Efficient Photorealistic Avatars using ML/AI

Group 1
Minh Khoa Đoàn
Cyrine Boudaya
Belinda Myteberi
Rebecca Charlotte Barth

Agenda

- Problem Statement
- Paper Review
- Roadmap
- Next Steps

Problem Statement

Goal: Rendering a photorealistic avatar with

- Monocular camera input
- Using an optimized neural radiance fields with state of the art input encoding
- Displaying the fourth dimension in terms of facial expressions and emotions

Problem Statement

Why:

- Animated human avatars can be used cross-applicational
- VR/AR technology in gaming
- Teleconferencing
- Healthcare sector
- Human computer interaction

Paper Review

Paper review: Face Reconstruction based on a Morphable Model

Objective :

To learn face models without using any pretrained models.

Method:

1) Face modeling : using PCA
2) Face Reconstruction : reconstruction are limited to the pre-defined 3DMM space
3) Joint Modeling and Reconstruction : The learning occurs in a self-supervised manner

Input :

Frames of a video

Dataset:

VoxCeleb and Emotionet

Results:

The Training was was implemented in Tensorflow and was done over three stages

Link to the paper : Learning Complete 3D Morphable Face Models From Images and Videos (thecvf.com)

Paper review: Face Reconstruction based on a Morphable Model

Objective :

To learn face models without using any pretrained models.

Method:

1) Face modeling : using PCA
2) Face Reconstruction : reconstruction are limited to the pre-defined 3DMM space
3) Joint Modeling and Reconstruction : The learning occurs in a self-supervised manner

Input :

Frames of a video

Dataset:

VoxCeleb and Emotionet

Results:

The Training was was implemented in Tensorflow and was done over three stages

Link to the paper : Learning Complete 3D Morphable Face Models From Images and Videos (thecvf.com)

Paper review: Neural Scene Representation Networks Nerf and optimization

NeRF: (https://arxiv.org/pdf/2003.08934.pdf)

Able to:

- overcomes the prohibitive storage costs of discretized voxel grids when modeling complex scenes at high resolutions

Method :

1) march camera rays through the scene to generate a sampled set of 3D points,
2) use those points and their corresponding 2D viewing directions as input to the neural network to produce an output set of colors and densities
3) use classical volume rendering techniques to accumulate those colors and densities into a 2D image.

Optimization:

- multiresolution hash encoding, which is adaptive and efficient, independent of the task.
- Unlike prior work, no structural updates to the data structure are needed at any point during training
- (https://nvlabs.github.io/instant-ngp/assets/mueller2022instant.pdf)

Paper review: FLAME-in-NeRF : Neural control of Radiance Fields for Free View Face Animation

Objective: Combine FLAME 3DMM with NeRF

Method:

- Condition the NeRF with the expression parameters from FLAME
- Disentangle background with FLAME silhouette rendering

Result:

- High fidelity in expressions compared to pure NeRF solution

Problems:

- Large head movements

Link: https://arxiv.org/abs/2108.04913v1

Paper review: Neural Head Avatars from Monocular RGB Videos

Neural Head Avatars learned from a **monocular RGB Portrait Video**

Able to:

- Accurately extrapolate to unseen poses and viewpoints
- Generate Natural Expressions while providing **sharp texture details**

Hybrid representation consisting of :

- a morphable model (FLAME-MESH)
- two feed-forward networks

Texture Network:

- Synthesis the appearance of the avatar by predicting a photorealistic texture
- Conditioned on the pose, expression and patches of surface normals

Output:

Avatar Articulation:

- Controlled via pose and expression parameters of the Face Mesh or by using an extracted driving sequence of them

Animation is consistent

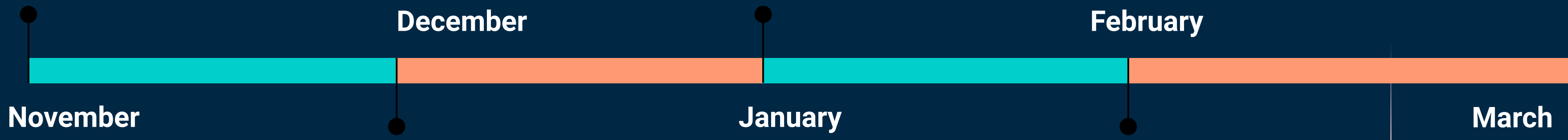
Links: https://arxiv.org/pdf/2112.01554.pdf
https://github.com/philgras/neural-head-avatars

Concept Solution

Methodology:

- Combine implicit and explicit representation
- Using the benefits of FLAME MESH + Nerf + Texture Network
- New input encoding with multi resolution Hashencoding
- Texture Network for spatial consistency and generalization to unseen poses/expressions (Video synthesis)
- Bonus: Train an emotion recognition network

Road Map



Agree on research based methods

Set-up tech stack and methods based on conducted research.

First Proof of Concept
Functioning code base with minimal features and simple reconstruction.

First prototype
Error free render of a static photorealistic Avatar.

Final Prototype & Draft of the Paper

Additional dynamic face expression and emotions.

Create a draft for the Paper

Final Product & Paper

Fully efficiently rendered 4D photorealistic avatar with monocular video input integrated in web platform.

Next Steps

- Decide on tech stack and set up working environment (read.me)
- Explore the state-of-art image processing in a video input domain
- Discuss available Data-Sets we will use
- Develop a first proof of concept
- Iterate on the proof of concept

Thank You!

(Additional Slide) Paper review: EMOCA: Emotion Driven Monocular Face Capture and Animation

Objective: To better reflect emotions

Method:

- Train an emotion recognition network
- ResNet-50, pre-trained on AffectNet dataset
- Add the network as expression encoder to existing model

Result:

- Finer details with highly emotional input

Problems:

- Emotion network difficult to optimize
- Usage of pre-trained network not optimal

Link: https://arxiv.org/abs/2204.11312
https://github.com/radekd91/emoca