ITU-T AI/ML IN 5G CHALLENGE

Shared Experience Using 5G+AI

Nitish Kumar Singh

October 29, 2020

CONTENTS

1 BACKGROUND

VR can help realize the utopian environment where distance disappears and we interact as richly with friends, family, and colleagues around the world as we do with those around us.This could also allow us to have natural interactions with those who cannot travel to meet physically. In order to create such rich experiences, there are certain challenges we need to overcome. The alienation that many of us already observe in connection with smartphones and the social fabric is only made more acute by a technological experience that fully absorbs our senses and that significantly reduces our ability to communicate directly with those near to us. Working out the obstacles to a truly shared experience is perhaps the most important challenge confronting virtual technologies and their broader cultural acceptance.

2 PROBLEM STATEMENT

Design a 3D Augmented + Virtual Reality based Immersive shared experience between friends for themes including but not limited to:

- Birthday Party
- Watching Cricket / Games
- Tourist site visit Monument visit:
  - Taj Mahal
  - Hill station visit: Rohtang Pass
  - Activity such as paragliding

3 Introduction

Twentieth century philosopher Merleau-Ponty said humans are fundamentally related to space and based on their capacity to perceive through their body, they make meaning of space. Of all digital technologies, nowhere are the notions of ‘being’ and ‘space’ of more consequence than in virtual reality (VR) as characterized by the fundamental concepts of embodiment and presence. Both ‘being’ and ‘space’ contribute to a user’s sense of presence defined as “...the strong illusion of being in a place in spite of the sure knowledge that you are not there”

A user is immersed in VR in two ways: first, through the representation of computer generated surroundings displayed from a first person point of view, and second, through a match between proprioceptive signals about the movements of their body with those of a corresponding virtual body. In prac- tice, a virtual body is often replaced by disembodied hands, which creates a conflict between proprioceptive data which tells the user their body is there, and sensory data in VR where the body does not exist. Embodiment is an attempt to reduce the conflict by providing a body representation where virtual body movements correspond with real body movements. Studies have shown that reported presence is higher if the match between proprioception and sensory data is high. Because of this match, natural walking is a desired feature in VR applications, and has repeatedly been shown to be superior to other navigation methods such as flying or using game controllers.

3.1 Proposed Approach

In this project our goal is to create experiences where users interact as richly over distances as they do with those in the same room. Our approach is novel in that prior work supporting multiuser interactions in VR does not take into account the size and shape of each individual user’s track- ing space. By mapping the different room-scale spaces into a single shared virtual space, our system allows each user to move in the shared virtual environment by making movements in their physical space.

4 System Design

The primary goal of the project is to build a usable system that can be easily accessed by user. Not only can the system be used to create different personal virtual experiences, but it can also help us to dynamically collect the audio and profile of the user. Which means, instead of a user having to manually enter his videos and get shared experience based on those. The system can help us to automatically create the user's avatar or video using Artificial Intelligence(AI) algorithm and libraries. Overall, we will make a minimum viable product where we will create video profiles of users and use them to create a virtual experience for the user using 3D model of assets(In this case models have been imported from google poly). In future, we can explore more possibilities. This section provides a brief overview about the theoretical and technical design aspects of the system.

4.1 Theoretical Aspects

The primary focus of this project is **"Audio-driven Talking Face Video Generation with Learning-based Personalized Head Pose"**. In real-world scenarios, natural head movement plays an important role in high-quality communication and human perception is very sensitive to subtle head movement in real videos. In fact, human can easily feel uncomfortable in communication by talking with fixed head pose.

To output a high-quality synthesized video of the target person with personalized head pose when speaking the input audio signal of source person, our system reconstructs 3D face animation and re-renders it into video frames. Given a light-weight rendering engine with limited information, these rendered frames are often far from realistic.

This generated video is then embedded on 3d asset to create virtual experience. AR frameworks like ARCore, VUforia, wikitude etc can be used for proper experience.

4.2 Technical Aspects

Following are the primary technical components that must be taken care of while designing the system proposed in the project:

- **Code Editor or IDE :** Google Colab is the preferred IDE for python running ML Algorithms. Google colabe provides free GPU support which helps in faster processing of ML algorithms.
- **Input Data :** Input data required are:
  - Video : A talking face video that satisfies:
    - contains a single person
    - 25 fps
    - longer than 12 seconds
    - without large body translation (e.g. move from the left to the right of the screen).
  - Audio : Audio for talking face.
- **Unity3D :** Unity is used for rendering videos of person generated on environment on which we want to create virtual experience.
- **Poly Toolkit :** Poly toolkit is provided by Google. It contains many 3D models and scenes which can be easily used in Unity.
- **AR Framework :** AR frameworks like ARcore, Vuforia, wikitude etc are used for building new augmented reality experiences that seamlessly blend the digital and physical worlds. Wikitude is used in this project.
- **Chromakey Shader :** This is a custom shader for unity downloaded from 'Holistic3D.com/resources'. This is used for making video background transparent

5 PROJECT - FUNCTIONAL FLOW

The project is divided into mainly two sections.

- Talking face video generation with the help of AI algorithms
- AI generated video is then rendered with 3D scenes or models in unity for reality experiences.

This section provides detailed overview about the steps followed.

5.1 TALKING FACE VIDEO GENERATION

- Proposed framework for talking face generation

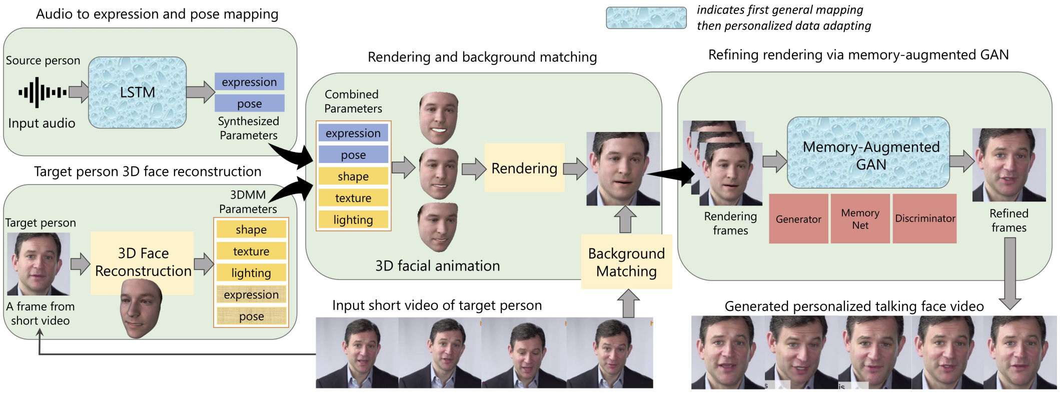

Figure 5.1: Pipeline

- create a virtual env, and install all the dependencies mentioned in requirements_colab.txt
- Download pre-trained models from Google Drive and copy to corresponding subfolders (Audio, Deep3DFaceReconstruction, render-to-video).
- Fine-tune on a target person's short video
  - Prepare a talking face video that satisfies: 1) contains a single person, 2) 25 fps, 3) longer than 12 seconds, 4) without large body translation
  - Extract frames and lanmarks
  - Conduct 3D face reconstruction
  - Fine-tune the audio network.
  - Fine-tune the gan network
- Test on a target person
  - Place the audio file (.wav or .mp3) for test under Audio/audio/. Run with poses from short video.

5.2 Render video in Unity

Talking face video generated from above section is then rendered with a 3d model or video background for realistic experience. Steps followed are as follows :

- Craete a new 3D project in Unity hub.
- Integrate poly toolkit with unity project.
- Download wikitude unity package and import it under assets folder.
- Add a 3D asset of Taj Mahal from Poly Toolkit.
- Import generated video under assets folder.
- Create a quad for video player and place it properly alongwith 3D asset for smooth experience. Also scale it properly.
- Use chromakey shader to make background transparent.

6 Resources

- Google Colab Notebook
- Pre-trained models
- Wikitude SDK
- Chromakey Shader
- Input Audio
- Final Output
- UnityProjects

7 REFERENCES

- Deepali Aneja, Daniel McDuff, Shital Shah,A High-Fidelity Open Embodied Avatar with Lip Syncing and Expression Capabilities.
- Audio-driven Talking Face Video Generation with Learning-based Personalized Head Pose (Ran Yi, Zipeng Ye, Juyong Zhang, Member, IEEE, Hujun Bao, Member, IEEE, and Yong-Jin Liu, Senior Member, IEEE)
- Misha Sra,Aske Mottelson,Pattie Maes: Your Place and Mine: Designing a Shared VR Experience for Remotely Located Users