Summary of Meshtalk WIP

Jay
CAU EAI Lab
Lab address Here
https://yukinyaa.github.io

Data Class	Resolution	Frame rate
Face Video	80 cameras	30 FPS
Face Mesh	1,672 vertices	30 FPS
Audio	16kHz	-
Mel Spectrogram	80 Dimension	10ms(100 FPS)

Table 1. Captured, then processed Datasets

Abstract

The article summarizes Meshtalk[3]. The goal of this practice is to get familiar with LaTeX and technical writing(hopefully).

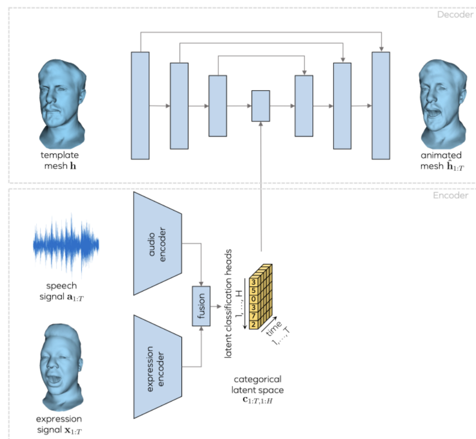

Figure 1. The network diagram[3].

1. Introduction to Meshtalk

Meshtalk is a generic method for generating full facial mesh animation from speech. It can generate lip-sync animation from a single frame of generic human facial mesh, and also can mix in emotional information from mesh animation.

1.1. Dataset

In total, there are 1.4 million frames of 13 hours equivalent audio-visual dataset from 250 subjects, reading 50 phonetically balanced sentences.

For training dataset, first 40 sentences out of 50 *about* 200 out of 250 subjects, total 40×200 dataset is used for training. For evaluation, remaining 10 sentences of 50 subjects are used.

Mesh Dataset

Face motion is captured with synchronized cameras, and processed into high-detail mesh, including eyelid, hairstyle, etc.

Audio Dataset

Audio data is recorded in 16kHz as shown on Table 1. For every mesh, 600ms of audio snippet is processed info Mel spectrogram, every 10ms, in 80 dimension. Hence, $\mathsf{a}_t \in \mathbb{R}^{60\times 80}$

1.2. Network Design

The network resembles Variational Autoencoder with multiple latent space as shown in Figure 1. Target mesh $\hat{h}$ is estimated from template mesh h and latent space c by computing function $\mathcal{D}$.

$$\hat{h}_{1:T} = \mathcal{D}(h, \mathsf{c}_{1:T,1:H}) \tag{1}$$

Sequence of latent space $\mathsf{c}_{1:T}$ is derived from audio sequence $\mathsf{a}_{1:T}$ and expression signal mesh sequence $\mathsf{x}_{1:T}$. c and a are first mapped to $T * H * C$ dimensional latent space, then passed through Gumbel-softmax[2] over every classification head.

$$\mathsf{c}_{1:T,1:H} = [\mathsf{Gumbel}(\mathsf{enc}_{t,h,1:C})]_{1:T,1:H} \tag{2}$$

$$\mathsf{enc}_{1:T,1:H,1:C} = \tilde{\xi}(x_{1:T}, a_{1:T}) \tag{3}$$

1.3. Training

The solution uses a novel cross-modality loss for calculating loss function for the network.

$$\begin{aligned}\mathcal{L}_{\mathsf{x}Mod} &= \sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(\mathsf{upper})}(||\hat{h}_{t,v}^{(\mathsf{expr})} - x_{t,v}||) \\ &+ \sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(\mathsf{mouth})}(||\hat{h}_{t,v}^{(\mathsf{audio})} - x_{t,v}||)\end{aligned} \tag{4}$$

Where $\hat{h}_{t,v}^{(\mathsf{expr})}$ is estimated with correct expression signal and random speech signal, and $\mathcal{M}^{(upper)}$ is a mask that assigned a higher weight to vertices around the mouth, and low weight to others. Correspondingly, $\hat{h}_{t,v}^{(\mathsf{audio})}$ is estimated with random expression signal and random speech signal. $\mathcal{M}^{(mouth)}$ is a mask that assigns a lower weight around the mouth, and vice versa. Loss for eye is defined as following.

$$\mathcal{L}_{\mathsf{eyelid}} = \sum_{t=1}^{T}\sum_{v=1}^{V}\mathcal{M}_v^{(\mathsf{eyelid})}(||\hat{h}_{t,v}^{(\mathsf{expr})} - x_{t,v}||) \tag{5}$$

$\mathcal{M}^{(eyelid)}$ is a specific eye loss, which was crucial.

Final Loss term is defined as $\mathcal{L} = \mathcal{L}_{\mathsf{x}Mod} + \mathcal{L}_{\mathsf{eyelid}}$, which gives both term equal weight.

2. Evaluation

To evaluate the network, the network is compared with networks that lacks speech signal and uses ℓ_2 loss. [1]

References

[1] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[2] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax, 2017. 1

[3] A. Richard, M. Zollhoefer, Y. Wen, F. de la Torre, and Y. Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement, 2021. 1