**Project Report**

*CS 6323 Computer Animation & Gaming - Fall 2021*

**Team Members**

**Topic**

Face NeRF: Audio-driven talking face generation based on dynamic NeRF.

**Problem Summary**

The aim is to generate high fidelity talking head video with input audio

sequence. The method proposed here is different from existing methods

which rely on 3D face shape or 2D landmarks to encode the facial image.

Current model-based methods require expertise involvement in

establishing audio semantic and lip motion relationships thereby not being

suitable for general applications. On the other hand, many data-driven

methods have been proposed to generate photo-realistic talking-head.

However, they require far more training samples compared to the approach

being tested here.

**Description of work**

**What has been implemented:**

Obama:

We have implemented the code as provided in the paper.

We however we have tried to create the model using the same data as what they had trained on.

Unlike them however we trained on a small timeframe of the original image to create the model (~ 1 min)

We trained for 13,000 iterations

Train1:

Given the input video file, we have created the same model as the one above, the

only difference being that the video used is not the one that the original author

used when they wrote the paper.

We trained on a smaller timeframe of the train1.mkv data provided (~ 1min) .

The reason behind us training on a smaller timeframe of the entire

train1.mkv data is due a combination of issues:

1) Errors: It is to our belief that the file provided has an error. The reason being

is that we continuously received a 'PowBackward0' error which we never faced when

we used the Obama.mp4 file. This is further explained in the "Problems Faced" section

of the paper

2) Time Required: The time required to generate and train the files for the entire

video is taking unusually long. Even generation of the configuration file was taking

well above 8 hours. We ended up aborting the task and did the same for a shorter

timeframe. This is further explained in the "Problems Faced" section

of the paper.

3)Train\_1 Timeframe 30secs:

When we tried to create model using a timeframe of 30secs to create the model, we trained

on 6000 iterations and we were unable to get a satisfactory result. Our model did

not show a proper face. So, we tried to do the same for a time period of 1 minute

**Problems Faced**

1) Windows:

Environment does not run locally on windows.

The environment creation that should happen on running the environment.yml file

fails as supported versions of some dependencies are not available using anaconda.

On manually installing using pip, some install librabries are incompatible with the

dependencies

2)Google Colab:

In google colab, the changes made are session-dependent and so if the session ends or if the screen is idle for an extended period of time (during training or during the running of process\_data.sh), the session ends and all of our work is lost.

Another issue faced is that on running in colab, depending on the area where we work from, certain steps take far longer to execute the same command

eg: running process\_data.sh for a 1 minute clip took 97mins when running from one location, whereas earlier the same command took over hours while running from a different location.

The only difference between the 2 locations was the available internet speeds

although given that the notebook is being executed on the cloud, we are unsure

whether that should be a factor.

3)Train1.mkv:

The train1.mkv file is not of the same video format as the original Obama.mp4.

Passing a video of this format raises an error.

On converting the video to the mp4 video format, after running the Process\_data.sh command

and obtained athe config file, we started to train.On training we faced a 'PowBackward0' error.

We kept on facing this error multiple times for different portions of the video.

Finally on selecting the range 1:55 - 2:56 range, we did not encounter the issue.

Our assumption is that the video has some errors and we got lucky in selection a

range that was error free.

**Results**

Obama

**Chart

Description automatically generated**

**Graph 1.0: Shows the iteration vs loss during the Train Head Phase**

**Chart

Description automatically generated**

**Graph 1.1: Shows the iteration vs loss during the Train Torso phase**

|  |  |
| --- | --- |
| **Iteration** | **Loss** |
| 10000 | 0.00121142 |
| 10100 | 0.000894912 |
| 10200 | 0.000973368 |
| 10300 | 0.001391755 |
| 10400 | 0.00062118 |
| 10500 | 0.001182042 |
| 10600 | 0.001328444 |
| 10700 | 0.00202092 |
| 10800 | 0.000919932 |
| 10900 | 0.001750853 |
| 11000 | 0.001061188 |
| 11100 | 0.000973303 |
| 11200 | 0.002272087 |
| 11300 | 0.001172229 |
| 11400 | 0.00246223 |
| 11500 | 0.00122793 |
| 11600 | 0.001606605 |
| 11700 | 0.00173767 |
| 11800 | 0.001637076 |
| 11900 | 0.000784099 |
| 12000 | 0.001130277 |
| 12100 | 0.000704273 |
| 12200 | 0.00140432 |
| 12300 | 0.00079604 |
| 12400 | 0.001167741 |
| 12500 | 0.001005876 |
| 12600 | 0.000594645 |
| 12700 | 0.001275917 |
| 12800 | 0.001172659 |
| 12900 | 0.001583686 |
| 13000 | 0.001036798 |

Table 1.1: Iteration vs Loss for Train Torso

|  |  |
| --- | --- |
| **Iteration** | **Loss** |
| 10000 | 0.023144953 |
| 10100 | 0.020705651 |
| 10200 | 0.013548402 |
| 10300 | 0.015913408 |
| 10400 | 0.025985524 |
| 10500 | 0.018750243 |
| 10600 | 0.010461519 |
| 10700 | 0.014495319 |
| 10800 | 0.020720717 |
| 10900 | 0.021214172 |
| 11000 | 0.017449478 |
| 11100 | 0.016930372 |
| 11200 | 0.021312114 |
| 11300 | 0.020306107 |
| 11400 | 0.015493808 |
| 11500 | 0.02375748 |
| 11600 | 0.017082632 |
| 11700 | 0.016273607 |
| 11800 | 0.023538854 |
| 11900 | 0.019626461 |
| 12000 | 0.01857917 |
| 12100 | 0.015178712 |
| 12200 | 0.019252904 |
| 12300 | 0.015399571 |
| 12400 | 0.019968208 |
| 12500 | 0.022204315 |
| 12600 | 0.0170257 |
| 12700 | 0.017232127 |
| 12800 | 0.018847268 |
| 12900 | 0.015973259 |
| 13000 | 0.02074529 |

Table 1.0: Iteration vs Loss for Train Head

**Images Rendered:**

**A person in a suit

Description automatically generated with medium confidenceA person in a suit

Description automatically generated with medium confidence**

**Head Images: Variation of Mouth Orientation**

**A picture containing person, smiling, green

Description automatically generatedA person looking at the camera

Description automatically generated with medium confidenceA picture containing person, person

Description automatically generated**

**Head Images: Variation of Eye Orientation**

**A picture containing person, smiling, green

Description automatically generatedA person smiling for the camera

Description automatically generated with low confidenceA person smiling with his eyes closed

Description automatically generated with low confidence**

**Trial 1:**

**Company name

Description automatically generated**

**Graph 2: iteration vs loss for Trial\_1 Head**

**A picture containing text, antenna

Description automatically generated**

**Graph 2.2 Iteration vs loss for Trial\_1 Torso**

Table 2.1: Trial\_1 Torse Iteration vs Loss

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Iterations** | **Loss** | **Iterations** | **Loss** | **Iterations** | **Loss** |
| 100 | 0.051457681 | 3000 | 0.053397648 | 5900 | 0.054249104 |
| 200 | 0.06269528 | 3100 | 0.059558213 | 6000 | 0.052746087 |
| 300 | 0.060169093 | 3200 | 0.056129597 | 6100 | 0.05519671 |
| 400 | 0.053257897 | 3300 | 0.056168359 | 6200 | 0.055936031 |
| 500 | 0.055834614 | 3400 | 0.051819578 | 6300 | 0.049354151 |
| 600 | 0.054465711 | 3500 | 0.056170151 | 6400 | 0.050669573 |
| 700 | 0.059115402 | 3600 | 0.052556217 | 6500 | 0.048542093 |
| 800 | 0.055947732 | 3700 | 0.063352689 | 6600 | 0.050700393 |
| 900 | 0.04945007 | 3800 | 0.058823269 | 6700 | 0.058003366 |
| 1000 | 0.056784712 | 3900 | 0.058495231 | 6800 | 0.063117594 |
| 1100 | 0.053936403 | 4000 | 0.051069319 | 6900 | 0.04810844 |
| 1200 | 0.061780937 | 4100 | 0.059728324 | 7000 | 0.0515007 |
| 1300 | 0.056498833 | 4200 | 0.059138492 | 7100 | 0.057287972 |
| 1400 | 0.06070637 | 4300 | 0.048242997 | 7200 | 0.063517541 |
| 1500 | 0.060971029 | 4400 | 0.054659735 | 7300 | 0.060322121 |
| 1600 | 0.05103622 | 4500 | 0.064929858 | 7400 | 0.049000882 |
| 1700 | 0.054375261 | 4600 | 0.067758575 | 7500 | 0.057954691 |
| 1800 | 0.048494965 | 4700 | 0.058860295 | 7600 | 0.060727775 |
| 1900 | 0.049850285 | 4800 | 0.063244067 | 7700 | 0.063387878 |
| 2000 | 0.054771781 | 4900 | 0.056125116 | 7800 | 0.047954362 |
| 2100 | 0.064945444 | 5000 | 0.055962738 | 7900 | 0.047618449 |
| 2200 | 0.055373281 | 5100 | 0.061934035 | 8000 | 0.059623599 |
| 2300 | 0.052784681 | 5200 | 0.048215222 | 8100 | 0.061444737 |
| 2400 | 0.057831179 | 5300 | 0.059739307 | 8200 | 0.052540731 |
| 2500 | 0.046829082 | 5400 | 0.060665112 | 8300 | 0.06821236 |
| 2600 | 0.050270669 | 5500 | 0.051059507 | 8400 | 0.05450359 |
| 2700 | 0.056096863 | 5600 | 0.055713229 | 8500 | 0.057222985 |
| 2800 | 0.048576176 | 5700 | 0.061390799 | 8600 | 0.060723048 |
| 2900 | 0.068589211 | 5800 | 0.055793863 | 8700 | 0.052145101 |
|  |  |  |  | 8800 | 0.049537055 |
|  |  |  |  | 8900 | 0.057404786 |
|  |  |  |  | 9000 | 0.058656566 |

Graph 2.1 : Trial\_1 Head Iteration vs Loss

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Iteration** | **Loss** | **Iteration** | **Loss** | **Iteration** | **Loss** |
| 100 | 0.041607425 | 3100 | 0.01410557 | 6100 | 0.012148965 |
| 200 | 0.026763674 | 3200 | 0.012058235 | 6200 | 0.006760534 |
| 300 | 0.01781185 | 3300 | 0.006752055 | 6300 | 0.011311894 |
| 400 | 0.015751379 | 3400 | 0.008670259 | 6400 | 0.009450678 |
| 500 | 0.01634904 | 3500 | 0.012311361 | 6500 | 0.007802862 |
| 600 | 0.016277533 | 3600 | 0.011968647 | 6600 | 0.009101514 |
| 700 | 0.010144245 | 3700 | 0.020367038 | 6700 | 0.009678159 |
| 800 | 0.012415413 | 3800 | 0.008667568 | 6800 | 0.009142492 |
| 900 | 0.011411111 | 3900 | 0.013292667 | 6900 | 0.015699811 |
| 1000 | 0.010447975 | 4000 | 0.014571786 | 7000 | 0.012174774 |
| 1100 | 0.0081666 | 4100 | 0.013646023 | 7100 | 0.008641826 |
| 1200 | 0.01312589 | 4200 | 0.00784385 | 7200 | 0.008808279 |
| 1300 | 0.008598428 | 4300 | 0.011399669 | 7300 | 0.008473348 |
| 1400 | 0.016072456 | 4400 | 0.00932524 | 7400 | 0.008086795 |
| 1500 | 0.009165147 | 4500 | 0.009041414 | 7500 | 0.007492753 |
| 1600 | 0.008715248 | 4600 | 0.013888923 | 7600 | 0.01182754 |
| 1700 | 0.012247166 | 4700 | 0.015208317 | 7700 | 0.007077898 |
| 1800 | 0.013583098 | 4800 | 0.007215842 | 7800 | 0.011631515 |
| 1900 | 0.011941254 | 4900 | 0.00617135 | 7900 | 0.011448504 |
| 2000 | 0.008736827 | 5000 | 0.00959686 | 8000 | 0.012089105 |
| 2100 | 0.012160868 | 5100 | 0.010272931 | 8100 | 0.012400381 |
| 2200 | 0.008810984 | 5200 | 0.008949444 | 8200 | 0.006071062 |
| 2300 | 0.010353294 | 5300 | 0.010284036 | 8300 | 0.006521791 |
| 2400 | 0.017938148 | 5400 | 0.008308444 | 8400 | 0.009230751 |
| 2500 | 0.008052841 | 5500 | 0.015228853 | 8500 | 0.009804416 |
| 2600 | 0.007687974 | 5600 | 0.010014318 | 8600 | 0.006686902 |
| 2700 | 0.011755411 | 5700 | 0.009885155 | 8700 | 0.011094477 |
| 2800 | 0.010997721 | 5800 | 0.010129781 | 8800 | 0.012124082 |
| 2900 | 0.010074388 | 5900 | 0.008837676 | 8900 | 0.009476879 |
| 3000 | 0.010177198 | 6000 | 0.009144846 | 9000 | 0.012115523 |

**Analysis of work**

*Ablation Study*

Audio Condition

The NeRF based talking head model is directly conditioned on audio as we wish to limit

information loss. We also want to avoid training cost that comes with additional intermediate

modalities.

Training Individual Neural Radiance Fields for head and torso:

Training separate head and torso neural radiance fields has an advantage. It

helps avoid pixel mismatches around the upper body which. If we do not train them separately but

train one, then the torso area (including neck and shoulders are transformed by the estimated

head pose matrices. This can result in the above-mentioned mismatches.

Comparison of Work:

Obama:

Compared to the original image, our trained model does not create a torso. Ideally our model would also have a torso but we suspect that we need to train for a lot more iterations

A person looking at the camera

Description automatically generated with medium confidenceA person looking at the camera

Description automatically generated with medium confidenceA person in a suit

Description automatically generated with medium confidence

**Figure 1.2: Left and Centre – Our model’s generated image, Right – Original image**

**Github**

**https://github.com/sopam/Graphics\_Final\_Project.git**