High-Fidelity Facial and Speech Animation for VR HMDs

Kyle Olszewski*‡ Joseph J. Lim† Shunsuke Saito*‡ Hao Li*‡§

*University of Southern California †Stanford University ‡Pinscreen
§USC Institute for Creative Technologies

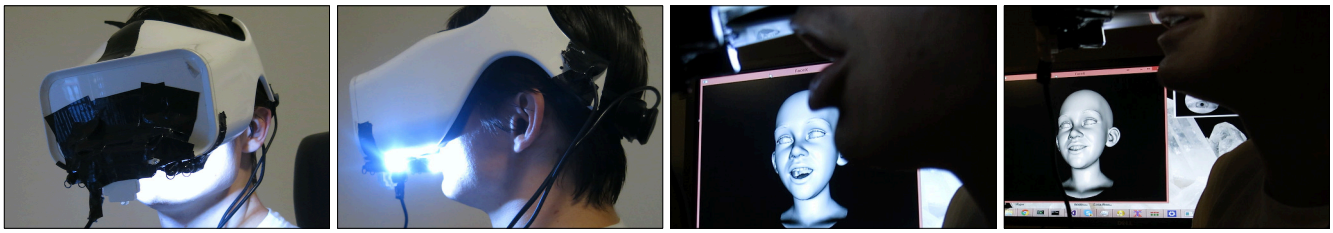

Figure 1: *A live demonstration of our system. We are able to obtain high-fidelity animations of the user's facial expressions in real-time using convolutional neural net regressors. Left: a user wearing our prototype system, which uses cameras attached to the HMD to track the user's eye and mouth movements. Right: a digital avatar controlled by the user.*

Abstract

Significant challenges currently prohibit expressive interaction in virtual reality (VR). Occlusions introduced by head-mounted displays (HMDs) make existing facial tracking techniques intractable, and even state-of-the-art techniques used for real-time facial tracking in unconstrained environments fail to capture subtle details of the user's facial expressions that are essential for compelling speech animation. We introduce a novel system for HMD users to control a digital avatar in real-time while producing plausible speech animation and emotional expressions. Using a monocular camera attached to an HMD, we record multiple subjects performing various facial expressions and speaking several phonetically-balanced sentences. These images are used with artist-generated animation data corresponding to these sequences to train a convolutional neural network (CNN) to regress images of a user's mouth region to the parameters that control a digital avatar. To make training this system more tractable, we use audio-based alignment techniques to map images of multiple users making the same utterance to the corresponding animation parameters. We demonstrate that this approach is also feasible for tracking the expressions around the user's eye region with an internal infrared (IR) camera, thereby enabling full facial tracking. This system requires no user-specific calibration, uses easily obtainable consumer hardware, and produces high-quality animations of speech and emotional expressions. Finally, we demonstrate the quality of our system on a variety of subjects and evaluate its performance against state-of-the-art real-time facial tracking techniques.

Keywords: real-time facial performance capture, virtual reality, communication, speech animation, eye tracking, head-mounted display

Concepts: •**Computing methodologies** → **Animation; Motion capture;**

SA '16 Technical Papers, December 05-08, 2016, Macao
ISBN: 978-1-4503-4514-9/16/12
DOI: http://dx.doi.org/10.1145/2980179.2980252

1 Introduction

Science fiction authors have excitedly envisioned immersive technologies that allow us to project our own digital avatars into captivating virtual worlds. Dramatic advancements in computer graphics and mobile display technologies have led to a remarkable revival of virtual reality, with the introduction of low cost consumer head-mounted displays, such as the Oculus Rift [Oculus VR 2014], the HTC Vive [HTC 2016], and the Google Cardboard [Google 2014]. Beyond immersive gaming and free-viewpoint videos, virtual reality is drawing wide interest from consumers and pushing the boundaries of next-generation social media platforms (e.g., High Fidelity, AltSpaceVR). We could mingle, discuss, collaborate, or watch films remotely with friends all over the world in a shared online virtual space. However, a truly immersive and faithful digital presence is unthinkable without the ability to perform natural face-to-face communication through personalized digital avatars that can convey compelling facial expressions, emotions, and dialogues.

State-of-the-art facial tracking methods commonly use explicitly tracked landmarks, depth signals in addition to RGB videos, or humans-in-the-loop. However, approaches directly using tracked landmarks to recover the full facial motion [Li et al. 2015] often suffer from occlusions. A tongue is invisible in many motions, and a large portion of the lips become invisible when an user bites her/his lips. In another approach, artists manually draw contours for all frames, and then solve a complex 3D model to fit the data [Bhat et al. 2013]. This is a very computationally intensive process and also suffers in the case of occluded regions.

While readily available body motion capture and hand tracking technologies allow users to navigate and interact in a virtual environment, there is no practical solution for accurate facial performance-sensing through a VR HMD, as more than 60% of a typical face is occluded by the device. Li and coworkers have recently demonstrated the first prototype VR HMD [Li et al. 2015] with integrated RGB-D and strain sensors, that can capture facial expressions with comparable quality to cutting edge optical real-time facial animation systems. However, it requires a tedious calibration process for each session, and the visual quality of lip motions during speech is insufficient for virtual face-to-face communication. Furthermore, while a personalized digital avatar model can be manually prepared by an artist, automatically digitizing an accurate 3D representation of a person's head is still very challenging.

Our objective is to enable natural face-to-face conversations in an

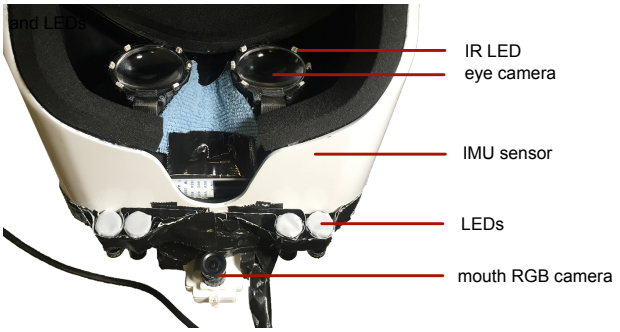

Figure 2: *VR HMD prototype with integrated eye and mouth cameras.*

immersive setting by animating highly accurate lip and eye motions for a digital avatar controlled by a user wearing a VR HMD. At the same time, we seek to improve on the ergonomic form factor of existing sensor-integrated headsets using lightweight hardware components, and eliminate the per-user calibration process present in existing methods.

To this end, we introduce a new learning framework for user-driven speech animation suitable for virtual reality. We employ a prototype (shown in Figure 2), which we modified based on a commercially available VR HMD (FOVE) with built-in eye cameras. This new prototype is more ergonomic and simpler compared to previously introduced designs [Oculus VR 2014; Google 2014; Li et al. 2015]. Our HMD device has attached cameras that see direct views of the user's mouth and eyes, and reduces the variation in the input caused by changes in the ambient illumination and the user's pose. Thus, our learning method can focus on specific subspace of facial motions with a strong prior, as the data obtained by our device will be from a lower dimensional subspace.

Recently, deep learning algorithms have shown promising results in many classification and regression tasks in computer vision [Krizhevsky et al. 2012; Toshev and Szegedy 2014]. One key strength of deep learning methods is that they are capable of learning and optimizing high-dimensional functions and are robust to various appearance changes. In this paper, we present a deep learning framework that can extract high fidelity motions from videos in real-time. Unlike traditional approaches using tracked faces, our method learns a direct embedding function of a high-dimensional image to lower-dimensional animation controls of a 3D rigged character. Additionally, our framework is designed so that it can take multiple frames as input. This enables holistically aggregating temporal and spatial cues from a sequence of frames. Our directly learned mapping enables extracting high fidelity facial motions in real time, while being robust to partial occlusion of regions of interest (such as the tongue), changes in the ambient lighting, and variations in the positioning of the HMD on the user's head.

A crucial aspect of our framework is the manner in which we collect the data used to train the deep convolutional networks in our system. This data consists of images of various users performing salient expressions with their mouths and eyes, and the corresponding animation parameters required to control a digital avatar. Our training data explicitly includes not only simple expressions used to convey basic emotions, but also the more complex mouth expressions required to produce plausible speech animation.

Our results demonstrate that this framework can efficiently produce such animations with a high degree of fidelity unmatched by any existing real-time performance-based facial animation techniques suitable for emerging use cases such as interaction and communication in virtual environments.

We thus present the following contributions:

- a regression method using deep convolutional networks that produces realistic facial expressions and visual speech animations. We introduce important techniques for improved accuracy and robustness.
- a set of training data that we use for this learning approach, acquired using the data collection framework described below, making use of images of eye and mouth FACS expressions, viseme samples from the Harvard psychoacoustic sentences, and the corresponding animation parameters. We intend to make this dataset publicly available for research purposes.
- a lightweight VR HMD prototype system with integrated mouth and eye cameras. Our solution does not require any user-specific calibration and offers a deployable solution for compelling and fully immersive face-to-face communication.

2 Previous Work

For over two decades, facial performance capture techniques have been developed in the graphics and vision community to facilitate the production of compelling 3D facial animation [Parke and Waters 1996; Pighin and Lewis 2006]. While striving for increased tracking fidelity and realism, production-level methods often rely on complex capture equipment [Guenter et al. 1998; Zhang et al. 2004; Li et al. 2009; Weise et al. 2009; Bradley et al. 2010; Beeler et al. 2011; Bhat et al. 2013; Fyffe et al. 2014] and intensive computations [Garrido et al. 2013; Shi et al. 2014; Suwajanakorn et al. 2014]. Even though several of these method are adopted by leading visual effects studios, the final animations are often reinforced with manual key-framing. These fine adjustments are mostly performed around mouth and eye regions, which provide critical emotional cues.

Eye and mouth animation. Motivated by these requirements, Bermano et al. [2015] have recently developed a dynamic reconstruction framework for eyelids using a deformation model that can reproduce self-occlusions due to intricate skin folds. While very convincing results are possible, this offline technique requires a multi-view camera setup and manual steps. Eye gaze directions also form a critical component for expressive facial expressions. A state of the art gaze estimation technique using a deep neural regressor is presented in [Zhang et al. 2015]. Despite its importance for verbal communication and numerous research advances in lip tracking [Basu et al. 1998], it is still very challenging to produce convincing speech animations because of the complex lip and tongue interactions. To achieve realistic lip-syncing and co-articulation effects during speech, the use of audio signals have been explored extensively to drive visual facial control parameters as an alternative to optical sensing [Bregler et al. 1997; Brand 1999; Massaro et al. 1999; Ezzat and Poggio 2000; Chuang and Bregler 2005; Wang et al. 2006; Deng et al. 2006; Xie and Liu 2007; Wampler et al. 2007; Taylor et al. 2012; Fan et al. 2015]. However, it is impossible to capture non-verbal mouth expressions using only audio signals.

Real-time facial animation. Facial tracking from pure RGB input is the most widely deployed technique for capturing performances. Data-driven methods based on active appearance models (AAM) [Cootes et al. 2001] or constrained local models [Cootes et al. 2001] have been introduced to detect 2D facial landmarks in real-time. Fully automatic techniques that do not require any user-specific training such as the regularized landmark meanshift method of [Saragih et al. 2011] or the supervised descent algorithm

of [Xiong and De la Torre 2013] have been recently proposed. While the mapping of sparse 2D facial features to the controls of complex 3D facial models has been explored [Chai et al. 2003], only coarse facial expressions can be recovered. More recently, Cao et al. [Cao et al. 2013] developed a real-time system that can produce compelling 3D facial animations through a 3D shape regression technique from RGB videos. By directly regressing head motion and facial expressions, rather than regressing 3D facial landmarks and then computing the pose and expression from this data, [Weng et al. 2014] were able to attain high tracking performance and accuracy, allowing for implementations running in real-time on mobile devices. While both of the aforementioned techniques rely on user-specific data to train their shape regressors, Cao et al. further improved their technique to eliminate the extra calibration step [Cao et al. 2014] and increased the tracking fidelity to capture wrinkle-level details [Cao et al. 2015].

Long before consumer-level depth sensors such as Microsoft's Kinect were available, Weise et al. [2009] demonstrated the first high-fidelity facial animation system with real-time capabilities using a structured light system. With RGB-D cameras becoming mainstream, a long line of research has followed this seminal work, improving the utility of low quality depth maps using motion priors [Weise et al. 2011; Faceshift 2014], removing the need for an extra facial model construction stage [Li et al. 2013; Bouaziz et al. 2013], and increasing the tracking fidelity using 2D facial features detected in the RGB channels [Li et al. 2013; Chen et al. 2013; Hsieh et al. 2015]. Though these RGB and depth sensor-based performance capture techniques can be integrated in non-occluded regions of a VR headset [Romera-Paredes et al. 2014; Li et al. 2015], none of them support compelling facial speech animation, despite the adoption of per-vertex Laplacian deformers in [Li et al. 2013; Hsieh et al. 2015] for improved expression tracking. Lately, Liu et al. [Liu et al. 2015] presented a state of the art real-time facial animation framework based on RGB-D and audio input using a speaker-independent acoustic deep neural network model. Even in the presences of background noise, they demonstrated superior animation output than audio and video signals alone, but the resulting lip motions are still far from production-level fidelity.

Wearable facial sensing systems. Facial performance capture using wearable, contact-based sensors began in the 1990s [Character Shop 1995]. Since VR HMDs occlude a large part of the upper face, contact-based sensors are potentially viable solutions for facial capture with highly constrained visibility. Non-optical measurement devices such as electroencephalograms (EEG) sensors have been used in [McFarland and Wolpaw 2011] to record brain activities to detect facial expressions and emotions, but extensive training and user concentration is required. Lucero and Munhall [1999] developed a system based on non-invasive electromyograms (EMG) to map muscle contractions to a physically-based facial model [Terzopoulos and Waters 1990]. Gruebler and Suzuki [2014] integrated EMGs into a wearable device to detect coarse emotional facial expressions for therapeutic purposes. EMG signals typically suffer from muscular crosstalk and their reliability depend on the placement of sensor locations and the subject's fat tissue. Tactile methods based on piezoelectric sensors have also been incorporated in smart glasses for facial expression recognition [Scheirer et al. 1999], but are not suitable for driving facial controls, since static states, such as keeping the mouth open, cannot be measured.

Facial sensing cameras that are mounted on HMDs have been first introduced for eye-gaze measurements [Huang et al. 2004; Steptoe et al. 2010]. Eye gaze tracking cameras that are directly integrated inside VR headsets have been recently deployed as commercial solutions [Fove 2015; SMI 2014] for an alternative input interface with virtual environments and realistic foveated rendering. Our system uses the Fove HMD [2015] to map video recordings of the eye region to animation controls of an avatar. While we animate movements of the upper face region such as squints and brow movement, gaze tracking is beyond the scope of this work. We note, however, that our system could be used in conjunction with the aforementioned solutions for gaze tracking to animate the movements of the eyes themselves.

Romera-Parades et al. [Romera-Paredes et al. 2014] attempted to track the entire face with a head-mounted display with partially-observing inward looking cameras. Their training data are collected using a separate RGB-D sensor and linear blendshape models obtained from the facial animation software Faceshift [2014]. As in this work, they also adopt a deep learning framework for regression, but only poor results could be demonstrated even with user-specific training data. Instead, our system generates highly compelling speech animation by mapping the video input from cameras directed at the regions of interest directly to the appropriate facial expression controls. Furthermore, we remove the need for user-specific training data by training our system on users of varying appearance.

Recently, Li et al. [2015] proposed the first VR HMD system to enable fully immersive face-to-face communication. While the deformations in the upper face regions are captured using ultra-thin flexible electronic materials that are mounted on the foam liner of the headset, the mouth performance is captured using a depth sensor mounted on the HMD and a cutting edge facial animation framework [Hsieh et al. 2015]. For every incoming frame, the expression shapes of a blendshape model are optimized to fit the user's lower face and the coefficients are transferred to a target avatar model for animation. While realistic facial expressions can be obtained, the system produces poor lip motions during speech and requires a complex calibration procedure for each subject.

3 Overview

System Prototype. Our system is based on a prototype of the FOVE VR HMD, with integrated eye tracking cameras and our custom mounted camera for mouth tracking. The HMD contains infrared (IR) cameras directed at the user's eyes and 6 IR LEDs (940nm wavelength) surrounding each eye, allowing the cameras to observe the user's eyes despite the occlusion from ambient illumination. The cameras runs at 60 fps at a resolution of 320x240. The cameras' field of view allows for tracking movements such as blinks as well as movements of the region surrounding the eye, such as squints and movement of the eyebrows. The HMD is also equipped with a gyroscope, allowing for the tracking of the user's head orientation.

The user's mouth is recorded with a Playstation Eye, modified to use a 3.8mm lens and enclosed in a mount attached to the underside of the HMD, placed approximately 7 cm from the user's mouth. This camera was used to record 640×480 RGB images of the user's mouth at 30 fps. Despite the close range, the field of view of this camera allows for tracking the user's full lip region even when the mouth is wide open. 2 Streamlight Nano LED lights, each producing 10 lumens of light, are attached on each side of the camera and directed at the mouth. Diffusion cloth is attached to the front of each light to reduce the intensity and sharpness of the emitted light. With a length of 1.47 inches and weighing 10.2 grams, they add little to the weight and size of the HMD, while reducing variation in the lighting and thus removing the dependence on ambient illumination. The system works both in fully illuminated rooms and in complete darkness, as demonstrated in the accompanying video.

We note that our system is more ergonomic than that of [Li et al. 2015], which required a depth camera to be placed on a mount attached to the front of the HMD at a significant distance from the

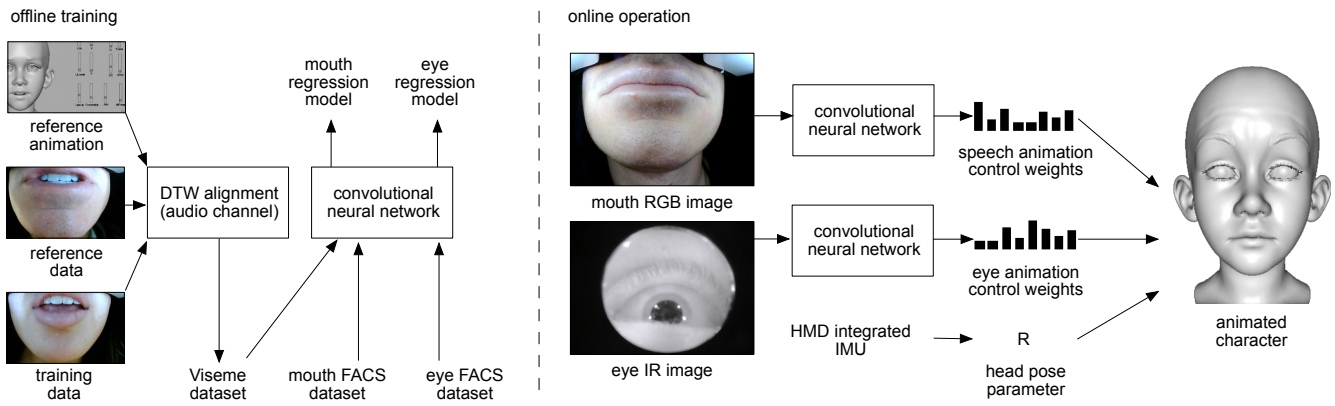

Figure 3: *An overview of our system. Left: Regressors used to control the avatar's movements are trained using recorded sequences from the internal and external camera and the corresponding blendshape weights. Right: These networks are then used to obtain appropriate blendshape weights for the upper and lower face regions in real-time using live input from the cameras.*

user's mouth in order to obtain usable depth data from the sensor's structured light pattern.

Facial Animation Pipeline. Our pipeline is illustrated in Figure 3. We use a standard blendshape model to control the avatar's expressions. During online operation, 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 the lower face region. A separate CNN regresses images from the eye camera to obtain blendshape weights for the eye region. The orientation of the head is obtained from the HMD's IMU.

To train the mouth region regressor to produce these blendshape weights, videos were recorded of several subjects reciting the same set of phonetically balanced training sentences. 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) [Sakoe and Chiba 1978] on the audio signal from these recordings, the other training sentence videos were aligned to this user's performance, allowing the blendshape weights from this animation sequence to be used for all subjects. This alleviates the need for producing high-quality speech animation sequences for each subject. Each user was also asked to perform a small set of relatively simple facial expressions based on the Facial Action Coding System. Animation data was provided by the artists for each of these sequences.

The eye animation dataset was obtained in a similar manner, with subjects performing some simple movements in the eye region also corresponding to FACS expressions. Corresponding animations were generated by the artists.

These eye and mouth images and their corresponding blendshape weights were then used to train the regressors for the separate facial regions, as described in Section 5.

While the audio signal is used to align the training images to the animated sequences, it is not used as input to directly train our system or during online operation to generate the animation sequences. This allows our system to remain fully functional in the presence of silence, such as during non-verbal expressions, or in the presence of common occurrences such as extreme background noise or crosstalk. In the supplementary video, we demonstrate the use of our system under such conditions.

Note that, for most of the subsequent figures, we show the lower and upper facial expressions separately and without the rigid motion obtained from the HMD's IMU to allow for easier evaluation of the resulting expressions. Several example images of full facial expressions with rigid head motion can be found in Section 6. Fully animated video sequences, including live demos under a variety of ambient lighting conditions, are in the supplementary video.

4 Data Collection

Creating a system that allows arbitrary users to control a digital avatar with no user-specific calibration and using only video input requires an appropriate collection of training data to be used as a prior for regressing a user's facial expression to the corresponding expression of the avatar. This data must account for variations in the facial expressions made by a given user and the appearances of different users. For our system, we thus sought to obtain training images of HMD users making a variety of salient mouth expressions, including those that are poorly captured by existing optical tracking techniques, and the parameters required to produce an expression corresponding to each image for the avatar.

Visual Speech Dataset. To produce truly plausible speech animation, the training set to be used must model the effects of coarticulation, which cause the facial expressions of one pronouncing a given phoneme (referred to as a *viseme*) to vary based on the context, as a user subconsciously prepares for the following phoneme and recovers from the previous one [Taylor et al. 2012]. As such, a model trained by simply having users recite a set of phonemes in isolation would not properly account for the subtle variations in users' expressions that are necessary for speech animation.

To this end, we 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. The video was recorded by the mounted camera, while the audio was captured using an external microphone at 48 KHz. These sentences were chosen from the Harvard sentences [Harvard 1969], a list of sample sentences in which phonemes appear at roughly the same frequency as in the English language. The sentences were chosen such that each of these phonemes appeared in a variety of contexts (prior and subsequent phonemes) such that a variety of coarticulation effects were captured.

Each subject was also asked to perform a set of 21 facial expressions with their mouth based on the Facial Action Coding System [Ekman and Friesen 1978]. As it is difficult for many subjects to perform

these actions in complete isolation from the others, the subjects were given the flexibility to perform actions only roughly corresponding to each of these expressions. They were recorded as they performed 2 iterations of each expression, going from a roughly neutral expression to the given expression, holding it for roughly 1 second, and then returning to the neutral expression. To make our training data more robust to variance in local illumination, we allowed subjects to move their heads arbitrarily while being recorded.

Animation Parameters. As described above, producing high-quality animation sequences based on captured data is a challenging task. Even state-of-the-art, computationally expensive offline processes fail to produce speech animation sequences of sufficient quality for our needs. Furthermore, we note that even state-of-the-art multi-view stereo performance capture methods such as [Beeler et al. 2011] focus on capturing the surface of the face, and thus fail to capture important details such as tongue motion that are crucial for speech animation, as the interior of the mouth is either partially or completely occluded from each viewpoint. As such, manual assistance from animators familiar with the full dynamics of facial movement, including tongue motion, is always required to produce truly high-quality speech animation sequences based on actual performances.

We employed 3 professional animators to create animation sequences for a blendshape model of a digital character corresponding to each subject's performance of each of these FACS-based eye and mouth region expressions. As most of these expressions could be animated using only a few of the blendshape weights available in the rig we employed, this process required little time and effort from the animators. The individual video images of the subject performing these expressions and their corresponding blendshape weights provided us with a set of training data for the regressors used in our system (Section 5).

Label Transfer via Audio Alignment. Speech animation is a particularly challenging and time-consuming task for an artist, given the subtle details that are required to create a sequence of mouth movements that plausibly correspond to an audio recording. The artists we employed reported that the time required to animate the entire series of 21 FACS-based mouth expressions for a given subject was comparable to that required to animate a single sentence lasting roughly 2-3 seconds. As such, creating individual animation sequences for each of the 300 recordings obtained from the subjects was intractable. However, given that synchronized video and audio of each subject was recorded speaking the same set of training sentences, we were able to exploit the coherence between these recordings to expand and generalize our training set using audio-based alignment of these sequences.

The animators produced high-quality animation sequences corresponding precisely to the mouth motion of a single subject (referred to as the "reference subject") for each frame of their recitation of the Harvard sentences. Given that these weights correspond to utterances of the same sentences spoken by other subjects, they can be used as the animation parameters for the other subjects as well. However, given that the rate at which the subjects spoke these sentences varied significantly, directly transferring these weights to the other subjects' sequences is not possible.

We thus employed a technique based on dynamic time warping, an algorithm used to measure the similarity of 2 time series, $a = \{a_0, a_1, \cdots, a_M\}$ and $b = \{b_0, b_1, \cdots, b_N\}$, that contain similar content but that may vary in speed and duration. It is commonly employed in the area of speech recognition to account for the variation in the rate at which utterances are spoken. By using some measure of distance between the segments of one series with each segment of the other, a local similarity matrix can be constructed. Dynamic programming can then be used to find a path through this matrix (i.e., starting at (a_0, b_0) and stepping forward in either or both series at each iteration until (a_M, b_N) is reached) which minimizes the cost of this traversal.

For each subject's utterance of each training sentence, a Short-time Fourier Transform (STFT) is performed on overlapping regions of the audio signal. The cosine distance between the STFT magnitudes of this utterance and those of the audio signal for the reference subject's utterance of this sentence are used to construct the similarity matrix. The minimum-cost path through this matrix then provides a mapping that can be used to interpolate the blendshape coefficients of each frame of the reference subject's recording such that they are appropriately aligned with the frames of the other subject's recording. Examples images from video sequences aligned using this approach and the corresponding reference animation data can be seen in Figure 4.

We note that this alignment process does not capture the entire variation in the appearance of each subject as they recite these sentences, caused by their unique physiology and the specific dynamics of their performance, such as how much they open their jaw when pronouncing a word. However, we observed that this approach can be used to produce appropriate training data for a system to produce natural and plausible speech animation for a digital avatar controlled by an arbitrary user.

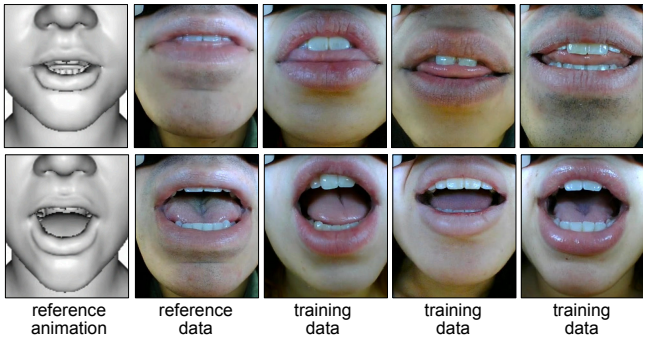

Figure 4: *Automatic alignment of training data to a reference sequence with corresponding animation curves.*

Eye Region Dataset. To obtain training data for controlling the upper face motion of our avatar, we 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. The animators then produced corresponding animations of the character. These image sequences and the corresponding blendshape weights were then used in as training data for the eye expression regressor (Section 5).

Facial Target Rig. The animation rig used to control the avatar seen in our experiments is controlled by a total of 57 blendshapes for the upper and lower face. 29 of these correspond to specific mouth motions commonly made during speech, including 5 shapes controlling the movement of the tongue. 21 correspond to the FACS-based mouth expressions the subjects were asked to perform during data collection, while 7 were used to control the movements of the eye region. In our results, we demonstrate the use of the blendshape values produced by our system to control rigs for other characters with corresponding blendshapes.

5 Deep Learning Model for Facial Expression

Our goal is to recover detailed 3D facial expressions from video frames. In this paper, we address this problem by representing a face as a set of facial blendshape meshes. Then, our algorithm determines the blendshape weights that best correspond to each frame. More concretely, let us assume we are given a generic blendshape model as a set of meshes $\mathbf{b} = \{\mathbf{b}_0, \mathbf{b}_1, \cdots, \mathbf{b}_N\}$. Our target expression of frame I^t at time t can be formulated as:

$$f^t = \mathbf{b}_0 + \sum_i^N \mathbf{w}_i^t(\mathbf{b}_i - \mathbf{b}_0), \tag{1}$$

where $\mathbf{b}_0$ is the neutral face expression, and $\mathbf{w}^t \in [0,1]^N$ is a corresponding blendshape weight vector. Then, our goal is to determine the value for $\mathbf{w}$ that best corresponds to a given image or video frame of a human face. In this paper, we formulate this as learning a mapping function ψ, which predicts a blendshape weight vector of image I^t, that minimizes

$$L(\psi) = \sum_t \|\psi(I^t) - \mathbf{w}^t\|_2^2. \tag{2}$$

For the rest of this section, we omit t wherever clear.

Despite many works [Cao et al. 2013; Li et al. 2015] on mapping 3D blendshapes to facial expression, there are many challenges that prevent us from obtaining very detailed facial expressions. For example, we need to find a non-linear mapping between a facial expression and blendshape weights. Defining and optimizing a high-dimensional non-linear function is a non-trivial task. It is especially tricky because the method has to handle large variations caused by occlusions (e.g. the tongue), user identities, personal appearance (e.g. growing a beard), jittering due to HMD movement, and environmental changes (e.g. lighting). Furthermore, we want to eliminate any type of calibration per user and maintain high fidelity for speech expressions.

In this work, we capitalize on recent developments in deep convolutional neural networks (CNN) [Krizhevsky et al. 2012; Simonyan and Zisserman 2014a; LeCun et al. 1998]. CNNs have attained impressive results for numerous classification and regression tasks in computer vision and robotics. They excel at optimizing high-dimensional non-linear functions and are robust to large variations including occlusions. Hence, we formulate the facial motion correspondence problem as a regression task using a CNN learning framework. In short, the CNN framework will find function parameters such that it optimizes a loss function (e.g. Equation 2) for all training images.

Here, we use a multi-frame CNN model. This is a common choice compared to other approaches, such as recurrent neural networks, for action recognition and video processing [Simonyan and Zisserman 2014b; Wang et al. 2015], because it has only a minor impact on performance while substantially reducing the training complexity.

In order to enable high fidelity facial speech animation, our CNN-based method builds upon two main ideas: (1) exploiting a sequence of frames to capture temporal signals, and (2) explicitly dampening an estimation toward the neutral face when necessary because humans are sensitive to neutral expressions. Our model thus contains two sub-networks. The expression network is a regression function that estimates facial expression weights from a sequence of frames. On the other hand, the neutral network is a classification function that determines whether the expression in the frame is neutral or not. The output of this network is used for dampening the estimation toward the neutral expression. The final output of our model, $\psi(\cdot)$ combines two network outputs to predict the final blendshape weights. As

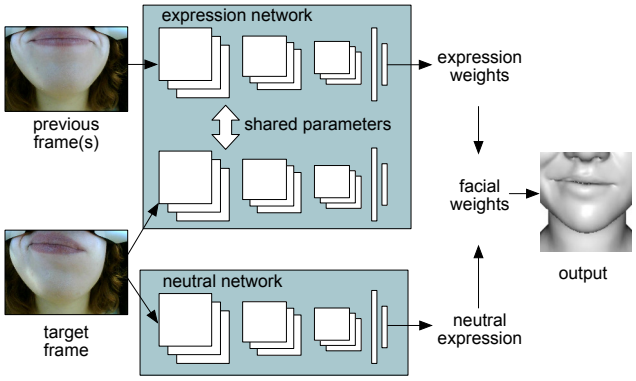

Figure 5: *The model trained to regress blendshape weights for the lower face. Our model has two sub-CNNs so it can incorporate temporal signals and control the sensitivity to important expressions.*

addressed in [Shalev-Shwartz and Shashua 2016; Lake et al. 2016], the two main advantages of compositing semantically abstracted subnetworks (i.e. expression and neutral) are the ability to train the model with an exponentially smaller amount of training data and explicit interpretability. Our full model is illustrated in Figure 5.

Expression Network. The goal of the expression network is to regress the blendshape weights that correspond to the target input frame. More concretely, our goal is to learn a function, $\psi_E(I^t; \theta_E)$, with the network parameters (or filter values), θ_E, that maps an image to its best corresponding blendshape weight vector $\mathbf{w}^{t*}$. We use the L_2 loss function for obtaining optimal network parameters:

$$\arg\min_\theta \sum_t \|\psi_E(I^t; \theta) - \mathbf{w}^t\|_2^2 \tag{3}$$

In fact, one advantage of using a frame sequence is that we can exploit temporal information. To achieve this, we follow a common practice in action recognition and further modify the model so that it takes more than one frame as an input. Hence, our CNN architecture will be trained to find θ_E such that it optimizes:

$$\arg\min_\theta \sum_t \|\psi_E(I^t, I^{t-k}; \theta) - \mathbf{w}^t\|_2^2. \tag{4}$$

In our experiment, we use two frames with $k = 3$, which showed significantly superior results compared to single-frame input.

Each stream of our network takes two multi-scaled frames as an input. Each input frame is scaled to 65×58 and 33×29. Using multiple image resolutions allows our network to learn parameters that correspond to features at different scales in the input images. Then, each stream applies a chain of layer operations to scaled images, which can be described as: 6×6 convolutional, ReLU, 2×2 max pooling, 6×6 convolutional, and ReLU layers. Then, the outputs from the two images are concatenated and again are applied to two 1000-dimensional fully-connected layers. We use this architecture for both the expression networks used to animate the mouth and eye regions (though the mouth expression network takes 3-channel RGB images as input, while single-channel grayscale images from the internal IR camera are used for the eye expression network).

Neutral Face Network. We found that humans are sensitive to neutral expressions on faces. Hence, we designed our model to pay extra attention to neutral expressions in the mouth region. As our dataset primarily contains images of the subjects' faces while in motion or making various facial expressions, neutral faces form a very small portion of the training set used for the mouth expression network. While we experimented with the inclusion of additional samples of neutral faces in the training set, we found that the best results (accurate depictions of emotional and speech expressions as well as neutral faces) were attained using the combination of two networks. We train a separate CNN-based function, $\psi_N(\cdot)$ that detects whether a facial expression in the target frame is neutral or not. The architecture of the neutral network is similar to that of the expression network above. One key difference is that the neutral network finds the network parameters θ_N by optimizing the softmax loss function (i.e. classification):

$$-\sum_t \Big[y_t \log(\psi_N(I^t;\theta)) + (1-y_t)\log(1-\psi_N(I^t;\theta))\Big], \quad (5)$$

where y_t indicates whether I^t is neutral or not.

The final output of our full model, $\psi(\cdot)$, is:

$$\psi(I^t) = \frac{1}{1+e^{-\alpha(\psi_N(I^t;\theta_N)+\beta)}}\ \psi_E(I^t, I^{t-k};\theta_E) \quad (6)$$

where α and β are free logistic function parameters that control the weight between neutral and expression networks. In our experiments we fix these parameters to the default values (i.e. $\alpha = 1$ and $\beta = 0$).

Training. We use the data generated in Section 4 to train our networks. A standard stochastic gradient descent algorithm can learn the function ψ and all parameters θ_E, θ_N, α, and β by optimizing Equations 2 and 6. However, we leave α and β as control parameters, in order to provide users the freedom to control the dampening factor based on each individual's sensitivity.

Our model parameters are then optimized by back-propagation in a distributed online implementation. For each mini-batch of size 64, adaptive gradient updates are computed. The learning rate starts from 0.01 and it gradually decreases to 0.0001.

Data Augmentation. When training CNNs, data augmentation is necessary to train with a relatively small dataset. We augment the data using a large number of random translations, rotations, and rescalings. Note that we do not apply left-right flips because blendshape weights are not symmetric.

6 Results

Our networks were implemented using the Caffe framework [Jia et al. 2014], which provides tools facilitating the design, training, and use of CNNs, as well as the use of GPUs to accelerate the training process. The system was tested with a variety of subjects under different circumstances, including some used in the training set and others who were not. For some tests, the user was asked to recite sentences from sets of the Harvard sentences that were not used in the original training set. For others, users were asked to improvise a variety of facial expressions or statements, or to have a dialogue with another person. The system was tested in a typical office environment with standard ambient illumination as well as in a dark room in which the HMD LED lights were the only source of illumination. Subjects were able to use the system one after another with no user-specific calibration in between sessions.

Some sample images demonstrating the variety of mouth and eye expressions our system is able to animate can be seen in Figure 6. The first and second rows portray users who were included in the training set, while the users in the third and fourth rows were not. The results demonstrate that our system is able to animate a wide variety of facial expressions related to speech and emotional expressions, including important expressive details such as a user smiling while talking. Other subtle details that are not attainable with existing real-time facial tracking techniques, such as motion of the tongue, are also seen in these images.

We also note that our system is robust to some variation in the positioning of the HMD and the user's appearance. In some cases, the user removed and replaced the HMD with slightly different positioning, or was asked to manually perturb the orientation of the HMD (Figure 6, first row), only to have the operation continue as before. Though no images of users with facial hair were included in the original training set, our system produced high-quality results on a user with a significant amount of facial hair, as seen in the first row of Figure 6. In another session, comparable results were achieved when the same subject used the system with no facial hair.

The last row of Figure 6 shows sample images of the results achieved with our eye tracker. We note that it is able to animate salient movements such as squinting, lowering and raising of the brow, and blinks with high fidelity. Thus, our system allows for animating the full face of an HMD user, despite the occlusion introduced by the display, which would make such animation infeasible for systems relying on a single camera to track the full face.

While these images provide examples of specific expressions produced by our system, as our primary goal is expressive speech animation, the results are best demonstrated on continuous facial motion. Figure 7 shows the results of our system on a sequence of images. Though no temporal smoothing was applied to the blendshape weights obtained from our regressors, the system smoothly animates the avatar's face an expressive manner corresponding to the user's expressions. This figure also demonstrates the results obtained with our system on the challenging sticky lip, an effect which is difficult to attain with existing real-time facial animation techniques.

For examples on longer sequences including speech with the corresponding audio, as well as live demonstrations of the use of our system with rigid head tracking, we ask the reader to refer to the supplementary video.

Comparisons. We evaluate our system by comparing it to several state-of-the-art methods for real-time facial tracking and animation. Figure 10 shows a sample of how our results compare to those obtained using an implementation of [Li et al. 2015], a system designed for tracking the face of an HMD user. A user wore each system while reciting the same set of sentences. We applied the blendshape weights obtained from each system to the avatar and compared the resulting expressions for similar expressions corresponding to the same points in the test sentences. Since both animations are produced using two different capture settings, we synchronize the input using dynamic time warping [Sakoe and Chiba 1978]. Our method does not use depth data, yet it produces results that are much closer to the user's expressions than those obtained when using a method requiring RGB-D data from a mounted depth camera.

We also compare our method to the original implementation of [Cao et al. 2014]. As this method is designed to work with unoccluded views of the user's full face, we asked the subjects used for data collection to recite the same training and testing sentences without wearing the HMD while their full face is recorded with an external camera at a resolution of 1280×720 at 30 fps. The subject's head remains fixed relative to the camera with negligible rigid motion.

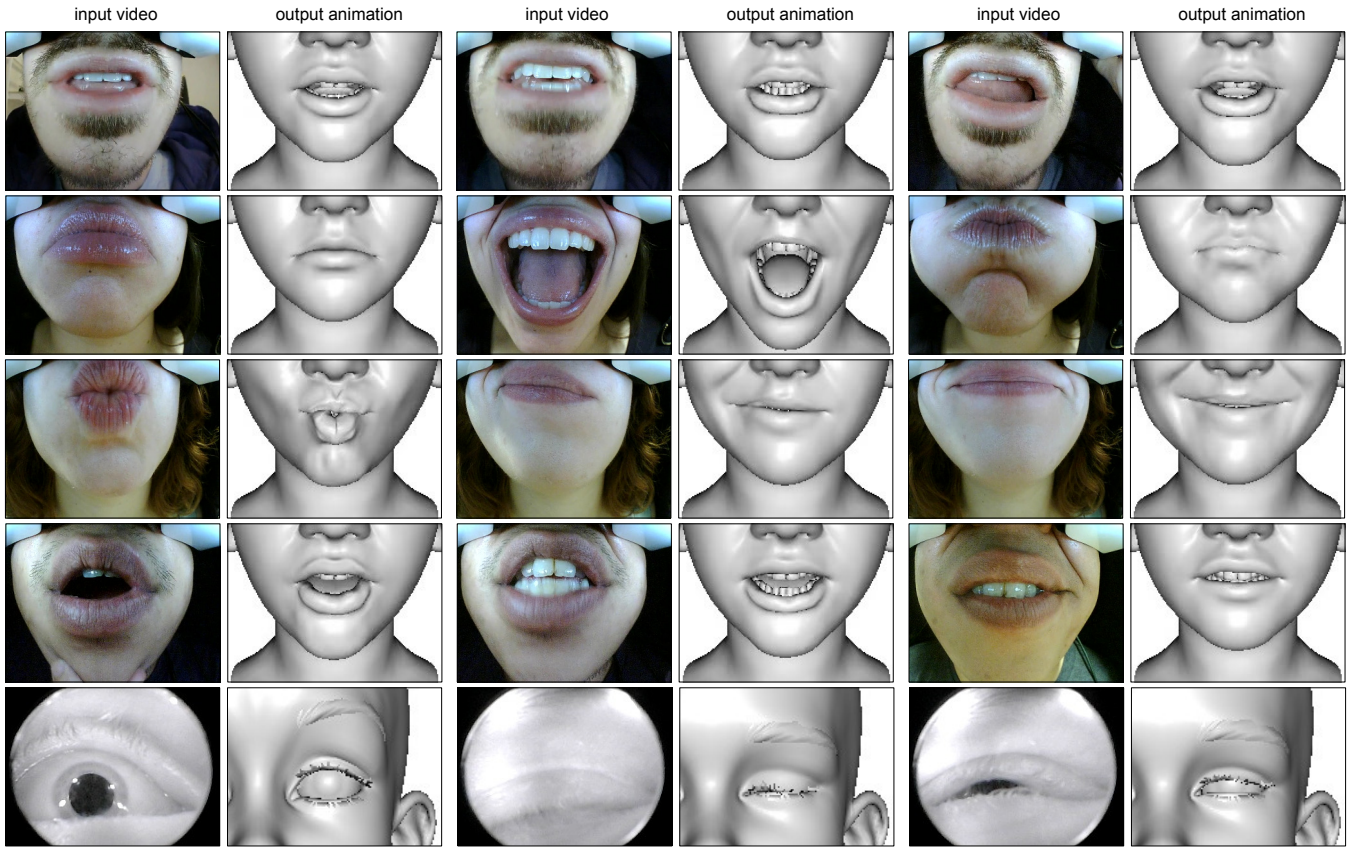

Figure 6: *Results obtained with our system on a variety of users.*

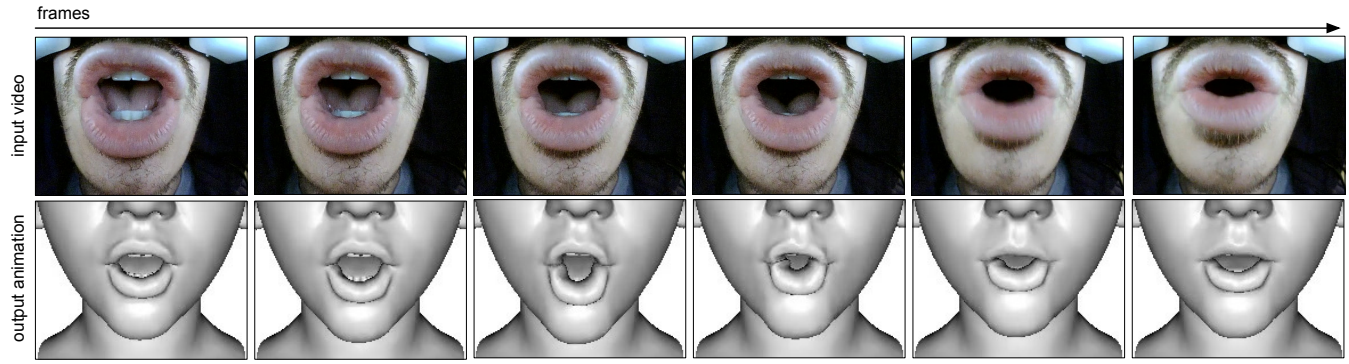

Figure 7: *Here we show performance capture results for the sticky lip, a deformation that challenges most performance capture techniques.*

We use the audio signal to align this data with the training animation sequences. Using a cropped region (856×642, resized to our network's input dimensions) of the video around the user's mouth, we trained our mouth expression and neutral face networks on this input data.

Figure 11 shows several example images of the results obtained using both methods on several test sequences. As the blendshape model used by [Cao et al. 2014] differs from that used in our system, we display their results using a personalized model of the user generated from FaceWarehouse, while ours are displayed using a model generated using the method of [Hsieh et al. 2015], with blendshapes corresponding to our original model generated using example-based facial rigging [Li et al. 2010]. For fairness, we display the results both with and without the mouth interior, to demonstrate the results both with and without the partially occluded tongue motion that we are able to animate using our system. These results demonstrate that our system produces expressions that more closely match the user's facial expressions on these speech sequences.

We note, however, that [Cao et al. 2014] has several advantages compared to our system. For example, it can track the rigid pose of the user's head as its position changes drastically relative to an external camera using only the video sequence, while our method requires that the user's head position varies little relative to the camera position and requires the IMU on the HMD for rigid pose tracking. Though the use of sparse facial landmarks for tracking limits the overall fidelity of the expressions that their method can

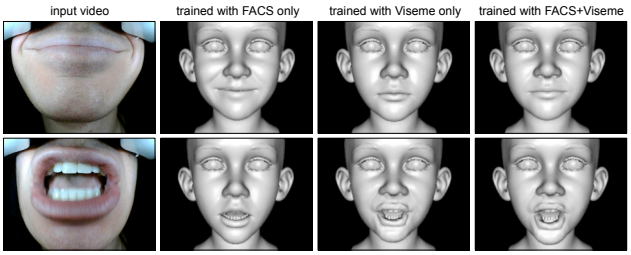

Figure 8: *We show that the use of FACS and viseme dataset achieves the best performance.*

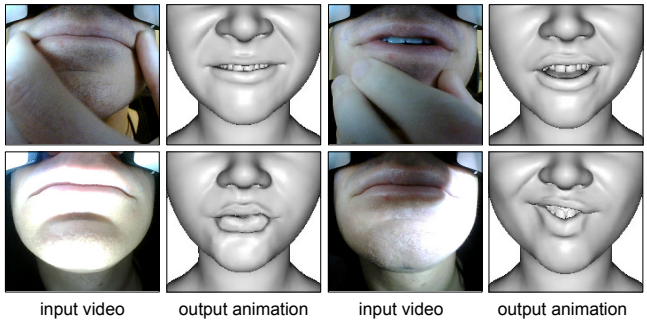

Figure 9: *We evaluate the robustness of our system. Extreme occlusions and illumination can cause the system to fail to produce appropriate animations.*

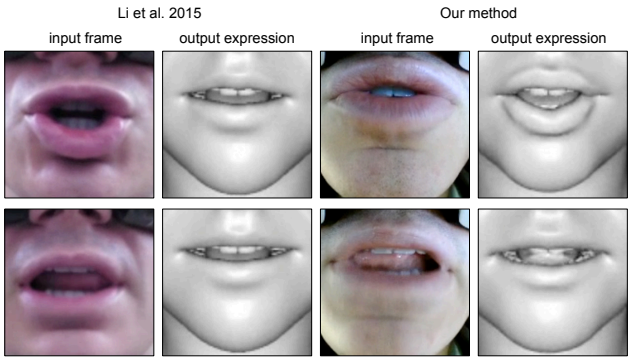

Figure 10: *Comparison with [Li et al. 2015]. The user wore each system and recited a set of sentenced that were aligned using Dynamic Time Warping. We show the resulting expressions for frames corresponding to the same utterance.*

track, it is more computationally efficient, allowing for real-time implementations on modern mobile devices. Our approach, however, uses a GPU to perform convolutions on the input images at runtime, making it more computationally intensive. Furthermore, while our use of multiple partial views of the user's face allows for animating the upper and lower facial expressions of HMD wearers whose faces are partially occluded from either internal or external cameras, their method can track the user's upper and lower facial expressions using a single video sequence of a face, provided that it is mostly unoccluded from the camera.

As the blendshape models and training data used in our experiments differ from those used in the aforementioned works, we performed further evaluations in which this data was used to train alternate regressors that were then tested on the same input. The results

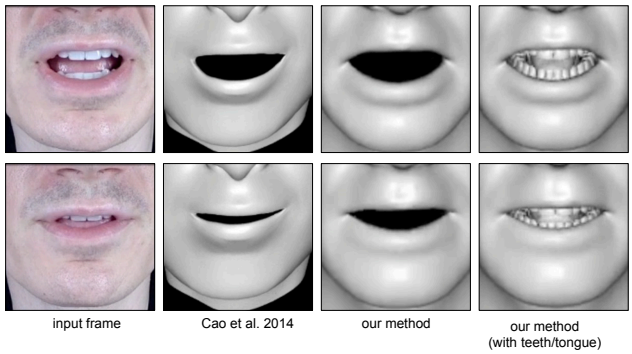

Figure 11: *Comparison with the original implementation of [Cao et al. 2014]. For fairness, we show examples both with and without the mouth interior to compare our results to theirs, which can animate the face surface but not tongue motion.*

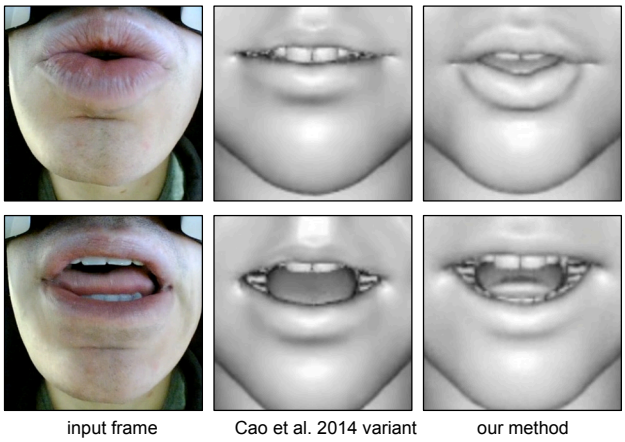

Figure 12: *Comparison with a modified implementation of [Cao et al. 2014], adapted to only track the mouth expressions given a fixed head pose and user identity.*

demonstrate that our approach is able to attain superior results compared to state-of-the-art real-time facial animation techniques.

Figure 12 demonstrates how our method compares with the DDE regressor [Cao et al. 2014] modified to track only the lower face for a specific person. A customized model corresponding to the user's identity was obtained in a preprocessing step. Then a person specific regressor was trained with 45 images containing the same FACS and viseme expressions as ours and 18 manually annotated landmark points on the mouth, following the original implementation of [Cao et al. 2014] for the training parameters. We found adding more images for training results in no improvement in terms of tracking accuracy. The user identity parameters, camera focal length and rigid head pose were fixed, while the expression blendshape weights were regressed for each frame. These weights were then applied to our model, and the resulting expressions were compared to those obtained using our system. The former are comparable with those obtained using [Li et al. 2015], while ours more closely match the subject's expression.

To provide a more direct comparison between these approaches to regressing blendshape parameters, we performed further evaluations in which both our system and the modified implementation of [Cao et al. 2014] were trained using the same training images and

reference animation data for the depicted user. For the [Cao et al. 2014] implementation, mouth contour landmarks in each training image were labeled by Amazon Mechanical Turk users, and the corresponding artist-generated blendshape values for each image were used as ground truth values to train the regressor to recover the expression parameters, with the rigid motion and identity parameters fixed as described above. Sample images and more details on the training process can be found in the supplementary document, and sample video sequences can be seen in the supplementary video. Our method is able to capture salient details missed by the [Cao et al. 2014] implementation, as well as the user's tongue motion.

Furthermore, we note that regressing directly to the animation parameters as in our approach does not suffer from a significant drawback of approaches that explicitly track the user's face based on depth data, facial landmarks, or a combination thereof, namely that a well-trained CNN can still produce plausible animation parameters even in the case of significant occlusion of the regions of interest. For example, our system can produce plausible animation parameters for the avatar's tongue even when the user's tongue is largely occluded by the user's lips and teeth, as seen in Figure 6. These occlusions present serious problems for approaches that assume that the regions of interest remain visible during tracking.

For further comparisons with the aforementioned approaches on video sequences, please consult the supplementary video.

Figure 13: *Effect of incorporating temporal information into our network structure by using the first and third previous frame as input. Using only the current frame as input causes transient details such as tongue motion to be missed or underrepresented.*

Figure 14: *Effect of the use of the neutral face network along with our mouth expression network. Using both together causes both neutral faces and expressions to be accurately animated.*

Network Structure Evaluation. We evaluated the efficacy of our chosen network structure by comparing the results against several variants. We found that overall these alternative approaches produced reasonable results for many cases, yet none of them consistently produced results comparable to those of our chosen network structure.

We evaluated our expression network against one using only the current frame as input, rather than using an additional earlier frame as described above. Figure 13 shows several examples of the effect this has on the resulting expressions. The structure of the network used for this single-frame architecture is otherwise identical to the aforementioned expression network, with the output of the stream being fed directly into the fully-connected layers without being concatenated with the output of the stream for the earlier frame. The results demonstrated that, while it typically produced reasonable results, the lack of temporal information caused this variant of the network to produce erroneous results for many frames, missing important but transient details such as quick tongue motion. It thus produced less plausible animations than our chosen architecture.

Figure 14 shows the results when we use our mouth expression network to animate the user's mouth expressions both with and without using the neutral face network. They demonstrate that we are able to capture neutral expressions more accurately using the combination of the two networks.

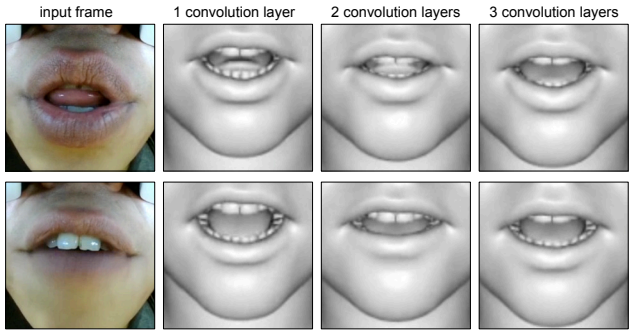

Figure 15: *Effect of using only different numbers of convolution layers for each input stream.*

We also evaluated our network's performance compared to variants using different numbers of convolution layers (Figure 15). Using only one convolution layer for each stream produced generally inferior results, capturing the overall expressions reasonably well but missing subtle details. This indicates that such a simple and shallow network is insufficient to recognize these small but crucial details.

Using an additional convolution layer for each stream also produced results that were worse than those seen when using only two convolution layers. Given the relatively small amount of data needed to train our network, using too many convolution layers may result in overfitting of the network parameters to the training data, leading to erroneous results on testing data. This may account for the inferior results produced by this variant of the network.

Comparisons of the results of these evaluations on video sequences can be found in the supplementary video.

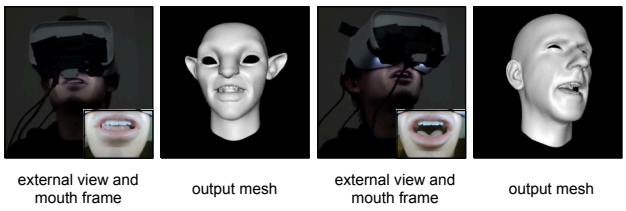

Figure 16: *Example results produced using retargeting to other characters. The left column contains external views of the user from a video roughly synced to the output animation.*

Retargeting. To demonstrate our approach on additional characters besides the one used to create the original animations, we used the example-based facial rigging method of [Li et al. 2010], with 10 training samples, to generate blendshape expressions based on additional models: an alien character, and one resembling the user (generated using the depth sensor-based modeling approach of [Hsieh et al. 2015]). Figure 16 and the supplemental video show examples from animation sequences using these characters.

Figure 17: *All test frames are sorted according to their predicted blendshape weights, $\mathbf{w}_k^t$. On the left, we show a particular blendshape we picked. On the right, the images are sorted by estimated weights of the picked blendshape in ascending order. This visualization displays a reasonable order, which indicates that our model can reliably predict the weights.*

Parameter Estimation. In addition to directly evaluating the rendering results, we also analyzed the estimated blendshape weights. First, we sort frames by their estimated magnitude from our model along one specific blendshape weight, $\mathbf{w}_k$. Figure 17 shows results when they are sorted by the openness of the jaw according to our model's predictions. It clearly shows that our method reasonably predicts how much the jaw is open (otherwise, the sorted results would appear in a random order).

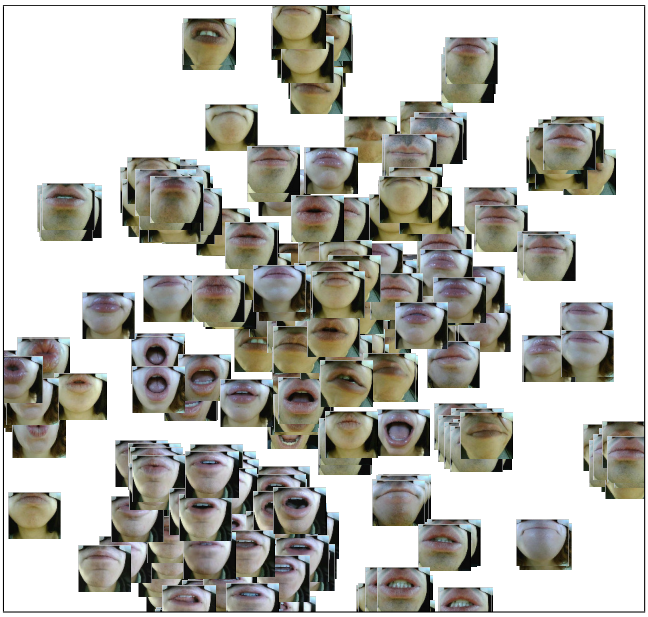

Figure 18: *We embed the estimated blendshape weight vector $\mathbf{w}^t$ of each frame I^t in a 2D space using the t-SNE method. We can observe that images are clustered by their mouth shapes. This shows that our method estimates stable and appropriate blendshape weights.*

Similarly, given estimated weights $\mathbf{w}^t$ for each frame I^t, we apply t-SNE [van der Maaten and Hinton 2008] to embed the high-dimensional vector in a 2D space. Figure 18 clearly shows that similar mouth shapes are clustered together. Both results provide strong evidence that our method learns meaningful network parameters and thus can predict appropriate blendshape weights.

User Study. Finally, in order to quantify the advantage provided by our approach in capturing subtle mouth motion, we ran a user study that compares 3 variants. Users were shown 4 synced videos: one input video containing facial motions and audio, and corresponding animation result videos, which were generated using 3 different methods. We compare our approach against 2 other training methods. One was trained using only the FACS data recorded from the 10 subjects, while the other was trained using only the data from their recitations of the Harvard sentences. Some examples of the differences between these 3 approaches are illustrated in Figure 8. The positions of the result videos were also swapped to remove any location bias. We then asked users to pick the result that recovered the facial motion most accurately. We queried 300 different Amazon Mechanical Turk users per video. The user study results are shown in Table 1. Our full model was significantly favored: more than 70% of users thought our method was the best, and our method had about 3 times as many votes as the second most favored method (FACS). Hence, we can conclude that combining these different types of data to train our model leads to superior results.

Sentence	FACS	Our full model
7.0%	21.8%	71.2%

Table 1: *We performed a user study in which viewers chose the best animation results among those generated using 3 different training methods. Each column shows how often a particular method was preferred over the others. Our method was substantially favored over the 2 other methods. More than 70% of users thought our final model was the best, and our method had about 3 times as many votes as the second most favored method (FACS).*

Limitations. Despite the impressive results demonstrated by our system, it does have several limitations. To produce high-quality results requires training the system with animation sequences that correspond very well with the associated training videos. We note that, while our system could be trained with animation sequences captured using more automated methods, such as those using multi-view stereo, using artist-generated data as we do allows for animating subtle but crucial expressions, such as tongue motion, that cannot be reliably captured using such techniques.

We demonstrate the feasibility of our approach using consumer-grade cameras such as the Playstation Eye. However, very fast mouth motion can lead to motion blur in the images captured using this camera, leading to artifacts in the result animations. We note that this problem could be alleviated using professional-grade high-speed cameras, although this would increase the system cost.

As we collected data from only a limited number of subjects, the system may not be robust to users whose appearance varies significantly from those in the training set. While we demonstrate that the system works on a user with some facial hair, despite the absence of such training data, larger variations in the appearance of a user, such as the growth of a full beard, may impair its accuracy. However, we expect that collecting more data from a wider variety of subjects would allow the CNN to adapt to address these issues.

While we demonstrate that minor hand-to-face interaction such as touching the chin does not interfere with the performance of our system (see Figure 6), our system was neither trained nor designed to handle extreme occlusions such as when much of the mouth

is covered by the hand, and thus our system fails to produce the correct mouth shape in these cases, as seen in Figure 9. While we demonstrate its ability to perform in standard ambient illumination and total darkness, extremely bright illumination (e.g., a powerful flashlight directed at the face) also causes erroneous results.

Performance. Our results were rendered using a framework running on an Alienware Area-51 equipped with an NVIDIA Titan X GPU, 16 GB RAM and a 6-core Intel Core i75930K Processor, running Windows 8.1. As Caffe does not officially support Windows, training and running each network was performed an identical system with 3 Titan X GPUs running Ubuntu 14.04. The blendshape coefficients were streamed to the rendering system using UDP.

Training each CNN takes approximately 24 hours using a single GPU. Each of our trained CNNs can process an image during online operation in no more than 1.6 ms. Our system's online performance is limited by our rendering framework, which is capable of achieving framerates up to 38 fps, although we used a video capture and rendering framerate of 30 fps for the results seen in our video.

7 Conclusion

We have presented a method for animating a digital avatar in real-time based on the facial expressions of an HMD user. Our system is more ergonomic than existing methods such as [Li et al. 2015], makes use of more accessible components, and is more straightforward to implement. Furthermore, it achieves higher fidelity animations than can be achieved using existing methods, and requires no user-specific calibration. As such, it makes a significant step towards enabling compelling verbal and emotional communication in VR, an important step for fully immersive social interaction through digital avatars.

Our approach regresses images of the user directly to the animation controls for a digital avatar, and thus avoids the need to perform explicit 3D tracking of the subject's face, as is done in many existing methods for realistic facial performance capture. Our system demonstrates that plausible real-time speech animation is possible through the use of a deep neural net regressor, trained with animation parameters that not only capture the appropriate emotional expressions of the training subjects, but that also make use of an appropriate psychoacoustic data set.

Future work. We hope to apply our techniques for speech animation to more general face tracking and animation scenarios. However, this introduces new challenges that must be addressed, such as changes in the head pose relative to the camera, wider variation in the lighting of the environment, and the occlusions introduced in less constrained settings.

Though we have achieved a significant degree of fidelity and expressiveness using only video as input, making this a very cost-effective solution, it is possible that superior results could be attained using a combination of input from various sources, such as depth and audio data captured in real-time. Techniques relying on sensor fusion, such as those explored in [Liu et al. 2015] may allow for achieving superior results when video data alone is insufficient to animate a speaking character, such as when occlusions or extreme lighting conditions make image-based tracking less reliable.

An open question to be addressed is how viable this system is as a tool for automatically capturing the emotional state of the user. It is possible that a system such as this could be used as an interface for communicating in a virtual environment with an AI agent which may react to the user's emotional expressions. It could thus be a powerful tool for many applications ranging from gaming to psychology.

Acknowledgements

The authors would like to thank David Rodriguez, Peter Birdsall, Matt Furniss, and Lyz Holder for their modeling and animation efforts. We would also like to thank Kun Zhou and Chen Cao, for their assistance in producing the comparisons; Qixing Huang, for his help in revising the paper; Kim Libreri of Epic Games, for providing the Kite Boy model; Yuka Kojima and Lochlainn Wilson from FOVE, for providing the hardware; Eric Whitmire and others at Oculus Research, for fruitful discussions and technical assistance; and the various subjects used for our video and audio recordings. This research is based upon work supported in part by Adobe, Oculus VR and Facebook, Sony, Pelican Imaging, Panasonic, Embodee, Huawei, the USC Integrated Media System Center, the Google Faculty Research Award, The Okawa Foundation Research Grant, the Office of Naval Research (ONR) / U.S. Navy, under award number N00014-15-1-2639, the Office of the Director of National Intelligence (ODNI), and Intelligence Advanced Research Projects Activity (IARPA), under contract number 2014-14071600010. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

BASU, S., OLIVER, N., AND PENTLAND, A. 1998. 3d modeling of human lip motion. In *ICCV*, 337–343.

BEELER, T., HAHN, F., BRADLEY, D., BICKEL, B., BEARDSLEY, P., GOTSMAN, C., SUMNER, R. W., AND GROSS, M. 2011. High-quality passive facial performance capture using anchor frames. In *ACM SIGGRAPH 2011 Papers*, ACM, New York, NY, USA, SIGGRAPH '11, 75:1–75:10.

BERMANO, A., BEELER, T., KOZLOV, Y., BRADLEY, D., BICKEL, B., AND GROSS, M. 2015. Detailed spatio-temporal reconstruction of eyelids. *ACM Trans. Graph. 34*, 4 (July), 44:1–44:11.

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.

BOUAZIZ, S., WANG, Y., AND PAULY, M. 2013. Online modeling for realtime facial animation. *ACM Trans. Graph. 32*, 4, 40:1–40:10.

BRADLEY, D., HEIDRICH, W., POPA, T., AND SHEFFER, A. 2010. High resolution passive facial performance capture. In *ACM SIGGRAPH 2010 Papers*, ACM, New York, NY, USA, SIGGRAPH '10, 41:1–41:10.

BRAND, M. 1999. Voice puppetry. In *SIGGRAPH'99*, 21–28.

BREGLER, C., COVELL, M., AND SLANEY, M. 1997. Video rewrite: Driving visual speech with audio. In *SIGGRAPH '97*, 353–360.

CAO, C., WENG, Y., LIN, S., AND ZHOU, K. 2013. 3d shape regression for real-time facial animation. *ACM Trans. Graph. 32*, 4, 41:1–41:10.

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.

CAO, C., BRADLEY, D., ZHOU, K., AND BEELER, T. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (TOG) 34*, 4, 46.

CHAI, J.-X., XIAO, J., AND HODGINS, J. 2003. Vision-based control of 3d facial animation. In *SCA '03*, 193–206.

CHARACTER SHOP, 1995. Facial waldo. http://www.character-shop.com/waldo.html.

CHEN, Y.-L., WU, H.-T., SHI, F., TONG, X., AND CHAI, J. 2013. Accurate and robust 3d facial capture using a single rgbd camera. In *ICCV*, IEEE, 3615–3622.

CHUANG, E., AND BREGLER, C. 2005. Mood swings: Expressive speech animation. *ACM Trans. Graph. 24*, 2, 331–347.

COOTES, T. F., EDWARDS, G. J., AND TAYLOR, C. J. 2001. Active appearance models. *IEEE TPAMI 23*, 6, 681–685.

DENG, Z., NEUMANN, U., LEWIS, J. P., 0002, T.-Y. K., BULUT, M., AND NARAYANAN, S. 2006. Expressive facial animation synthesis by learning speech coarticulation and expression spaces. *IEEE Trans. Vis. Comput. Graph. 12*, 6, 1523–1534.

EKMAN, P., AND FRIESEN, W. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto.

EZZAT, T., AND POGGIO, T. 2000. Visual speech synthesis by morphing visemes. *International Journal of Computer Vision 38*, 1, 45–57.

FACESHIFT, 2014. http://www.faceshift.com/.

FAN, B., WANG, L., SOONG, F. K., AND XIE, L. 2015. Photo-real talking head with deep bidirectional LSTM. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 4884–4888.

FOVE, 2015. http://www.getfove.com/.

FYFFE, G., JONES, A., ALEXANDER, O., ICHIKARI, R., AND DEBEVEC, P. 2014. Driving high-resolution facial scans with video performance capture. *ACM Trans. Graph. 34*, 1 (Dec.), 8:1–8:14.

GARRIDO, P., VALGAERT, L., WU, C., AND THEOBALT, C. 2013. Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph. 32*, 6, 158:1–158:10.

GOOGLE, 2014. Google cardboard. https://www.google.com/get/cardboard/.

GRUEBLER, A., AND SUZUKI, K. 2014. Design of a wearable device for reading positive expressions from facial emg signals. *Affective Computing, IEEE Transactions on 5*, 3, 227–237.

GUENTER, B., GRIMM, C., WOOD, D., MALVAR, H., AND PIGHIN, F. 1998. Making faces. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, ACM, New York, NY, USA, SIGGRAPH '98, 55–66.

HARVARD, 1969. Harvard sentences. http://www.cs.columbia.edu/~hgs/audio/harvard.html.

HSIEH, P.-L., MA, C., YU, J., AND LI, H. 2015. Unconstrained realtime facial performance capture. In *CVPR*, to appear.

HTC, 2016. HTC Vive. https://www.htcvive.com/.

HUANG, H., ALLISON, R. S., AND JENKIN, M. 2004. Combined head-eye tracking for immersive virtual reality. *ICAT'2004 14th International Conference on Artificial Reality and Telexistance*.

JIA, Y., SHELHAMER, E., DONAHUE, J., KARAYEV, S., LONG, J., GIRSHICK, R., GUADARRAMA, S., AND DARRELL, T. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

LAKE, B. M., ULLMAN, T. D., TENENBAUM, J. B., AND GERSHMAN, S. J. 2016. Building machines that learn and think like people. *arXiv preprint arXiv:1604.00289*.

LECUN, Y., BOTTOU, L., BENGIO, Y., AND HAFFNER, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (November), 2278–2324.

LI, H., ADAMS, B., GUIBAS, L. J., AND PAULY, M. 2009. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2009) 28*, 5 (December).

LI, H., WEISE, T., AND PAULY, M. 2010. Example-based facial rigging. *ACM Trans. Graph. 29*, 4, 32:1–32:6.

LI, H., YU, J., YE, Y., AND BREGLER, C. 2013. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph. 32*, 4, 42:1–42:10.

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).

LIU, Y., XU, F., CHAI, J., TONG, X., WANG, L., AND HUO, Q. 2015. Video-audio driven real-time facial animation. *ACM Trans. Graph. 34*, 6 (Oct.), 182:1–182:10.

LUCERO, J. C., AND MUNHALL, K. G. 1999. A model of facial biomechanics for speech production. *The Journal of the Acoustical Society of America 106*, 5, 2834–2842.

MASSARO, D. W., BESKOW, J., COHEN, M. M., FRY, C. L., AND RODRIGUEZ, T. 1999. Picture my voice : Audio to visual speech synthesis using artificial neural networks. In *Proceedings of International Conference on Auditory-Visual Speech Processing*, 133–138. QC 20100507.

MCFARLAND, D. J., AND WOLPAW, J. R. 2011. Brain-computer interfaces for communication and control. *Commun. ACM 54*, 5 (May), 60–66.

OCULUS VR, 2014. Oculus Rift DK2. https://www.oculus.com/dk2/.

PARKE, F. I., AND WATERS, K. 1996. *Computer Facial Animation*. A. K. Peters.

PIGHIN, F., AND LEWIS, J. P. 2006. Performance-driven facial animation. In *ACM SIGGRAPH 2006 Courses*, SIGGRAPH '06.

ROMERA-PAREDES, B., ZHANG, C., AND ZHANG, Z. 2014. Facial expression tracking from head-mounted, partially observing cameras. In *IEEE International Conference on Multimedia and Expo, ICME 2014, Chengdu, China, July 14-18, 2014*, 1–6.

SAKOE, H., AND CHIBA, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoust., Speech, and Signal Process. ASSP 26*, 43–49.

SARAGIH, J. M., LUCEY, S., AND COHN, J. F. 2011. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision 91*, 2, 200–215.

SCHEIRER, J., FERNANDEZ, R., AND PICARD, R. W. 1999. Expression glasses: A wearable device for facial expression recognition. In *CHI EA '99*, 262–263.

SHALEV-SHWARTZ, S., AND SHASHUA, A. 2016. On the sample complexity of end-to-end training vs. semantic abstraction training. *arXiv preprint arXiv:1604.06915*.

SHI, F., WU, H.-T., TONG, X., AND CHAI, J. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph. 33*, 6, 222:1–222:13.

SIMONYAN, K., AND ZISSERMAN, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556*.

SIMONYAN, K., AND ZISSERMAN, A. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 568–576.

SMI, 2014. Sensomotoric instruments. http://www.smivision.com/.

STEPTOE, W., STEED, A., ROVIRA, A., AND RAE, J. 2010. Lie tracking: Social presence, truth and deception in avatar-mediated telecommunication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '10, 1039–1048.

SUWAJANAKORN, S., KEMELMACHER-SHLIZERMAN, I., AND SEITZ, S. M. 2014. Total moving face reconstruction. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, 796–812.

TAYLOR, S. L., MAHLER, M., THEOBALD, B.-J., AND MATTHEWS, I. 2012. Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, SCA '12, 275–284.

TERZOPOULOS, D., AND WATERS, K. 1990. Physically-based facial modelling, analysis, and animation. *The Journal of Visualization and Computer Animation 1*, 2, 73–80.

TOSHEV, A., AND SZEGEDY, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

VAN DER MAATEN, L., AND HINTON, G. E. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*.

WAMPLER, K., SASAKI, D., ZHANG, L., AND POPOVIĆ, Z. 2007. Dynamic, expressive speech animation from a single mesh. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, SCA '07, 53–62.

WANG, G.-Y., YANG, M.-T., CHIANG, C.-C., AND TAI, W.-K. 2006. A talking face driven by voice using hidden markov model. *J. Inf. Sci. Eng. 22*, 5, 1059–1075.

WANG, L., XIONG, Y., WANG, Z., AND QIAO, Y. 2015. Towards Good Practices for Very Deep Two-Stream ConvNets. *ArXiv e-prints* (July).

WEISE, T., LI, H., GOOL, L. V., AND PAULY, M. 2009. Face/off: Live facial puppetry. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer animation (Proc. SCA'09)*, Eurographics Association, ETH Zurich.

WEISE, T., BOUAZIZ, S., LI, H., AND PAULY, M. 2011. Realtime performance-based facial animation. *ACM Trans. Graph. 30*, 4, 77:1–77:10.

WENG, Y., CAO, C., HOU, Q., AND ZHOU, K. 2014. Real-time facial animation on mobile devices. *Graphical Models 76*, 3, 172–179.

XIE, L., AND LIU, Z.-Q. 2007. A coupled hmm approach to video-realistic speech animation. *Pattern Recogn. 40*, 8 (Aug.), 2325–2340.

XIONG, X., AND DE LA TORRE, F. 2013. Supervised descent method and its application to face alignment. In *CVPR*, IEEE, 532–539.

ZHANG, L., SNAVELY, N., CURLESS, B., AND SEITZ, S. M. 2004. Spacetime faces: High-resolution capture for modeling and animation. In *ACM Annual Conference on Computer Graphics*, 548–558.

ZHANG, X., SUGANO, Y., FRITZ, M., AND BULLING, A. 2015. Appearance-based gaze estimation in the wild. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 4511–4520.