Buccal: Low-Cost Cheek Sensing for Inferring Continuous Jaw Motion in Mobile Virtual Reality

Richard Li, Gabriel Reyes
School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA
{lichard49, greyes}@gatech.edu

ABSTRACT

Teleconferencing is touted to be one of the main and most powerful uses of virtual reality (VR). While subtle facial movements play a large role in human-to-human interactions, current work in the VR space has focused on identifying discrete emotions and expressions through coarse facial cues and gestures. By tracking and representing the fluid movements of facial elements as continuous range values, users are able to more fully express themselves. In this work, we present *Buccal*, a simple yet effective approach to inferring continuous lip and jaw motions by measuring deformations of the cheeks and temples with only 5 infrared proximity sensors embedded in a mobile VR headset. The signals from these sensors are mapped to facial movements through a regression model trained with ground truth labels recorded from a webcam. For a streamlined user experience, we train a user independent model that requires no setup process. Finally, we demonstrate the use of our technique to manipulate the lips and jaw of a 3D face model in real-time.

ACM Classification Keywords

H.5.2. [Information Interfaces and Presentation]: User Interfaces — Input Devices and Strategies.

Author Keywords

Mobile; virtual reality; VR; jaw motion; proximity; sensing; machine learning

INTRODUCTION

Recent developments in computer graphics and mobile hardware have led VR headsets to rapidly grow in the commercial and consumer space. However, despite the potential for what VR spaces could do to support and enhance human-to-human interactions, the latest VR experiences mostly isolate users rather than connect them. While development has mostly focused on improving the graphics and display components of VR headsets, approaches to sensing the user's (e)motions and behaviors are still lacking, unable to support rich interactions.

Furthermore, many of these recent improvements require bulky and resource-intensive hardware. Using facial gestures and expressions for input has been previously explored, typically using camera-based solutions. Research work has modified off-the-shelf virtual reality enclosures with additional sensors for the purposes of enhancing the input capabilities, typically focused on the upper part of the face (e.g., eye tracking). For VR headsets to ultimately become a fully mobile platform capable of facial input, our work titled *Buccal* (relating to the cheeks) demonstrates a low cost and simple yet effective approach to tracking the continuous range of jaw motions, with the eventual goal of supporting full facial reconstruction.

RELATED WORK

Facial Expression Detection

Facial expressions and gestures are one of the most natural ways of communicating with other people. In some situations (e.g., virtual reality, motor impairments, remote collaboration, etc.), computers can assist the user in capturing facial expressions as a way to mediate interactions with the digital world or others around us. Many people with severe motor impairments can control only a single switch, triggered by a muscle that has some mobility (e.g., cheek twitch or eye movement) [1]. Head tracking has been widely utilized as a way to move a cursor and other input controls [8]. Tracking and detecting facial expressions can provide a much more expressive method of expressing emotional state and intent. The Kinovea software, used in our work, has been previously used to develop a tool for measuring and assessing the reliability of bidimensional facial movements [2].

AffectiveWear [5] focuses on facial expression detection with smart glasses for use in everyday life. The technique uses infrared (IR) proximity sensors embedded in the rim of the glasses to categorize 7 states by measuring the distance between the glasses and the person's face. While the work contributes the classification of various expressions, it does so as a classification problem treating facial expressions as discrete. Our technique uses regression to reconstruct jaw and lip motion to capture movements in a more natural way. EarFieldSensing [6] explored the use of various electric sensing technologies to capture facial muscle movements by placing electrodes inside the ear canal. The technique was capable of

ISWC '18, October 8–12, 2018, Singapore, Singapore
© 2018 ACM. ISBN 978-1-4503-5967-2/18/10. . . $15.00
DOI: https://doi.org/10.1145/3267242.3267265

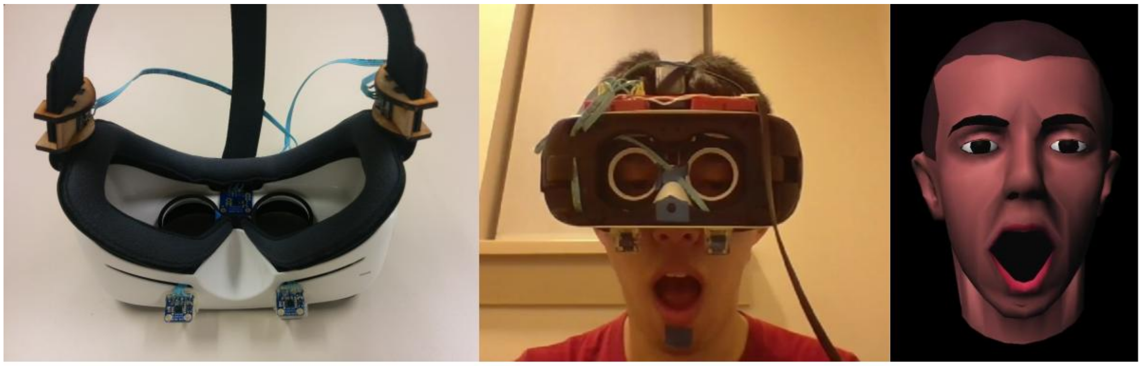

Figure 1. Buccal is a head-mounted display (HMD) augmented with IR proximity sensors to infer jaw motions that can be reflected on a 3D face model in order to enhance face-to-face interactions in virtual environments.

classifying over 20 facial-related gestures and required custom earbuds. The custom hardware design enables capturing gross head and neck movements, as well as certain facial expressions, but currently blocks the ear canal, limiting any auditory user experience.

Tracking Jaw Motion Using HMDs

Augmenting the VR enclosure with additional sensing capabilities is one way to enable the capture and analysis of facial expressions and jaw movements. Li et al. [4] attached strain gauges all around the foam lining of an Oculus Rift head-mounted display (HMD) to measure the deformation of the foam when certain facial movements were performed. By combining this technique with a depth camera attached to the HMD and oriented at the mouth, the researchers were able to reconstruct a 3D model of the user's face, enabling participants in a multi-player game to see each other's facial expressions. Olszewski et al. [7] also presented a method for animating a digital avatar in real-time by instrumenting the HMD using an RGB camera pointed at the mouth. The system uses a deep neural net regressor on images of the mouth to directly control the avatar. While these camera-based systems provide high-resolution data, they do suffer from a number of disadvantages — very fast mouth motion can lead to motion blur, the presence of facial hair may affect capture, and the techniques are susceptible to ambient noise producing erroneous results. Camera-based approaches also require greater cost and computing power. Our technique recreates jaw motion by indirectly sensing at the cheeks and temples, using low-cost proximity sensors and efficient regression modeling. It supports fast mouth motion while talking and is not affected by facial hair.

Perhaps most closely related to our work, Kawahara et al. [3] proposes an attachment with photoreflectors on an Oculus Rift monitoring the presence of the mouth cavity for detecting mouth movements. The approach requires a mask-like attachment covering the entire face. Suzuki et al. [9] presented a technique for recognizing and mapping facial expressions of VR HMD users to an avatar using embedded IR photoreflective sensors. The technique only focused on five basic facial expressions and required 16 IR photo reflective sensors. The main drawback is that the system did not reproduce the wearer's mouth movements when speaking or moving the lips, mainly focusing on tracking the upper part of the face and inferring motions based on emotion state changes. Furthermore, we contend that our system, which reproduces observed jaw and lip motion by regressing with a handful of sensors inside and around the HMD, enables more expressivity.

SYSTEM PROTOTYPE

An off-the-shelf Samsung Gear VR HMD was used as a base physical device. Five VCNL4020 fully-integrated proximity sensors with infrared emitters were attached to the HMD (see Figure 1a). Two sensors were placed over the cheeks, mounted on the bottom of the HMD with laser-cut triangle brackets. Two more sensors were placed over the temples on the sides of the head, attached with laser-cut hooks that fit over the strap of the HMD. The final sensor was placed over the bridge of the nose. The sensors are strategically positioned to directly capture cheek motion and indirectly detect the motion of the jaw bone from the side of the head through the temporalis muscle. All five of these sensors were connected to a Teensy 3.2 microcontroller through a 1-to-8 I2C multiplexer. The signals were sampled at 50 Hz by the microcontroller and forwarded to a laptop over a USB cable for further processing.

DATA COLLECTION PROCEDURE

A user study with 6 participants (5 male, 1 female; ages 18-24) was conducted to collect data for training and evaluating the machine learning model. Each participant was asked to wear the HMD while sitting in front of a laptop computer with a webcam. A piece of square, blue masking tape was adhered to each participant's chin during the ground truth data collection process in order to provide more robust tracking of the face and headset with the webcam. Each study consisted of a training session to become acquainted with the data collection tool, followed by 5 recorded sessions. Offline analysis was conducted post-hoc and participants did not receive feedback on their performance. Between each session, the users were asked to remove the device, walk around the room, return to the laptop, and wear the device again. This step simulates independent sessions and creates more variance within the dataset.

During each session, participants received a series of instructions displayed one at a time on the laptop screen prompting the user to either: a) perform a jaw gesture, or b) speak a sentence. The instructions are listed in the Appendix, along with the number of times each instruction was to be repeated. The gestures were designed to diversify the dataset by asking the user to open and close their jaw at a variety of speeds. On the other hand, the sentences were selected from "The Tortoise and the Hare" from Aesop's Fables to contain a variety of rhythms (i.e., different numbers of syllables) and inflections (i.e., statements versus questions). Each task was presented to the participant in random order. Note that while a smiling event was included in the original procedure, the data was not considered in the final evaluation due to a lack of trackable ground truth features in the video.

In each recorded session, the signals from the five sensors were saved to the laptop, and a video of the user wearing the HMD and performing the tasks was captured at 20 frames per second (FPS) using the webcam. The sensor signals and videos were ensured to be time synchronized.

Labeling the Ground Truth

To obtain the ground truth of how wide the jaw has opened, the video data was annotated with Kinovea, a sports analysis software that uses pattern recognition algorithms to track preselected regions of interest. The paths are returned as relative coordinates from the originally labeled region. For each video, the chin with the blue marker, the nose (with a blue marker on the HMD), and the two eye sockets were labeled.

To roughly find correspondence between the pixel unit movements given by Kinovea and real physical units (millimeters), the vertical distances between each eye hole of the HMD and the nose were measured. This value was empirically determined to be 31.75 mm on the physical headset. For each video frame, this distance was measured in pixels from each eye hole, and then averaged to account for slight tilting of the head. The scaling factor was then determined to be the measured distance divided by the average. More concretely:

$$Scaling\ Factor = \frac{31.75\ mm}{(\Delta left\ px + \Delta right\ px)/2}$$

where $\Delta left$ and $\Delta right$ refer to the vertical distances between the left and right eye holes from the nose, respectively. Finally, for each video frame, the distance between the nose and the chin was also measured, and this scaling factor was applied. These values will be used in future sections as the ground truth signal.

INFERRING JAW MOTIONS

Preprocessing and Regression

The frequency of the ground truth labels from the videos was increased using a one dimensional linear interpolation in order to match the frequency of the sensor signals. A first order low-pass Butterworth filter with a cut-off frequency of 1 Hz was used to remove the resulting noise in the signal. The same filter was also applied to the raw sensor signals.

After the preprocessing step, both the sensor and the ground truth signals were downsampled using a sliding window of 5 samples, sliding by 2 samples, and averaging each window. The downsampled signals were then used directly to train the linear regression model. The ElasticNet linear model, as implemented in the Scikit-learn toolkit in Python, was chosen for its ability to make use of large numbers of samples with a small number of features. The output of the regression model is a prediction of how wide the mouth has opened in millimeters, given proximity sensor data.

Subject	1	2	3	4	5	6
RMSE	4.026	6.412	5.314	5.342	22.729	5.328

Table 1. Average root mean squared error (RMSE) in millimeters for 6 participants.

Evaluation Metrics

The root mean squared error (RMSE), a commonly used measure of similarity between the values predicted by a model and the values actually observed, was used as an objective evaluation metric. The RMSE was evaluated as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}$$

where $\hat{y}_t$ is the predicted value at time t and y_t is the observed value from the ground truth video at time t. In this evaluation, the results are given as the square root of squared error in millimeters.

RESULTS AND DISCUSSION

A user independent model was used for the evaluation. In testing on a participant's data, all of the data from the rest of the participants was used to train the model, which in turn was used to predict the test participant's mouth openness. This predicted result was then compared to the ground truth signal obtained from the video using RMSE. These results are reported in Table 1.

Since RMSE does not account for the length of the sequence, it is plausible to accumulate more error given longer sessions. Across the thirty sessions of data from six subjects, with an average session length of 5 minutes and 39 seconds, the average RMSE obtained was 8.192. This overall average is skewed upward with P5 considered an outlier, and the average RMSE without P5 is 5.285. Participant 5's results were considerably worse due to more significant head movement as compared to other subjects. The participant also scratched their nose with their hand on a few instances, momentarily obscuring part of their face from the webcam.

Disregarding P5 as an outlier, the small variance in RMSE scores demonstrates the robustness of our approach. While our results were obtained in a laboratory setting, there is a potential concern regarding ambient light interfering with the infrared proximity sensors. However, despite phone-based VR systems being mobile platforms, they are generally used within a controlled environment – indoors, where there is minimal influence from the sun's infrared rays, and users are generally stationary without being exposed to continuously changing lighting conditions.

3D FACE MODEL MANIPULATION

We use the third party tool called Facade, as animating a 3D model of the face is not our core contribution. The tool

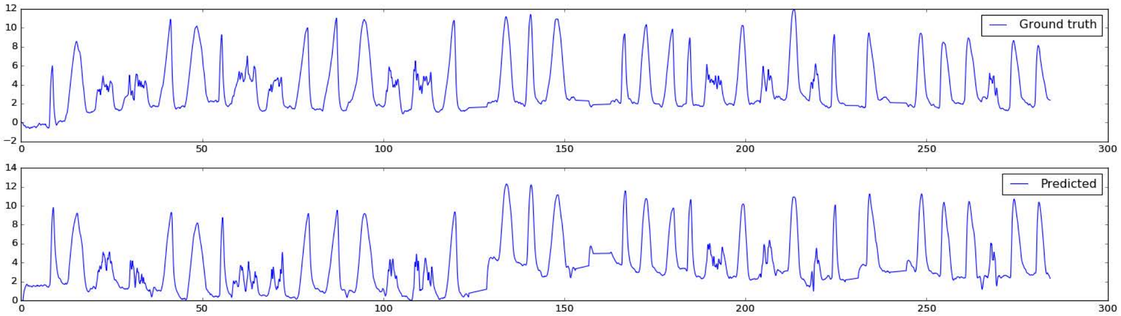

Figure 2. Comparison of the results of the ground truth labels (top) with the output of the regression model (bottom). Units in millimeters.

can manipulate various facial features, including the angle of the jaw's rotation. To map the values from the distance between the nose and the chin to an angle of the jaw, we use the arctangent of the predicted distance value divided by an assumed jaw length of 44 millimeters (empirically determined during pilot testing). With a maximum distance of around 70 millimeters from the entire ground truth dataset, the possible resulting angles fall in the range of 0 to 60 degrees. Due to the simple nature of our system, the minimal processing overhead (especially when compare to camera-based systems) allows the 3D face model can be manipulated in real-time with no visible impact on the fluidity of the 3D model's movements. The sensor signals from the HMD are read from the serial port, fed into a pre-trained regression model, and converted into the appropriate units to pass into Facade. However, due to the smoothing applied by the sliding window, sudden movements might be dampened.

CONCLUSION AND FUTURE WORK

The Buccal system prototype infers jaw motions in a continuous range using only 5 infrared proximity sensors in an HMD. A user-independent model was trained using ground truth labels from webcam videos. The inferences are used to manipulate a 3D model to mimic the user's jaw motions in real time, enabling fluid interactions such as speaking. While we were able to obtain positive results in inferring mouth opening and closing, it seems promising that with additional work we could capture even more motions including asymmetrical jaw motions (e.g., crooked smiling). We plan to further this technique to track other parts of the face (e.g., eyebrows) with the goal of reconstructing the entire face. Our system provides a non-invasive sensing approach to animate the lower face of an avatar and further enhance remote communication in VR.

REFERENCES

1. Kohei Arai. 2015. Relations between Psychological Status and Eye Movements. *International Journal of Advanced Research in Artificial Intelligence (IJARAI)* 4, 6 (2015). DOI: **http://dx.doi.org/10.14569/IJARAI.2015.040603**
2. Marjolaine Baude, Emilie Hutin, and Jean-Michel Gracies. 2015. A Bidimensional System of Facial Movement Analysis Conception and Reliability in Adults. (2015). DOI:**http://dx.doi.org/10.1155/2015/812961**
3. Keisuke Kawahara, Mose Sakashita, Amy Koike, Ippei Suzuki, Kenta Suzuki, and Yoichi Ochiai. 2016. Transformed Human Presence for Puppetry. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology (ACE '16)*. ACM, New York, NY, USA, 38:1–38:6. DOI: **http://dx.doi.org/10.1145/3001773.3001813**
4. Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial Performance Sensing Head-mounted Display. *ACM Trans. Graph.* 34, 4 (July 2015), 47:1–47:9. DOI: **http://dx.doi.org/10.1145/2766939**
5. Katsutoshi Masai, Yuta Sugiura, Masa Ogata, Katsuhiro Suzuki, Fumihiko Nakamura, Sho Shimamura, Kai Kunze, Masahiko Inami, and Maki Sugimoto. 2015. AffectiveWear: Toward Recognizing Facial Expression. In *ACM SIGGRAPH 2015 Emerging Technologies (SIGGRAPH '15)*. ACM, New York, NY, USA, 4:1–4:1. DOI:**http://dx.doi.org/10.1145/2782782.2792495**
6. Denys J. C. Matthies, Bernhard A. Strecker, and Bodo Urban. 2017. EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input Through Facial Expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1911–1922. DOI: **http://dx.doi.org/10.1145/3025453.3025692**
7. Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity Facial and Speech Animation for VR HMDs. *ACM Trans. Graph.* 35, 6 (Nov. 2016), 221:1–221:14. DOI: **http://dx.doi.org/10.1145/2980179.2980252**
8. G. C. D. Silva, M. J. Lyons, S. Kawato, and N. Tetsutani. 2003. Human Factors Evaluation of a Vision-Based Facial Gesture Interface. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, Vol. 5. 52–52. DOI:**http://dx.doi.org/10.1109/CVPRW.2003.10055**
9. K. Suzuki, F. Nakamura, J. Otsuka, K. Masai, Y. Itoh, Y. Sugiura, and M. Sugimoto. 2017. Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display. In *2017 IEEE Virtual Reality (VR)*. 177–185. DOI: **http://dx.doi.org/10.1109/VR.2017.7892245**