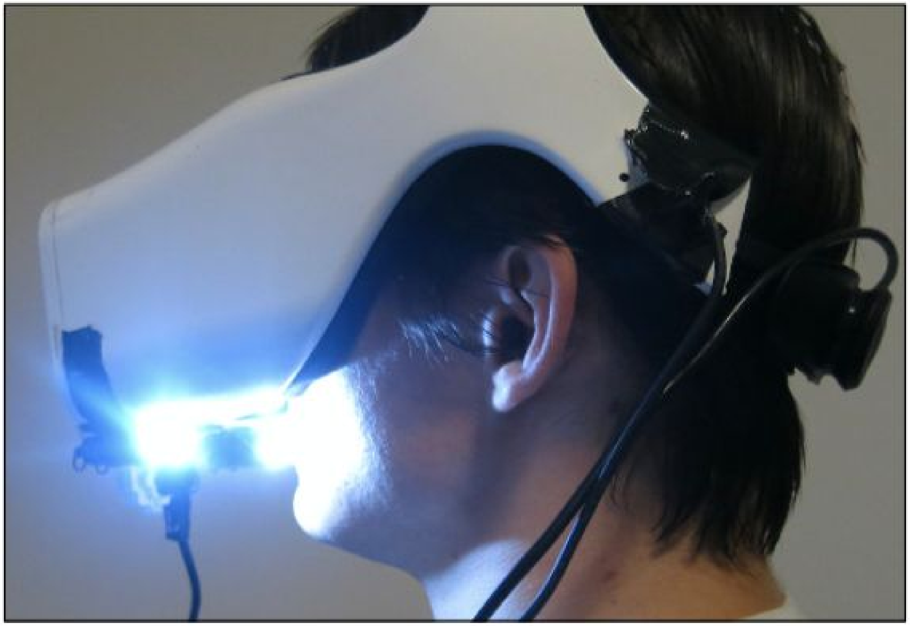
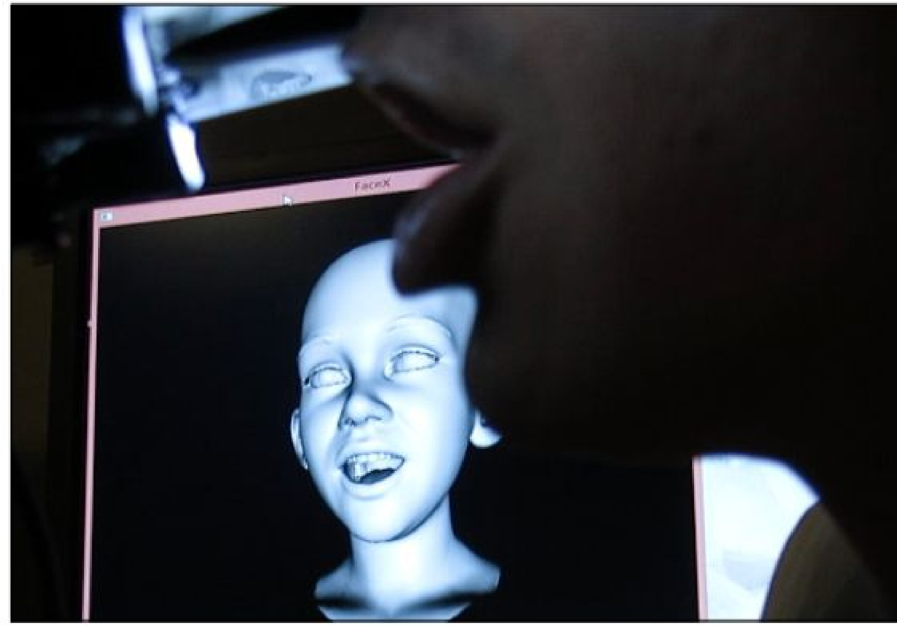
High-Fidelity Facial and Speech Animation for VR HMDs

Lukas Lipp

Forecast





- facial recognition with
  - Head-Mounted Display (HMD)
  - two small cameras

Olszewski et al. 2016

Motivation

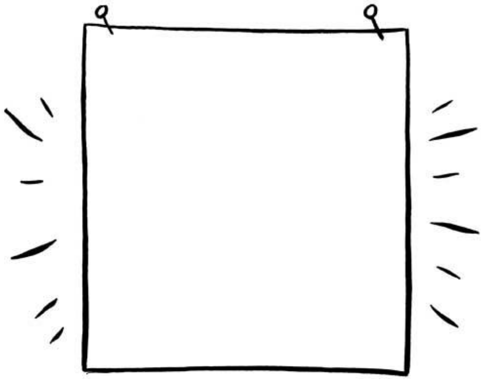
- reconstruction of the user’s face
  - user is wearing an HMD
  - covers up much of that person’s face
- reconstruction to be sent to another place
  - telepresence between two distant people
  - both are wearing an HMD.

Papers

- BHAT, K. S., GOLDENTHAL, R., YE, Y., MALLET, R., AND KOPERWAS, M. 2013. High fidelity facial animation capture and retargeting with contours. In SCA '13, 7–14.
- CAO, C., HOU, Q., AND ZHOU, K. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. 33, 4, 43:1–43:10.
- LI, H., TRUTOIU, L., OLSZEWSKI, K., WEI, L., TRUTNA, T., HSIEH, P.-L., NICHOLLS, A., AND MA, C. 2015. Facial performance sensing head-mounted display. ACM Transactions on Graphics (Proceedings SIGGRAPH 2015) 34, 4 (July).
- Olszewski, K., Lim, J. J., Saito, S., and Li, H. 2016. High-fidelity facial and speech animation for VR HMDs. ACM Transactions on Graphics (TOG), vol. 35, no. 6, p. 221.
- TOSHEV, A., AND SZEGEDY, C. 2014. Deeppose: Human pose estimation via deep neural networks. In IEEE Conference on Computer Vision and Pattern Recognition.
- WENG, Y., CAO, C., HOU, Q., AND ZHOU, K. 2014. Real-time facial animation on mobile devices. Graphical Models 76, 3, 172–179.

Outline

- Introduction of the topic
- State of the Art Technology
- Research and previous work
- Content of Paper

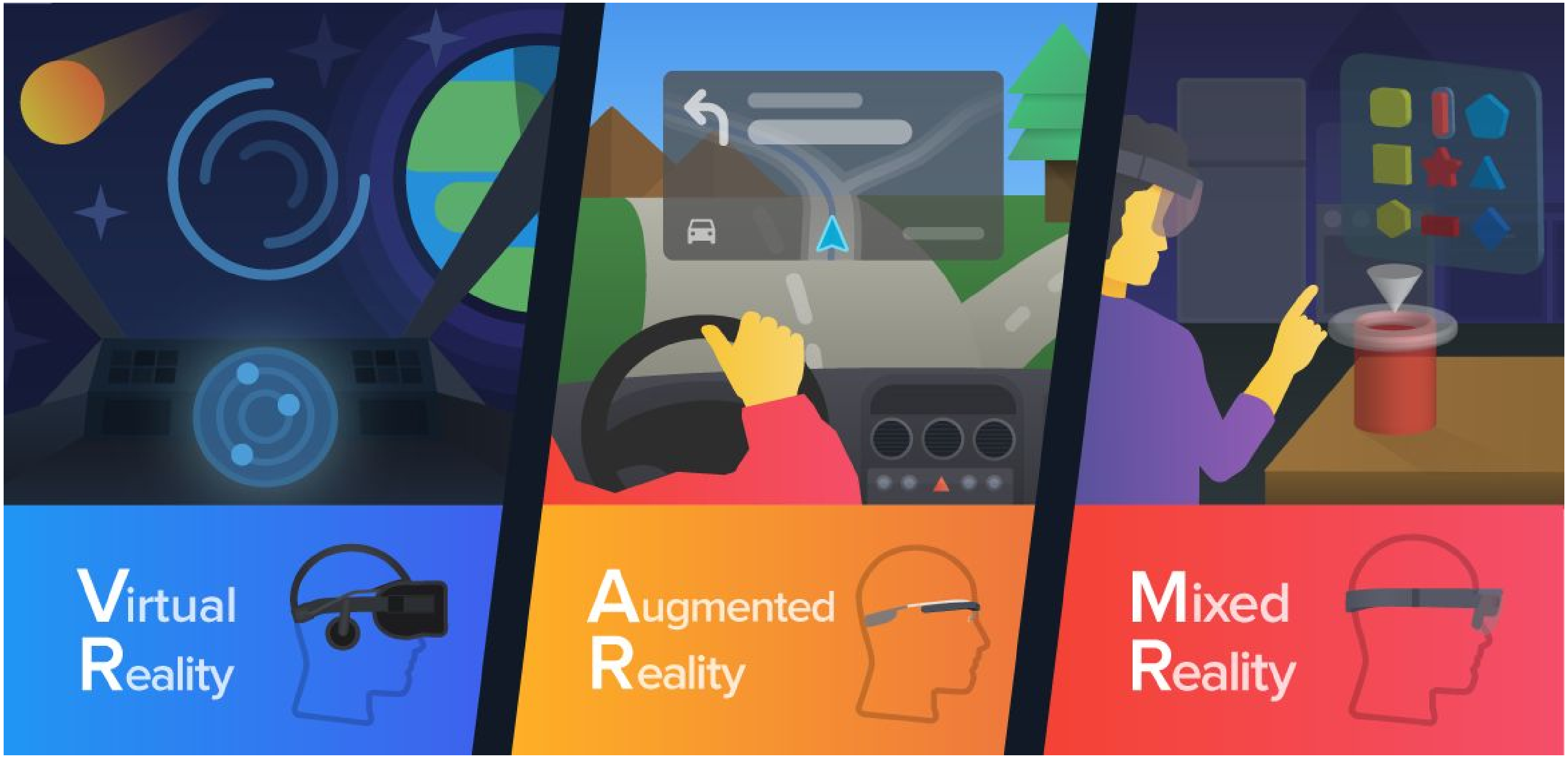


[P]

Outline Paper

- Overview
- Strategy
- Data Collection
- Deep Learning Model
- Evaluation and Comparison to other Approaches
- Results / Conclusion

Virtual Reality



[1]

- Virtual reality ≠ Augmented reality ≠ Mixed Reality

Virtual Reality

- first systems already in 1950
- VR could never establish itself...
- ...since Oculus Rift appeared
- goals
  - immersive
  - easy to use
  - cheap

[D]

State of the art technology

Oculus Rift [2]

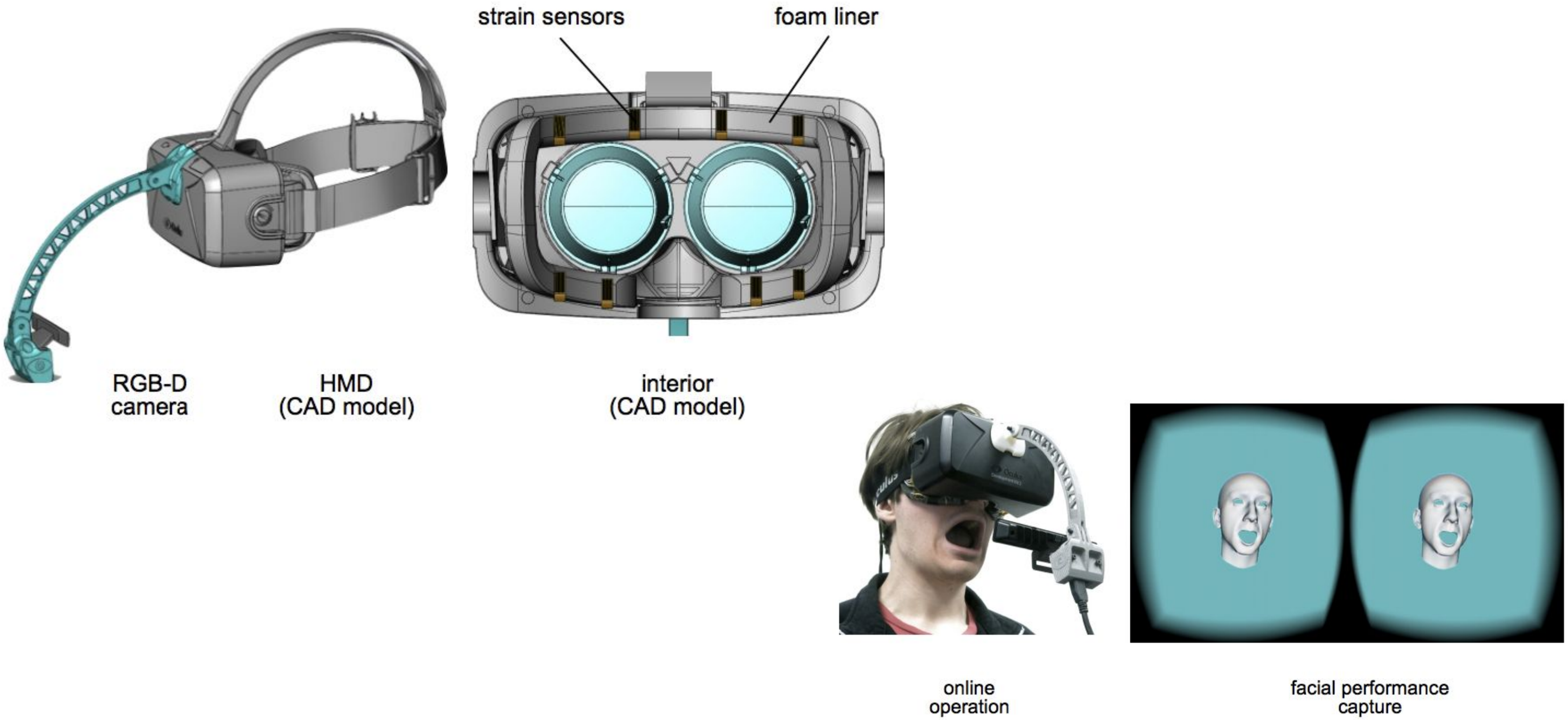
HTC Vive [3]

State of the art facial tracking methods

- used Techniques
  - tracked landmarks
  - depth signals
  - RGB videos
  - humans-in-the-loop

- common problems
  - occlusion
  - tongue not visible
  - portion of lips sometimes hidden

Facial Performance Sensing Head-Mounted Display



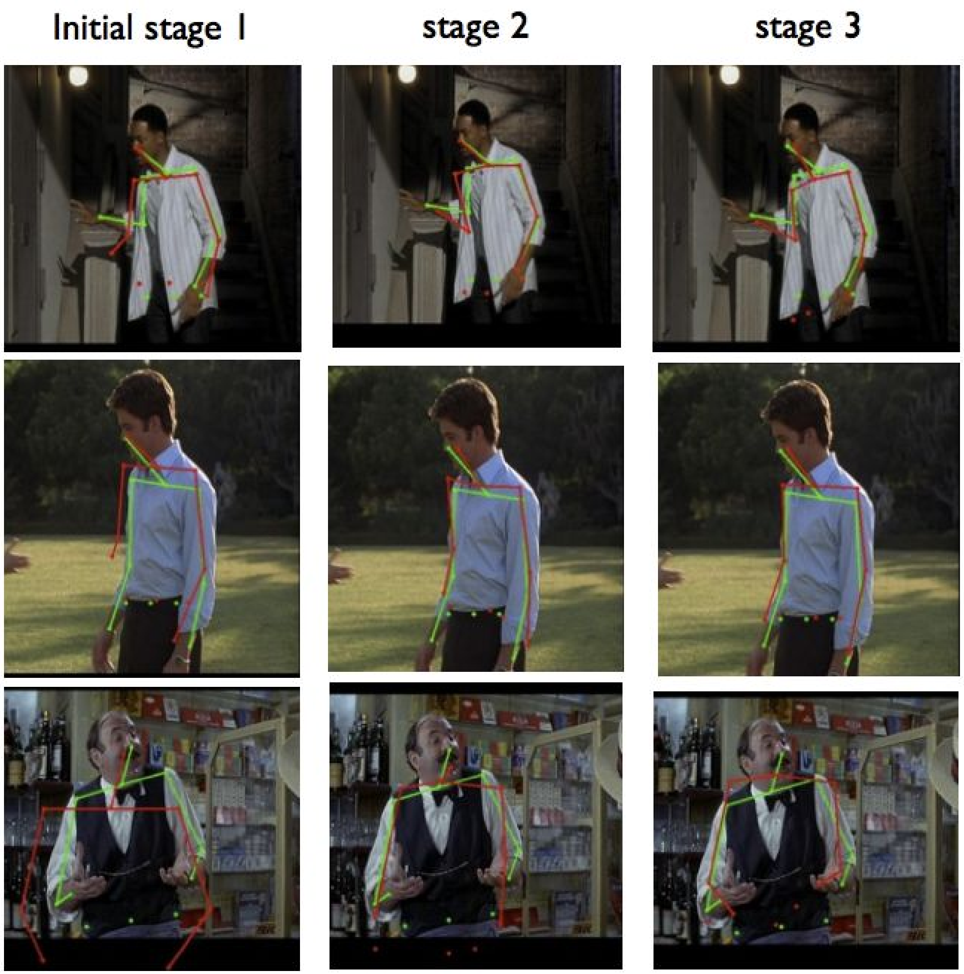
- calibration before each use
- RGB-D camera and strain sensors in foam

Li et al. 2015

- tracking with contour points

Bhat et al. 2013

DeepPose: Human Pose Estimation via Deep Neural Networks



- (green) ground truth
- (red) prediction

Toshev and Szegedy 2014

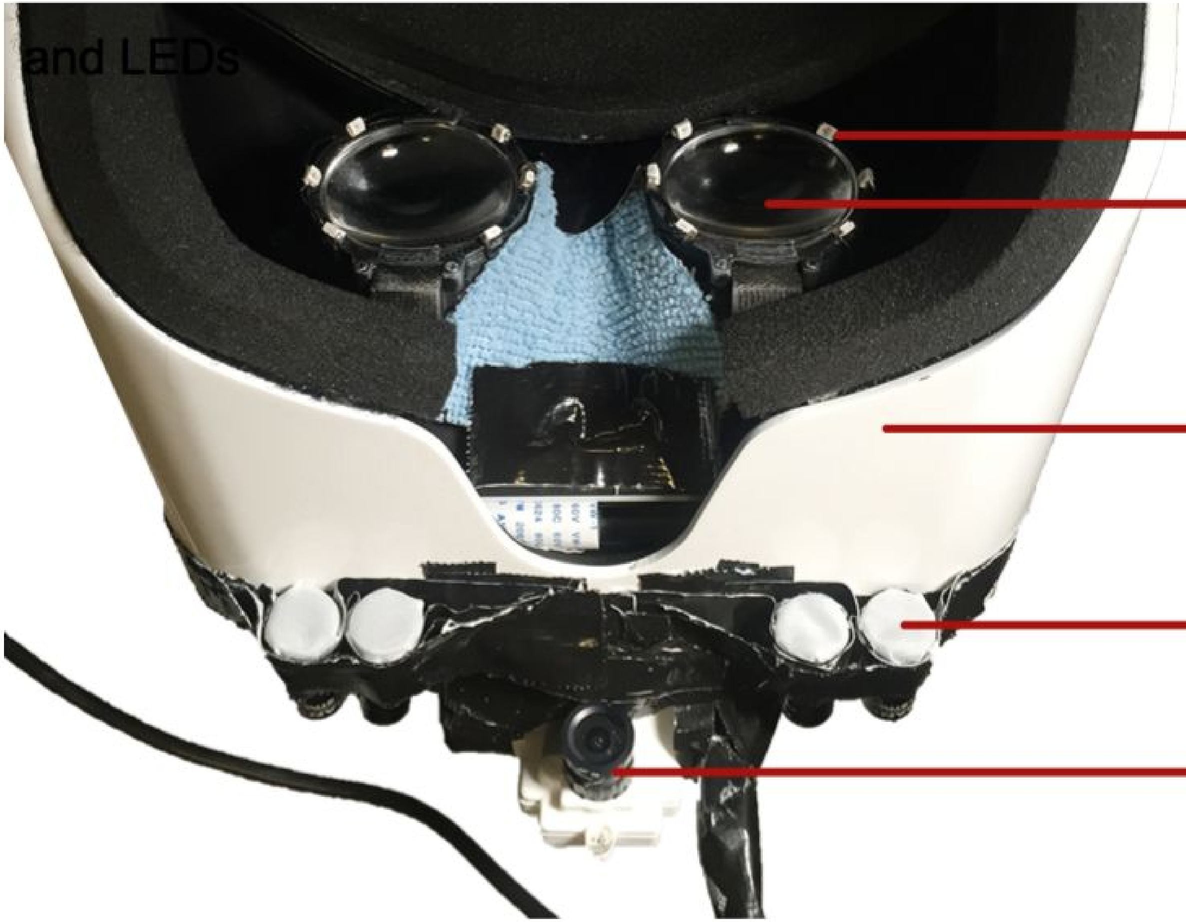
Previous Work

- many methods require
  - complex capture equipment
  - intensive computations

- eye and mouth animation
  - audio exploring and optical sensing
- real-time facial animation
  - 2D facial features detected in the RGB
  - RGB with depth sensor
- wearable facial sensing systems
  - eye tracking cameras inside VR headsets

Olszewski et al. 2016

System Prototype



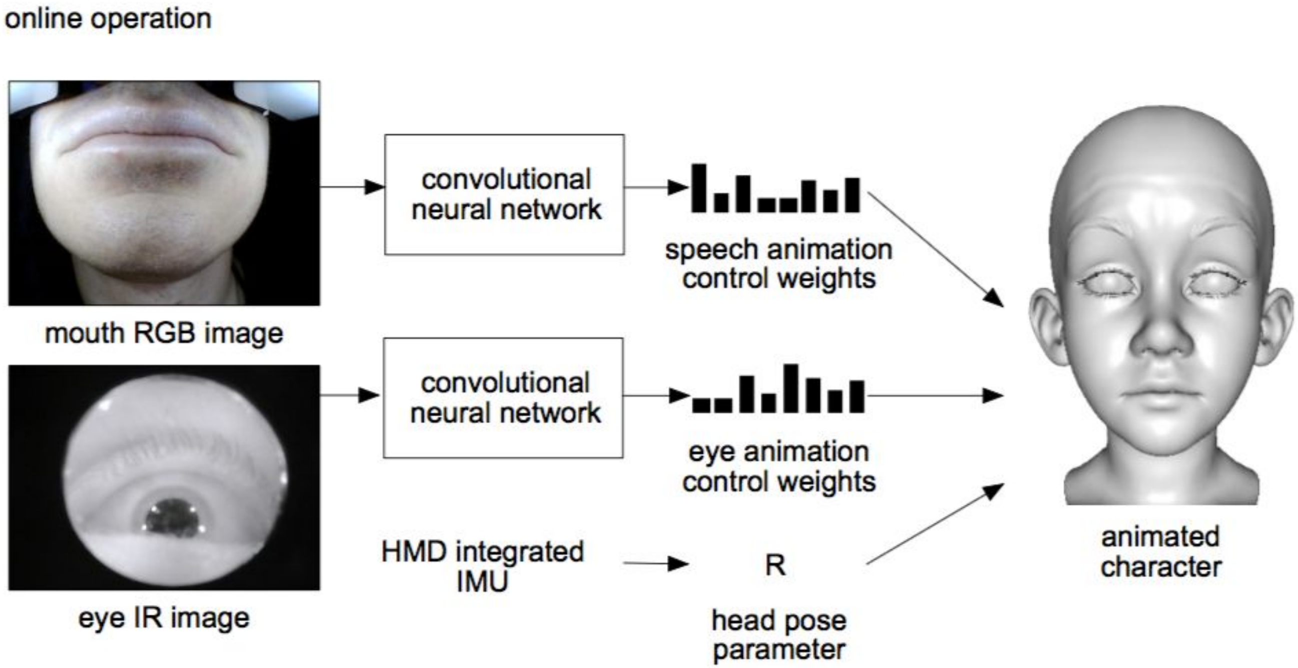
Olszewski et al. 2016

System Prototype

- HMD
  - eyes: infrared cameras
    - 6 IR LEDs
    - 60 fps with 320 x 240
  - head movement: gyroscope
  - mouth: Playstation Eye, modified with a 3.8mm lens
    - 30 fps with 640 x 480
    - 2 Streamlight Nano LED

Olszewski et al. 2016

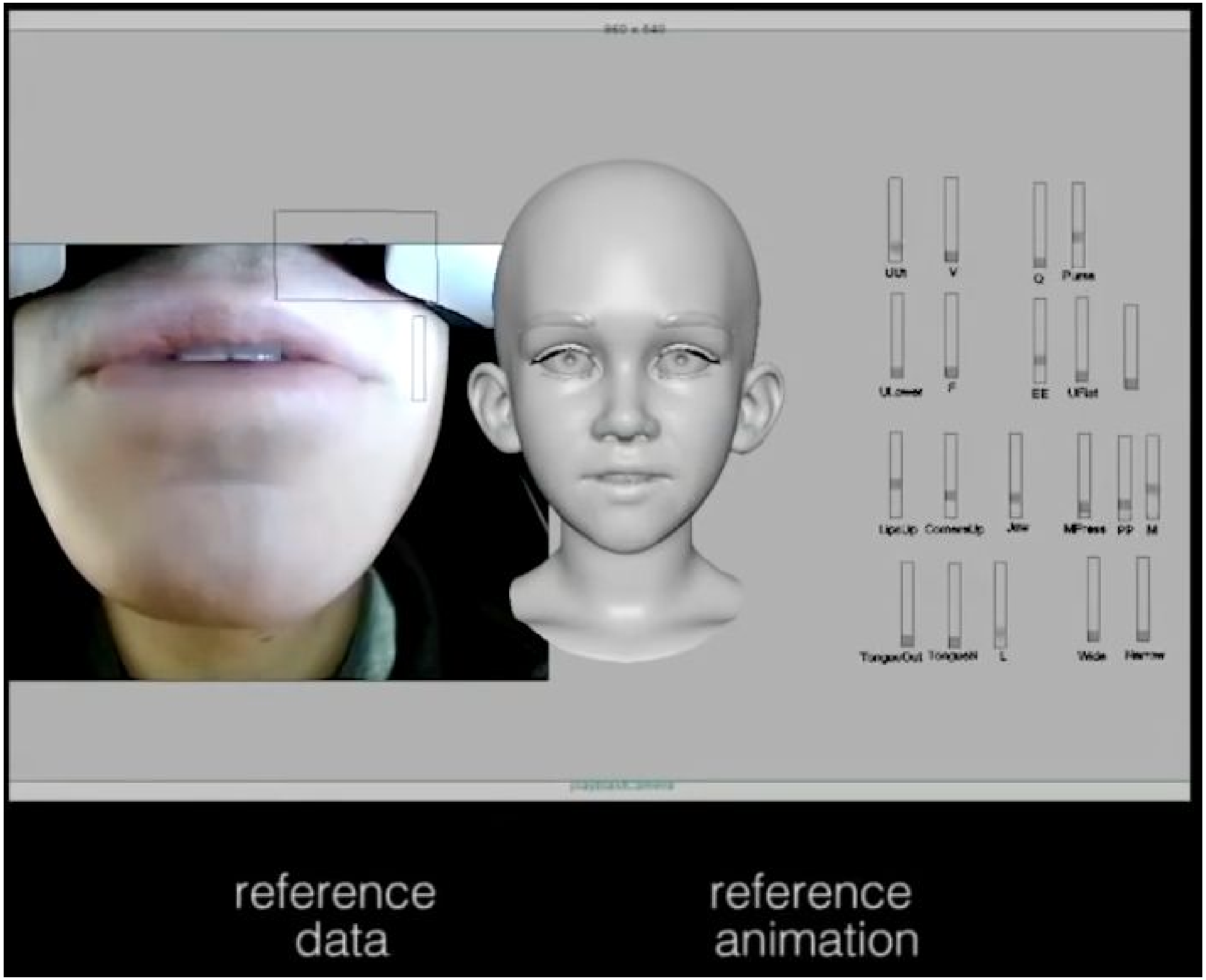
Strategy



- deep learning model (blendshape weights)
  - removes the need to set up the system for each user
  - trained with the use of recorded sequences

Olszewski et al. 2016

Strategy



[Y]

Olszewski et al. 2016

Data Collection

- video and audio recordings of 10 subjects
  - 30 sentences
- harvard sentences list
  - phonemes roughly same frequency as in the english language
- subjects perform 21 facial expressions based on the Facial Action Coding System (FACS)
- neutral expression to given expression and back
  - 2 iterations
- professional animation for each subject
  - upper and lower face

Olszewski et al. 2016

Deep Learning Model

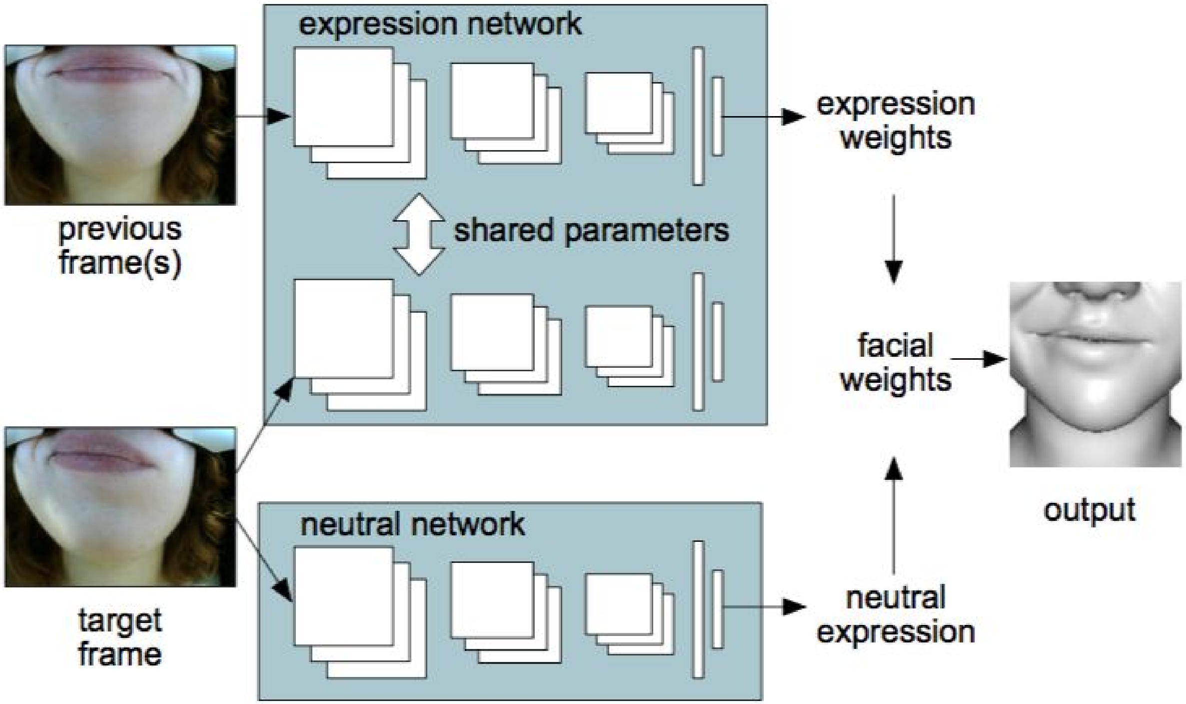
- blendshape model as meshes b = {b0,b1,··· ,bN}

$$f^t = \mathbf{b}_0 + \sum_i^N \mathbf{w}_i^t(\mathbf{b}_i - \mathbf{b}_0)$$

- defining a high-dimensional non-linear function is difficult
  - method has to handle large variations
    - occlusions
    - user identities
    - personal appearance
    - jittering
    - environmental changes

Olszewski et al. 2016

Deep Learning Model



- combination of neutral and non-neutral expression
  - less training data needed
  - interpretation from two different subnetworks

Olszewski et al. 2016

Deep Learning Model

input frame | without neutral network | with neutral network

- neural network essential for a stable result
- unexpected inputs can be better processed

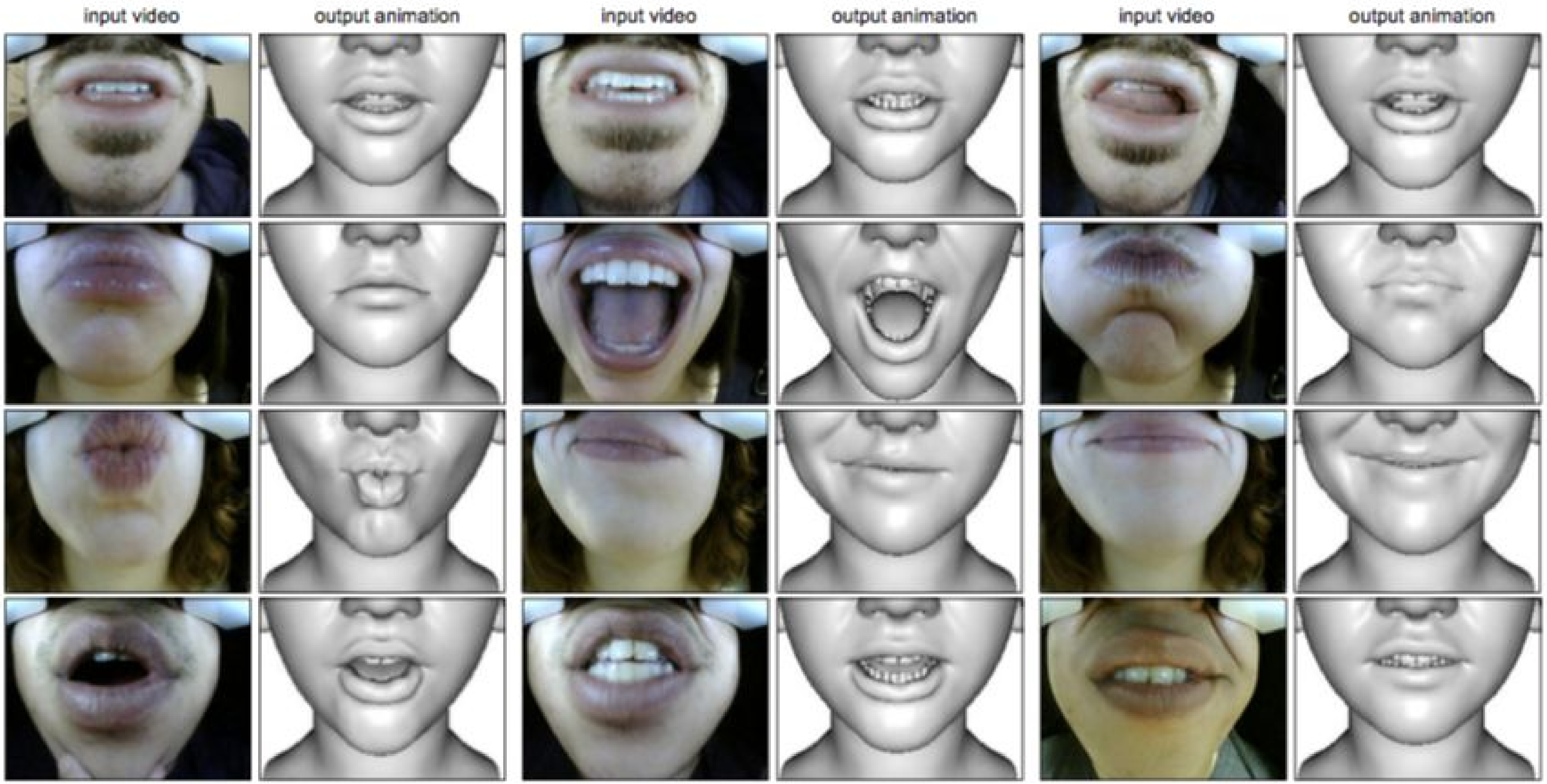
Olszewski et al. 2016

Results Test

- test scenarios
  - circumstances which were not in the training set
    - e.g. facial hair
  - sentences that were not covered
  - improvised facial expressions
    - sticky lip
  - different lighting/location
    - e.g. dark scene
  - sequence of images

Olszewski et al. 2016

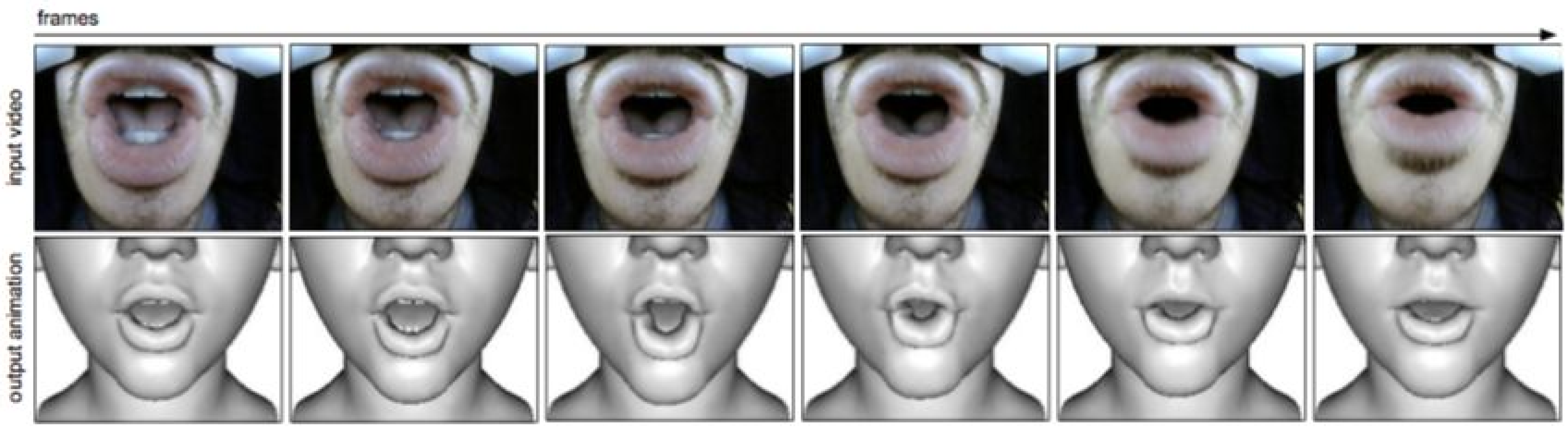
Results Test



- first row perturb orientation of HMD
- third and fourth row were not included in training set

Olszewski et al. 2016

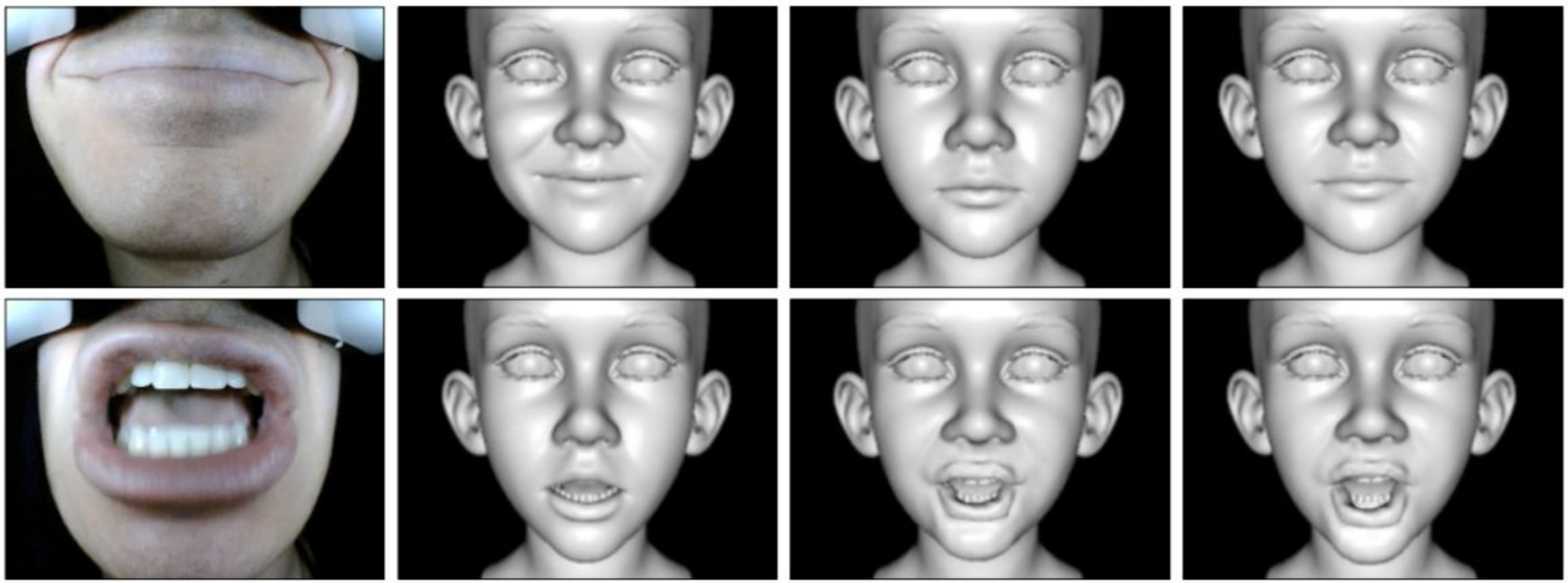
Results Test



- sequence of images of expression
- sticky lip, a deformation that challenges most performance capture techniques

Olszewski et al. 2016

Results User Study

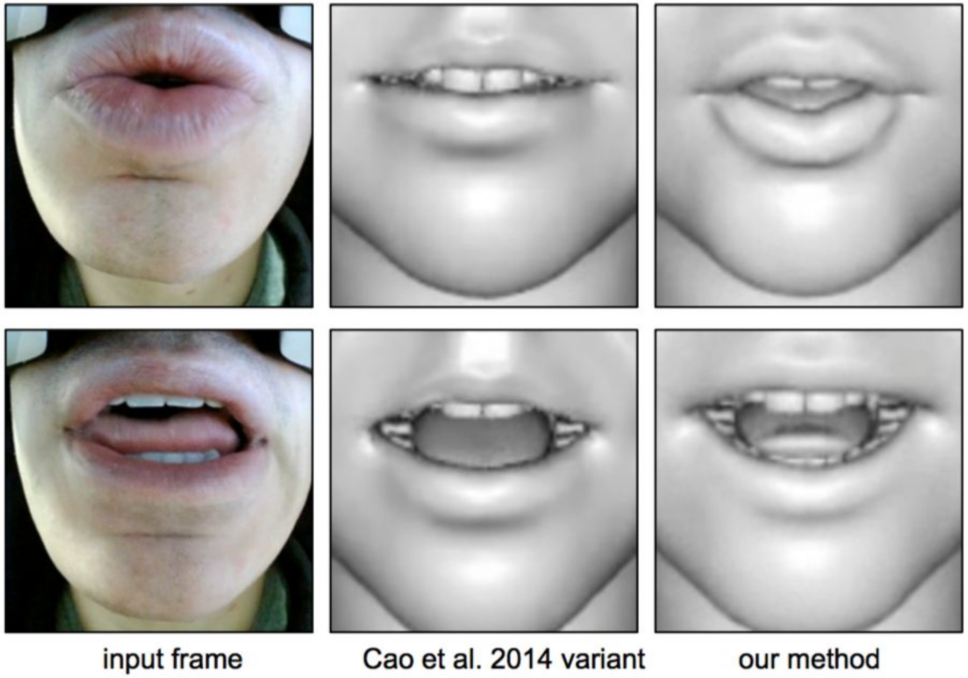


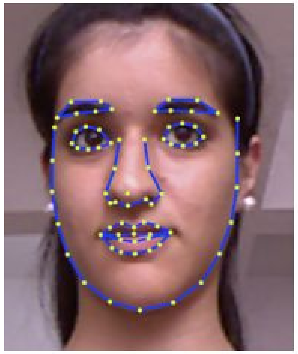
Sentence	FACS	Our full model
7.0%	21.8%	71.2%

- users were shown 4 synced videos
- combination leads to superior results

Olszewski et al. 2016

Results compared to other Applications





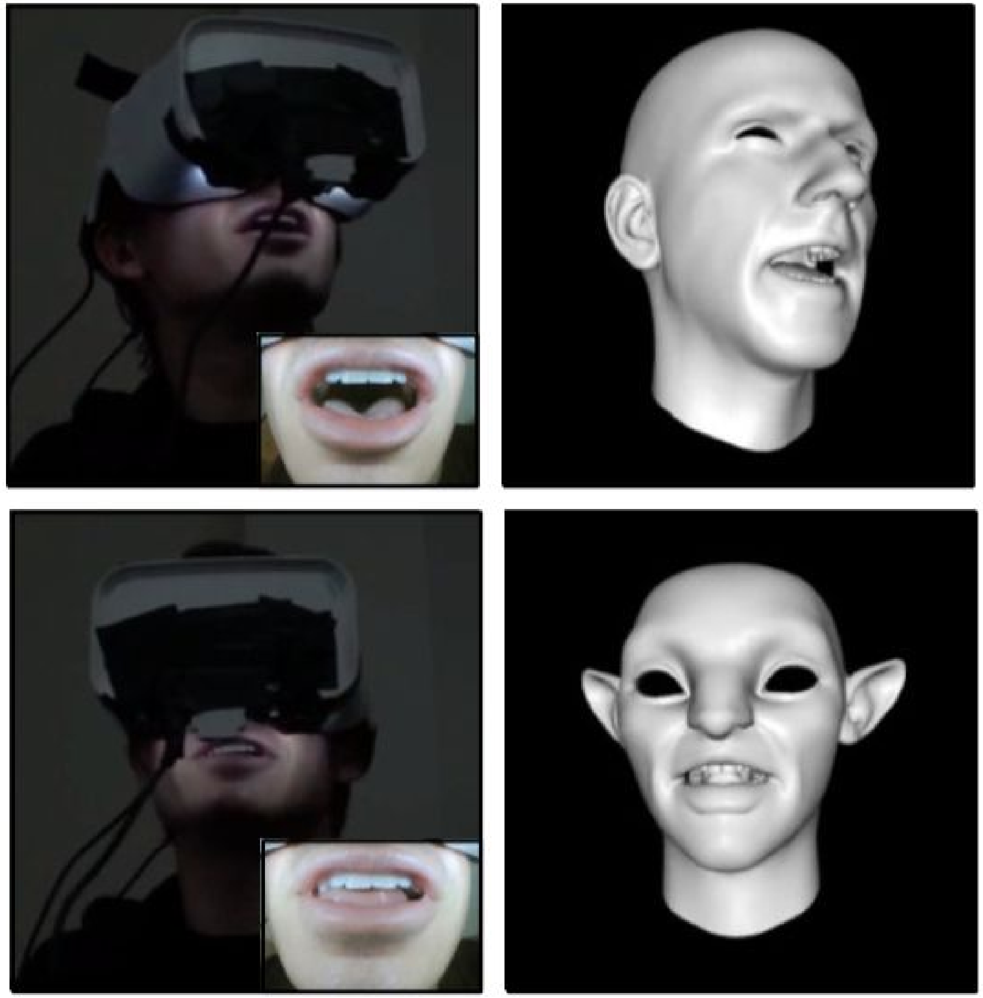
Cao et al. 2014

- Cao et al: needs no calibration for each user
  - uses only a single video camera

- training set re recorded with camera at a resolution of 1280 × 720 at 30 fps

Olszewski et al. 2016

Results



external view and mouth frame

output mesh

Olszewski et al. 2016

Conclusion

- system does not generate mesh from users face
  - maps facial expression to 3D mesh
- 3D artist needs to create users face
  - Avatar Digitization From a Single Image For Real-Time Rendering from Hu et al. 2017
  - otherwise usable for random characters
- extreme expressions?
  - (sticking out tongue etc.)

Olszewski et al. 2016

Conclusion

- requires no user-specific calibration
- cheap methode
- achieves high fidelity animations
- tongue is also tracked
- significant step towards enabling compelling verbal and emotional communication in VR

Olszewski et al. 2016

Conclusion

Olszewski et al. 2016

Reference

- [1] https://8ar.appearition.com/wp-content/uploads/2017/03/blog_illustations_VRDifferences_2016-06-_VR_AR_MR_-_Horizontal_Feature.png
- [2] https://www.oculus.com/rift/
- [3] https://www.vive.com/eu/
- [P] https://www.howtodrawit.com/img/cartoons25.jpg
- [D] https://www.vrs.org.uk/images/virtual-reality-immersion-1.jpg
- [Y] https://www.youtube.com/watch?v=eOjzC_NPCv8&t=510s
- BHAT, K. S., GOLDENTHAL, R., YE, Y., MALLET, R., AND KOPERWAS, M. 2013. High fidelity facial animation capture and retargeting with contours. In SCA ’13, 7–14.
- CAO, C., HOU, Q., AND ZHOU, K. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. 33, 4, 43:1–43:10.
- LI, H., TRUTOIU, L., OLSZEWSKI, K., WEI, L., TRUTNA, T., HSIEH, P.-L., NICHOLLS, A., AND MA, C. 2015. Facial performance sensing head-mounted display. ACM Transactions on Graphics (Proceedings SIGGRAPH 2015) 34, 4 (July).
- Olszewski, K., Lim, J. J., Saito, S., and Li, H. 2016. High-fidelity facial and speech animation for VR HMDs. ACM Transactions on Graphics (TOG), vol. 35, no. 6, p. 221.
- TOSHEV, A., AND SZEGEDY, C. 2014. Deeppose: Human pose estimation via deep neural networks. In IEEE Conference on Computer Vision and Pattern Recognition.
- WENG, Y., CAO, C., HOU, Q., AND ZHOU, K. 2014. Real-time facial animation on mobile devices. Graphical Models 76, 3, 172–179.
- Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. 2017. Avatar digitization from a single image for real-time rendering. ACM Trans. Graph. 36, 6, Article 195 (November 2017), 14 pages

Thank you