Seminar Presentation

ON

META TALK: LEARNING TO DATA-EFFICIENTLY GENERATE AUDIO-DRIVEN LIP-SYNCHRONIZED TALKING FACE WITH HIGH DEFINITION

BY

AJITH D
(PRP19CS008)

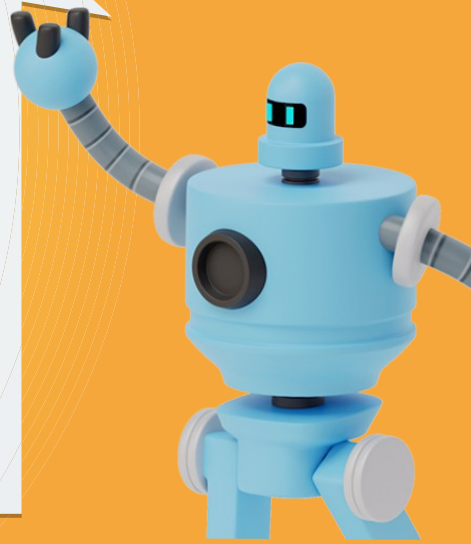
Contents

- Introduction
- Scope
- Literature Survey
- Training Phase
- Testing Phase
- Analysis
- References

INTRODUCTION

This paper proposes data-efficient audio-driven talking face generation method, which uses just a short target video to produce both lip-synchronized and high-definition face video driven by arbitrary audio in the wild.

In this work, the original target character's face images are decomposed into 3D face model parameters including expression, geometry, illumination,etc.



Scope

- Virtual Education
- Videoconferencing
- Game
- Entertainment
- Film and Television Animation

Literature Survey

Year	Authors	Title
2017	Supasorn Suwajanakorn,Steven M. Seitz,Ira Kemelmacher-Shlizerman	Synthesizing Obama: learning lip sync from audio
2020	Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen,and Shi-Min Hu	Photorealistic audio-driven video portraits
2019	Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu	Hierarchical cross-modal talking face generation with dynamic pixel-wise loss
2020	K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar	A lip sync expert is all you need for speech to lip generation in the wild

Year	Authors	Title
2020	Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li	Makelttalk
2021	Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu	Pirenderer: Controllable portrait image generation via semantic neural rendering
2021	Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li,Ming Zeng,Madhukar Budagavi, and Xiaohu Guo	3d talking face with personalized pose dynamics
2018	Triantafyllos Afouras, Joon Son Chung, Andrew Senior,Oriol Vinyals, and Andrew Zisserman	Deep audio-visual speech recognition

Year	Authors	Title
2019	Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen,Yunde Jia, and Xin Tong	Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set
2009	Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter	A 3d face model for pose and illumination invariant face recognition
2019	Yudong Guo, juyong zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng	Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images
2014	Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou	Facewarehouse: A 3d facial expression database for visual computing

Year	Authors	Title
2001	R. Ramamoorthi and P. Hanrahan	An efficient repre-sentation for irradiance environment maps
2018	Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A.Efros	Image-to-image translation with conditional adversarial networks

Let’s Start

What is meant by Lip Synchronization?

Lip Synchronization

It is a technical term for matching a speaking or singing person's lip movements with sung or spoken vocals.

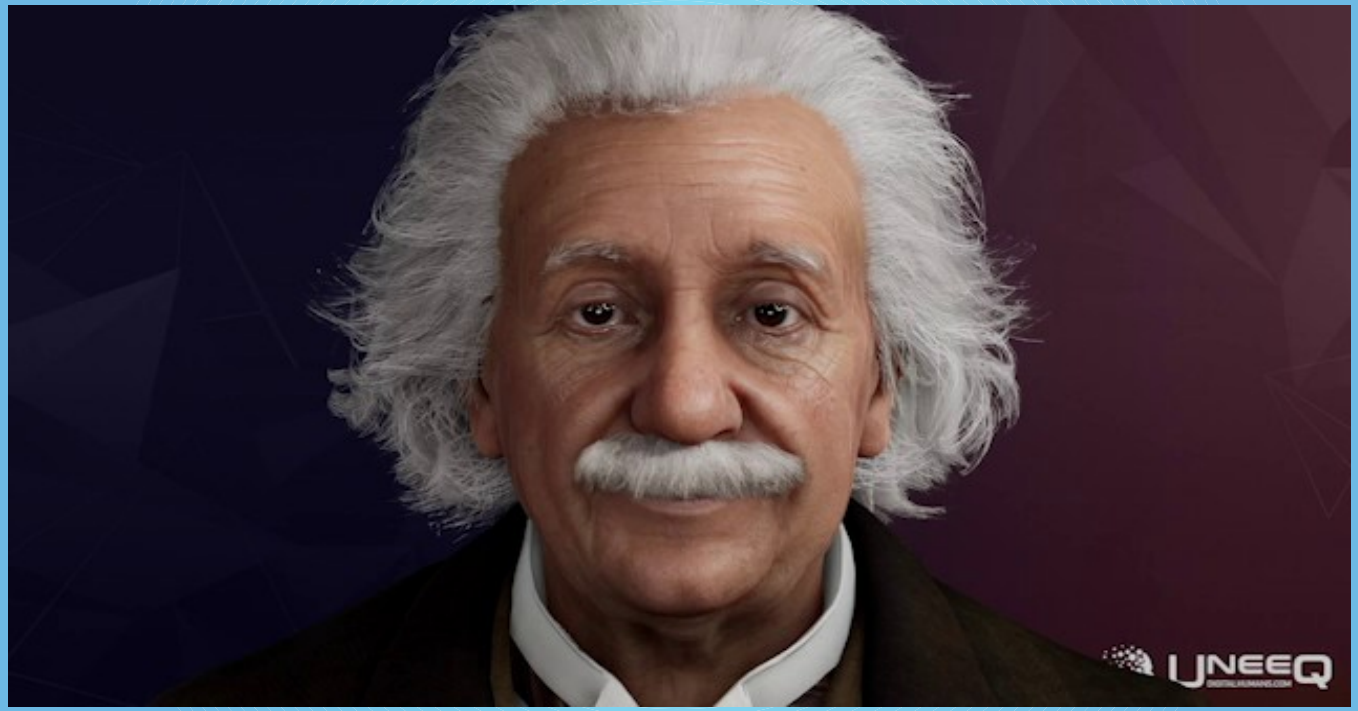

AE
Ah
BMP
ChJ
EE
Er
FV
Ih
KGHNG
Oh
R
SZ
TLDN
Th
WOO

Let’s get in to the Paper

- Talking face generation has attracted intensive attention in the field of multi-modal human-computer interaction. Its powerful communication mode makes it widely used in virtual education,videoconferencing,virtualanchor,game, entertainment, film and television animation production and other fields.

- Accurate lip movements during lip-syncing and realistic video portraits are key to better user experience feedback , natural head poses and eye blinks can also enhance user experience in these application scenarios.

Example:



Training Phase

Step 1:

Crop the original target video into the target face video, which is then resized to be low-resolution to generate a low-definition talking face video with LRS2 audio using the pre-trained model Wav2Lip.

Output: a pseudo video with excellent lip synchronization.

- **LRS2**

  Lip Reading Sentences 2 (LRS2) Dataset.The dataset consists of thousands of spoken sentences from BBC television.

- **Wav2Lip**

  Wav2Lip is a neural network that adapts video with a speaking face for an audio recording of the speech

Step 2:

Then, the 3DMM parameters are used to re-render the synthetic facial images in the target video. The facial 3D morphable model (3DMM) parameters including expression, geometry, texture, pose, illumination coefficients are extracted from each frame of them.

Output: Synthetic face

- **3DMM**

  A 3D Morphable Face Model is a generative model for face shape

Step 3:

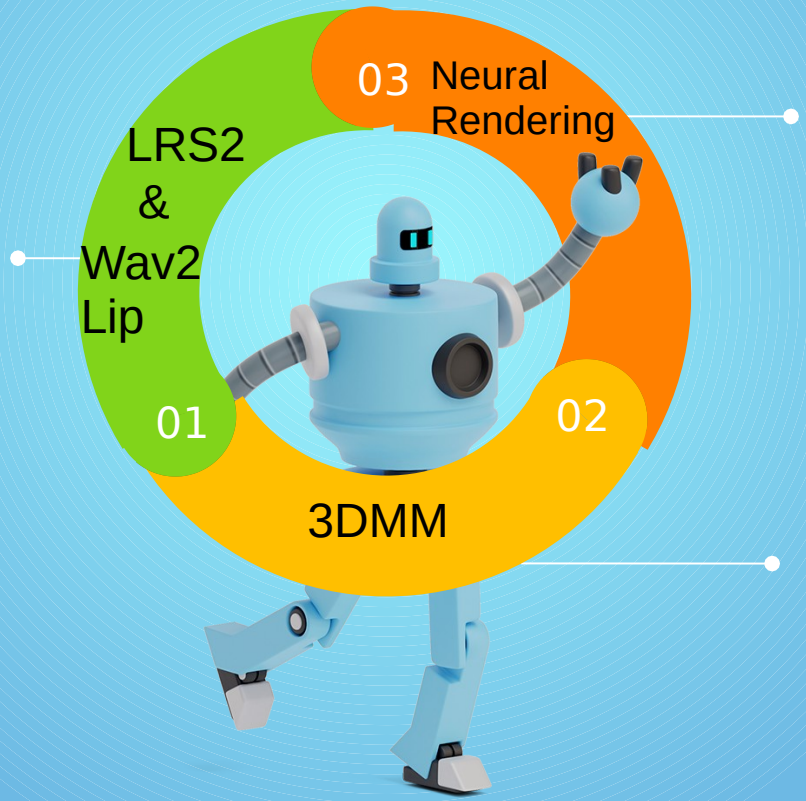
Finally, we train a neural rendering network with the lower half of synthetic and real target faces to generate a high-definition photo-realistic talking face video.

- **Neural rendering**

  It is a method, based on deep neural networks , which can create novel images and video footage based on existing scenes

Our process is easy

A pseudo video with excellent lip synchronization



high-definition photo realistic talking face video.

Synthetic face

Testing Phase

- Input : Trained Model
- We test the current model using a Audio to Expression network(A2E) to guarantee the accurate lip motion.
- **A2E**

  A neural network model used to map audio to expressions

Analysis

synchronization

A Best Quality Lip synchronization is obtained

Efficiency

Achieving data-efficient training

virtual education, videoconferencing, game,entertainment, film and television animation,etc

OPPORTUNITIES

Generate a high-definition video

Quality

References

- “Synthesizing obama: Learning lip sync from audio,”ACM Transactions on Graphics, vol. 36, no. 4CD, pp.95.1–95.13, 2017.
- Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen,and Shi-Min Hu, “Photorealistic audio-driven video portraits,” IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 12, pp. 3457–3466,2020.
- Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu, “Hierarchical cross-modal talking face generation with dynamic pixel-wise loss,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7824–7833.

- K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar, “A lip sync expert is all you need for speech to lip generation in the wild,” in Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, 2020, MM ’20, p.484–492, Association for Computing Machinery.
- Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li,“Makelttalk,” ACM Transactions on Graphics, vol. 39,no. 6, pp. 1–15, Nov 2020.
- Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu, “Pirenderer: Controllable portrait image generation via semantic neural rendering,” 2021.

- Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li,Ming Zeng, Madhukar Budagavi, and Xiaohu Guo, “3d talking face with personalized pose dynamics,” IEEE Transactions on Visualization and Computer Graphics,pp. 1–1, 2021.
- Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng,Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo, “FA-CIAL: synthesizing dynamic talking face with implicit attribute learning,” CoRR, vol. Abs/2108.07938, 2021.
- Triantafyllos Afouras, Joon Son Chung, Andrew Senior,Oriol Vinyals, and Andrew Zisserman, “Deep audio-visual speech recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2018.

- Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen,Yunde Jia, and Xin Tong, “Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), 2019, pp. 285–295.
- Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter, “A 3d face model for pose and illumination invariant face recognition,” in 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, 2009, pp. 296–301.

- Yudong Guo, juyong zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng, “Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 6, pp. 1294–1307, 2019
- Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou, “Facewarehouse: A 3d facial expression database for visual computing,” IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 3, pp.413–425, 2014.

- R. Ramamoorthi and P. Hanrahan, “An efficient repre-sentation for irradiance environment maps,” Proceed-ings of the 28th annual conference on Computer graph-ics and interactive techniques, 2001.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A.Efros, “Image-to-image translation with conditional adversarial networks,” 2018.

Thank You