Stick Figure Dance Generation Project Proposal

James Wood, Mathew Varughese, Patrick Gardner
CS 1699

1. What do you propose to do?

We propose to utilize deep learning in order to generate believable stick figure movements, specifically dance movements. The minimum viable product of our project will be along the lines of generating a dancing stick figure based on a user's input of dance style or music type. If we succeed in this, we would like to explore the ability to create a dancing stick figure that moves based on any input of sound, interpreting the most "human" way to dance.

2. What have others attempted in this space, i.e. what is the relevant literature?

Researchers out of UC Berkeley have created a speech to gesture model based on analysis of video from people speaking, mostly talk show hosts. This is used to generate synthetic video of movement based on audio input - https://people.eecs.berkeley.edu/~shiry/projects/speech2gesture/index.html

UCF Researchers have developed a dataset of 50 human action categories of YouTube videos with appropriate labels - https://www.crcv.ucf.edu/data/UCF101.php

A research article coming from Georgia Tech titled "Let's Dance: Learning From Online Dance Videos". The researchers collected a dataset of 1000 videos featuring 10 dance styles. (100 videos per style.) They attempt to classify online dance videos using deep learning. They use different deep learning methodologies and show their relative effectivity. Particularly, they mention that adding a temporal element to the input data drastically improved their success rates. (https://www.cc.gatech.edu/cpl/projects/dance/paper.pdf)

Deep Video Generation, Prediction and Completion of Human Action Sequences (http://openaccess.thecvf.com/content_ECCV_2018/html/Chunyan_Bai_Deep_Video_Generation_ECCV_2018_paper.html)

Generation of Character Illustrations from Stick Figures Using a Modification of Generative Adversarial Network https://ieeexplore.ieee.org/abstract/document/8377853

3. Why is what you are proposing interesting?

This is interesting because it will delve into the realm of mimicking human motion. Additionally, dancing itself is a human artform that is hard to understand. Seeing if a computer can learn the mechanics behind it seems like a challenging and interesting problem. Additionally, since this deals with generating motion, rather than strictly static images, it seems like there is a lot of space to learn about cutting edge deep learning research.

4. Why is it challenging?

Deep Learning has been used extensively on computer vision related to images, but there is less detailed research on generating videos and learning from videos. It would be challenging to try to understand different ways to explore this problem space. It also has connections to motion and other movement-related physics topics, such as optical flow. Additionally, since the three of us are new to deep learning, the basics of neural networks, and generative networks, there will be a lot of challenging material to learn and experiment with. Also james has not taken Linear Algebra

5. Why is it important?

Although generating stick figures dancing itself sounds silly, analyzing motion and video has numerous applications to the real world. This ranges from motion detection, to analyzing surroundings in autonomous vehicles. Further, it helps demonstrate the bounds of what is possible for machine learning, and, more specifically, deep learning. When people mention artificial intelligence potentially mimicking entire humans, motion is one of the key limitations in achieving this.

6. What data do you plan to use?

We have a couple of data sets that we are interested in. Primarily, we have been focusing on the dataset provided by the "Let's Dance" research paper by Georgia Tech researchers - https://www.cc.gatech.edu/cpl/projects/dance/. This data set includes video frames of dancers online. It also conveniently provides the extracted skeletons and optical flow. Each video is tagged with one of the following dance styles: Ballet, Flamenco, Latin, Square, Tango, Breakdancing, Foxtrot, Quickstep, Swing, and Waltz.

There is also a Recognition Data Set from UCF that has 13320 videos from 101 action categories. We think perhaps looking at more research done with this data set will inspire us - https://www.crcv.ucf.edu/data/UCF101.php.

7. What is your high-level idea of how your method will work?

For our idea, we plan on using coordinates of each part of the stick figure we would like to know the locations of (joints, facial features, etc.). We will use a set of coordinates for each frame of input video as well as generated output video. These sequences of coordinates will allow us to draw the frames which form the video. If we move onto inputting songs/audio files, we will represent them as .wav files in order to represent them as bitstreams.

We are going to use a generative model, most likely with a recurrent neural network (RNN) in order to make the dancing more contiguous as RNNs utilize previous outputs (frames in our case) as successive inputs.

8. In what ways is this method novel?

This method is novel because we have not encountered the use of RNNs in video motion analytics and generation. Further, we have not seen much work in generating full-body human motion. Using RNNs will hopefully aid in the issue of regular frame-by-frame methods causing overfitting of models due to lack of general motion recognition as seen in the "Let's Dance" research paper.

9. How will you evaluate the method, i.e. what metrics are you going to use, and what baselines are you going to compare to?

Since this method has relatively little work and applicable research, we do not have many hard metrics to use. Being a purely generative task, our project does not have a standard other than our and other people's views on its effectiveness in dancing in a human manner. Also, this project is very innovative in the sense that it is exploratory rather than replicative of other people's efforts; with that in mind, we hope to simply be able to generate a tangible, acceptable result.

One thing we think might be useful is creating an adversarial neural network that has the ability to categorize generated stick figure dances as real or not real. If the adversarial network cannot identify the difference between real and fake videos, it's a good sign for our generated dances. We hope to learn more about generative methods as we progress through our project to get a better sense of better ways to evaluate our method.

10. Give a (1) conservative and (2) an ambitious schedule of milestones for your project.

Conservative Timeline

February 14th - Have a good understanding of neural network basics and how they can be applied to our project specifically. (i.e. a generative model).

Feb 28th - Make a stick figure model that has joints that are capable of moving in a 'natural' human way. (AKA Joints can move in a way that is human-like, but may be able to move in more ways. Deep Learning Network will filter out unhuman-like behavior later.) We want the joints to be in proportion to an actual human body

March 5th - Have the start of a network that can determine if a generated stick figure dance video is "fake" or not that can be used in improving the generation of stick figures.

March 26th - Have a model we are happy with that can somewhat generate stick figures that can move.

April 14th - Our Deep Learning Network will be able to generate stick figure dances of a variety of different styles. We will indicate the style of dance for the stick figure to perform

Ambitious Timeline

February 14th - Make a stick figure model that has joints that are capable of moving in a 'natural' human way. (AKA Joints can move in a way that is human-like, but may be able to move in more ways. Deep Learning Network will filter out unhuman-like behavior later.) We want the joints to be in proportion to an actual human body.

Feb 28th - Have the start of a network that can determine if a generated stick figure dance video is "fake" or not that can be used in improving the generation of stick figures.

March 5th - Our Deep Learning Network will be able to generate stick figure dances of a variety of different styles. We will indicate the style of dance for the stick figure to perform

March 26th - Our project is able to decode music files fed into the system as an input.

April 14th - Our Deep Learning Network will be able to recognize a song’s beat and style and, from that, generate a stick figure that dances to that style of music on beat.