Synthesis-Based Low-Cost Gaze Analysis

Zhuoqing Chang(✉), Qiang Qiu, and Guillermo Sapiro

Electrical and Computer Engineering, Duke University, Durham, NC, USA
{zhuoqing.chang,qiang.qiu,guillermo.sapiro}@duke.edu

Abstract. Gaze analysis has gained much popularity over the years due to its relevance in a wide array of applications, including human-computer interaction, fatigue detection, and clinical mental health diagnosis. However, accurate gaze estimation from low resolution images outside of the lab (in the wild) still proves to be a challenging task. The new Intel low-cost RealSense 3D camera, capable of acquiring submillimeter resolution depth information, is currently available in laptops, and such technology is expected to become ubiquitous in other portable devices. In this paper, we focus on low-cost, scalable and real time analysis of human gaze using this RealSense camera. We exploit the direct measurement of eye surface geometry captured by the RGB-D camera, and perform gaze estimation through novel synthesis-based training and testing. Furthermore, we synthesize different eye movement appearances using a linear approach. From each 3D eye training sample captured by the RealSense camera, we synthesize multiple novel 2D views by varying the view angle to simulate head motions expected at testing. We then learn from the synthesized 2D eye images a gaze regression model using regression forests. At testing, for each captured RGB-D eye image, we first repeat the same synthesis process. For each synthesized image, we estimate the gaze from our gaze regression model, and factor-out the associated camera/head motion. In this way, we obtain multiple gaze estimations for each RGB-D eye image, and the consensus is adopted. We show that this synthesis-based training and testing significantly improves the precision in gaze estimation, opening the door to true low-cost solutions.

1 Introduction

Gaze tracking is the process of analyzing human eye movement. There has been a great increase of interest in this area over the last decade due to its relevance in a wide range of applications, including human-computer interaction, fatigue detection, and clinical mental health diagnosis. In many of these applications, gaze tracking needs to be performed either from a distance or in a non-intrusive manner, therefore limiting the resolution of the acquired eye images. In this common low-resolution scenario, appearance-based methods, which use the eye appearance image to estimate gaze directly, are more popular compared to model-based methods, which use derived geometric features as input.

The advantage of appearance-based methods is that no small-scale feature need to be extracted since the high-dimensional eye image is directly mapped

C. Stephanidis (Ed.): HCII 2016 Posters, Part II, CCIS 618, pp. 95–100, 2016.
DOI: 10.1007/978-3-319-40542-1_15

to the low-dimensional gaze direction, allowing the use of low resolution images. However, the biggest limitation of appearance-based methods is that variation in eye images arise from factors other than gaze, head motion being a major contributor. In general, appearance-based methods require a large amount of training data to cover the eye appearance space. For example, Zhang *et al.* [8] collected more than 200,000 images over a 3-month period to train a deep neural network.

In this paper, we propose an appearance-based approach for gaze estimation using synthesized eye images to efficiently address the problem of acquiring large amounts of data. Sugano *et al.* [5] proposed a synthesis approach to generate multiple views of the eye from a 3D reconstruction of the face. However, their method is only able to synthesize eye appearance changes due to head motion and not the (gaze-related) eye movement itself. In addition, they use eight industry-level cameras for the multi-view reconstruction requiring complex calibration and high cost. In contrast, we use one Intel RealSense 3D camera [1], a low-cost commercial RGB-D camera similar to Microsoft's Kinect but better in short range capabilities, to acquire 3D surface data of the face, from which multiple views of the eye are synthesized. We show how the RealSense can be used to accurately align and segment eye images automatically. In addition, we propose a novel approach to synthesize eye appearance changes due to eye movements using a linear method. Finally, we demonstrate that a test-by-synthesis approach is able to further improve the gaze estimation performance.

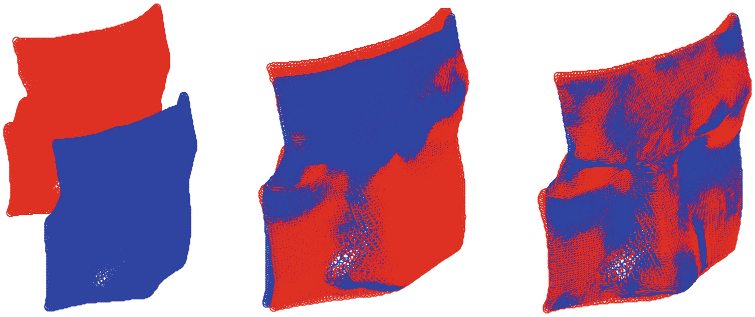
2 Methods

Given the RealSenses pre-registered RGB images and 3D point clouds, our synthesis-based gaze estimation approach can be summarized in four steps: face alignment, head pose synthesis, eye movement synthesis, and random forest regression. Face alignment normalizes the head pose to a canonical viewpoint, from which multiple pose eye images are generated by projecting the 3D eye surface in different directions. Within each pose, a linear approach is used to synthesize eye appearance changes to further increases the number of training (and then testing) samples. A random forest is trained for each pose to learn the mapping from the eye appearances to the gaze directions.

2.1 Face Alignment

The main goal of face alignment is to rectify the head pose to a canonical viewpoint and scale such that the eye appearance variation is not affected by head motion. Accurate alignment also allows for consistent segmentation of the eye region, which is critical in appearance-based methods.

Our method first locates facial landmarks of the face in the RGB image using Intraface [7], from which 3D landmark points are derived by mapping the 2D points to the 3D point cloud. With the 3D landmark points, the region between the forehead and mouth is cropped as the face model. We chose to use this region since it is robust to occlusion and expression changes. A frontal facing frame is used as a reference model and all other frames are aligned to it using a two step approach. First a rigid transformation is obtained by aligning 11 facial landmarks to the corresponding reference model landmarks. The rigid transformation is used to approximately align the facial point cloud to the reference model, after which Iterative Closest Point (ICP) [4] is used to further refine the alignment. Since the point clouds are from the same person, we achieve near perfect alignment using this method. The alignment process is shown in Fig. 1.



(a) Before alignment (b) Rigid alignment (c) ICP alignment

Fig. 1. Face alignment process. (Color figure online)

2.2 Head Pose Synthesis

We use an approach similar to [5] to generate different views of the eye by projecting the point cloud in different directions. The projection angle was chosen to be from −10 to 10 degrees with 5 degree intervals in both the horizontal and vertical direction, yielding a total of 25 different views. This range was chosen such that eye appearance changes due to head pose variation were clearly noticeable, but at the same time not distorting the eye images to an extent where gaze information is lost. A fixed eye region could be defined for each projection angle since the projections are aligned. Eye images are then sampled within the region at a fixed resolution of 24×40 pixels (Fig. 2a). Our proposed method is able to handle scaling and multi-resolution images by tuning the sampling resolution within the defined eye region.

2.3 Eye Movement Synthesis

We synthesize different eye appearances for a given head pose using a linear approach. Lu *et al.* [3] showed that gaze positions and eye appearances have a similar manifold when the head is static, and that gaze could be computed as a linear combination of a sparse set of training samples. We reverse this idea and use the linearity of gaze directions to synthesize eye images.

For a set of eye images $\{\boldsymbol{x}_i\} \in \mathbb{R}^d$, their corresponding 3D gaze direction is denoted as $\{\boldsymbol{g}_i\} \in \mathbb{R}^3$. Let $\{\tilde{\boldsymbol{g}_j}\} \in \mathbb{R}^3$ denote the set of gaze directions for which we wish to synthesize eye images $\{\tilde{\boldsymbol{x}_j}\} \in \mathbb{R}^d$. The idea is to solve

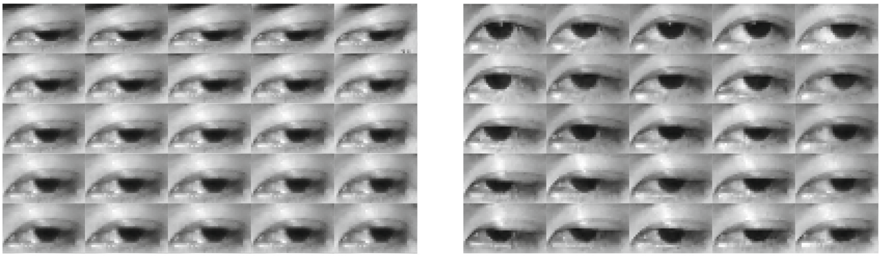
$$\min |\boldsymbol{\alpha}_j|_1 \quad s.t. \quad \tilde{\boldsymbol{g}_j} = D_g \cdot \boldsymbol{\alpha}_j \tag{1}$$

for the sparse weight vector $\boldsymbol{\alpha}_j \in \mathbb{R}^n$. $D_g = [\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_n]$ is the dictionary of gaze directions. The synthesized eye images are then computed by

$$\tilde{\boldsymbol{x}_j} = D_x \cdot \boldsymbol{\alpha}_j \tag{2}$$

where $D_x = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ is the dictionary of eye images. Samples of synthesized eye images are shown in Fig. 2b.

The range of gaze directions to synthesize $\tilde{\boldsymbol{g}_j}$ was chosen to be the same as the range of the original set $\boldsymbol{g}_i$. Typically the range is about 40 degrees in the horizontal direction and 20 degrees in the vertical direction. The interval was set to 0.2 degrees in each direction which results in approximately 20,000 synthesized images. Note that the synthesis gaze range mentioned here and the synthesis head pose range in Sect. 2.2 are two different concepts. We solve Eq. 1 using Orthogonal Matching Pursuit (OMP) [6]. The eye movement synthesis process takes less than a second using a Matlab implementation.



(a) Head pose synthesis

(b) Eye movement synthesis

Fig. 2. Synthesized eye images using (a) head pose synthesis, and (b) eye movement synthesis

2.4 Gaze Estimation Using Random Forests

We chose to use random forests [2] to learn the mapping between eye appearance and gaze direction because it has been shown to work well in related tasks [5,8]. Each eye image is converted to grayscale, raster-scanned to form a feature vector and normalized to have unit length. The gaze direction is defined as a unit vector that points from the center of the eye, an arbitrary fixed point relative to the face model, to the visual target.

We do not cluster samples according to head poses as in [5,8] since we only have 25 discrete head poses after face alignment and head pose synthesis. We instead train a random forest for each head pose, which we then use to independently estimate the corresponding synthesized testing image. The gaze output for each random forest is compensated by the head pose synthesis angle and the face alignment angle before being aggregated to get the final estimation.

3 Experiments

3.1 Data Collection

The data collection system consists of a 14 inch laptop with a 1600×900 display and a RealSense camera. RGB images captured by the RealSense were set to 1920×1080 pixels while the point cloud resolution was limited to 640×480. The two devices have been pre-calibrated such that pixel coordinates on the laptop screen can be mapped to its 3D coordinate in the camera reference system.

During recording sessions, 3 participants, 2 male and 1 female, were asked to sit approximately 40 cm in front of the laptop and look at 60 visual targets that appeared randomly on the screen. Participants were allowed to have there head relaxed and natural but were told to refrain from making large head movements. After post processing, 10 frames were kept for each visual target per participant. Half of the data is randomly selected as the training data and the other half as the testing data.

3.2 Results

We evaluated the performance of our proposed method by comparing three sets of experiments: no-synthesis, half-synthesis, and full-synthesis. The no-synthesis experiment is used as a baseline comparison, where only the normalized eye images are used for training and testing. The half-synthesis experiment uses eye movement synthesis on the training set to train a larger regression forest on which the testing set is tested. The full-synthesis experiment uses both eye movement and head pose synthesis on the training set to train 25 regression forests corresponding to different head poses. Head pose synthesis is also used

Table 1. Comparison of gaze estimation error

Participant	No-synthesis	Half-synthesis	Full-synthesis
1	$1.32 \pm 1.54°$	$1.18 \pm 1.37°$	$0.85 \pm 0.98°$
2	$1.40 \pm 1.73°$	$1.25 \pm 1.48°$	$1.07 \pm 1.29°$
3	$1.25 \pm 1.51°$	$0.91 \pm 1.09°$	$0.77 \pm 0.93°$
Average	$1.32 \pm 1.59°$	$1.11 \pm 1.31°$	$0.90 \pm 1.07°$

on the testing set to generate different views of each testing image which is then tested using the corresponding regression forest. After compensating the output of each regression forest with its associated pose angle, the median is adopted as the final gaze angle. The estimation errors are given in Table 1.

4 Conclusion

We proposed a novel synthesis-based approach for gaze analysis using a low-cost commercial 3D camera. We synthesize eye appearance changes due to both head motion and eye movement to dramatically increase the amount of training data. In the testing stage, we use synthesis to obtain multiple gaze estimates. Moreover, our approach works with low resolution images and is able to handle slight natural head motion. The reported results demonstrate that synthesis can be used to compensate the lack of calibration data and greatly improve gaze estimation accuracy and stability, justifying the feasibility of low-cost gaze analysis.

Acknowledgments. Work partially supported by NSF, ONR, ARO, NGA, and AFOSR.

References

1. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3D gaze estimation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1821–1828, June 2014. doi:10.1109/CVPR.2014.235
2. Intel RealSense 3D Camera. http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html. Accessed 07 Apr 2016
3. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
4. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4511–4520 June 2015
5. Rusinkiewicz, S., Levoy, M.: Efficient variants of the ICP algorithm. In: 3-D Digital Imaging and Modeling, pp. 145–152. IEEE (2001)
6. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Adaptive linear regression for appearance-based gaze estimation. IEEE Trans. Pattern Anal. Mach. Intell. **36**(10), 2033–2046 (2014)
7. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inf. Theory. **53**(12), 4655–4666 (2007)
8. Breiman, L.: Random forests. Mach. learn. **45**(1), 5–32 (2001)