High-Fidelity Facial and Speech Animation for VR HMDs

Olszewski, K., Lim, J. J., Saito, S., & Li, H. (2016). High-fidelity facial and speech animation for VR HMDs. ACM Transactions on Graphics (TOG), 35(6), 221.

Introduction

- Their objective is to enable natural face-to-face conversations in an immersive setting by animating highly accurate lip, tongue and eye motions for a digital avatar controlled by a user wearing a VR HMD.

- Using a monocular camera attached to an HMD with an internal infrared (IR) camera to enable full facial tracking, they record multiple subjects performing various facial expressions as video training data

- They apply a convolutional neural network (CNN) framework to regress images of a user's eye and mouth regions to the parameters that control a animated avatar.

System Prototype

- Based on FOVE VR HMD

- With integrated eye tracking camera and a custom mounted camera for mouth tracking

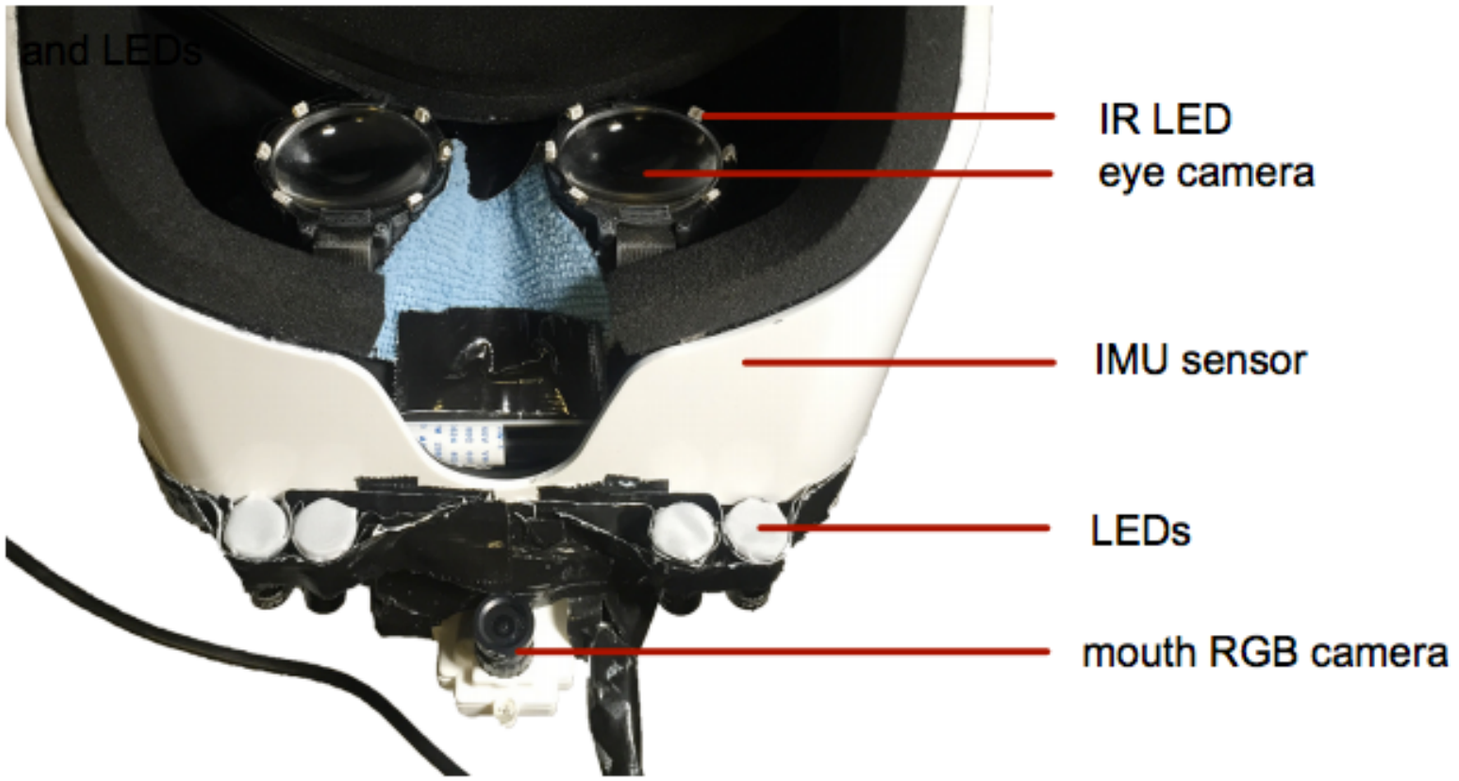


Figure 2: *VR HMD prototype with integrated eye and mouth cameras.*

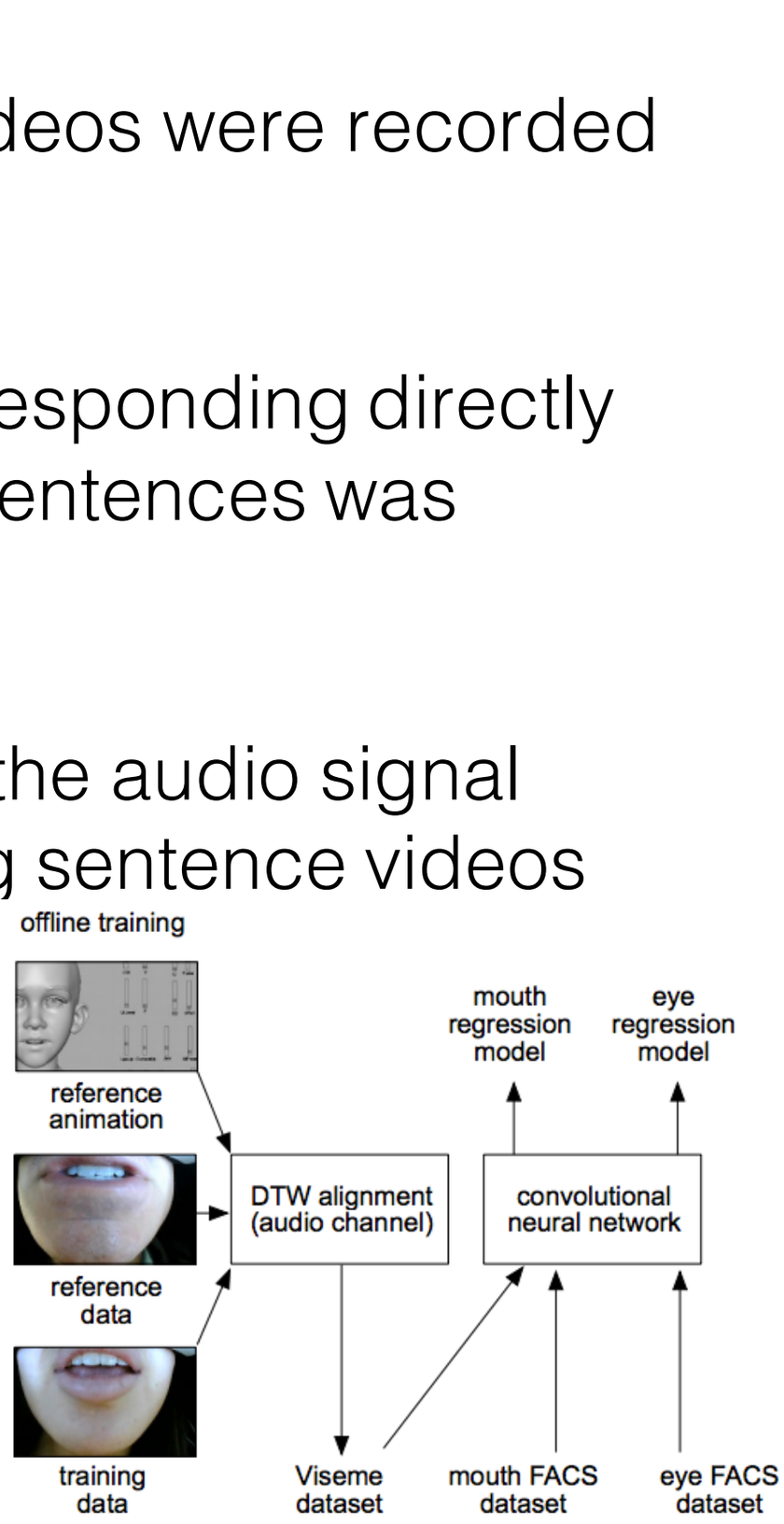
Facial Animation Pipeline

- To train the mouth regression model, videos were recorded of several subjects.

- A high-quality animation sequence corresponding directly to one subject's performance of these sentences was created by professional animators.

- Using dynamic time warping (DTW) on the audio signal from these recordings, the other training sentence videos were aligned to this user's performance

- Video [04:44 -]

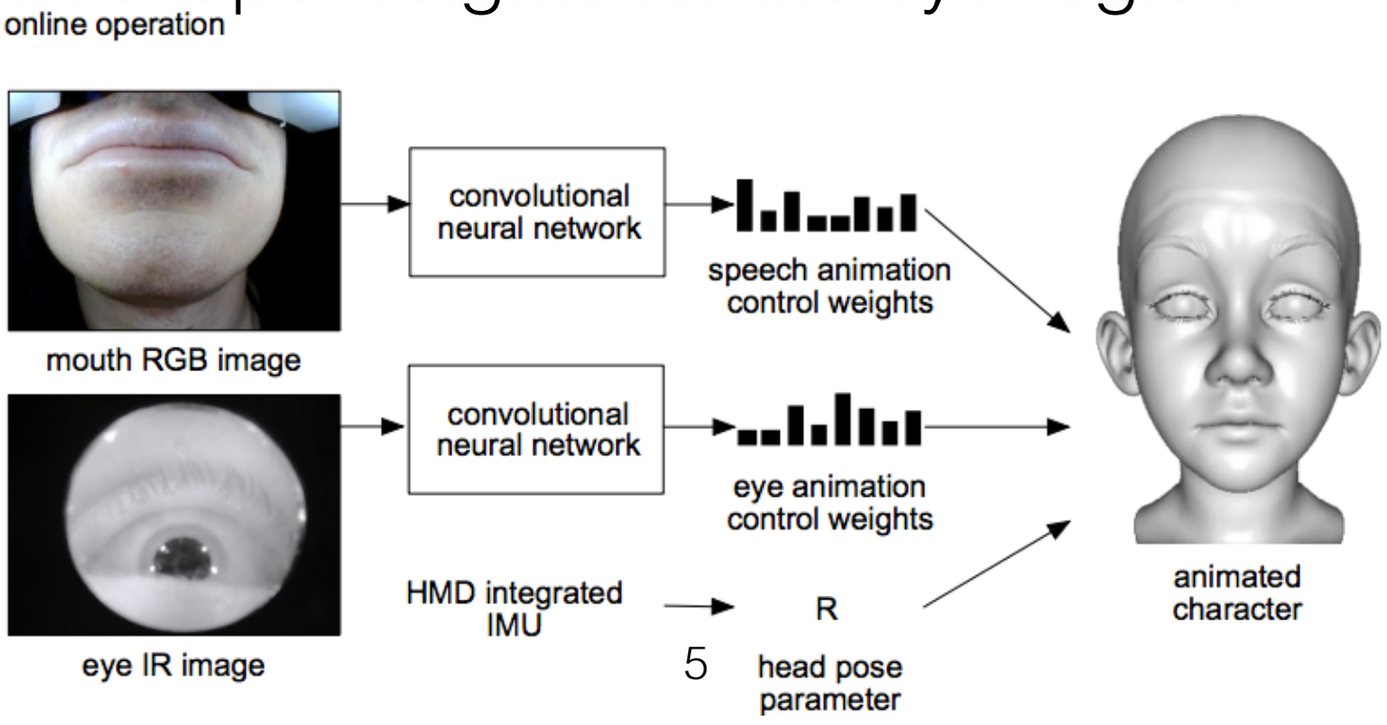


Facial Animation Pipeline

- Images are captured from the mouth and eye tracking cameras

- The mouth images are passed through a CNN regressor which outputs the appropriate blendshape weights for speech animation

- A separate CNN regresses images from the eye camera to obtain blendshape weights for the eye region



Data collection - Visual Speech Dataset

- They first collected synchronized video and audio recordings of 10 subjects (5 male, 5 female) each reciting a list of 30 sentences while wearing the HMD and maintaining a roughly neutral expression.

- These sentences were chosen from the Harvard sentences, a list of sample sentences in which phonemes appear at roughly the same frequency as in the English language.

- Each subject was also asked to perform a set of 21 facial expressions with their mouth based on the Facial Action Coding System.

Data collection - Eye Region Dataset

- They recorded sequences of the subjects performing a variety of movements with their upper face, including squints, blinks, and eyebrow movements, using the IR camera within the HMD.

Deep Learning Model

- They use a multi-frame CNN model.
  (CNN framework have attained impressive results for numerous classification and regression tasks in computer vision and robotics.)

- The goal of an expression network is to regress the blendshape weights that corresponds to the target input frame

- An neutral network detects whether a facial expression in the target frame is neutral or not.

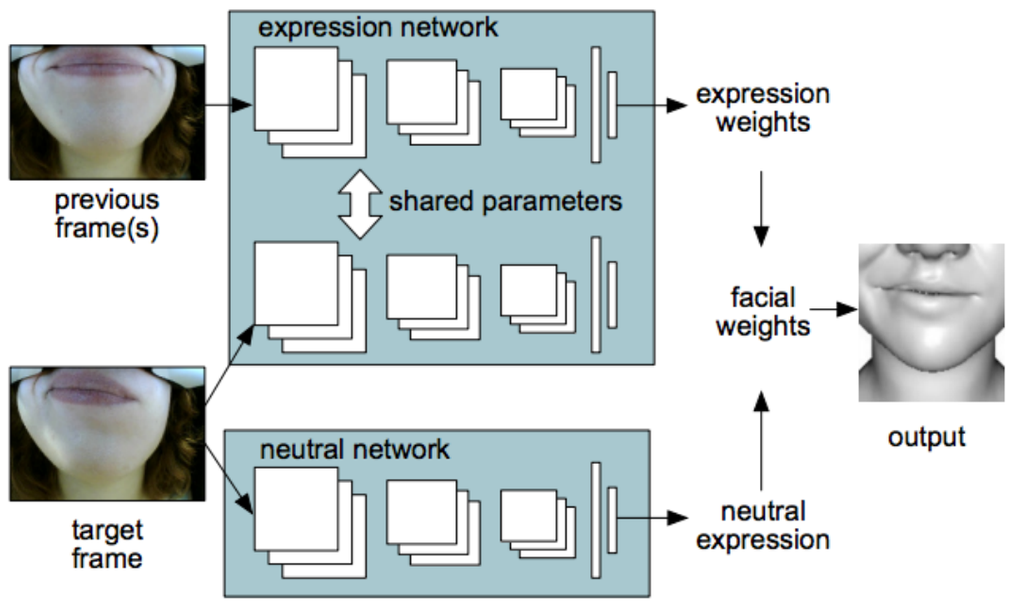


Figure 5: *The model trained to regress blendshape weights for the lower face. Our model has two sub-CNNs so it can incoporate temporal signals and control the sensitivity to important expressions.*

Results

- High-Fidelity Facial and Speech Animation for VR HMDs (SIGGRAPH Asia 2016)
  https://www.youtube.com/watch?v=eOjzC_NPCv8

- Video [02:03 -]

- Evaluation [04:56 -]

Conclusion and Limitations

- They have presented a CNN model for animating a digital avatar in real-time based on the facial expressions of an HMD user.

- However, very fast mouth motion can lead to motion blur in the captured images. Currently, their way to alleviating this problem is to use professional-grade high-speed cameras and conversely increase their system cost.

- In the eye region of digital avatar, they only take account of squints, blinks, and eyebrow movements except eyeball movements.