HMD-Based Facial Performance Capture

USC ICT GFY 15 Seedling Proposal

1 Abstract

Impending advances in the quality and affordability of Head-Mounted Displays (HMDs) such as the Oculus Rift and the Sony Morpheus indicate that the delivery of virtual training may be done increasingly effectively through such devices. As in many simulation applications including ICT projects, the user in the HMD is represented as an avatar present in the virtual environment, which can be seen by other people experiencing the simulation either as audience members or as other active participants. Such avatars would be able to mirror the facial expressions naturally and appear to mouth the words being spoken by the person wearing the HMD. Sony's SOEmote system (https://www.soe.com/soemote/) uses a web camera and a system similar to the Image Metrics LiveDriver (http://www.image-metrics.com/livedriver/) to drive avatars with the user's facial expressions as they participate in traditional multi-user games. However, for users experiencing a simulation with an HMD, the face is largely obscured from observation by a facial performance capture system. In this project, we hypothesize that a multi-sensory approach can be used to record an HMD user's facial performance intent and map it realistically onto a virtual character. The multisensory approach will include invisible light close-focus cameras observing the eyes from within the HMD, pressure and motion sensors around the facial interface to the HMD, and a close-focus camera at the bottom of the HMD to observe the mouth. We will train the system using data from an unoccluded facial performance capture system, and evaluate its effectiveness for generating facial performance parameters for novel facial expressions.

2 Background, Motivation, and Opportunity

Background A great deal of work has been done on facial performance capture [Williams 1990]. Modern systems successfully use a variety techniques, including depth sensors [Weise et al. 2011], passive multi-view stereo [Beeler et al. 2011], and monocular web cameras [Cao et al. 2013]. Some techniques can leverage data from cameras mounted directly to the head in the form of "head-cam" devices, used on movies such as Avatar and research projects such as [Jones et al. 2010], allowing the user to physically walk within an environment. However, to our awareness, there is no system available which can enable facial motion capture for a user wearing a head-mounted display where the face is largely occluded from view. One facial performance capture technique which could be applied in this circumstance is electroencephalography (EEG) to detect muscle activations from electrical currents sensed with electrodes in contact with the skin. Such a system was used

experimentally on the Sony Imageworks' movie *Beowulf* for eye gaze estimation [Robertson 2007], but the results were somewhat unreliable and required carefully placing electrodes with high sensitivity on the face. Nonetheless, the problem is important because sensing the user's facial motion is crucial to effectively representing them as an expressive avatar in a virtual training simulation.

Motivation and Opportunity The opportunity we will leverage is to augment an HMD with four different types of sensors: eye cameras, motion sensors, a mouth camera, and a microphone, together with a training procedure to map the detected data to facial performance actions. This will allow HMD wearing participants to appear in a virtual environment in a way that their avatar faithfully mirrors their facial expressions and performance, including being able to make eye contact virtually with co-participants.

Fig. 1. (Left) An HD Oculus Rift HMD taken apart to show the 6" flat-panel display and lens optics. (Center) The facial interface of the Oculus rift, showing where motion sensors may be placed. (Right) View of an HMD wearer's mouth while speaking from a camera attached below the back for edge of the HMD.

Addressing S&T Required Capabilities This seedling project will enhance the S&T required capability of (T-4) Enhanced Gaming Capability by allowing multiple training game participants use believable facial communication which using low-cost head-mounted displays for experiencing the environment. It will also help provide (T-7) Virtual Human Capabilities of providing sophisticated persona capable of natural interaction with live participants since the facial expressions of the live participants using HMDs will be available to the virtual human AI algorithms. In addition, the virtual human avatars of the live participants will have enhanced photorealistic capability (T-7) by being able to mirror the facial performance intent and eye gaze of their human counterparts.

3 Research Objective, Approach, and Milestones

Our objective is to show how it may be possible to record facial performance in an effective manner while a user is wearing an HMD. The research will be conducted will be in two parts. The first is to design the hardware which is capable of recording the facial motion information to derive the necessary performance data from the HMD device. The second is to design a process for interpreting the data so that it can be used to drive a digital face model of the avatar matching the wearer's facial performance intent. With these

hardware and software systems in place, we will be able to test the hypothesis that data which can be sensed by an HMD augmented with additional hardware can be used to record the intended facial performance of the user.

The hardware will be designed to record the performance intent of the face from data sensible from the HMD platform. We say performance *intent* since we recognize that facial movement will be somewhat limited by contact with the HMD, namely, the rim of the structure forming a shaded space between the eyes and the lenses of the HMD. The parts of the facial motion intent we need to sense are the mouth, the gaze direction, the skin around the eyes, the forehead, and the cheeks.

The mouth is the easiest to record since it not covered by the HMD and can be observed by a camera. The right panel of Fig. 1 shows the view of the mouth recordable from a small camera sensor attached to the bottom of the HMD fitted with a miniature fisheye lens. The entire mouth is viewable, and data from such imagery could be used to drive facial animation using, for example, an Active Appearance Model (AAM) [Cootes et al. 1998]. The image of the mouth will be affected by ambient illumination, which will change as the HMD user turns their head to look around or physically walks within a space. To counteract this, active illumination from white LEDs could provide a constant illumination condition on the mouth. Additionally, the LEDs could be pulsed brightly at the frame rate of the camera, with the camera's shutter set to a very short exposure time, to ensure that the active illumination on the mouth will dominate the ambient illumination, as shown in [Jones et al. 2010].

The motion of the skin near the eyes is more challenging to record since it is inside the shade of the HMD, close to the HMD lenses, in a relatively dark environment, and visible active illumination would distract from the imagery viewed by the user. Fortunately, thanks to miniaturization from the mobile communications and medical industries, cameras are quite small and several could be placed within hood of the HDM looking at the skin around the eyes. To avoid visible distraction, we propose to use infrared LEDs to light the skin and infrared-only cameras to record the reflections. Since skin exhibits a greater degree of translucency in the infrared spectrum due to subsurface scattering, we will see if parallel-polarization to accentuate specular reflections helps to generate a strong enough detectable signal for recording the motion of the skin beneath the surface. This subsystem may resemble the technique employed by infrared laser mouse pointer devices, which project an infrared laser pattern onto the surface below the mouse and use simple image processing algorithms to detect the direction and the amount of motion of the surface under the mouse.

Detecting the eye gaze of a user wearing an HMD has been previously explored for the purpose of concentrating rendering computation where the user happens to be looking. One technique employed in [Miyashita et al. 2008] is electrooculography (EOG), where electrode sensors placed around the eye sense the electrical stimuli to the eye muscles in order to determine how the gaze is being directed. However, this was only found to be

69% effective at determining the gaze between eight positions in a grid, which is insufficient for producing the few degrees of accuracy [Chen 2002] needed for proper eye contact for a virtual avatar. Better results were reported by [Lee and Park 2007], which used a corneal model and colluminated infrared LEDs and a camera to detect gaze direction. We propose to implement such a system using the same cameras and illumination with which we plan to detect the motion of the skin around the eyes. Also, we will test to see whether the eye's reflection of the images displayed by the HMD in the visible spectrum can be used in determining the eye gaze direction.

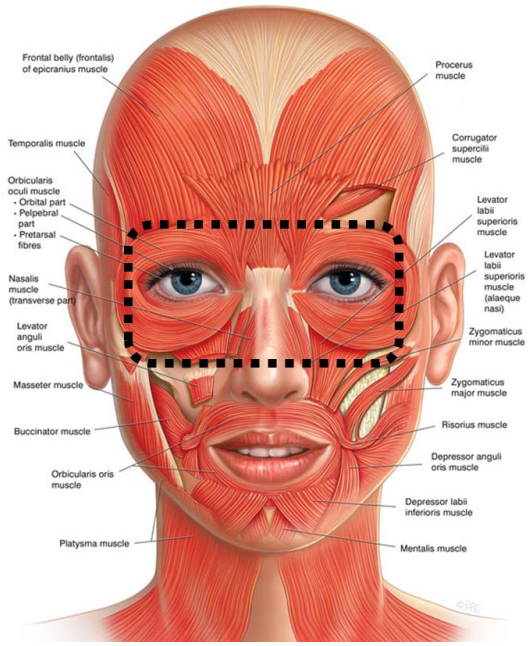

Fig. 2. The principal muscles of the face, and the approximate area of the skin in direct contact with a typical HMD device.

We have now proposed methods to acquire data within an HMD for eye gaze direction, the motion of the skin around the eyes, and mouth animation. The part of the face not yet observed includes the forehead and cheeks, whose motions are also part of the expressive repertoire. As seen in Fig. 2, significant parts of these areas are directly beneath the padded interface between the HMD and the face. Thus, to detect the motion of these areas, we propose to physically detect the motion of the skin using pressure sensors. We will test that multiple dimensions of motion (left-right, up-down, and in-out) can be usefully detected at various points around the rim of the HMD to determine the muscle activations of the frontalis and corrugator muscles of the forehead (providing raised eyebrows and knitted brows), the lateral portions of the orbicularis occuli (providing additional information about the motion of the skin around the eye), and the nasalis muscles (responsible for scrunched noses).

The second major area of inquiry of this project is to train a mapping so that the multisensor performance data can be efficiently mapped into blendshape facial performance data able to drive a realistic avatar. We note that the HMD will likely, to some extent, restrict the motion of the face due to the contact points. Thus, the mapping we wish to determine is from the detected facial activations to the *intended* facial expression the user is trying to make.

In this proposal, we will test and evaluate the following approach to determining this mapping. For a number of sample subjects, we will guide the subject to make a series of facial expressions based on the Facial Action Coding System (FACS) [Ekman and Friesen 1978] both while wearing and while not wearing the HMD. While not wearing the HMD, the user will be prompted through images of the facial expressions shown on a monitor, and the HMD itself will show the expression images again for the HMD capture. While outside the HMD, we will use an image-based facial performance capture system such as FaceShift (www.faceshift.com) based [Weiss et al. 2011] to determine the facial blendshape parameters for each expression. We will then collect the corresponding data from the multimodal HMD sensors sensing the facial muscle motion and the visual appearance of the mouth and the skin around the eyes. (We propose that the eye gaze will be driven directly from the eye gaze measurement subsystem.) From this training data, we will determine the facial performance mapping between the data sensed from the HMD and the blendshape parameters from the traditional facial performance capture system.

4 Evaluation Criteria

Our metric of success is whether the facial motion parameters detected with the sensors HMD can be provide comparable facial performance capture data to an unoccluded facial performance capture approach. For a number of subjects, after training the system using a subset of the FACS poses, we will ask the subjects to create novel facial expressions not in the FACS training set. These poses will also be taken from randomly chosen moments of an extended facial performance. We will then be able to plot the degree of correspondence between the facial blendshape poses detected by each of the systems. We will report which facial poses are detected most consistently by the proposed system, and which facial poses are detected least successfully, and draw conclusions regarding the success of the system and avenues for future work on improvements.

We will also evaluate the effectiveness of the system qualitatively by asking users to compare the facial motion performance synthesized from data recorded of the HMD wearer to that of similar performances recorded with the unoccluded facial capture system. A series of users will be able to rate the expressiveness, realism, and engagement of virtual character facial performances recorded from both the HMD-based capture system and the unoccluded capture system. The results will be tabulated and the performance of the two systems will be compared.

The anticipated benefit will be the development and demonstration of a practical technique for a person to more fully inhabit and communicate as an avatar within a head-mounted virtual environment system.

5 Research Staff (list only)

Andrew Jones (principal), Xueming Yu, Jay Busch, Graham Fyffe, Thai Phan, and advisors Mark Bolas and Paul Debevec

6 ROM cost estimate

$80K for equipment (cameras, pressure sensors, computer), materials, and effort

7 References

[Beeler et al. 2011] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. 2011. High-quality passive facial performance capture using anchor frames. In *ACM SIGGRAPH 2011 papers* (SIGGRAPH '11), Hugues Hoppe (Ed.). ACM, New York, NY, USA, , Article 75 , 10 pages. DOI=10.1145/1964921.1964970 http://doi.acm.org/10.1145/1964921.1964970

[Cao et al. 2013] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 2013. 3D shape regression for real-time facial animation. *ACM Trans. Graph.* 32, 4, Article 41 (July 2013), 10 pages. DOI=10.1145/2461912.2462012 http://doi.acm.org/10.1145/2461912.2462012

[Chen 2002] Milton Chen. 2002. Leveraging the asymmetric sensitivity of eye contact for videoconference. In*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '02). ACM, New York, NY, USA, 49-56. DOI=10.1145/503376.503386 http://doi.acm.org/10.1145/503376.503386

[Cootes et al. 1998] T.F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. ECCV, 2:484–498, 1998

[Ekman and Friesen 1978] P. Ekman and W. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, 1978.

[Jones et al. 2010] Andrew Jones, Graham Fyffe, Xueming Yu, Alex Ma, Jay Busch, Mark Bolas, and Paul Debevec. 2010. Head-mounted photometric stereo for performance capture. In *ACM SIGGRAPH 2010 Emerging Technologies* (SIGGRAPH '10). ACM, New York, NY, USA, , Article 14 , 1 pages. DOI=10.1145/1836821.1836835 http://doi.acm.org/10.1145/1836821.1836835

[Lee and Park 2007] Eui Chul Lee and Kang Ryoung Park. 2007. A study on eye gaze estimation method based on cornea model of human eye. In *Proceedings of the 3rd international conference on Computer vision/computer graphics collaboration techniques* (MIRAGE'07), André Gagalowicz and Wilfried Philips (Eds.). Springer-Verlag, Berlin, Heidelberg, 307-317.

[Miyashita et al. 2008] Hiromu Miyashita, Masaki Hayashi, and Ken-ichi Okada. Implementation of EOG-based Gaze Estimation in HMD with Head-tracker. 18th International Conference on Artificial Reality and Telexistence, 2008.

[Robertson 2007] Barbara Robertson. Beowulf Effects. http://www.cgsociety.org/index.php/CGSFeatures/CGSFeatureSpecial/beowulf_effects . 2007.

[Sifakis et al. 2005] Eftychios Sifakis, Igor Neverov, and Ronald Fedkiw. 2005. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Trans. Graph.* 24, 3 (July 2005), 417-425. DOI=10.1145/1073204.1073208 http://doi.acm.org/10.1145/1073204.1073208

[Weiss et al. 2011] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. *ACM Trans. Graph.* 30, 4, Article 77 (July 2011), 10 pages. DOI=10.1145/2010324.1964972 http://doi.acm.org/10.1145/2010324.1964972

[Williams 1990] Lance Williams. 1990. Performance-driven facial animation. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques* (SIGGRAPH '90). ACM, New York, NY, USA, 235-242. DOI=10.1145/97879.97906 http://doi.acm.org/10.1145/97879.97906