Real-time 3D Face Reconstruction and Gaze Tracking for Virtual Reality

Shu-Yu Chen[1,2] Lin Gao[1,*] Yu-Kun Lai [3] Paul L. Rosin [3] Shihong Xia[1,*]

[1] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences

[2] University of Chinese Academy of Sciences

[3] School of Computer Science & Informatics, Cardiff University

ABSTRACT

With the rapid development of virtual reality (VR) technology, VR glasses, a.k.a. Head-Mounted Displays (HMDs) are widely available, allowing immersive 3D content to be viewed. A natural need for truly immersive VR is to allow bidirectional communication: the user should be able to interact with the virtual world using facial expressions and eye gaze, in addition to traditional means of interaction. Typical application scenarios include VR virtual conferencing and virtual roaming, where ideally users are able to see other users' expressions and have eye contact with them in the virtual world. Despite significant achievements in recent years for reconstruction of 3D faces from RGB or RGB-D images, it remains a challenge to reliably capture and reconstruct 3D facial expressions including eye gaze when the user is wearing VR glasses, because the majority of the face is occluded, especially those areas around the eyes which are essential for recognizing facial expressions and eye gaze. In this paper, we introduce a novel real-time system that is able to capture and reconstruct 3D faces wearing HMDs and robustly recover eye gaze. We demonstrate the effectiveness of our system using live capture and more results are shown in the accompanying video.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Virtual reality; Human-centered computing—Human computer interaction (HCI)—Interaction techniques—Gestural input

1 INTRODUCTION

With the rapid development of computing technology, virtual reality (VR) is becoming increasingly mature. Various types of portable head-mounted displays (HMDs), a.k.a. VR glasses, have been developed. Current commercial VR systems such as Microsoft HoloLens and HTC Vive achieve interaction in the game settings using additional sensors to capture the pose and motion of the user. However, arguably the most natural form of interaction between human subjects is via facial expression and eye contact. Therefore, capturing facial expressions and eye gaze while wearing HDMs is one of the urgent problems to be solved in VR technology.

Reconstruction of facial expression has received attention recently. Wang et al. [7] developed the first system that is able to simultaneously perform 3D facial expression reconstruction and eye gaze tracking with a web-camera. However, their method cannot deal with occlusion. The methods proposed by Li et al. [4] and Olszewski et al. [5] are able to reconstruct 3D facial expressions while wearing VR glasses. Unlike such works, we propose to directly *capture* occluded facial features within VR glasses by using 3 infrared (IR) cameras. To reconstruct 3D facial expressions, we detect feature points from the output of the IR cameras, and use them jointly to drive 3D facial models. We further propose a new eye gaze tracking method based on sampling and correlation of 3D directions with captured 2D IR images of eyes. While eye tracking in VR has been addressed by commercial products such as FOVE, we are not aware of published academic work for this problem. Our real-time system captures subtle expressions and allows free head movement. Our method is robust to occlusion of faces and eye blinking.

*Corresponding authors.

E-mail addresses: gaolin@ict.ac.cn (Lin Gao), xsh@ict.ac.cn (Shihong Xia).

Figure 1: Our hardware setup: VR glasses fitted with three infrared (IR) cameras. (a) The cameras shown in the red and green boxes capture the left and right eye images respectively. (b) The camera shown in the magenta box is used to capture unoccluded facial motion.

2 SYSTEM OVERVIEW

To address the facial occlusion by VR glasses, and considering that VR content can be enjoyed when the environment is dark, we use three infrared cameras shown in Fig. 1, two inserted in the VR glasses for eye images, and one for facial images. Seven infrared lamps are fitted internally to each side of the device wall. Camera locations are carefully chosen to avoid affecting the user's views.

In order to obtain the posture of the head, we use the sensor of a cell phone fitted to the VR glasses to get the orientation information. When wearing the device, the relative spatial relation between the phone and the head is fixed, so the cell phone's orientation sensor provides the rotation information of the head. We use the OmniVision OV7675 image sensors which produce 640×480 infrared images at 30 fps. The three cameras are synchronized by activating each CCD's VSYNC pin simultaneously. The overall workflow of our method is summarized in Fig. 2.

Detection of Feature Points. Our method for 3D facial expression reconstruction starts with detecting feature points. Since we use three infrared cameras simultaneously capturing the images of different facial areas, we detect feature points in each individually. IR images have different characteristics with grayscale and color images, and the IR images are also captured from an unusual view. We are not aware of existing annotated facial databases with infrared images, and so we captured images and labeled them manually. We captured IR images for 30 subjects and altogether manually labeled 6000 images of eyes and 3000 images of occluded faces. 14 feature points are landmarked in each eye image, out of which 6 are on the eyebrow, 8 are around the eye. 29 feature points are chosen on the face excluding the eyes and eyebrows. Our method thus has a total of 67 feature points.

We employ the method [3] to extract feature points, due to its efficiency, accuracy and reliability. Given the labeled training data, the method takes a set of triplets involving an input image, initial feature point positions, and an update vector from the landmarked datasets and learns a set of cascaded regression trees using gradient

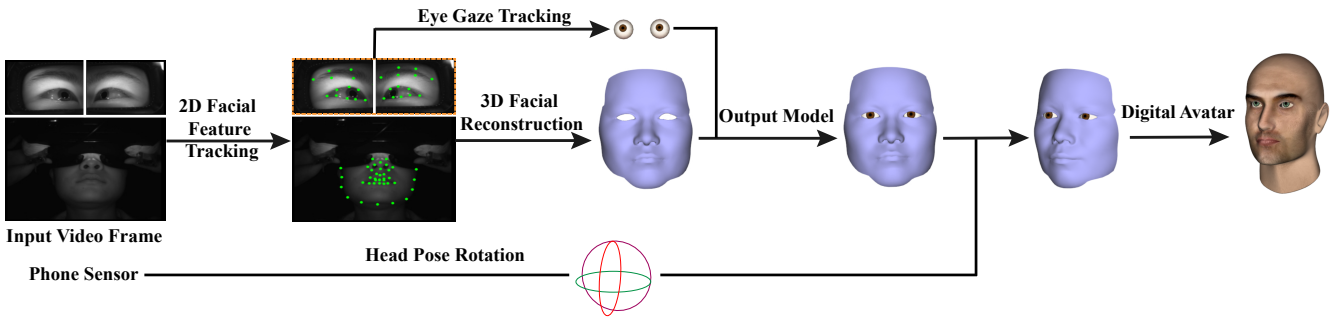


Figure 2: Overview of the workflow of our method.

tree boosting. In the training stage, we build three models which detects a set of feature points from each camera.

3D Reconstruction using Multilinear Fitting. After detecting the feature points, we reconstruct the 3D facial expression using a multilinear model [2] which contains 150 subjects, each with 47 expressions. A new 3D face can be synthesized by specifying the identity parameters and the expression parameters. Traditional methods only use one camera projection matrix. Since there are three cameras in our system, we need to put three feature point sets into a consistent coordinate system. For this, we use a multi-camera self-calibration method [6] to get the intrinsic and extrinsic parameters.

We formulate the 3D facial reconstruction problem as optimizing model parameters and a transformation such that the transformed 3D face model when projected to individual camera views has landmarks as close as possible to the detected feature points. We use Google-Ceres [1] which can be used to model and solve large-scale non-linear least squares problems. The problem is optimized by alternating the following steps iteratively: 1) optimizing the global transformation; 2) optimizing identity parameters, and 3) optimizing expression parameters. After a short duration, assuming the identity is correctly identified and kept unchanged, we keep identity parameters and the global transformation fixed, and only optimize the expression parameters for the remaining frames. This helps to further improve the efficiency of our method, achieving real-time performance. The 3D reconstruction obtained using the optimization above generally works well. However, it may produce slight jittering which may not be visually attractive. To address this, we add a smoothness constraint to improve temporal coherence.

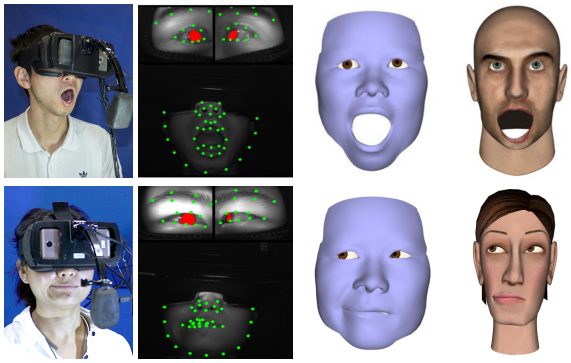
3D Eye Gaze Tracking. To facilitate eye contact in VR settings, our method also recovers 3D eye gaze, including the position of the eye ball center c, the radius of the iris and pupil region r and the eye gaze direction $d = (\phi, \theta)$ using 3D spherical coordinates. c is recovered using multilinear fitting by augmenting each training model with the eye ball center location, r is estimated by analyzing edges from the eye IR image in the initial step, and d is obtained by sampling and choosing a direction such that the rendered eye region has the highest Pearson correlation with the eye IR image.

3 Experimental Results

We perform qualitative evaluation as shown in Fig. 3. Since there is no ground truth, for each example we show a color image of the setup using an extra RGB camera (Fig. 3(a)), the three IR images captured (Fig. 3(b)) and the constructed 3D face and eye gaze (Fig. 3(c)), and use it to drive a 3D avatar (Fig. 3(d)). Note that due to space restriction of VR glasses, the infrared images for the eyes are taken from views at an angle rather than frontal (see the hardware setup in Fig. 1). More results, especially live demos, are provided in the accompanying video.

4 Conclusions

In this paper, we introduce a novel method that can robustly reconstruct 3D facial expressions and eye gaze in real time. Our equipment is easily obtained by fitting simple VR glasses with small infrared



(a) (b) (c) (d)

Figure 3: 3D facial expression reconstruction and eye gaze tracking. (a) The picture captured by an extra RGB camera to show the setup. (b) The three captured IR images. (c) The reconstructed 3D face and eye gaze. (d) An avatar driven by the captured 3D face.

LEDs and cameras and is cheap to build, whereas Li et al. [4] and Olszewski et al. [5] developed their methods on expensive HMD equipment (Oculus and Fove) with built-in sensors. Our device also captures the entire eyes and eyebrows, and uses all the information for accurate 3D face reconstruction. In the future, we will investigate combining our technique with advanced HMDs. Many potential applications in virtual reality can benefit from real-time 3D reconstruction in this work. We will investigate these in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61502453, No. 61772499 and No. 61611130215), Royal Society-Newton Mobility Grant (No. IE150731), the Science and Technology Service Network Initiative of Chinese Academy of Sciences (No. KFJ-STS-ZDTP-017), and the Knowledge Innovation Program of the Institute of Computing Technology of the Chinese Academy of Sciences under Grant No. ICT20166040.

References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. `http://ceres-solver.org`, 2017.

[2] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3D shape regression for real-time facial animation. *ACM Trans. Graph.*, 32(4):41, 2013.

[3] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.

[4] H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls, and C. Ma. Facial performance sensing head-mounted display. *ACM Trans. Graph.*, 34(4):47, 2015.

[5] K. Olszewski, J. J. Lim, S. Saito, and H. Li. High-fidelity facial and speech animation for VR HMDs. *ACM Trans. Graph.*, 35(6):221, 2016.

[6] T. Svoboda, D. Martinec, and T. Pajdla. A convenient multicamera self-calibration for virtual environments. *PRESENCE: teleoperators and virtual environments*, 14(4):407–422, 2005.

[7] C. Wang, F. Shi, S. Xia, and J. Chai. Realtime 3D eye gaze animation using a single RGB camera. *ACM Trans. Graph.*, 35(4):118, 2016.