Summary of Meshtalk WIP

Anonymous ICCV submission

Paper ID ****

Data Class	Resolution	Framerate
Face Video	80 cameras	30 FPS
Face Mesh	1,672 verticies	30 FPS
Audio	16kHz	-
Mel Spectogram	80 Dimension	10ms(100 FPS)

Table 1. Captured, then processd Datasets

Abstract

The article summerises Meshtalk[2]. The goal of this practice is to get familiar with LaTeX and technical writing(hopefully).

1. Introuction to Meshtalk

Meshtalk is a generic method for generating full facial mesh animation from speech. It can generate 'lipsync' 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 equivelent audio-visual dataset from 250 subjects, reading 50 phonetically balanced sentences.

Mesh Dataset

Face motion is captured with synchronized cameras, and processd 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 spectogram, every 10ms, in 80 dimention. Hence $\mathsf{a}_t \in \mathbb{R}^{60\times 80}$

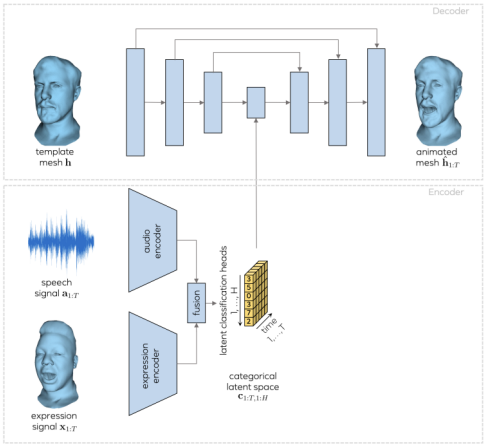

Figure 1. The network diagram[2].

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 $T*H*C$ dimentional latent space , then passed through Gumbel-softmax[1] 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. Cross-

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

and loss for eye is defined as following.

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

$\mathcal{M}^{(upper)}$ is a mask that assignes a higer weight to verticies around the mouth, and low weight to other. Correspondingly, $\mathcal{M}^{(mouth)}$ is a mask that assignes a lower weright around the mouth, and vice versa. $\mathcal{M}^{(eyelid)}$ is a specific eye loss, which was crucial.

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

References

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

[2] 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