# Facial Performance Sensing Head-Mounted Display

Hao Li[†*] Laura Trutoiu[‡*] Kyle Olszewski[†*] Lingyu Wei[†*]
Tristan Trutna[‡] Pei-Lun Hsieh[†] Aaron Nicholls[‡] Chongyang Ma[†]

[†] University of Southern California [‡] Oculus & Facebook

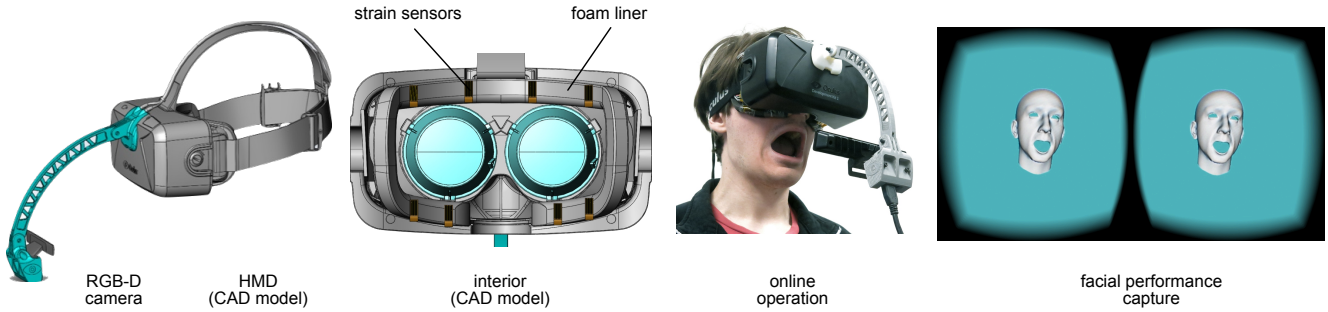



**Figure 1:** *To enable immersive face-to-face communication in virtual worlds, the facial expressions of a user have to be captured while wearing a virtual reality head-mounted display. Because the face is largely occluded by typical wearable displays, we have designed an HMD that combines ultra-thin strain sensors with a head-mounted RGB-D camera for real-time facial performance capture and animation.*

## Abstract

There are currently no solutions for enabling direct face-to-face interaction between virtual reality (VR) users wearing head-mounted displays (HMDs). The main challenge is that the headset obstructs a significant portion of a user's face, preventing effective facial capture with traditional techniques. To advance virtual reality as a next-generation communication platform, we develop a novel HMD that enables 3D facial performance-driven animation in real-time. Our wearable system uses ultra-thin flexible electronic materials that are mounted on the foam liner of the headset to measure surface strain signals corresponding to upper face expressions. These strain signals are combined with a head-mounted RGB-D camera to enhance the tracking in the mouth region and to account for inaccurate HMD placement. To map the input signals to a 3D face model, we perform a single-instance offline training session for each person. For reusable and accurate online operation, we propose a short calibration step to readjust the Gaussian mixture distribution of the mapping before each use. The resulting animations are visually on par with cutting-edge depth sensor-driven facial performance capture systems and hence, are suitable for social interactions in virtual worlds.

**CR Categories:** I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual reality;

**Keywords:** real-time facial performance capture, virtual reality, depth camera, strain gauge, head-mounted display, wearable sensors

[*] Authors on the first row have contributed equally.

**ACM Reference Format**
Li, H., Trutoiu, L., Olszewski, K., Wei, L., Trutna, T., Hsieh, P., Nicholls, A., Ma, C. 2015. Facial Performance Sensing Head-Mounted Display. ACM Trans. Graph. 34, 4, Article 47 (August 2015), 9 pages.
DOI = 10.1145/2766939 http://doi.acm.org/10.1145/2766939.


SIGGRAPH '15 Technical Paper, August 09 – 13, 2015, Los Angeles, CA.

DOI: http://doi.acm.org/10.1145/2766939

## 1 Introduction

Recent progress towards mass-market head-mounted displays (HMDs) by Oculus [Oculus VR 2014] and others, has led to a revival in virtual reality (VR). VR is drawing wide interest from consumers for gaming and online virtual worlds applications. With the help of existing motion capture and hand tracking technologies, users can navigate and perform actions in fully immersive virtual environments. However, users lack a technological solution for face-to-face communication that conveys compelling facial expressions and emotions in virtual environments. Because a user's face is significantly occluded by the HMD, established methods for facial performance tracking, such as optical sensing technologies, will fail to capture nearly the entire upper face.

To address this need, we develop a prototype HMD around an existing device. We augment the system with eight ultra-thin strain gauges (flexible metal foil sensors) placed on the foam liner for surface strain measurements and an RGB-D camera mounted on the HMD cover to capture the geometry of the visible face region. Aside from a slight increase in weight, our design unobtrusively integrates the sensors without further constraining user performance as compared to any standard virtual reality HMD.

Complex anatomical characteristics, such as individual facial tissue and muscle articulations, challenge the low dimensionality of our surface measurements across subjects. To map the input signals to a tracked 3D model in real-time, we first train a regression model by detaching the cover from the HMD to maximize visibility while the strain gauges are recording. This procedure is only performed once for each individual, and each subsequent use does not require unmounting the cover. Because of slight misplacements as well as the additional weight of the cover and the RGB-D camera, the sensitivity and measured surface locations can differ greatly between the training session and online operation (when the display is attached). For subsequent wearings by the same person, we propose a short calibration step that readjusts the Gaussian mixture distributions of the mapping [Gales 1998].

Like many real-time facial animation systems, our method uses linear blendshape models to produce output animations based on FACS expressions [Ekman and Friesen 1978]. The semantics of each blendshape mesh can be conveniently used for facial performance

retargeting to arbitrary digital characters. Although the upper face region is largely covered by the HMD and only a few statically integrated strain gauges are used, our system can produce 3D facial animations comparable to state of the art monocular real-time tracking systems based on RGB cameras [Cao et al. 2014; Cao et al. 2013] or depth sensors [Weise et al. 2011; Li et al. 2013; Bouaziz et al. 2013; Chen et al. 2013].

While non-optical sensing technologies (e.g., acoustic, electroencephalogram, electromyogram, piezoelectric) have been extensively explored in psychology, affective computing, and facial animation synthesis, our system is the first to demonstrate high-fidelity facial performance capture while wearing a VR headset. Our wearable device is also unique in that strain gauges are used to measure occluded face regions. Compared to other sensors, direct measurements of surface strain have low latency, do not suffer from complex muscular crosstalk, are suitable for non-verbal communication, and can accurately reproduce the linearity of human facial expressions. Like RGB-D cameras, strain gauges are affordable and can be integrated into existing HMDs without altering the system's ergonomics and user experience.

**Contributions.** We propose the first system that jointly uses optical and non-optical signals to produce compelling 3D facial performance capture while wearing an HMD. We present the first usage of strain gauges for facial performance capture, while wearing VR head mounted display. The use of signals obtained from flexible electronic materials combined with an RGB-D camera presents a unique innovative and ergonomic solution for real-time facial performance sensing on wearable devices. Our system also introduces a practical framework for facial expression mapping that decouples a single-instance offline training process from a repeatable and fast online calibration step.

## 2 Previous Work

Performance-driven facial animation has been dominated by optical capture systems that use cameras or depth sensors. Hundreds of works on facial representations, tracking, mapping, and animation have been developed, greatly impacting film and game production [Parke and Waters 1996; Pighin and Lewis 2006]. Because we focus on lightweight wearable systems with real-time capabilities, we highlight the latest optical systems for real-time facial animation that are designed for non-studio settings. We also provide an overview of non-optical sensors to motivate our choice to use strain gauges for occlusion-free sensing.

**Optical Systems.** In the entertainment industry, facial performance capture is an established approach to improve the efficiency of animation production by reducing manual key-framing tasks and minimizing complex physical simulations of facial biomechanics [Terzopoulos and Waters 1990; Sifakis et al. 2005]. To achieve the highest possible facial tracking fidelity, marker-based solutions, hand-assisted tracking, and multi-camera settings are still commonly employed [Bhat et al. 2013; Bickel et al. 2007; Pighin and Lewis 2006], while often requiring intensive computation.

Deployable and real-time techniques that require only a monocular RGB camera have been popularized by data-driven algorithms, such as active appearance models (AAM) [Cootes et al. 2001] and constrained local models (CLM) [Cristinacce and Cootes 2008], where pre-labeled sequences of specific subjects are used to track sparse 2D facial features. Lately, several extensions have demonstrated fully automatic and highly accurate methods that do not require any user-specific training, such as the 2D landmark prediction approach [Saragih et al. 2011] or the supervised descent method [Xiong and De la Torre 2013]. These 2D landmarks can be used as low dimensional signals to drive the control parameters of a more complex face rig [Chai et al. 2003].

Recently, 3D face priors, dense flow, and shape from shading methods have been introduced to achieve higher fidelity 3D tracking from unconstrained monocular videos [Garrido et al. 2013; Shi et al. 2014]. Using dimension-reduced linear models and sufficient prior facial data, a single camera can be used to generate compelling facial animation in real-time without any calibration [Cao et al. 2014].

Lightweight RGB-D cameras (e.g., Microsoft Kinect, Intel RealSense, etc.) can improve modeling and tracking accuracy, as well as robustness because the depth information is acquired at the pixel level. Weise et al. [2011] use a large database of hand-animated blendshape curves to stabilize the low-resolution and noisy input data from the Kinect sensor. An example-based rigging approach [Li et al. 2010] is introduced to further reduce a tedious user-specific calibration process. Recently, calibration-free methods where the the shape of the expressions are learned during the tracking process have been presented [Li et al. 2013; Bouaziz et al. 2013]. The combination of sparse 2D facial features in the RGB channels and dense depth maps also improves the tracking quality [Li et al. 2013; Chen et al. 2013].

So far, compelling facial animations can be generated in real-time with minimal hardware, however, the captured faces must be fully visible, which is not possible with any existing virtual reality HMDs. Nevertheless, a solution for eye gaze tracking under an HMD has been recently introduced by a commercial solution [SMI 2014] using additional infrared cameras integrated into the headset. Hsieh et al. [2015] has presented a technique for real-time facial performance tracking in the presence of occlusions. Occluding objects are explicitly segmented as outliers from both the RGB and depth input. Their method focuses on tracking by using the segmented face regions (i.e., inliers), but does not provide a solution to generate the animation of the full face.

**Non-Optical Sensors.** Audio signals have been used to drive control parameters of facial models as a faster alternative to traditional motion capture. Although particularly effective for lip-syncing and producing co-articulation phenomena, audio signals are less effective in capturing the upper portion of the face and, in general, non-verbal expressions [Bregler et al. 1997; Brand 1999].

Unlike audio signals and optical systems, contact-based systems are not affected by occlusions and can be potential candidates for sensing behind the HMD. Similar to using exoskeletons for body tracking, rigid mechanical systems have been suggested for positional measurements for the face, but are rather uncomfortable and difficult to integrate into HMDs.

Other advancements in non-optical sensors include bioelectric signals and piezoelectric sensors. Bioelectric signals measure change in electric potential across tissues and organs via sensors placed on the skin. For example, electroenceophalograms (EEG) [McFarland and Wolpaw 2011] record brain activities and have been used to detect basic facial expressions and emotional responses. Unfortunately, EEGs require extensive training and concentration from the user. Non-invasive electromyograms (EMG) are widely used by psychologists to measure muscle contractions achieved by poses based on FACS [Ekman and Friesen 1978]. Lucero and Munhall [1999] have demonstrated a system that uses EMG signals to directly control a physics-based facial rig [Terzopoulos and Waters 1990], but requires anatomically compatible placements of electrodes across individuals. Gruebler and Suzuki [2014] proposed a more ergonomic solution of an EMG-based wearable device, but could only reliably detect a few expressions. EMG signals are

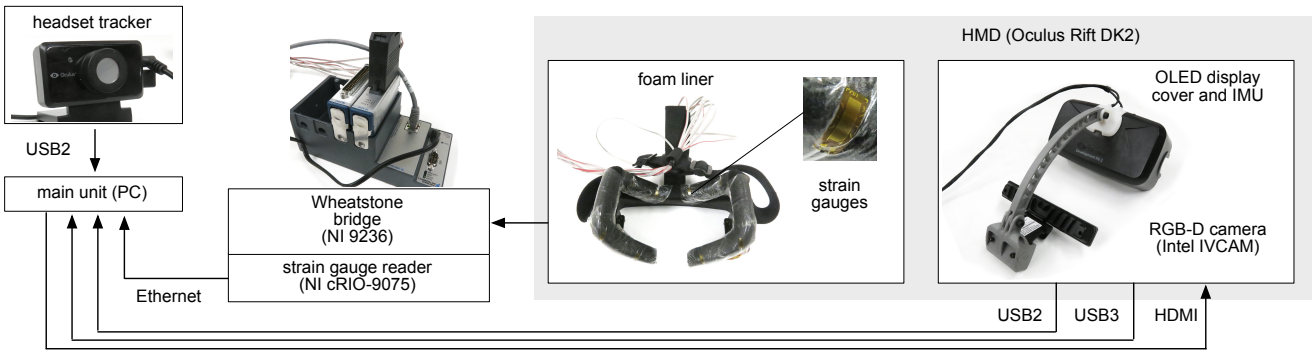



**Figure 2:** *Our system combines signals from eight strain gauge sensors and an RGB-D camera.*

generally very sensitive of their sensor locations, depend highly on the subject's fat tissue, and suffer from muscular crosstalk problems.

Piezoelectric sensors are tactile devices that measure strain rate and have been explored for smart glasses [Scheirer et al. 1999] in affective computing to recognize facial expressions. However, because static states such as a holding up eye brows for an extended period cannot be measured, they are not suitable for controlling facial models. Ultra-thin strain gauges [Sekitani et al. 2014], on the other hand, are tiny flexible electronic materials that directly measure strain on a surface as its electrical resistance changes when bent. In the context of motion capture, they have only been used so far in datagloves for hand tracking [Kramer and Leifer 1988]. We show in this work that strain gauges are particularly reliable for facial performance measurements when placed on the foam liner of the HMD because they can conform to any surface shape.

## 3 System Prototype

### 3.1 Hardware Components

As shown in Figures 1 and 2, our framework consists of a standard HMD (the Oculus Rift DK2 [Oculus VR 2014]) and additional sensing hardware, which are seamlessly integrated together. We attach an Intel RealSense 3D Camera (Bell Cliffs) [Intel 2014] to the HMD cover with a 3D printed mount. The RGB-D camera is directly connected to the main computer. In addition, we place eight strain gauges on the foam liner of the headset (Figure 1). The strain signals first go to a Wheatstone bridge that sits on a strain gauge reader. A Wheatstone bridge, depicted in Figure 3, is a resistor configuration optimized for measuring small changes in resistance as a deviation from zero volts at the output. These voltages are then recorded by the main unit via Ethernet. To recover the global rigid head pose, we use an off-the-shelf near-IR 60Hz CMOS sensor and integrated 1kHz nine-degree-of-freedom IMU of the Oculus Rift. Our hardware design maximizes optical acquisition of visible face regions and ensures reliable performance capture of facial deformations behind the headset without sacrificing ergonomic constraints.

**Strain Gauges.** While strain gauges are typically used as planar stretch sensors, we use them to detect relative changes in the radius of curvature over the length of the gauge. We use eight Omega 6mm KFH-series 350$\Omega$ strain gauges for capturing the change in shape of the foam liner of the HMD [Omega 2014; NI 2014]. The gauges are metal foil-based strain gauges on polyimide with a nominal resistance of 350$\Omega$ and a gauge factor (GF) of two (the ratio of change in resistance to mechanical strain). The gauge factor of the gauges varies less than $0.05\%$ from $-10^\circ$ to $45^\circ$C, and the fatigue life of the gauges is expected to exceed $10^7$ cycles for the small strains expected within our system [Omega 2014]. The gauges are bent over the edge and securely attached with a thin stretchable adhesive tape as depicted in Figure 5. This configuration allows the strain gauges (total thickness $0.05$mm) to flex with the underlying foam without significantly impacting the feel or function of the foam liner.

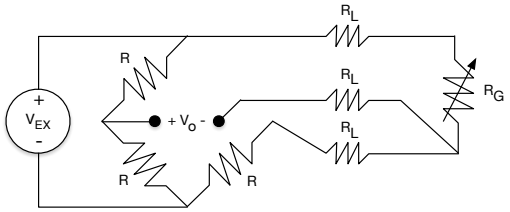



**Figure 3:** *Three-wire Wheatstone bridge configuration*

**Strain Gauge Data Acquisition and Synchronization.** The strain gauges are measured using a three-wire Wheatstone bridge configuration depicted in Figure 3. This arrangement allows the lead wire resistance $(R_L)$ to be neglected for maximum sensitivity. In this configuration the measured voltage $(\mathbf{V}_o)$ will be related to the resistance in the strain gauge $(R_G)$, reference resistors $(R)$, and the $3.3$V excitation voltage $(\mathbf{V}_{\text{EX}})$:

$$\mathbf{V}_o = \left(\frac{1}{2} - \frac{R_G}{R_G + R}\right)\mathbf{V}_{\text{EX}} \tag{1}$$

The data is acquired by a National Instruments 9075 CompactRIO (cRIO) real-time computer via a NI 9236 quarter-bridge module. The NI 9236 module comprises a set of eight Wheatstone bridges with 24-bit ADCs allowing for high precision strain simultaneous measurement. The cRIO polls the strain gauges at 100Hz and outputs the voltages over UDP to the main unit, also at 100Hz.

**Head-Mounted RGB-D Camera.** We attach an Intel RealSense 3D Camera on the cover of the HMD using a customized 3D printed camera mount as shown in Figure 1. We choose this RGB-D sensor due to its remarkable close range sensing capabilities (15cm) of the lower face region. This sensor consists of an IR and color camera, as well as an IR coded light projection and has a maximum depth map resolution of $640 \times 480$ pixels, where only $320 \times 240$ are used (both RGB and depth channels) for increased performance. To ensure synchronization between the RGB and the depth channels we use a capture rate of 30fps.

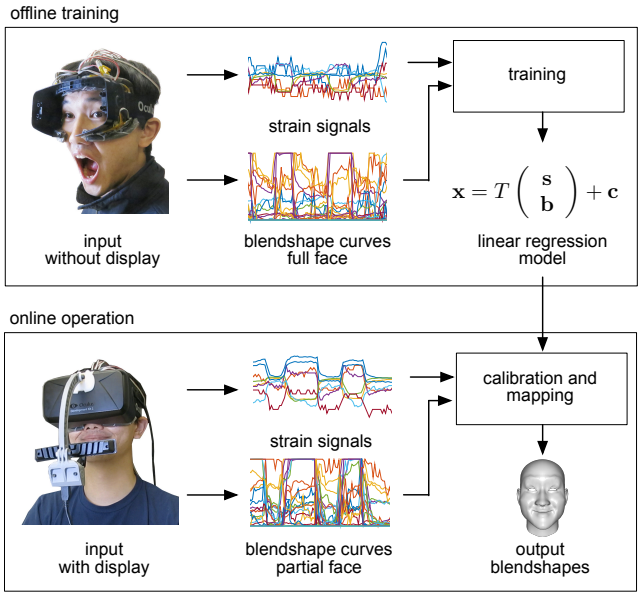


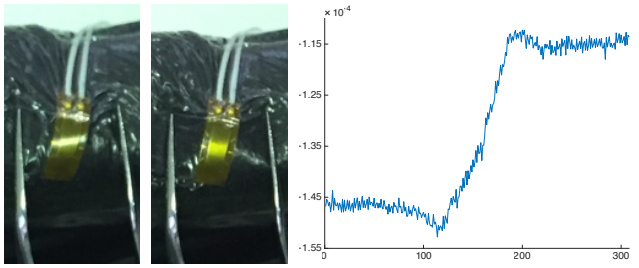

**Figure 4:** *Facial performance capture pipeline. Our pipeline consists of a training phase, where the linear mapping between mouth blend shapes and strain gauges is learned. This mapping is then applied during online operation to generate a complete facial blend shape.*

### 3.2 Performance Capture Pipeline

As illustrated in Figure 4, our facial performance capture pipeline consists of an offline training stage and an online operation stage.

**Offline Training.** To build a personalized tracking model and to train a regression model, we ask the user to detach the RGB-D camera from the 3D printed mount and remove the OLED display from the HMD to ensure the full face is visible. Only the foam liner of the HMD containing the strain gauges is worn during this stage. By externally mounting the RGB-D camera aimed at the face, we build a personalized blendshape model as described in Section 4.2 and go through a complete sequence of 20 FACS-based expressions. We simultaneously track the face using the blendshape model and, for each frame, obtain the corresponding strain signals as input to our training algorithm (see Section 5.1). The offline training session is only performed once per user.

**Online Operation.** After building the regression model, we reattach the RGB-D camera and the display to the HMD for the online operation: the system can now be reused by the same user without the full training. Because of misplacements of the wearable device as well as the additional weight of the camera and the display, the recorded signals may vary slightly with each use. Accordingly, we run a short calibration procedure to normalize and adapt the strain signals before each use, as described in Section 5.2. After calibration, we capture facial performance using the adjusted strain signals in the occluded regions and the RGB-D acquisition of the lower part of the face. We adopt a reduced set of blendshapes from the personalized model that only consists of expressions of the mouth region to capture extra signals in the form of blendshape coefficients. Both strain and RGB-D measurements are directly mapped to the personalized blendshape model of the entire face via the regression model.




(a) (b) (c)

**Figure 5:** *One strain gauge adhered to the foam interface of the HMD. From left to right: (a) undeflected foam; (b) deflected foam; (c) the measured voltage before and after the deflection.*

## 4 Signal Processing

In order to create a facial animation, the information from the strain gauges and the RGB-D camera needs to be meaningfully combined. In our system, the eight strain signals and the blendshape coefficients of the jaw region are used as input for training and mapping (discussed in Section 5).

### 4.1 Strain Sensing

As described in Section 3.1, changes in the resistance of a strain gauge as a result of physical deformation are measured as voltages across a Wheatstone bridge. Figure 5 depicts a single strain gauge adhered to the foam interface of the HMD. In Figure 5a, the strain gauge follows the relatively sharp curve of the edge of the foam. In Figure 5b, the edge of the foam is flattened slightly by applying pressure with tweezers and as a result the attached gauge is also flattened.

In the depicted configuration, the strain in the system can be estimated by modeling the strain gauge as an Euler-Bernoulli beam in simple bending, because the foam substrate and bending stiffness are several orders of magnitude less than the tensile stiffness of the polyimide gauge substrate. In simple bending, the strain ($\epsilon$) at the surface of the beam is given as the ratio of the distance from the neutral bending axis ($z$) and the radius of curvature ($\rho$) of the beam–for strain at the surface of thin beams, $z$ can be approximated as half of the total beam thickness ($t$) [Hibbeler 2005]:

$$\epsilon = -\frac{z}{\rho} = -\frac{t/2}{\rho} \tag{2}$$

The bending strain and thus the radius of curvature in the strain gauge can be related to the Wheatstone bridge voltage and the gauge factor by combining Equations 1 and 2:

$$\frac{\mathbf{V}_o}{\mathbf{V}_{\text{EX}}} = -\frac{\text{GF}\cdot\epsilon}{4}\left(\frac{1}{1+\text{GF}\cdot\epsilon/2}\right) = \frac{\text{GF}\cdot t}{8\rho}\left(\frac{1}{1-\text{GF}\cdot t/4\rho}\right)$$

Examining the response in Figure 5c, the voltage read from the Wheatstone bridge converges towards zero as the strain gauge is flattened under load ($\rho$ increasing). This basic operating principle can be used to infer a range of coupled mechanical deformations of the foam under load from the HMD, face, and facial expressions.

### 4.2 RGB-D Sensing

We use the RGB-D camera to capture the motion of the mouth region, which is not occluded by the HMD. The output consists

of blendshape curves that correspond to blendshape expressions of the lower part of the face using a simplified version of the tracking pipeline presented in [Li et al. 2013]. For tracking, we use both the sparse 2D lip features that describe the lip contours obtained from [Xiong and De la Torre 2013] and the input depth map to solve for blendshape coefficients $\mathbf{x}$ of a linear blendshape model with vertices $\mathbf{v}(\mathbf{x}) = \mathbf{b}_0 + \mathbf{B}\mathbf{x}$, where $\mathbf{b}_0$ is the neutral shape, the columns of $\mathbf{B}$ the expression meshes, and $\mathbf{x} \in [0,1]^N$ the blendshape coefficients ($N = 20$). Our linear blendshape model only needs to be constructed once for each user and can be reused over again.

**Blendshape Personalization.** To build the personalized blendshape model $(\mathbf{b}_0, \mathbf{B})$, we capture the depth data of the person in neutral face pose without the HMD. We then use the iterative closest point (ICP) algorithm [Rusinkiewicz and Levoy 2001] to estimate the initial pose between a statistical mean face model and the input depth data. Next, we warp this model to fit the captured face geometry using a linear PCA model of faces [Blanz and Vetter 1999] to determine the neutral pose $\mathbf{b}_0$ [Li et al. 2013]. The personalized blendshape expressions $\mathbf{B}$ are obtained by transferring the deformations from a database of generic lower face region expressions to $\mathbf{b}_0$ via deformation transfer [Sumner and Popović 2004].

**Rigid Motion Estimation.** Every time a person puts on the headset, we perform five iterations of ICP at the beginning to adjust the rigid motion between the personalized blendshape model and the HMD. Since the wearable device is tightly attached to the face, we lock the local head pose of the user. The global head motion is obtained from the system's IMU and the external headset tracker from the Oculus Rift.

**Blendshape Tracking.** To solve for the coefficients $\mathbf{x} \in [0,1]^N$, we fit the blendshape $\mathbf{v}(\mathbf{x})$ to the input depth map and the detected sparse 2D lip features by minimizing the following energy term:

$$\min_{\mathbf{x}} \sum_i c_i^S(\mathbf{x}) + w \sum_j c_j^F(\mathbf{x}) \,, \tag{3}$$

where $c_i^S(\mathbf{x})$ is the depth map term, $c_j^F(\mathbf{x})$ the 2D lip feature term, and $w = 5 \cdot 10^{-5}$ its weight. The depth map term describes a point-to-plane energy:

$$c_i^S(\mathbf{x}) = \left(\mathbf{n}_i^\top \left(\mathbf{v}_i(\mathbf{x}) - \bar{\mathbf{v}}_i\right)\right)^2 \,, \tag{4}$$

where $\mathbf{v}_i$ is the $i$-th vertex of the mesh, $\bar{\mathbf{v}}_i$ is the projection of $\mathbf{v}_i$ to the depth map and $\mathbf{n}_i$ the surface normals of $\bar{\mathbf{v}}_i$. The lip feature term describes a point-to-point energy:

$$c_j^F(\mathbf{x}) = \|\pi(\mathbf{v}_j) - \mathbf{u}_j\|_2^2 \,, \tag{5}$$

where $\mathbf{u}_j$ is the position of a tracked 2D facial feature and $\pi(\mathbf{v}_j)$ its corresponding mesh vertices projected into camera space. We solve this bounded linear optimization using the fast iterative projection method of Sugimoto et al. [1995].

# 5 Facial Expression Control

## 5.1 Training

To model the mapping between the input data and facial expressions, we train a linear regression model in an offline training step. During this training session, the user wears the foam interface with the display removed. This enables sufficient visibility for tracking facial expressions in the regions that would be normally occluded by the HMD. Full facial tracking is achieved using the method presented in Li et al. [2013].

We formulate the mapping between the input data and the blendshape coefficients $\mathbf{x} \in \mathbb{R}^{28}$ of the full face as follows:

$$\mathbf{x} = T \begin{pmatrix} \mathbf{s} \\ \mathbf{b} \end{pmatrix} + \mathbf{c} \tag{6}$$

where $\mathbf{s}$ is the vector of the eight observed strain signals, $\mathbf{b} \in \mathbb{R}^{20}$ the blendshape coefficients of the lower face part, $T$ the linear mapping, and $\mathbf{c}$ a constant. The matrix $T \in \mathbb{R}^{28\times 28}$ and the constant $\mathbf{c} \in \mathbb{R}^{28}$ are optimized using the least squares method by minimizing the mean squared error between the observed coefficients and the prediction.

## 5.2 Calibration

In practice, the strain signals captured during the training session may be inconsistent with signals captured during the online operation when the full framework is in use. This inconsistency is due to two factors: (1) variations in HMD fit between wearings, and (2) the extra weight from the display and mounted depth camera. As a result, directly using the mapping obtained during the training session (Section 5.1) will not result in accurate facial expressions. To resolve this issue, we introduce an extra calibration step at the beginning of each online operation that maps the strain gauge signals captured during the online operation back to the distribution space of signals captured during the offline training session.

Specifically, at the end of the training session, we ask the user to perform a short sequence of specific expressions. During the calibration session, the user performs the same sequence of expressions. We assume that there is a correlation between the distribution of strain gauge signals captured at the end of the training session and the distribution for the calibration session.

We first learn a Gaussian mixture model (GMM) $\mathcal{M}_0 = \{w_m, \mu^{(m)}, \Sigma^{(m)}\}_{m=1,\ldots,M}$ to describe the distribution $S$ of the signals captured at the end of the training session:

$$S \sim \sum_{m=1}^{M} w_m \mathcal{N}(\mu^{(m)}, \Sigma^{(m)}) \tag{7}$$

where $w_m$ is the weight of each component, and $\mu^{(m)} \in \mathbb{R}^8$ and $\Sigma^{(m)} \in \mathbb{R}^{8\times 8}$ is the mean and covariance of its $m$-th component, respectively. We use $M = 8$ for all of our examples.

Next we model the correlation as an affine transformation $(H, \mathbf{h})$, where $H \in \mathbb{R}^{8\times 8}$ and $h \in \mathbb{R}^8$. For the sequence of strain gauge signals $\bar{\mathbf{s}}$ captured during the online calibration, we compute the transformed signal $\hat{\mathbf{s}}$ as follows, which is used in Equation 6 to get the blendshape coefficients of the full face:

$$\hat{\mathbf{s}} = H\bar{\mathbf{s}} + \mathbf{h} \tag{8}$$

We obtain the optimal transformation $(H, \mathbf{h})$ by maximizing the log-likelihood that the signals captured during the calibration session follow the distribution $S$ in Equation 7 under this transformation. We solve this optimization problem using an EM algorithm by iteratively alternating between an E-step and an M-step [Gales 1998].

**Discussion.** The sequence used for calibration does not have to be the full sequence of FACS-based expressions used for training. A short sequence of facial expressions (e.g., smiling, surprise, and sadness) is sufficient, because we only map the distribution of the signals between two sequences in the calibration session.

**Figure 6:** *Our system can track and animate a wide range of facial expressions for different users and with arbitrary head motion.*

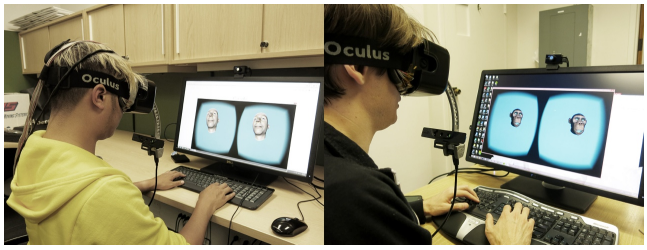


**Figure 7:** *Remote face-to-face communications between two users.*

**Figure 8:** *Optical input of our system. From left to right: RGB input; depth input; extracted 2D lip feature points; output mesh overlaid with the RGB input.*

## 6 Results

As shown in Figure 6, our system can accurately capture a wide range of facial expressions of different users when they are wearing the HMD. This capability enables multi-party face-to-face communication, with each user having their own 3D avatar and wearing a prototype of our system, as shown in Figure 7 and the accompanying video. The communication is achieved by exchanging the blendshape coefficients computed during online operation.

Figure 8 demonstrates the optical input for our system. Throughout our paper draft, the curves of strain gauge signals are sampled at 100Hz, while the curves of blendshape coefficients are sampled at 30Hz if not otherwise mentioned. Figure 9 shows the eight input signals from the strain sensors overlaid with the sequence of coefficients for the blendshape representing a raised left eyebrow, which is computed based on the optical input. As can be seen from Figure 9, although for a given blendshape coefficient, we cannot find a strain signal that has a one-to-one mapping to it, we can easily find a combination of the signals that are related to it. This motivated our use of a linear regression model to learn the mapping.

**Evaluation.** The strain gauge signal is robust and the foam shape deformation is repeatable over time, cycles, and temperature variation. To illustrate the robustness and accuracy of the strain gauge signal, in Figure 10 we show the voltage values from one strain gauge at the beginning and end of a five minute trial. The

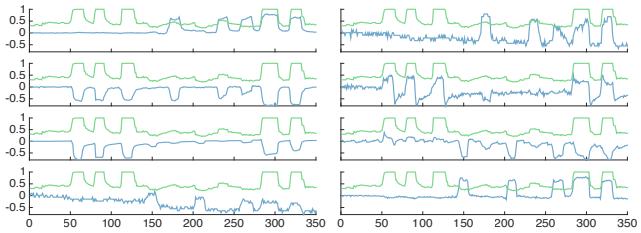


**Figure 9:** *The eight strain signals (blue) and the coefficient of the blendshape representing a raised left eyebrow (green). The strain signals are scaled and the blendshape coefficients are resampled at* 100*Hz for visualization.*

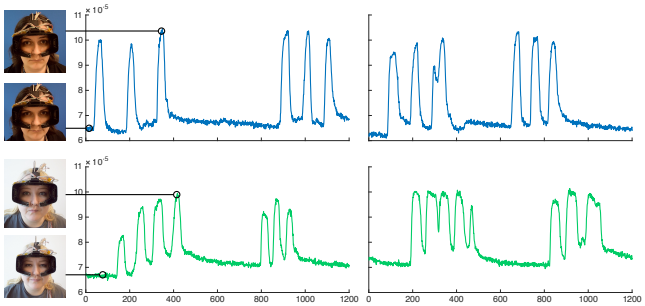


**Figure 10:** *Strain gauge signal over extended use for two users (five minute trial). Twelve seconds of repeated eyebrow raises at the beginning of the trial (left) and end of the trial (right).*

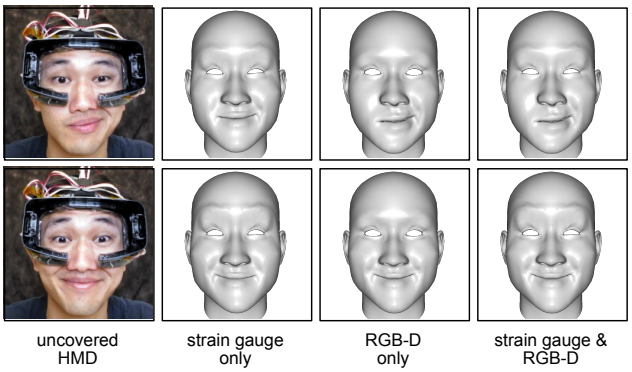



**Figure 11:** *The combination of strain gauge and RGB-D sensors yield the best results.*

users were instructed to repeat the same facial expression (eyebrow raises) several times during the length of the trial. As expected, we observed some variability in the baseline value (DC offset) of the strain gauge when the HMD is removed and replaced. Due to this variability, we incorporated the calibration step at the beginning of each trial.

Figure 11 illustrates different performance capture results when only the RGB-D camera is activated, only the strain signals, or both together. Although the strain signals can capture some of the larger mouth motions such as smile or mouth open, the best results are obtained when both the RGB-D input and the strain gauge signals are used.

To show the importance of per-user training, we first demonstrate the effect of using a person-specific regression model compared

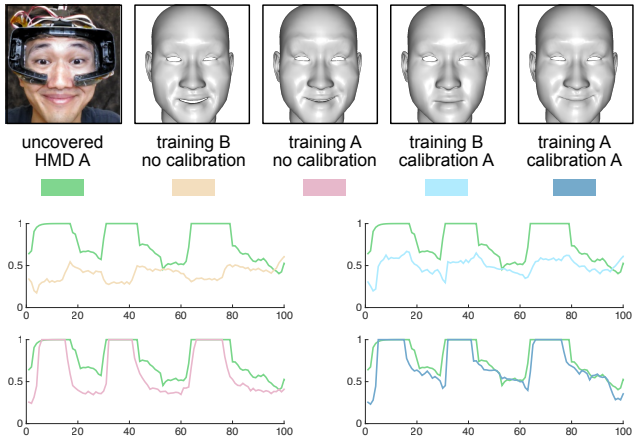



**Figure 12:** *Assessment of the effect of both user-specific training and per-session calibration on the blendshape coefficient for the right eyebrow. The green line represents the ground truth, obtained by observing the user with the display removed. Each test was performed by Subject A, while the training data of another user (Subject B) was used for several tests. Thus, the tan and light blue lines show the result of Subject A's use of Subject B's training data.*

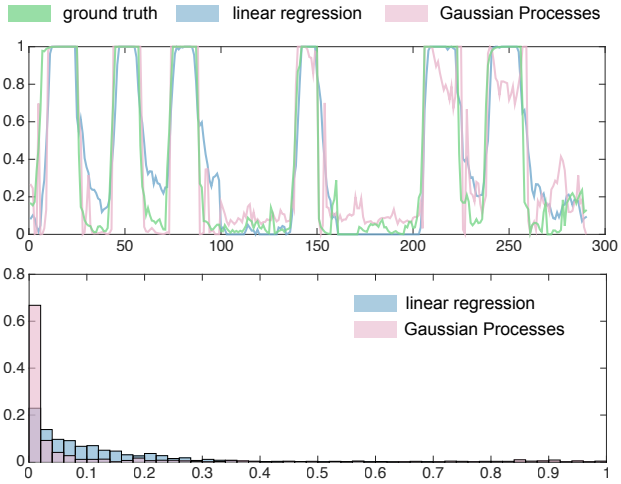



**Figure 13:** *Comparison between the error distributions of Gaussian Processes and linear regression.*

to one built from another subject. As illustrated in Figure 12, geometric and anatomical differences between the subjects can cause different strain measurements and thus make the single-instance training session necessary for accurate tracking even with the short calibration step from Section 5.2. Figure 12 shows how the calibration step improves the facial expression mapping between different sessions when the HMD is taken off and put on again.

Our use of linear regression for facial expression mapping is simple, efficient, and robust. We compare this approach with a more sophisticated but offline non-linear regression method based on Gaussian processes (GP) [Rasmussen and Williams 2005]. While the overall probability distribution of the prediction is more accurate using GP, our method produces visually comparable results, as shown in Figure 13. Furthermore, due to the sensitive variation of strain signals when the HMD is worn again for the same user, the non-linear components of the GP computed regression model tend

**Figure 14:** *Comparison between our approach and state of the art real-time facial performance capture method. From left to right: input; [Li et al. 2013]; our result.*

to become less robust.

**Comparison.** To validate the accuracy of our tracking when the HMD is worn and occluding the entire upper face, we create an experiment where the face is largely visible by removing the HMD display (offline training mode), with only the strain sensors in contact with the face. By keeping the head pose fixed, we then place the depth sensor at a position equivalent to that when mounted to the headset. We then track the face using the state of the art real-time performance capture method of [Li et al. 2013] without occlusions as ground truth while simultaneously capturing the facial performance using our input signals. To simulate a worn headset, we synthetically occlude the upper face region as input to our method similar to the previous evaluations. Comparisons are shown in Figure 14.

**Limitations.** From a hardware design standpoint, even though the full training session is only required once per user, our system still requires a short per-trial calibration. We believe that an ideal system should allow instantaneous tracking for any user and deduce the mapping from the shapes of the tracking model and strain measurements during online operation. Because pressure distribution varies with HMD placement and head orientation, our strain signal measurements can drift or slightly decrease in accuracy. In practice, we observe that our mapping is sufficiently robust due to the additional signals obtained from RGB-D.

We have tested our prototype by capturing seven subjects, including five males and two females. Depending on a user's facial features and how well adjusted the HMD straps are, there are situations in which the strain gauges are not in tight contact with the face. In such cases, it is difficult to accurately recover the expressions around eyebrows, while the rigid motion of the head and the expressions of the mouth still can be faithfully captured. Furthermore, subtle expressions such as eye blinks or squints are difficult to capture using the sparsely placed strain gauges on the foam liner.

**Performance.** Real-time performance and low latency is essential for virtual reality enabled face-to-face communication. Our system reaches 30fps on a 3.7GHz quad-core Intel Core i7-4820K with 32GB RAM and a GeForce GTX 980 graphics card. Due to the very low latency of the strain gauge measurements, the latency of our system is only bounded by the Intel IVCAM depth sensor acquisition, which has a latency of 60ms. While typical video conferencing requires a one-way latency to not exceed 150ms, 20ms is deemed acceptable for virtual reality applications [Abrash 2012]. During online operation, we measure 3ms for facial feature detection, 5ms for blendshape optimization and 3ms for the mapping.

## 7 Conclusion

We have developed a system that augments an HMD with ultra-thin strain gauges and a head-mounted RGB-D camera for facial performance capture in virtual reality. Our hardware components are easily accessible and integrate seamlessly into the Oculus Rift DK2 headset [Oculus VR 2014], without drastically altering the ergonomics of the HMD, except for a negligible increase in weight. For manufacturers, strain gauges are attractive because they can be fabricated using traditional high volume production techniques.

Even though the face is largely occluded by the HMD, our system is able to produce accurate tracking results indistinguishable from existing RGB-D based real-time facial animation methods where the face is fully visible [Weise et al. 2011; Li et al. 2013]. Our system is easily accessible to non-professional users as it only requires a single training process per user, and the trained linear regression model can be reused for the same person. Given that the strain sensitivity can be slightly different between the offline training process (without the display) and the online operation, we also demonstrate the effectiveness of optimizing Gaussian mixture distributions to improve the tracking accuracy though a short calibration step between each use.

**Future Work.** While acceptable for a consumer audience, our current framework still requires a complete facial expression calibration process before each trial. In the future, we plan to eliminate this step by training a large database of face shapes and regression models. We believe that such a database can be used to determine a correct regression model without the need for calibration before each use. Similar to head mounted cameras used in production for motion capture [Bhat et al. 2013], we attach an RGB-D camera to record the mouth region of the subject, but still intend to design a more ergonomic solution with either smaller and closer range cameras, or even alternative sensors. To advance the capture capabilities of our system, we hope to increase the number of strain gauges and combine our system with other sensors such as HMD integrated eye tracking systems [SMI 2014].

**Broader Impact.** The ability to capture facial performances while wearing the HMD is the first step towards enabling compelling multi-way face-to-face interaction in virtual reality. Additionally, recording a user's facial expressions could enhance immersive gaming experiences by introducing mood sensitive intelligent agents. In film and game production, actors can also be immersed in digital film sets while their performances are being recorded.

## Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback. The authors would also like to thank Chris Twigg and Douglas Lanman for their help in revising the paper, Gio Nakpil and Scott Parish for the 3D models, Fei Sha for the discussions on machine learning algorithms, Frances Chen for being our capture model, Liwen Hu for helping with the results, and Ryan Ebert for his mechanical design. This project was supported in part by Oculus & Facebook, USC's Integrated Media Systems Center, Adobe Systems, Pelican Imaging, and the Google Research Faculty Award.

## References


ABRASH, M., 2012. Latency – the sine qua non of AR and VR. http://blogs.valvesoftware.com/abrash/latency-the-sine-qua-non-of-ar-and-vr/.

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.

BICKEL, B., BOTSCH, M., ANGST, R., MATUSIK, W., OTADUY, M., PFISTER, H., AND GROSS, M. 2007. Multi-scale capture of facial geometry and motion. *ACM Trans. Graph. 26*, 3.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99*, 187–194.

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

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.

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

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.

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

CRISTINACCE, D., AND COOTES, T. 2008. Automatic feature localisation with constrained local models. *Pattern Recogn. 41*, 10, 3054–3067.

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

GALES, M. 1998. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language 12*, 75–98.

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.

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.

HIBBELER, R. 2005. *Mechanics of Materials*. Prentice Hall, Upper Saddle River.

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

INTEL, 2014. Intel realsense ivcam. https://software.intel.com/en-us/intel-realsense-sdk.

KRAMER, J., AND LEIFER, L. 1988. The talking glove. *SIGCAPH Comput. Phys. Handicap.*, 39, 12–16.

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.

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.

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

NI, 2014. National instruments. http://www.ni.com/white-paper/3642/en/.

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

OMEGA, 2014. Omega strain gages. http://www.omega.com/Pressure/pdf/KFH.pdf.

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.

RASMUSSEN, C. E., AND WILLIAMS, C. K. I. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

RUSINKIEWICZ, S., AND LEVOY, M. 2001. Efficient variants of the icp algorithm. In *International Conference on 3-D Digital Imaging and Modeling*, IEEE, 145–152.

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.

SEKITANI, T., KALTENBRUNNER, M., YOKOTA, T., AND SOMEYA, T. 2014. Imperceptible electronic skin. *SID Symposium Digest of Technical Papers 45*, 1, 122–125.

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.

SIFAKIS, E., NEVEROV, I., AND FEDKIW, R. 2005. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Trans. Graph. 24*, 3, 417–425.

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

SUGIMOTO, T., FUKUSHIMA, M., AND IBARAKI, T. 1995. A parallel relaxation method for quadratic programming problems with interval constraints. *Journal of Computational and Applied Mathematics 60*, 12, 219 – 236.

SUMNER, R. W., AND POPOVIĆ, J. 2004. Deformation transfer for triangle meshes. *ACM Trans. Graph. 23*, 3, 399–405.

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

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

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