Deep Learning Frameworks for Human Motion Analysis

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science in Electronics and Communication Engineering by Research

by

NEERAJ BATTAN
201431194
neeraj.battan@research.iiit.ac.in

International Institute of Information Technology, Hyderabad
(Deemed to be University)
Hyderabad - 500 032, INDIA
MARCH 2021

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certi ed that the work contained in this thesis, titled “Deep Learning Frameworks for Human Motion Analysis” by Neeraj Battan, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Dr. AVINASH SHARMA

To my loved ones

Acknowledgments

First and foremost, I would like to thank my family for their love and support. It was their trust in me that kept me going in hard times. They ensured not to burden me with any kind of problems to allow me to pursue my dreams and kept lifting me up in all possible ways at all times.

Next, I show my gratitude towards my advisor, Dr. Avinash Sharma, who guided me to the right path through both bright and dark times. I joined as a dual degree with Dr. Avinash Sharma in my third year. I feel grateful that I got to work with him. We used to have daily discussions on the problems that we were working on that helped to develop my research skills and he used to bring in some philosophy along with the research discussion that helped me grow as a better person in life.

I would like to pay my special thanks to my dearest friends Abbhinav, Aditya, Anirudh, Hemant, Govinda, Joyneel, Manan, Nikita, Raj, Sanket, Shreya, Shreyash and many others. Working, playing and hanging out with them made the research seem way easier and fun. I will cherish the moments when we used to hang out at the canteen and juice center, play at the courts and swimming sessions at Gopichand academy.

I extend my gratitude to all the staff members of IIIT-H for helping me in administrative and nancial processes.

Abstract

The human body is a complex structure consisting of multiple organs, muscles and bones. However, in regard to modeling human motion, the information about xed set of joint locations (with xed bone length constraint) is suf cient to express the temporal evolution of poses. Thus, one can represent a human motion/activity as temporally evolving skeleton sequence. In this thesis, we primarily explore two key problems related to modeling of human motion, the rst is ef cient indexing & retrieval of human motion sequences and the second one is automated generation/synthesis of novel human motion sequences given input class prior.

3D Human Motion Indexing & Retrieval is an interesting problem due to the rise of several data-driven applications aimed at analyzing and/or re-utilizing 3D human skeletal data, such as data-driven animation, analysis of sports biomechanics, human surveillance, etc. Spatio-temporal articulations of humans, noisy/missing data, different speeds of the same motion, etc. make it challenging and several of the existing states of the art methods use hand-craft features along with optimization-based or histogram-based comparison in order to perform retrieval. Further, they demonstrate it only for very small datasets and few classes. We make a case for using a learned representation that should recognize the motion as well as enforce a discriminative ranking. To that end, we propose, a 3D human motion descriptor learned using a deep network. Our learned embedding is generalizable and applicable to real-world data - addressing the aforementioned challenges and further enables sub-motion searching in its embedding space using another network. Our model exploits the inter-class similarity using trajectory cues and performs far superior in a self-supervised setting. State of the art results on all these fronts is shown on two large scale 3D human motion datasets - NTU RGB+D and HDM05.

In regard to the second research problem of Human Motion Generation/Synthesis, we aimed at long-term human motion synthesis that can aid to human-centric video generation [9] with potential applications in Augmented Reality, 3D character animations, pedestrian trajectory prediction, etc. Long-term human motion synthesis is a challenging task due to multiple factors like long-term temporal dependencies among poses, cyclic repetition across poses, bi-directional and multi-scale dependencies among poses, variable speed of actions, and a large as well as partially overlapping space of temporal pose variations across multiple class/types of human activities. This paper aims to address these challenges to synthesize a long-term (>6000ms) human motion trajectory across a large variety of human activity

[1]Video for Human Motion Synthesis.
[2]Video for Human Motion Indexing and Retrieval.

classes (> 50). We propose a two-stage motion synthesis method to achieve this goal, where the first stage deals with learning the long-term global pose dependencies in activity sequences by learning to synthesize a sparse motion trajectory while the second stage addresses the synthesis of dense motion trajectories taking the output of the first stage. We demonstrate the superiority of the proposed method over SOTA methods using various quantitative evaluation metrics on publicly available datasets.

In summary, this thesis successfully addressed two key research problems, namely, efficient human motion indexing & retrieval and long-term human motion synthesis. In doing so, we explore different machine learning techniques that analyze the human motion in a sequence of poses (or called as frames), generate human motion, detect human pose from an RGB image, and index human motion.

Contents

List of Figures

Figure | Page

List of Tables

Chapter 1

Introduction

The human body is a complex structure consisting of multiple organs, muscles and bones. However, in regard to modeling human motion, the information about fixed set of joint locations (with fixed bone length constraint) is sufficient to express the temporal evolution of poses. Alternatively, surface models can be used to represent human body (e.g., meshes, voxels, etc.). In this thesis, we stick to former representation with human motion/activity modeled as temporally evolving skeleton sequence and attempt to solve two key problems related to human motion, namely, Human Motion Indexing & Retrieval and Long-term Human Motion Synthesis.

3D human motion retrieval is an emerging field of research due to several attractive applications such as data-driven animation, athletic training, analysis of sports bio-mechanics, human surveillance and tracking etc. Performing such analysis is challenging due to the high articulations of humans (spatially and temporally), noisy/missing data, different speeds of the same action etc. Recent research in pose estimation, reconstruction, as well as the advancement in motion capture systems has now resulted in a large repository of human 3D data that requires processing. Moreover, since the procurement of new motion data is a time-consuming and expensive process, re-using the available data is of primary importance. To that end, we solve the problem of 3D Human Motion Retrieval and address several of the aforementioned challenges, using, a 3D human motion descriptor learned using a deep learning model. While 3D human motion recognition is a commonly researched field, 3D human motion retrieval is much less explored. The task of human motion retrieval consists of two parts - building the feature representation and then the retrieval algorithm. Therefore, it requires recognizing the action as well as, importantly, enforcing a ranking i.e., a “low-dimensional”, “recognition-robust” and “discriminative” feature embedding that is capable of fast retrieval is desirable.

Recently, human motion synthesis is gaining ground as it is widely used to aid human-centric video generation with potential applications in Augmented Reality, 3D character animations, pedestrian trajectory prediction, etc. Human motion synthesis is a challenging task due to multiple factors like, long-term temporal dependencies among poses, cyclic repetition across poses, bi-directional and multi-scale dependencies among poses, variable speed of actions and a large as well as partially overlapping space of temporal pose variations across multiple class/types of human activities.

(a) CCTV surveillance (b) Athletic training

Figure 1.1: Applications of human motion indexing & retrieval

1.1 Motivation

In recent years due to an increase in the availability of 3D human pose data, it has become a cumbersome task to keep track and reuse of the existing data. This lead the researchers to look for a method to index a large amount of 3D data and retrieve it when needed. There are multiple other applications associated with human motion indexing & retrieval such as:

Data-driven Animation Data-driven animations are created using data collected from multiple data sources where data can be stored in different modalities. We require motion indexing & retrieval to ef ciently collect the data from these sources.

Athletic Training Generally the athletic trainers are highly quali ed, multi-skilled professionals who provide sports training to the athletes. Generally, they are not affordable for a large number of athletes. As shown in Fig. 1.1b 3D human motion indexing & retrieval can help in training the athletes by comparing his/her performance with a professional athlete and provide a score accordingly.

Human Surveillance Video surveillance involves the act of observing a scene and looking for speci c behaviors that are improper or that may indicate the emergence or existence of improper behavior as

(a) Virtual reality

(b) Generation robot motion

(c) Pedestrian trajectory prediction

(d) Bio mechanics [31]

Figure 1.2: Applications of human motion synthesis

shown in Fig. 1.1a. Motion indexing & retrieval can be used for the purpose of surveillance by comparing a given scenario with the past ones.

As 3D animation is an expensive and time-consuming process, researchers started to explore the directions along human motion synthesis to make the process cheaper and faster. Recently, human motion synthesis is gaining ground due to its potential applications such as

Augmented RealityAugmented reality is an interactive experience of a real-world environment as shown in Fig. 1.2a. Human motion synthesis is used to mimic different types of activities in augmented reality. It is also used in robotics to predict the trajectories of joints while performing an action.

Pedestrian Trajectory Prediction As shown in Fig. 1.2c pedestrian trajectory prediction is the problem of predicting the spatial coordinates of pedestrians in the vicinity of the vehicle. Human motion synthesis can be used to predict future poses and hence the location of the pedestrian.

Bio MechanicsHuman motion synthesis techniques can be used to predict anatomical symptoms, such as bone deformity and contracture, affect full-body movements. Simulation of dynamic human models can evaluate the effectiveness of prostheses also as shown in Fig. 1.2d.

1.2 Problem Formulations

1.2.1 Human motion indexing & retrieval

This problem deals with nding discriminative representation of human motion sequences so as ef ciently index and retrieve them. More speci cally, we aim to represent a sequence of poses as a xed dimensional feature vector and map it to a common discriminative subspace where similar activity sequences are project together. At test time we can calculate the feature vector for the query sequence and retrieve sequences similar to this using nearest neighbor approach in the feature space. Fig. 1.3 outline the construction of a motion sequence indexing & retrieval setup.

Figure 1.3: Human Motion Indexing & Retrieval Setup : Attempts to learn a discriminative feature space in order to project the input motion sequence for indexing and subsequently employ nearest neighbors in the feature space for retrieval task.

1.2.2 Long-term human motion generation

Here the aim is to generate the motion sequence in two stages: the rst stage deals with learning the long-term global pose dependencies in motion sequences by learning to synthesize a sparse motion trajectory, while the second stage addresses the generation of dense motion trajectories taking the output of the rst stage. Fig. 1.4 outline the typical setup for human motion synthesis.

Figure 1.4: Long-term human Motion Generation : Initial poses along with activity class label are given as input to predict a sequence of future poses.

Following are the major challenges associated to these two problems.

Length of an action: The length of action can vary depending on the performer or the situation. It becomes dif cult for a model to capture the information among poses when the same activity is spanned across different number of poses.

Speed of motion performed:Same activity can be performed at different speeds that make it even more complicated for a network to differentiate between different types of activities.

Ef cient indexing: It is necessary for indexing & retrieval method to be ef cient and capable of retrieving the results in real-time for it to be used as an application. Achieving this becomes challenging when we have to extract out the features of a sequence and then search for similar sequences over a large dataset.

Sub-activity retrieval: For a given sub-activity it is a challenging task to retrieve the activities that contain it. As the features of a sub-activity will lie in a different space than that of activity, it becomes impossible to compare the features of these two.

Noisy data: Training a network becomes more challenging in case of inaccurate 3D joint locations as shown in Fig. 1.5b. A typical pre-processing attempts to reduce such noise by smoothening the data across the temporal domain using the Savitzky Golay lter. Additionally, many times there are missing joints owing to severe self-occlusions or joint is not visible in the frame as shown in Fig. 1.5a.

Long-term dependencies among poses: It is a challenging task to capture the long-term dependencies among poses due to their ambiguous nature. They can be different within the same class and same across