interest.

5.3.3 *Gesture Resynchronization using cGAN*

The proposed experiments suggest that the 3D gesture produced in the first stage appears natural and in-sync with the audio. However, since the similarity metric of the k-NN serves as an approximation to the real audio-gesture correspondence, its predicted frames may not always lead to the most optimal solution. Furthermore, the use of window-based search at a regular interval may also limit the ability of the algorithm to consider longer correlations. To address these issues, the synthesis quality is enhanced by passing the output of the first stage into a learned conditional Generative Adversarial Network (cGAN). Adversarial-based generative models are known for their ability to produce high quality synthesis that closely matches the real data distribution, especially if they are also guided by a conditional input signal (Isola et al., 2017; Mirza and Osindero, 2014). To this end,

the proposed approach uses a generator network G which transforms an audio-gesture pair $(\bar{\mathbf{F}}, \mathbf{G}_{kNN})$ generated by the k-NN into another pair $(\bar{\mathbf{F}}, \mathbf{G}_{syn})$ which has similar characteristics to the real audio-gesture pairs $\{(\bar{\mathbf{F}}^m, \mathbf{G}^m_{real})\}_{m=0}^{M-1}$ in the training data. Every feature $\bar{\mathbf{f}}_t \in \mathbb{R}^{28}$ in a sequence $\bar{\mathbf{F}}$ is denoted as the concatenation of the MFCC feature $\mathbf{m}_t \in \mathbb{R}^{14}$ with its first derivative. A separate discriminator network D is trained to classify between the real audio-gesture sequence pairs from the real 3D gesture distribution and the fake audio-gesture pairs generated by the k-NN. Both networks are trained in an alternating fashion to compete with each other. Once the training converges, the generator is expected to produce more realistic 3D body and hand gestures given the conditioning 3D gesture input from the k-NN. As the task of the generator is to update the initial 3D gesture produced by the k-NN, experimental results suggest that using parent-relative representation for $\mathbf{G}_{real}$, $\mathbf{G}_{kNN}$, and $\mathbf{G}_{sync}$ leads to a more stable result. The Wasserstein GAN loss formulation (Arjovsky et al., 2017) with gradient penalty (Gulrajani et al., 2017) is used as the training objective for the model:

$$\mathcal{L}_{Adv}(G, D) = \mathbb{E}_{\bar{\mathbf{F}}, \mathbf{G}_{real}}\big[D(\bar{\mathbf{F}}, \mathbf{G}_{real})\big] - \mathbb{E}_{\bar{\mathbf{F}}, \mathbf{G}_{syn}}\big[D(\bar{\mathbf{F}}, \mathbf{G}_{syn})\big], \tag{5.1}$$

where $\mathbf{G}_{syn} = G(\bar{\mathbf{F}}, \mathbf{G}_{kNN})$. Furthermore, the gradient penalty is defined as follows:

$$\mathcal{L}_{GP}(G, D) = \mathbb{E}_{\mathbf{G}_{syn}}\big[(\parallel \nabla_{\mathbf{G}_{syn}} D(\mathbf{G}_{syn}) \parallel -1)^2\big]. \tag{5.2}$$

To ensure that the output of the generator can be guided by the 3D gesture produced by the k-NN, a reconstruction loss $\mathcal{L}_{Rec} = \mathcal{L}_1(\mathbf{G}_{kNN}, \mathbf{G}_{syn})$ is used to encourage gesture similarity between the k-NN and generator output.

$$\mathcal{L} = w_1 \cdot \mathcal{L}_{Rec} + w_2 \cdot \mathcal{L}_{Adv}(G, D) + w_3 \cdot \mathcal{L}_{GP}(G, D). \tag{5.3}$$

The architecture of the proposed network closely follows the architecture introduced in Chapter 4 with several minor tweaks. The generator takes as input MFCC features $\bar{\mathbf{F}} \in \mathbb{R}^{B \times C_m \times T_{match}}$ and the 3D gesture $\mathbf{G}_{kNN} \in \mathbb{R}^{B \times C_g \times T_{match}}$, where B is the batch size, while C_m and C_g are respectively the size of the audio and gesture features. For the generator G, the encoder of the network is comprised of 8 blocks of 1D convolution, 1D batch normalization (BN) (Ioffe and Szegedy,

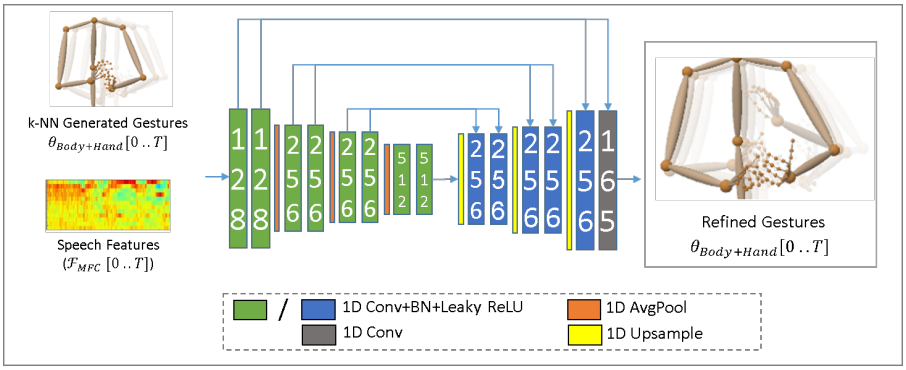
2015, and ReLU activation functions Nair and Hinton, 2010). A Max Pooling layer is used after every second block with the exception of the last. The decoder consists of 8 blocks mirroring the encoder, each of which contains [1D conv, 1D BN, ReLU] layers except for the last one, which uses just a single 1D convolution to produce the final resynchronized gesture $\mathbf{G}_{syn}$. The decoder blocks are interleaved with an upsampling layer after every second block. On the other hand, the discriminator takes as input the MFCC features $\bar{\mathbf{F}} \in \mathbb{R}^{B \times C_m \times T_{match}}$ and either the real $\mathbf{G}_{real} \in \mathbb{R}^{B \times C_g \times T_{match}}$ or generated $\mathbf{G}_{kNN} \in \mathbb{R}^{B \times C_g \times T_{match}}$ gesture sequence (see Equation 1). The discriminator D consists of 6 blocks of 1D convolution, 1D instance normalization, and a leaky ReLU activation function with an Average Pooling layer after every second block, followed by a fully-connected layer at the end to produce a scalar value that rates the similarity of the input with respect to the real audio-gesture distribution.

Figure 5.4 shows a detailed representation of the cGAN-based resynchronization network architecture. We use a similar architecture to Habibie et al. (2021b) with a modification to accommodate the input motion from the k-NN. The input channel of the first layer of the generator consists of 193 parameters (165 parameters for body+hand and 28 parameters for audio) instead of only 28 parameters. Since it does not predict facial expressions, the cGAN uses only one decoder which produces 165 parameters as output. The incorporation of the additional motion matched input has not been previously explored.

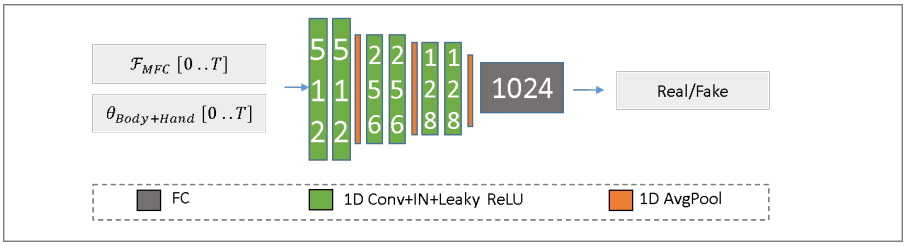
A standard WGAN-GP formulation is employed to train the method. To this end, the last sigmoid layer is removed from the discriminator. The generator is updated after every 5 iterations to ensure that the average of the combined real and fake critic training curve fluctuates around 0.

5.3.4 *Training Details*

The method is trained using the 3D annotated speech-to-gesture data introduced in Chapter 4. A new dataset is prepared to train the cGAN resynchronization network, which contains audio-gesture sequences with a strided overlap of 5 frames between each consecutive sample. Since the adversarial loss compares the real 3D gesture sequence $\mathbf{G}_{real}$ with the "fake" or k-NN-generated 3D gesture sequence $\mathbf{G}_{kNN}$, the



(a) Generator architecture.



(b) Discriminator architecture.

Figure 5.4: The proposed network for the cGAN gesture resynchronization. The generator takes as input the MFCC audio feature and the 3D gesture generated by the k-NN and produces a refined 3D gesture. The numbers in the blocks represent the number of feature channels output by the block. Since Wasserstein GAN formulation with Gradient Penalty is employed, the last layer of the discriminator or critic network does not include a sigmoid activation.

model also needs to generate the fake gesture "ground truth" $\mathbf{G}_{kNN}$. To achieve this, the k-NN algorithm is applied over the training sequences to generate $\mathbf{G}_{kNN}$ by using the training audio features as input. To ensure that the network can handle different gesture characteristics from different k-nearest neighbors, 50% of the data is sampled from $k = 1$ while the rest 50% is sampled uniformly from $k = 2$ to $k = 15$. In all experiments, the initial pose feature for the k-NN is generated by randomly sampling a feature frame from the database. The network is trained over 300,000 iterations using Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1e-4$. The hyperparameters w_1, w_2 and w_3 are set to be $0.1, 1$, and 100, respectively.

5.4 RESULTS

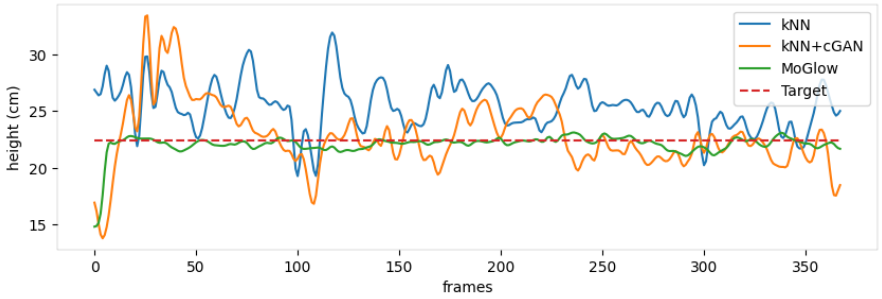
To verify the feasibility of the proposed approach, this section discusses the evaluation of the proposed method's performance with various

control signals. It also examines the versatility of the proposed design in achieving various types of control without the need for re-training and shows how the method can be extended to perform semantically meaningful gesture synthesis. Finally, in the absence of a control signal, the experiments also show that the proposed method achieves better performance than the prior state-of-the-art approach (Habibie et al., 2021b).

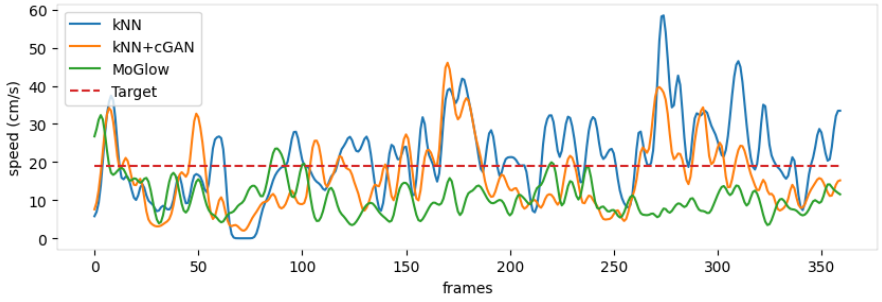
Since multiple motions may be correct for a given audio sequence, there are no well-established metrics for assessing performance. Hence, the proposed evaluation resort to user studies for performance evaluation which is the most standard evaluation protocol for many gesture synthesis tasks (Alexanderson et al., 2020b; Habibie et al., 2021b; Kucherenko et al., 2021; Yoon et al., 2019).

Because gesture style is known to be speaker-specific, most prior works train and test their models on the same speaker. To this end, a single speaker (John Oliver) of the 3D annotated version (Habibie et al., 2021b) of the in-the-wild Berkeley speech-gesture dataset (Ginosar et al., 2019b) is used to train and test the examined methods. In this case, using a single subject also makes it easier for the participants to recognize their speaking style.

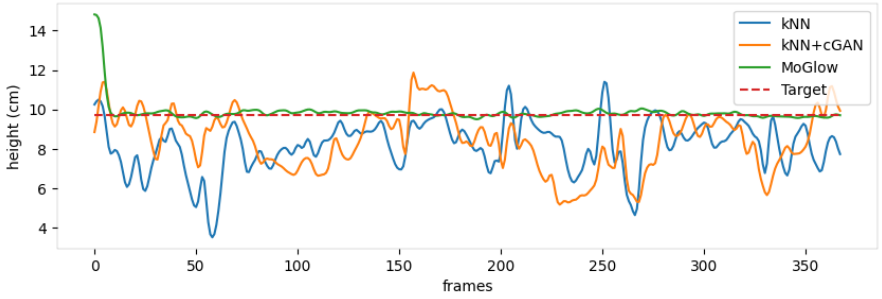
All user study participants were recruited from Amazon Mechanical Turk. Before the study, each user was shown two real video examples of the speaker along with the 3D face, body, and hand tracking results. The users were asked to ignore the synthesis of the facial expression, which always use 3D tracked ground truth keypoints. During each study, the 3D rendered gesture videos, along with their audio, were shown one-by-one to the user. The video playback control was disabled once the user clicked the play button, and the user was not able to proceed until the playback had been completed. At the end of every video, each user was asked to rate the quality of the gesture synthesis using a seven-point scale, ranging from 1 (lowest) to 7 (highest). The users were asked to rate each video based on two prompts: *1)* Does the clip appear natural and the gesture follows the speaking style of the speaker?, and *2)* Are the gesture and the audio well synchronized? Multiple preliminary tests were conducted to ensure that the objective of the study is well understood by the participants based on their feedback. All comparison videos are 24 seconds long and are uniquely and randomly sampled for each user from the original test dataset of Ginosar et al. (2019b).



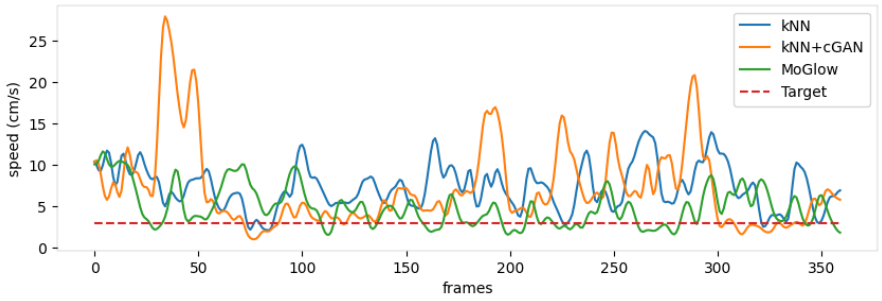
(a) Wrist position using “high hand" control



(b) Wrist position using “fast speed" control



(c) Wrist position using “low hand" control



(d) Wrist position using “slow speed" control

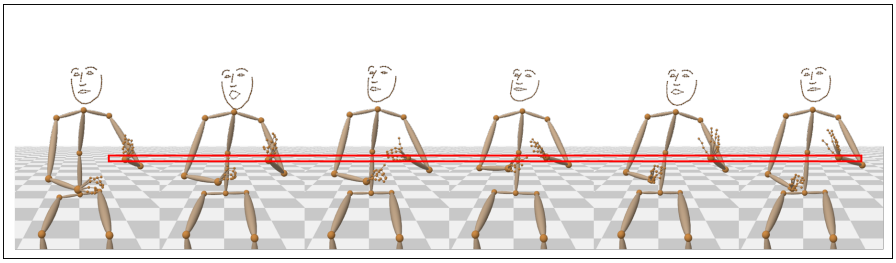
Figure 5.5: Control-based comparison of hight hand height (a), high hand velocity (b), low hand height (c), and low hand velocity (d) between k-NN (proposed, blue), k-NN+cGAN (proposed, orange), and MoGlow (Alexanderson et al., 2020b, green) over a test sequence. The larger variation produced by the proposed methods lead to more natural motion variations, unlike MoGlow, which could lead to a temporally static gesture w.r.t. the control signal.

Table 5.2: A user study for evaluating various control-based synthesis techniques. The proposed approach was consistently rated as more natural and more in-sync than MoGlow (Alexanderson et al., 2020b).

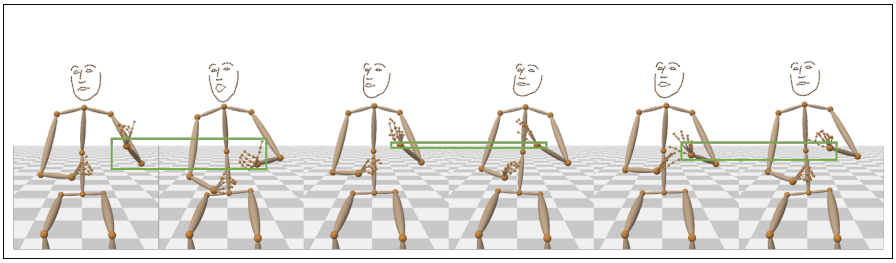
Method	**Naturalness** ↑	**Synchrony.** ↑
GT	5.95 ± 1.06	6.00 ± 1.17
Proposed Height	$\mathbf{5.25 \pm 1.26}$	$\mathbf{5.10 \pm 1.53}$
MoGlow Height	4.79 ± 1.45	4.71 ± 1.65
Proposed Speed	$\mathbf{5.33 \pm 1.36}$	$\mathbf{5.25 \pm 1.55}$
MoGlow Speed	5.20 ± 1.35	5.21 ± 1.36
Proposed Symmetry	$\mathbf{5.21 \pm 1.16}$	$\mathbf{5.33 \pm 1.12}$
MoGlow Symmetry	4.77 ± 1.58	4.70 ± 1.62

5.4.1 *Evaluation of High-level Gesture Control*

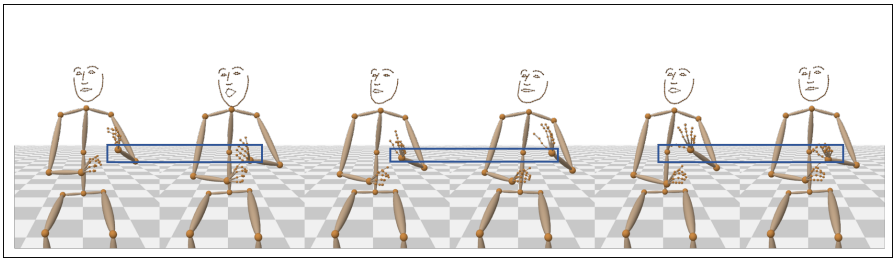
SUBJECTIVE EVALUATION The performance of the proposed k-NN+cGAN method is evaluated and compared against the state-of-the-art, audio-driven, control-based gesture synthesis approach of MoGlow (Alexanderson et al., 2020b). Here, three different control signals are examined: left wrist height, left wrist speed, and wrist height symmetry. To this end, the MoGlow model is re-trained on the 3D annotated version (Habibie et al., 2021b) of the Berkeley speech-to-gesture data (Ginosar et al., 2019b) based on their publicly available implementation. Gaussian smoothing is applied to the training data to ease the training process and use the best qualitative result after conducting a grid search over around 30 different parameter combinations. For each control category, two different results were synthesized, one on the higher (e.g., “high left wrist position") and one on the lower (e.g., “low left wrist speed") end of each control signal value, defined by the 85th and 15th percentile of the training data. For the proposed method, this effectively limits its search space to only 15% of the total training data. The evaluation also includes the ground truth as the topline comparison. Each user was shown one video from each control level. The audio track is randomly sampled from one of six possible test sequences. The user study involved 42 respondents. Table 5.2 summarizes the result of the study. The proposed k-NN+cGAN is consistently better at producing natural-looking and in-sync results compared to MoGlow. This result also suggests that the proposed search-based approach can produce a plausible synthesis even with a smaller search space created by conditioning.



(a) Moglow



(b) Proposed k-NN+cGAN



(c) Proposed k-NN+cGAN

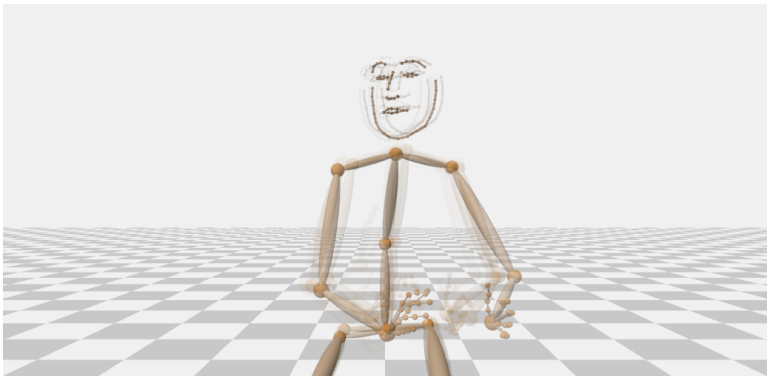
Figure 5.6: Frame-aligned synthesis comparison between MoGlow (a), the proposed k-NN (b), and the proposed k-NN+cGAN (c) when conditioned using "high left hand" control on the same speech input. While MoGlow generates motion that satisfies the given hand height value, it fails to produce natural-looking gestures due to the constant height of the generated hand. In contrast, the proposed method can satisfy the control signal while at the same time producing realistic gesture variation.

QUANTITATIVE EVALUATION Here, the performance of each method is quantitatively analyzed when subjected to a particular control signal. It should be noted that the proposed method and MoGlow use the control signals differently to produce the desired outcome. The proposed k-NN-based algorithm achieves gesture control by using the control value as a threshold to limit the search space, while MoGlow directly uses the control value as a regression target to modify a specific outcome of the gesture. Because of this, the control signal is treated differently for each method. To allow a higher gesture variation, the control is only enforced on the first and last frame of the gesture candidate. While this allows the output to vary outside the specified threshold, this ensures the average value of the controlled variable will

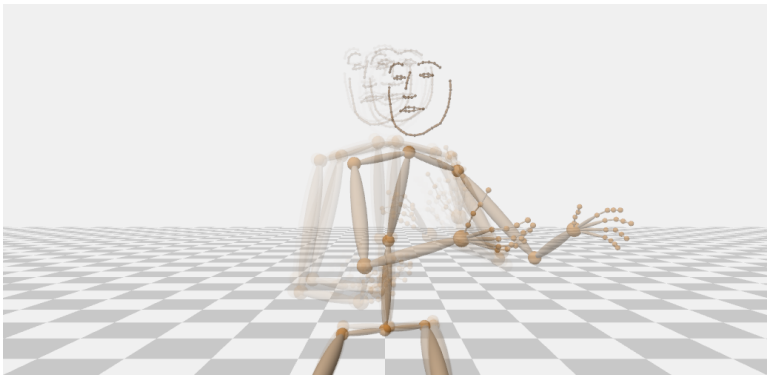
be close to the desired (threshold) value while at the same time ensuring greater motion variability. In contrast, MoGlow uses the control input directly as a target value that needs to be satisfied in the output space. Therefore, their output gesture often produces less variation over the motion space, although it generally stays closer to the intended control signal. For example, if the left hand is conditioned to be at a certain height, it is no longer able to produce body gestures with varying hand height. This, in many cases, makes the hand appear "stuck" at the given height. Correcting this would require significant manual labor. In contrast, the controlled results produced by the proposed method appear more realistic since combining real motion sequences from the database will likely induce natural modulation. Table 5.3 compares the predicted value with respect to the control signal of each method, and Figure 5.5 shows the quantitative behavior of each method when conditioned by the given control signal. The qualitative comparison shown in Figure 5.6 and Figure 5.7 demonstrates the efficacy of the proposed method to follow the provided control input.

5.4.2 *Synthesis with Complex and Low-level Control*

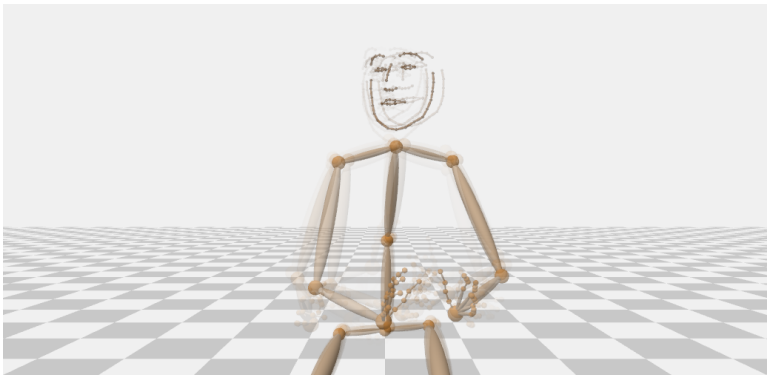
Unlike MoGlow (Alexanderson et al., 2020b), the proposed gesture control can be achieved without model re-training. Hence, various control signals can be given during test time at any particular frame window, enabling the user to perform far more complex motion control. This is particularly useful when generating gestures that reflect the emotional state of the speaker. For example, if the speech of the speaker reflects an emotional change from sad to angry, we may want to synthesize gestures with slow and low hand positions at the beginning, and progress towards fast and extended hand form at the end of the speech. Such a synthesis scenario can be achieved by the proposed framework in one pass without requiring any re-training. On the other hand, pure learning-based controlled synthesis methods will fail in such tasks since producing gestures with different control signals (e.g. "speed" vs. "height" control) requires other models with different training sets. In addition to the high-level control synthesis described above, the proposed formulation can also be extended to follow time-specific signals, including signals with semantically meaningful information. As an example, it can be experimentally shown



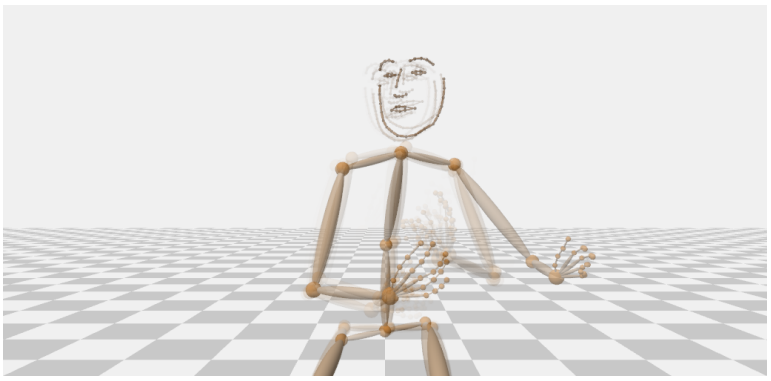
(a) Low left hand control result of k-NN



(b) High left hand control result of k-NN



(c) Low left hand control result of k-NN + cGAN



(d) High left hand control result of k-NN + cGAN

Figure 5.7: Qualitative comparison of the controlled synthesis of k-NN with low left hand signal (a), k-NN with high hand signal (b), k-NN + cGAN with low hand signal (c), and k-NN + cGAN with high hand signal (d) over a test sequence. The proposed k-NN and k-NN+cGAN produce a wide gesture variation even when constrained by the control signals.

Table 5.3: Quantitative comparison of control-based synthesis for left wrist height, speed, and symmetry. The proposed approach generates more natural looking gestures with larger motion variations. MoGlow, however, produces gestures with less variation which can be "stuck" at a given control signal, such as height, rendering unnatural-looking results.

Method	**Threshold/ Target**	**Mean**	**Deviation/ Variation ↑**
k-NN (wrist high)	22.2 cm	25.6 cm	3.4 cm
k-NN+cGAN (wrist high)	22.2 cm	23.4 cm	**4.2 cm**
MoGlow (wrist high)	22.2 cm	22.2 cm	1.1 cm
k-NN (wrist low)	9.7 cm	8.5 cm	2.1 cm
k-NN+cGAN (wrist low)	9.7 cm	9.1 cm	**2.9 cm**
MoGlow (wrist low)	9.7 cm	9.9 cm	0.6 cm
k-NN (wrist fast)	19.1 cm/s	22.8 cm/s	**12.0 cm/s**
k-NN+cGAN (wrist fast)	19.1 cm/s	17.5 cm/s	10.3 cm/s
MoGlow (wrist fast)	19.1 cm/s	11.6 cm/s	5.8 cm/s
k-NN (wrist slow)	3 cm/s	5.9 cm/s	3.8 cm/s
k-NN+cGAN (wrist slow)	3 cm/s	5.4 cm/s	**4.1 cm/s**
MoGlow (wrist slow)	3 cm/s	5.5 cm/s	3.0 cm/s
k-NN (asymm.)	10 cm	12.2 cm	3.0 cm
k-NN+cGAN (asymm.)	10 cm	10.4 cm	**3.9 cm**
MoGlow (asymm.)	10 cm	9.7 cm	1.6 cm
k-NN (symmetric)	0 cm	1.0 cm	1.2 cm
k-NN+cGAN (symmetric)	0 cm	2.1 cm	**2.0 cm**
MoGlow (symmetric)	0 cm	0.7 cm	0.5 cm

that the proposed framework can be used to produce a specific body gesture whenever a specific keyword is detected in the speech. The keywords can be inferred from speech by applying an off-the-shelf speech-to-text system to the input audio. When such a keyword is detected, instead of loading a gesture from the standard database, the gesture is selected from a separate database containing gestures that are semantically correlated with the keyword.

5.4.3 *Synthesis Evaluation without Control Signals*

Here, the result of the approach without the presence of any control signals is evaluated by comparing the performance of both of the proposed components against the ground truth and four different baselines. This includes two different versions of the proposed k-NN:

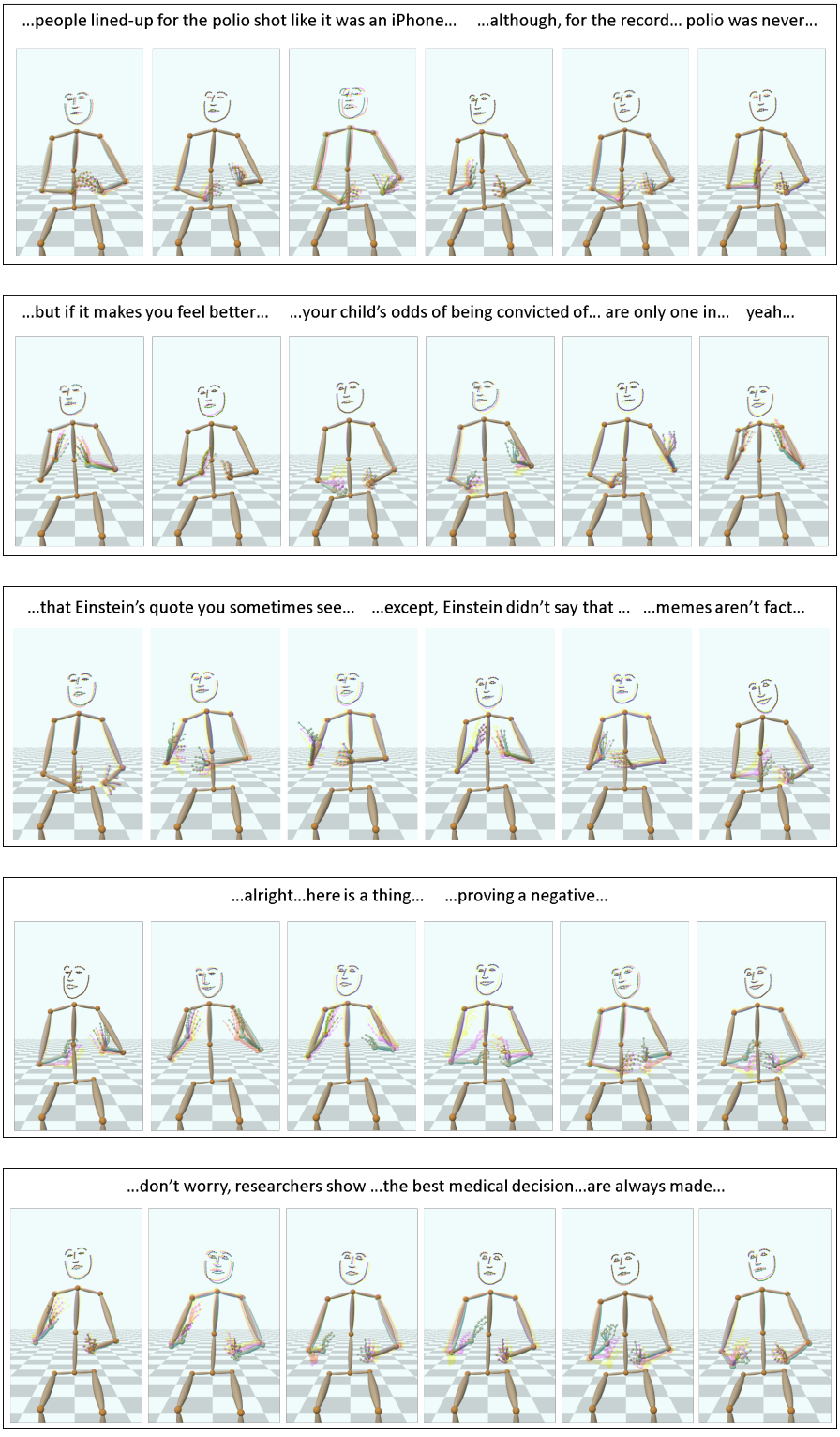


Figure 5.8: Qualitative results of the unconditional gesture synthesis using the proposed k-NN+cGAN approach. Even though it is mainly designed for controllable synthesis, the proposed method achieves competitive synthesis quality against the state-of-the-art for gesture generation without control signals. Motion visualization is based on the Harris shutter effect.

Table 5.4: User study results assessing the performance between synthesis methods in the absence of control signals. The proposed k-NN + cGAN outperforms other baselines both in terms of naturalness and synchronization.

Method	**Naturalness** ↑	**Synchrony.** ↑
Ground Truth	$\mathbf{6.26 \pm 1.02}$	$\mathbf{5.99 \pm 1.02}$
Mismatched audio-gesture	-	5.48 ± 1.34
Habibie et al. (2021b)	5.79 ± 1.16	5.66 ± 1.14
MoGlow (Alexanderson et al., 2020b)	4.84 ± 1.79	4.83 ± 1.65
k-NN pose-only similarity	4.57 ± 2.11	4.82 ± 1.93
Proposed kNN (Audio pre-selected)	5.73 ± 1.13	5.50 ± 1.21
Proposed kNN (Pose pre-selected)	5.55 ± 1.38	5.33 ± 1.34
Proposed kNN+cGAN	$\mathbf{5.83 \pm 1.26}$	$\mathbf{5.82 \pm 1.13}$

one where the candidates are first selected based on pose similarity (pose pre-selected k-NN), and another one where the pre-selection is performed based on audio similarity (audio pre-selected k-NN). The kNN-cGAN, which uses the pose pre-selected k-NN result as input, is also included in this evaluation.

The first baseline is a simple k-NN that only predicts the next gesture based on the 3D pose similarity at every frame T without considering audio similarity. Next, the proposed method is compared against randomly paired audio-gesture sequences. This baseline has been reported to perform strongly in previous studies (Kucherenko et al., 2021). Another baseline included in the evaluation is the recent GAN-based gesture regression approach by Habibie et al. (2021b). Finally, the proposed method is compared against Alexanderson et al. (2020b).

The evaluation is conducted through a user study involving 41 different respondents using the exact instructions discussed at the beginning of Sec. 5.4. During the survey, each respondent was shown 16 different synthesis videos from two different random audio tracks sampled from a total of 15 possible tracks.

The result of the study is shown in Table 5.4. The gesture refinement results produced by the proposed k-NN+cGAN achieved the highest score in terms of synchronization and naturalness, including the prior state-of-the-art method of Habibie et al. (2021b). Moreover,

unlike their approach, which directly predicts the gesture from the audio input, the proposed method can follow various control signals and generate different gesture sequences given the same audio. The proposed method also performed better than the control-aware 3D gesture synthesis approach of Alexanderson et al. (2020b) in terms of naturalness and synchronization. Overall, the results obtained by the proposed method show that it consistently outperforms the state-of-the-art at synthesizing both control-based as well as unconstrained gestures. Figure 5.8 shows several gesture synthesis examples of the proposed method under this unconditional setting.

5.5 CONCLUSION

This chapter presents an improvement over the gesture synthesis method introduced in Chapter 4. The presented approach allows for controllable speech-driven body gesture synthesis, which cannot be achieved using a direct speech-to-gesture mapping using only a neural network. To this end, the new approach utilizes database search together with adversarial learning to produce natural and synchronized gestures. Compared to prior work, it offers more diverse manipulation and does not require re-training for every control signal. Results show that the proposed approach outperforms the state-of-the-art both in terms of naturalness and audio-synchronicity, even in the absence of control.

Despite its benefits, the proposed approach also has several limitations. Currently, hand-designed criteria are used for extracting and estimating feature similarity. Hence, future work can investigate using a learning-based approach for extracting and measuring such feature similarity, akin to the work of Chung and Zisserman (2016) for the audio and lip sync alignment task. The proposed cGAN is currently not conditioned on the control signal, which may lead the result to deviate from the intended outcome, even though the experiments suggest that the deviation is tolerable. Another limitation of the proposed search-based algorithm is the potentially expensive computation time compared to single-pass inference approaches of the purely learning-based counterparts. Since the method searches through the whole database to find the closest candidate for every frame window, the time complexity grows quadratically with the number of sequences

in the database. A potential extension to remedy this issue is to train both stages of the method (k-NN and cGAN) in an end-to-end manner like the work of Holden et al. (2020).

6

CONCLUSION

This thesis explored novel ways to ease the creation of 3D human animation, specifically in the space of motion capture, synthesis, and control from multi-modal input signals. The common thread of the presented approaches is that they all leverage predictive algorithms that learn from a large amount of data. One of the main challenges of employing learning-based methods in this domain is the need for more accurate 3D annotations of human motion. This thesis incorporated several novel approaches to overcome this problem by introducing new techniques that enable motion capture and synthesis by using partially annotated and noisy data.

Chapter 3 introduces a novel approach to improve the accuracy of 3D human body pose estimation from in-the-wild monocular images. It is achieved by introducing a novel neural network architecture, as well as a combination of loss functions that can effectively use 2D pose labels, which are easier to obtain than studio-captured 3D pose annotations. The abundance of 2D labels from in-the-wild images provides an order of magnitude higher 3D pose variations, foreground appearance, background, and occlusions than one can achieve from a studio capture recording.

Based on the promising monocular estimation results, this thesis makes use of state-of-the-art monocular 3D face reconstruction, as well as 3D body and pose estimation approaches to collect a large dataset that can be used to train downstream motion synthesis tasks. In addition to the 3D hand pose estimator of Zhou et al. (2020) and the 3D face tracker of Garrido et al. (2015), the proposed solution uses the XNect (Mehta et al., 2020) pose estimation approach to capture the 3D body motion, as XNect is designed to be robust to partial occlusion scenarios. This thesis explores two novel algorithms for generating 3D body gestures from speech input, which is a challenging task due to the inherent ambiguity between the input and output space. To resolve the ambiguity issue, Chapter 4 of this thesis introduces an adversarial neural network to ensure the naturalness of the motion prediction

with respect to the speech input, thus preventing the generation from collapsing into a dampened motion.

In Chapter 5, this adversarial-based formulation is then extended to allow for a more controllable gesture synthesis that can be modulated from various control signals, such as hand height and velocity. A novel motion matching-inspired algorithm is introduced to enable this feature, allowing for a direct and efficient database search based on speech and previous pose similarity. Furthermore, the generated gesture is further synchronized with the input audio using a neural network.

6.1 INSIGHTS

6.1.1 *The Effectiveness of Automatically Annotated Data from 3D Monocular Estimation*

This thesis proposes the first large-scale publicly available dataset for the task of 3D gesture synthesis from speech that contains 3D dense face reconstruction, as well as 3D body and hand pose annotations. It consists of more than 33 hours of six different subjects recorded in the in-the-wild setting, which is unreasonable to obtain from a controlled studio setup. By design, this dataset can be easily extended on a virtually endless amount of human motion videos.

Two separate studies presented in this thesis demonstrated the usage of monocular 3D reconstructions, especially for 3D gesture synthesis, which often involves subtle gestures. Despite the ill-posed nature of the setup and various challenging occlusion scenarios, the approach produces high-quality monocular 3D pose estimation outputs that appear consistent throughout the dataset. As a result, such monocular approach enables hours of motion capture from in-the-wild that would be tedious to perform in a studio setting. As shown in Chapter 4 and Chapter 5 of this thesis, the annotation produced by 3D monocular estimation approaches can be used for the task of gesture synthesis from speech by leveraging a deep neural network or a motion matching-based approach that achieves competitive performances.

6.1.2 *Leveraging Incomplete Annotations*

The monocular pose estimation method presented in Chapter 3 employs a form of weak supervision as one of the key components of its training objective. While a weakly supervised approach does not have access to the complete information during training, it significantly benefits from a much larger 2D labeled training data that is significantly easier to annotate. The experiments presented in Chapter 3 show that a combination of weak supervision in the form of 2D pre-training, 2D-to-3D lifting, and 3D-to-2D projection can be used to improve 3D estimation quality by leveraging 2D data.

As shown in the ablation study of Chapter 3, adding each of the proposed terms gradually improves the final pose estimation result of the method. In addition, using multiple weak supervision strategies addresses the generalization performance of the model from different angles. Since 2D monocular pose estimation is a considerably easier task, pre-training our 3D pose estimator neural network on the 2D labeled in-the-wild images allows the model to learn a strong correlation between relevant parts of a human body on a wide range of foreground and background scene appearance variations. To ensure that the model is capturing such correlations, parts of the network's latent features are designed to explicitly capture 2D pose information. In this way, the 2D pose representation can be seamlessly combined another weak supervision strategy in the form of 2D-to-3D lifting. This lifting approach has been shown to achieve a competitive performance against direct 3D pose prediction (Martinez et al., 2017b). To accommodate the potential prediction errors due to the failure to disambiguate multiple plausible 3D mapping from the same 2D pose input, the proposed lifting approach combines both the explicit latent 2D pose representation with additional unconstrained latent space to learn the necessary information. Finally, the proposed 3D-to-2D projection allows the model to receive a weak supervision on its 3D prediction even when the 3D pose annotation is not available.

Considering the competitive performance achieved by combining multiple incomplete annotation through various loss functions and neural network design as shown in Chapter 3, it is highly probable that further performance gains can be accomplished by incorporating other forms weak supervisions.

6.1.3 *Combining Database Search and Deep Learning for Motion Generation*

While the proposed adversarial-based training in Chapter 4 has shown state-of-the-art performance for 3D gesture synthesis, it is not able to reproduce the stochastic and multi-modal nature of the problem where a single speech input of a speaker can be mapped into multiple correct 3D gestures. Furthermore, it is also often helpful to control the style of the gesture to match the speaker's mood or internal state. Unfortunately, style-based conditioning is often not feasible using pure learning-based approaches without the need for re-training for every type of control.

Chapter 5 shows that the aforementioned issues can be resolved by introducing a novel Nearest Neighbor-based synthesis approach inspired by the Motion Matching algorithm (Büttner and Clavet., 2015), which has recently become the method of choice for 3D animation in many AAA video games. Motion generation is performed by querying the most plausible sequence from a database with respect to the control signal using the nearest neighbor algorithm. Due to the search-based nature of the algorithm, adjusting the synthesis according to a particular style can be achieved by limiting the search space according to the control criteria.

However, unlike locomotion data, speech-to-gesture mapping is unique due to the ambiguous nature of the data. This thesis presents two key contributions that enable Motion Matching for the task of gesture synthesis. First, it proposes a new motion matching-based search algorithm that considers both the similarity of the audio and previously generated motion. The second contribution is re-purposing the adversarial loss to enhance the pure matching-based approach by comparing real motion samples with the gestures generated by the adopted Motion Matching algorithm. This step is crucial to improve synthesis quality, as the best matching gesture in the database does not always guarantee optimal synchronization with the speech. This hybrid solution can also be generalized to other motion generation tasks beyond the 3D speech-to-gesture domain, as ambiguity is a common problem in most motion control scenarios.

6.2 FUTURE DIRECTIONS

6.2.1 *Occlusion-robust Integrated Full-body Capture of 3D Face, Body, And Hands*

Chapter 3 of this thesis mainly focuses on a monocular 3D pose estimation approach for human body. However, Chapter 4 of this thesis also introduces a speech-to-gesture dataset that consists of 3D human face, body, and hand annotations predicted from monocular videos found on the internet. So far, each of the body-part annotations is predicted using separate algorithms without any information exchange between them. However, the motion between parts of human bodies are inherently correlated by nature. Integrating their inter-correlation may improve 3D monocular estimation quality, especially for challenging cases involving rarely seen poses or occluded parts. This issue can be addressed by employing a full-body model which explicitly incorporates the correlation between body parts, such as the work of Rong et al. (2021) and Zhou et al. (2020).

As also discussed in Chapter 4, one of the main bottlenecks in using monocular pose estimation approaches on RGB images and videos is the presence of severe body occlusions. While a good 3D pose estimator may be able to produce the most likely natural 3D pose to fill-in the missing body parts, the result may not always be consistent when the method is applied to a video input. To resolve the issue of severe body occlusion, one may resort to the use of optimization strategies that combine the 3D pose prediction with its temporal history as well as its 2D estimate (Mehta et al., 2020, 2017b; Xiang et al., 2019). Monocular models with explicit training on the temporal domain can also significantly improve visual stability to address the temporal noise due to monocular image prediction (Arnab et al., 2019; Dabral et al., 2018; Kanazawa et al., 2019; Kocabas et al., 2020).

6.2.2 *An End-to-end Learning-based 3D Animation Pipeline*

This thesis has shown the promising potential of using purely data-driven approaches to solve several independent problems in the domain of 3D human animation, specifically for the tasks of motion capture, synthesis, and control. Among the most exciting potential

future directions shown by this thesis is the possibility to design a new animation framework that combines the traditionally separate 3D motion capture and 3D motion synthesis systems into a unified pipeline, thus opening new possibilities for creatively producing 3D animation by solely leveraging monocular videos that can either be captured using affordable monocular cameras or collected from the internet.

So far, this thesis has explored the learning-based solutions of motion capture and motion synthesis in separation. However, the approach discussed in Chapter 4 have also shown that it is possible to use a learning-based 3D monocular capture method to collect massive amounts of 3D motion data to train a learning-based 3D motion synthesis model. Extending this pipeline into a unified learning-based 3D animation framework will not only result in a more streamlined process between the captured data and the synthesis algorithms, but also allows animators to combine different motion capture datasets to train a relevant synthesis task. One of the potential benefit is the access to pre-train a particular motion synthesis or control model when the training data of the desired task is scarce. Another benefit of having such framework is the ability to mix-and-match various types of motion control modalities, such as directional, audio, or text inputs.

Furthermore, the availability of such a unified pipeline can open novel ways of performing learning-based animation design. For example, instead of following the traditional approach of using the captured motion data to train a synthesis model, the pipeline can enable the use of synthesis models as a motion prior to achieve a more accurate 3D motion capture system. Rempe et al. (2021) presents a particular manifestation of this idea, in which a learned prior model trained on a large amount of motion data is used to improve the accuracy of 3D monocular pose estimation approaches. Different types of motion priors can be designed based of the size and types of modalities that are available in the framework. The next-generation 3D animation system may be constructed by coupling the aforementioned learning-based synthesis algorithms and motion priors into a generalized framework where the capture and synthesis components can be contiguously trained to reinforce each other. For example, a large amount of standing and walking poses in the motion capture data can be used to enhance the quality of 3D locomotion synthesis, and similarly, a powerful gesture synthesis models may be used to enhance the quality of 3D hand pose

estimation (see also Ng et al. (2021)). The existence of shared modules and datasets will also ease the design of synthesis methods from various control signals, including the recently popular text-based motion synthesis approaches (Ahuja and Morency, 2019; Ghosh et al., 2021; Guo et al., 2022; Tevet et al., 2022). Given the tedious and expensive nature of the traditional animation workflow, such frameworks may further democratize the use of 3D motion generation.

6.2.3 *A Quantitative Metric for Gesture Synthesis*

Due to the multi-modal nature of the speech gesture, one of the biggest challenges in analyzing the 3D gesture quality of speech-driven synthesis models is the lack of agreeable quantitative metrics to assess their performance. In general, a given speech input can have multiple correct gesture solutions. Measuring the alignment between the motion and the prosody signal is not trivial, as speech gestures are inherently stochastic, and a matching beat between the speech pulse and the gesture stroke may not always indicate a correct gesture. Furthermore, the naturalness of the gesture is also hard to quantify, as it is correlated with the speaking style and the internal state of the particular speaker.

While several studies have proposed various intuitive potential strategies to resolve this issue, most of them fail to demonstrate consistent evaluation compared to human-rated assessment. Chapter 4 briefly discusses an idea to quantitatively rate the synthesis quality that shows good consistency when using score gesture misalignment with the audio. Here, the misalignment score is computed by passing an audio and gesture pair into a network that has been trained to classify between pairs of aligned and misaligned audio-gesture data. Unfortunately, it fails to generalize its performance on synthesized data. Another promising solution is through the use of a likeliked-based approach, e.g., Alexanderson et al. (2020a). Probabilistic generative models such as Normalizing Flows (Dinh et al., 2015; Dinh et al., 2016; Germain et al., 2015; Rezende and Mohamed, 2015) rely on optimizing a likelihood value as way to measure of how well their predicted output matches the target probability distribution. This associated likelihood value can be used to guide performance selections between models. However, its consistency when compared to human-based ratings has yet to be

proven. Resolving this issue will undoubtedly ease the effort to design powerful speech-driven gesture synthesis models in the future.

BIBLIOGRAPHY

Ahuja, Chaitanya, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency (2020). "Style Transfer for Co-speech Gesture Animation: A Multi-speaker Conditional-Mixture Approach." In: *ECCV*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm.

Ahuja, Chaitanya and Louis-Philippe Morency (2019). "Language2pose: Natural language grounded pose forecasting." In: *2019 International Conference on 3D Vision (3DV)*. IEEE, pp. 719–728.

Alexanderson, Simon, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow (2020a). "Style-controllable speech-driven gesture synthesis using normalising flows." In: *Computer Graphics Forum*.

– (2020b). "Style-controllable speech-driven gesture synthesis using normalising flows." In: *Computer Graphics Forum*.

Andriluka, Mykhaylo, Leonid Pishchulin, Peter Gehler, and Bernt Schiele (2014). "2D Human Pose Estimation: New Benchmark and State of the Art Analysis." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andriluka, Mykhaylo, Stefan Roth, and Bernt Schiele (2009). "Pictorial structures revisited: People detection and articulated pose estimation." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Arikan, Okan and D. A. Forsyth (2002). "Interactive Motion Generation from Examples." In: *ACM Trans. Graph.*

Aristidou, Andreas, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir (2018). "Inverse Kinematics Techniques in Computer Graphics: A Survey." In: *Comput. Graph. Forum* 37.6, pp. 35–58.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein Generative Adversarial Networks." In: *Proceedings of the 34th International Conference on Machine Learning*.

Arnab, Anurag, Carl Doersch, and Andrew Zisserman (2019). "Exploiting temporal context for 3D human pose estimation in the wild." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3395–3404.

Baak, Andreas, Meinard Mueller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt (2011). "A Data-Driven Approach for Real-Time Full Body Pose Reconstruction from a Depth Camera." In: *IEEE International Conference on Computer Vision (ICCV)*.

Blanz, Volker and Thomas Vetter (1999). "A morphable model for the synthesis of 3D faces." In: *Proc. SIGGRAPH*. ACM Press/Addison-Wesley Publishing Co., pp. 187–194.

Bogo, Federica, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black (2016). "Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image." In: *Computer Vision – ECCV 2016*. Lecture Notes in Computer Science. Springer International Publishing, pp. 561–578.

Brau, Ernesto and Hao Jiang (2016). "3D Human Pose Estimation via Deep Learning from 2D Annotations." In: *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pp. 582–591.

Büttner, Michael and Simon Clavet. (2015). *Motion Matching*. https://www.youtube.com/watch?v=z_wpgHFSWss&.

Cao, Zhe, Tomas Simon, Shih-En Wei, and Yaser Sheikh (2017). "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." In: *CVPR*.

Cassell, Justine (2000). "Embodied Conversational Interface Agents." In: *Commun. ACM*.

Cassell, Justine, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone (1994). "Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents." In: *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*.

Cassell, Justine, Hannes Högni Vilhjálmsson, and Timothy Bickmore (2004). "BEAT: the Behavior Expression Animation Toolkit." In: *Life-Like Characters: Tools, Affective Functions, and Applications*.

Cha, Y., T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J. Frahm, and H. Fuchs (2018). "Towards Fully Mobile 3D Face, Body, and Environment Capture Using Only Head-worn Cameras." In: *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Chen, Ching-Hang and Deva Ramanan (2017). "3D Human Pose Estimation = 2D Pose Estimation + Matching." In: *2017 IEEE Conference*

on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 5759–5767.

Chen, Wenzheng, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen (2016). "Synthesizing Training Images for Boosting Human 3D Pose Estimation." In: *3D Vision (3DV)*.

Chiu, Chung-Cheng and Stacy Marsella (2011). "How to Train Your Avatar: A Data Driven Approach to Gesture Generation." In: *IVA*. Ed. by Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson.

– (2014). "Gesture Generation with Low-dimensional Embeddings." In: *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. AAMAS '14.

Chung, J. S. and A. Zisserman (2016). "Out of time: automated lip sync in the wild." In: *ACCV*.

Cudeiro, Daniel, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black (2019). "Capture, Learning, and Synthesis of 3D Speaking Styles." In: *CVPR*.

Dabral, Rishabh, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain (2018). "Learning 3D Human Pose from Structure and Motion." In: Lecture Notes in Computer Science.

Davis, Steven and Paul Mermelstein (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4, pp. 357–366.

Dinh, L, D Krueger, and Y Bengio (2015). "NICE: non-linear independent components estimation 3rd Int." In: *Conf. on Learning Representations, ICLR (San Diego, CA, USA,) Workshop Track Proc*.

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio (2016). "Density Estimation Using Real NVP." In: *ICLR 2017*.

Egger, Bernhard, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter (2020). "3D Morphable Face Models - Past, Present and Future." In: *ACM Transactions on Graphics (TOG)*.

Elhayek, A., E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt (2015). "Efficient ConvNet-based Marker-less Motion Capture in General Scenes with a Low

Number of Cameras." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Elhayek, Ahmed, Edilson de Aguiar, Arjun Jain, Jonathan Thompson, Leonid Pishchulin, Mykhaylo Andriluka, Christoph Bregler, Bernt Schiele, and Christian Theobalt (2016). "MARCOnI—ConvNet-based MARker-less motion capture in outdoor and indoor scenes." In: *IEEE transactions on pattern analysis and machine intelligence* 39.3, pp. 501–514.

FFmpeg Developers (2016). *FFMPEG. ffmpeg.org*. URL: ffmpeg.org.

Ferstl, Ylva and Rachel McDonnell (2018). "Investigating the Use of Recurrent Motion Modelling for Speech Gesture Generation." In: *IVA*.

Ferstl, Ylva, Michael Neff, and Rachel McDonnell (2019). "Multi-Objective Adversarial Gesture Generation." In: *Motion, Interaction and Games*.

Fragkiadaki, Katerina, Sergey Levine, Panna Felsen, and Jitendra Malik (2015). "Recurrent network models for human dynamics." In: *Proceedings of the IEEE international conference on computer vision*, pp. 4346–4354.

Ganapathi, Varun, Christian Plagemann, Daphne Koller, and Sebastian Thrun (2012). "Real-time human pose tracking from range data." In: *European conference on computer vision*. Springer, pp. 738–751.

Garrido, Pablo, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt (2015). "VDub - Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track." In: *Computer Graphics Forum (Proceedings of EUROGRAPHICS)*.

Garrido, Pablo, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt (2016). "Reconstruction of Personalized 3D Face Rigs from Monocular Video." In: *ACM Transactions on Graphics (TOG)* 35.3, p. 28.

Germain, Mathieu, Karol Gregor, Iain Murray, and Hugo Larochelle (2015). "Made: Masked autoencoder for distribution estimation." In: *International conference on machine learning*. PMLR, pp. 881–889.

Ghosh, Anindita, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek (2021). "Synthesis of compositional animations from textual descriptions." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1396–1406.

Ginosar, S., A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik (2019a). "Learning Individual Styles of Conversational Gesture." In: *CVPR*.

– (2019b). "Learning Individual Styles of Conversational Gesture." In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Girshick, Ross, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew Fitzgibbon (2011). "Efficient regression of general-activity human poses from depth images." In: *2011 International Conference on Computer Vision*.

Goldin-Meadow, Susan (1999). "The role of gesture in communication and thinking." In: *Trends in cognitive sciences*.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems*.

Guan, Peng, Alexander Weiss, Alexandru O. Bãlan, and Michael J. Black (2009). "Estimating human shape and pose from a single image." In: *2009 IEEE 12th International Conference on Computer Vision*.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville (2017). "Improved Training of Wasserstein GANs." In: *Advances in Neural Information Processing Systems*.

Guo, Chuan, Xinxin Zuo, Sen Wang, and Li Cheng (2022). "TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts." In: *ECCV*.

Haag, Kathrin and Hiroshi Shimodaira (2016). "Bidirectional LSTM Networks Employing Stacked Bottleneck Features for Expressive Speech-Driven Head Motion Synthesis." In: *IVA*.

Habibie, Ikhsanul, Mohamed Elgharib, Kripashindu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt (2022). "A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech." In: *SIGGRAPH '22 Conference Proceedings*.

Habibie, Ikhsanul, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura (2017). "A recurrent variational autoencoder for human motion synthesis." In: *28th British Machine Vision Conference*.

Habibie, Ikhsanul, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt (2021a). "Learning Speech-driven 3D Conversational Gestures from Video." In: *ACM International Conference on Intelligent Virtual Agents (IVA)*.

Habibie, Ikhsanul, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt (2021b). "Learning Speech-driven 3D Conversational Gestures from Video." In: *Proceedings of the International Conference on Intelligent Virtual Agents*.

Habibie, Ikhsanul, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt (2019). "In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng (2014). *Deep Speech: Scaling up end-to-end speech recognition*. arXiv: 1412.5567.

Hasegawa, Dai, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi (2018). "Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network." In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Henter, Gustav Eje, Simon Alexanderson, and Jonas Beskow (2020). "MoGlow: Probabilistic and controllable motion synthesis using normalising flows." In: *ACM Transactions on Graphics* 39.4, 236:1–236:14.

Holden, Daniel, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa (2020). "Learned Motion Matching." In: *ACM Trans. Graph.*

Holden, Daniel, Taku Komura, and Jun Saito (2017). "Phase-Functioned Neural Networks for Character Control." In: *ACM Trans. Graph.*

Holden, Daniel, Jun Saito, and Taku Komura (2016). "A Deep Learning Framework for Character Motion Synthesis and Editing." In: *ACM Trans. Graph.*

Holden, Daniel, Jun Saito, Taku Komura, and Thomas Joyce (2015). "Learning Motion Manifolds with Convolutional Autoencoders." In: *SIGGRAPH Asia 2015 Technical Briefs*.

Huang, Yinghao, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll (2018). "Deep Inertial Poser Learning to Reconstruct Human Pose from Sparse Inertial Mea-

surements in Real Time." In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 37.6, 185:1–185:15.

Insafutdinov, Eldar, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele (2016). "DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model." In:

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In: *Proceedings of the 32nd International Conference on Machine Learning*.

Ionescu, Catalin, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu (2014). "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7, pp. 1325–1339.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017). "Image-to-Image Translation with Conditional Adversarial Networks." In: *CVPR*.

Johansson, Gunnar (1973). "Visual perception of biological motion and a model for its analysis." In: *Perception & Psychophysics* 14, pp. 201–211.

Johnson, S. and M. Everingham (2011). "Learning effective human pose estimation from inaccurate annotation." In: *CVPR 2011*.

Johnson, Sam and Mark Everingham (2010). "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation." In: *Proc. BMVC*. doi:10.5244/C.24.12, pp. 12.1–11. ISBN: 1-901725-40-5.

Joo, Hanbyul, Tomas Simon, and Yaser Sheikh (2018). "Total capture: A 3d deformation model for tracking faces, hands, and bodies." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8320–8329.

Kanazawa, Angjoo, Michael J. Black, David W. Jacobs, and Jitendra Malik (2018). "End-to-end Recovery of Human Shape and Pose." In: *Computer Vision and Pattern Recognition (CVPR)*.

Kanazawa, Angjoo, Jason Y Zhang, Panna Felsen, and Jitendra Malik (2019). "Learning 3d human dynamics from video." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5614–5623.

Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen (2017). "Audio-driven Facial Animation by Joint End-to-end Learning of Pose and Emotion." In: *ACM Trans. Graph.*

Kendon, Adam (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization." In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun.

Kingma, Diederik P and Max Welling (2014). "Auto-encoding variational {Bayes}." In: *Int. Conf. on Learning Representations*.

Kocabas, Muhammed, Nikos Athanasiou, and Michael J. Black (2020). "VIBE: Video Inference for Human Body Pose and Shape Estimation." In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*. IEEE.

Kovar, Lucas, Michael Gleicher, and FrÃ©dÃ©ric Pighin (2002). "Motion Graphs." In: *Proceedings of SIGGRAPH '02*. San Antonio, TX.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, pp. 1097–1105.

Kucherenko, Taras, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström (2019). "Analyzing Input and Output Representations for Speech-Driven Gesture Generation." In: *IVA*.

Kucherenko, Taras, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström (2020). "Gesticulator: A Framework for Semantically-Aware Speech-Driven Gesture Generation." In: *Proceedings of the 2020 International Conference on Multimodal Interaction*.

Kucherenko, Taras, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter (2021). "A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA Challenge 2020." In: IUI '21.

Lamere, Paul, Philip Kwok, Evandro Gouvêa, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf (2003). *The CMU SPHINX-4 Speech Recognition System*.

Lee, G., Z. Deng, S. Ma, T. Shiratori, S.S. Srinivasa, and Y. Sheikh (2019a). "Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis." In: *ICCV*.

Lee, Jehee, Jinxiang Chai, Paul S. A. Reitsma, Jessica K. Hodgins, and Nancy S. Pollard (2002). "Interactive Control of Avatars Animated with Human Motion Data." In: *ACM Trans. Graph.*

Lee, Seunghwan, Moonseok Park, Kyoungmin Lee, and Jehee Lee (2019b). "Scalable Muscle-Actuated Human Simulation and Control." In: *ACM Trans. Graph.*

Lee, Yongjoon, Kevin Wampler, Gilbert Bernstein, Jovan Popović, and Zoran Popović (2014). "Motion Fields for Interactive Character Locomotion." In: *ACM Trans. Graph.*

Levine, Sergey, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun (2010). "Gesture controllers." In: *SIGGRAPH '10*.

Levine, Sergey, Christian Theobalt, and Vladlen Koltun (2009). "Real-time Prosody-driven Synthesis of Body Language." In: *ACM Trans. Graph.*

Li, Tianye, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero (2017). "Learning a model of facial shape and expression from 4D scans." In: *ACM Transactions on Graphics*.

Liu, Yilong, Feng Xu, Jinxiang Chai, Xin Tong, Lijuan Wang, and Qiang Huo (2015). "Video-audio Driven Real-time Facial Animation." In: *ACM Trans. Graph.*

Loper, Matthew, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black (Oct. 2015). "SMPL: A Skinned Multi-Person Linear Model." In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6, 248:1–248:16.

Luvizon, Diogo, David Picard, and Hedi Tabia (2018). "2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning." In: *CVPR*. Salt Lake City, USA.

Marcard, Timo von, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll (2018). "Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera." In: *European Conference on Computer Vision (ECCV)*.

Marcard, Timo von, Bodo Rosenhahn, Michael Black, and Gerard Pons-Moll (2017). "Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs." In: *Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*, pp. 349–360.

Mariooryad, S. and C. Busso (2012). "Generating Human-Like Behaviors Using Joint, Speech-Driven Models for Conversational Agents." In: *IEEE Transactions on Audio, Speech, and Language Processing*.

Marsella, Stacy, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro (2013). "Virtual Character Performance from Speech." In: *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. SCA '13.

Martinez, Julieta, Rayat Hossain, Javier Romero, and James J. Little (2017a). "A Simple yet Effective Baseline for 3D Human Pose Estimation." In: *ICCV*.

– (Oct. 2017b). "A simple yet effective baseline for 3d human pose estimation." In: *Proceedings IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ, USA: IEEE.

McNeill, David (2000). *Language and Gesture*. Language Culture and Cognition. Cambridge University Press.

Mehta, Dushyant, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt (2017a). "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision." In: *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE.

Mehta, Dushyant, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt (2020). "XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera." In: *ACM Trans. Graph.*

Mehta, Dushyant, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt (2017b). "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera." In: *ACM Transactions on Graphics* 36.4.

Mermelstein, Paul (1976). "Distance measures for speech recognition, psychological and instrumental." In: *Pattern Recognition and Artificial Intelligence*, pp. 374–388.

Mirza, Mehdi and Simon Osindero (2014). "Conditional Generative Adversarial Nets." In: *CoRR* abs/1411.1784.

Moon, Gyeongsik, Ju Yong Chang, and Kyoung Mu Lee (2018). "V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map." In: *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pp. 5079–5088.

Nair, Vinod and Geoffrey E. Hinton (2010). "Rectified Linear Units Improve Restricted Boltzmann Machines." In: *ICML*.

Neff, Michael, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel (2008). "Gesture Modeling and Animation Based on a Probabilistic Re-creation of Speaker Style." In: *ACM Trans. Graph.*

Newell, Alejandro, Kaiyu Yang, and Jia Deng (2016). "Stacked Hourglass Networks for Human Pose Estimation." In: *ECCV*.

Ng, Evonne, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo (2021). "Body2hands: Learning to infer 3d hands from conversational gesture body dynamics." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11865–11874.

Omran, Mohamed, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele (Sept. 2018). "Neural Body Fitting: Unifying Deep Learning and Model-Based Human Pose and Shape Estimation." In: *3DV*.

Pavlakos, Georgios, Xiaowei Zhou, and Kostas Daniilidis (2018). "Ordinal Depth Supervision for 3D Human Pose Estimation." In: *Computer Vision and Pattern Recognition (CVPR)*.

Pavlakos, Georgios, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis (2017). "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Pavllo, Dario, Christoph Feichtenhofer, David Grangier, and Michael Auli (2019). "3D human pose estimation in video with temporal convolutions and semi-supervised training." In: *CVPR*.

Pham, Hai Xuan, Yuting Wang, and Vladimir Pavlovic (2018). "End-to-end Learning for 3D Facial Animation from Speech." In: *ICMI*.

Pons-Moll, Gerard, David J. Fleet, and Bodo Rosenhahn (2014). "Posebits for Monocular Human Pose Estimation." In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Columbus, Ohio, USA, pp. 2345–2352.

Rao, K Sreenivasa and KE Manjunath (2017). *Speech recognition using articulatory and excitation source features*. Springer.

Rempe, Davis, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas (2021). "HuMoR: 3D Human Motion Model for Robust Pose Estimation." In: *International Conference on Computer Vision (ICCV)*.

Rezende, Danilo and Shakir Mohamed (2015). "Variational inference with normalizing flows." In: *International conference on machine learning*. PMLR, pp. 1530–1538.

Rhodin, Helge, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt (2016). "General automatic human shape and motion capture using volumetric contour cues." In: *European conference on computer vision*. Springer, pp. 509–526.

Rhodin, Helge, Nadia Robertini, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt (2015). "A Versatile Scene Model with Differentiable Visibility Applied to Generative Pose Estimation." In: *Proceedings of the 2015 International Conference on Computer Vision (ICCV 2015)*.

Rhodin, Helge, Mathieu Salzmann, and Pascal Fua (2018a). "Unsupervised Geometry-Aware Representation Learning for 3D Human Pose Estimation." In:

Rhodin, Helge, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua (2018b). "Learning Monocular 3D Human Pose Estimation from Multi-view Images." In: p. 16.

Robitza, Werner (2019). *ffmpeg-normalize. github.com/slhck/ffmpeg-normalize*. URL: https://github.com/slhck/ffmpeg-normalize.

Rong, Yu, Takaaki Shiratori, and Hanbyul Joo (2021). "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1749–1759.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation." In: *International Conference on Medical image computing and computer-assisted intervention*.

Sadoughi, Najmeh, Yang Liu, and Carlos Busso (2014). "Speech-Driven Animation Constrained by Appropriate Discourse Functions." In: *ICMI*.

– (2017). "Meaningful head movements driven by emotional synthetic speech." In: *Speech Communication*.

Safonova, Alla and Jessica K. Hodgins (2007). "Construction and Optimal Search of Interpolated Motion Graphs." In: *ACM Trans. Graph.*

Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen (2016). "Improved Techniques for Training GANs." In: *Advances in Neural Information Processing Systems*.

Saragih, Jason M., Simon Lucey, and Jeffrey F. Cohn (2011). "Deformable Model Fitting by Regularized Landmark Mean-Shift." In: *IJCV*.

Shlizerman, Eli, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman (2018). "Audio to body dynamics." In: *CVPR*.

Shotton, Jamie, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake (2011). "Real-time human pose recognition in parts from single depth images." In: *CVPR 2011*.

Sigal, Leonid, Alexandru Balan, and Michael Black (2007). "Combined discriminative and generative articulated pose and non-rigid shape estimation." In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc.

Smith, Harrison Jesse and Michael Neff (2017). "Understanding the impact of animated gesture performance on personality perceptions." In: *ACM Transactions on Graphics (TOG)* 36.4, pp. 1–12.

Starck, Jonathan and Adrian Hilton (2003). "Model-based multiple view reconstruction of people." In: *IEEE international conference on computer vision*, pp. 915–922.

Starke, Sebastian, He Zhang, Taku Komura, and Jun Saito (2019). "Neural state machine for character-scene interactions." In: *ACM Trans. Graph.* 38.6, pp. 209–1.

Stoll, Carsten, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt (2011a). "Fast Articulated Motion Tracking Using a Sums of Gaussians Body Model." In: *Proceedings of the 2011 International Conference on Computer Vision*. ICCV '11.

– (2011b). "Fast articulated motion tracking using a sums of Gaussians body model." In: *2011 International Conference on Computer Vision*.

Sun, Xiao, Jiaxiang Shang, Shuang Liang, and Yichen Wei (2017). "Compositional Human Pose Regression." In: Lecture Notes in Computer Science.

Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman (2017). "Synthesizing Obama: Learning Lip Sync from Audio." In: *ACM Trans. Graph.*

Takeuchi, Kenta, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi (2017a). "Speech-to-Gesture Generation: A Challenge in Deep Learning Approach with Bi-

Directional LSTM." In: *Proceedings of the 5th International Conference on Human Agent Interaction*.

Takeuchi, Kenta, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta (2017b). "Creating a Gesture-Speech Dataset for Speech-Based Automatic Gesture Generation." In: *HCI International – Posters' Extended Abstracts*. Ed. by Constantine Stephanidis.

Taylor, Graham W and Geoffrey E Hinton (2009). "Factored conditional restricted Boltzmann machines for modeling motion style." In: *Proceedings of the 26th annual international conference on machine learning*, pp. 1025–1032.

Taylor, Sarah, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews (2017). "A Deep Learning Approach for Generalized Speech Animation." In: *ACM Trans. Graph.*

Tekin, Bugra, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua (2016). "Structured Prediction of 3D Human Pose with Deep Neural Networks." In: *British Machine Vision Conference*.

Tevet, Guy, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or (2022). "MotionCLIP: Exposing Human Motion Generation to CLIP Space." In: *ECCV*.

Tomè, Denis, Chris Russell, and Lourdes Agapito (2017). "Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5689–5698.

Tompson, Jonathan, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler (2015). "Efficient object localization using Convolutional Networks." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 648–656.

Tompson, Jonathan, Arjun Jain, Yann LeCun, and Christoph Bregler (2014). "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation." In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'14. Montreal, Canada: MIT Press.

Toshev, Alexander and Christian Szegedy (2014). "DeepPose: Human Pose Estimation via Deep Neural Networks." In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '14.

Tzirakis, Panagiotis, Athanasios Papaioannou, Alexander Lattas, Michail Tarasiou, Björn W. Schuller, and Stefanos Zafeiriou (2019).

"Synthesising 3D Facial Motion from "In-the-Wild" Speech." In: *CoRR* abs/1904.07002.

Ubisoft (2020). *Introducing Learned Motion Matching*. https://montreal.ubisoft.com/en/introducing-learned-motion-matching/.

Usuyama, Naoto (2018). *github.com/usuyama/pytorch-unet*. URL: github.com/usuyama/pytorch-unet.

Van Mulken, Susanne, Elisabeth André, and Jochen Müller (1998). "The persona effect: How substantial is it?" In: *People and computers XIII*. Springer, pp. 53–66.

Wang, C., Y. Wang, Z. Lin, and A. Yuille (2018). "Robust 3D Human Pose Estimation from Single Images or Video Sequences." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wei, Shih-En, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh (2016). "Convolutional Pose Machines." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4732.

Xiang, Donglai, Hanbyul Joo, and Yaser Sheikh (2019). "Monocular total capture: Posing face, body, and hands in the wild." In: *CVPR*.

Yang, Wei, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang (2018). "3D Human Pose Estimation in the Wild by Adversarial Learning." In: *CVPR*.

Yasin, Hashim, Umar Iqbal, Björn Krüger, Andreas Weber, and Juergen Gall (June 2016). "A Dual-Source Approach for 3D Pose Estimation from a Single Image." In: *IEEE Conference on Computer Vision and Pattern Recognition 2016 (CVPR)*. Las Vegas, USA.

Ye, Mao, Yang Shen, Chao Du, Zhigeng Pan, and Ruigang Yang (2016). "Real-Time Simultaneous Pose and Shape Estimation for Articulated Objects Using a Single Depth Camera." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yoon, Y., W. Ko, M. Jang, J. Lee, J. Kim, and G. Lee (2019). "Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots." In: *ICRA*.

Yoon, Youngwoo, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee (2020). "Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity." In: *ACM Trans. Graph.*

Yoon, Youngwoo, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee (2019). "Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots."

In: *Proc. of The International Conference in Robotics and Automation (ICRA)*.

Zhang, He, Sebastian Starke, Taku Komura, and Jun Saito (2018). "Mode-adaptive neural networks for quadruped motion control." In: *ACM Transactions on Graphics (TOG)* 37.4, pp. 1–11.

Zhou, Xingyi, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei (2017). "Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach." In: *The IEEE International Conference on Computer Vision (ICCV)*.

Zhou, Xingyi, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei (2016). "Deep Kinematic Pose Regression." In: *ECCV Workshop on Geometry Meets Deep Learning*.

Zhou, Yuxiao, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu (2020). "Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data." In: *CVPR*.