A Neural Lip-Sync Framework for Synthesizing Photorealistic Virtual News Anchors

Abstract

Here we present a novel lip-sync framework specially designed for producing a virtual news anchor for a target person. A pair of Temporal Convolutional Networks are used to learn the seq-to-seq mapping from audio signals to mouth movements, followed by a neural rendering model that translates the intermediate face representation to a high-quality appearance. This fully-trainable framework avoids several time-consuming steps in traditional graphics-based methods, meeting the requirements of many low-delay applications.

Challenges

Two main problems in applying current methods to the virtual anchor projects.
1. The lack of enough resolution, visual consistency in details, and natural appearance in synthetic videos.
2. The lack of training, inference, and processing efficiency also prevent current methods from many low-delay application scenarios. The candidate frame selection in traditional graphics-based methods is laborious and time-consuming.

Our Solution

Our solution can be interpreted as two stages of work:
1. A pair of Temporal Convolutional Networks (TCN) learning the seq-to-seq mapping from audio signals to lip motion
2. An image-to-image translation-based neural renderer generating high-resolution and photorealistic texture from the synthetic face representation.

Mapping from Audio to Lip Motion

Different from RNN-based implementations, we employ a pair of Temporal Convolutional Networks (TCN) to learn the audio-to-mouth mapping, bringing the TCN's strength such as large perceptive field, stable gradients, and low memory requirements into the lip-sync task. A TCN-based generator learns the mapping from audio features to mouth features. It consists of four 1-D convolutions layers, two fully-connected layers, and a TCN block. The TCN block is wrapped with 1-D convolutions layers which downsample the rate of audio features to the video rate. We also devise a similar TCN-based discriminator to support the training of the mapping network. The discriminator takes the combination of audio and mouth sequences as input and outputs a real or fake label.

Neural Face Rendering

We devise a neural rendering module based on the hierarchical image-to-image translation model. We first synthesized the specially designed facial maps as an intermediate face representation. Then these facial maps are sent to the rendering network to generate high-resolution face appearance. For building Synthetic Facial Maps, we integrate the generated mouth into a face template. Generating continuous and accurate details is one of the main challenges for current rendering methods, especially for generating high-resolution videos. Instead of using the optical flow or temporal-consistent losses to improve visual consistency, we directly provide necessary information via the Canny edges from target frames.

Experiments

We evaluate the lip-sync framework at both the audio-to-mouth mapping and the rendering stages. We compare the audio-to-shape mapping performance between our model, two representative RNN-based baselines from recent lip-sync studies, and a basic TCN generator. Inter-Frame MSE measures the frame-wise velocity. We also compare the training time and inference times. The synthesized final frames show good visual compatibility and embouchure consistency, accurately capturing the mouth movements in the sound-source video while representing realistic facial expressions.