Real-time Retinal Localization for Eye-tracking in Head-mounted Displays

Chen Gong, Steven L. Brunton, Eric J. Seibel
Univerisity of Washington, Seattle

Laura Trutoiu, Brian T. Schowengerdt
Magic Leap Inc.

Abstract

Accurate and robust eye-tracking is highly desirable in head-mounted displays. A method of using retina movement videos for estimating eye gaze is investigated in this work. We localize each frame of the retinal movement video on a mosaicked large field of view search image. The localization is based on a Kalman filter, which embeds deep learning in the estimation process with image registration as the measurement. This algorithm is demonstrated on experiments, where the retinal movement videos are captured from a dynamic real phantom. The average localization error of our algorithm is 0.68°*, excluding the annotation error. The classic pupil-glint eye tracking method has an average error of* 0.5°-1°*, while using retina videos results in a tracking resolution of* 0.05° *per pixel, which is nearly 20 times higher than that of pupil-glint methods. The accuracy of our inherently robust method is expected to be improved with further development.*

1. Introduction

Head-mounted displays (HMD) have been explored for a wide range of applications in the fields of 3D virtual and augmented environments [3, 14]. Accurate and high-speed eye tracking is important to enable key scenarios in HMD, *e.g*. the field of view (FOV) and resolution trade-off through fovea-contingent display schemes and novel interactive interfaces for people [11, 10].

Eye-trackers embedded in HMD can be divided into invasive methods, *e.g*. scleral coil [17] and non-invasive video-based methods, the latter being more common [4]. Current video-based methods mainly use different features of the eyeball, such as iris, pupil and glint [4], and pupil-glint methods are the most widely used. These methods have an average tracking error of 0.5°-1°, while the tracking resolution of such features is around 0.7°-1° per pixel [17, 7]. It is not easy to further improve the accuracy beyond the tracking resolution.

Besides using features of the eye surface, retina images are also utilized for eye-tracking in medical field, such as eye-tracking scanning laser ophthalmoscopes (SLOs) [13]. They leverage the scanning distortion for retinal movement estimation in small FOV high-resolution images, however this technique is designed for small saccades and SLOs are not easily integrated into a HMD.

Retinal-based eye tracking in HMD has its own advantages: a higher tracking resolution without advanced sensors, linear gaze estimation models and direct localization of the fovea on the retina. Furthermore, retinal tracking provides a wide range of medical applications with the HMD.

In this paper, we present a real-time retinal localization method for eye-tracking based on retinal movement videos, where each frame is localized on a mosaicked search image. The schematic of the proposed method is shown in Fig. 1. The novelty of our method is using Kalman filter [16] to combine the performance of deep learning and the classic image registration method, where the result of deep learning is used to build the state transition model and image registration provides the measurement.

Our method is validated on the synthetic data and retinal movement videos imaged with the scanning fiber endoscope (SFE). The details of our dataset and its challenges are introduced in Section 2. Using the retina videos, the eye tracking resolution in our system is 0.05°/pixel. Our retinal localization method currently achieves 0.68° mean error not considering the annotation variation. Compared to the classic pupil-glint methods which have a low tracking resolution, we hypothesize that retinal-based eye tracking accuracy will greatly improve in future development.

2. Data Acquisition and Characteristics

Virtual retinal display (retinal scan display) for AR/VR has been proposed for a long time [5]. To maintain the compactness of the HMD system, the retinal imaging can share most of the optic path with the retinal scan display. VRD draws a scanning display directly onto the retina, thus the SFE imaging device with a scanning pattern is used since it is low cost and has a miniature probe tip [12]. SFE has a spiral scanning pattern from center to periphery and the full frame is imaged ring by ring. When the target is moving, the rings scanned at different time are from different regions, which creates movement distortions in the video frame. We take the imaging position on the retina as the ground truth when the frame is completed, thus the outer rings in each frame are closer to the ground truth.

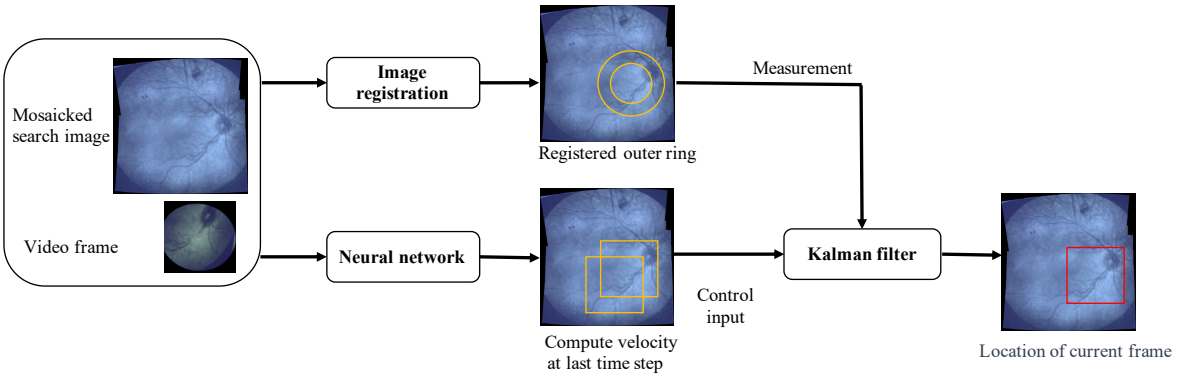

Figure 1: Schematic of proposed real-time localization method. Kalman filter is used to combine the performance of deep learning and the classic image registration method.

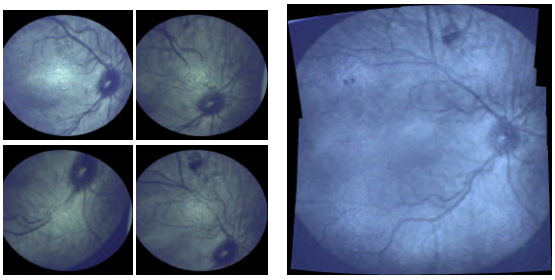

Figure 2: Example retina frames of the SFE and the mosaicked large FOV baseline image.

In data collection, we attach a retinal phantom and a laser pointer onto the tip of a robot arm (Meca500) for simulating retina movement and use a position sensitive detector (PSD) for real-time position recording of the laser beam. The PSD data can be the annotation for each frame after data preprocessing. The annotation in the current setup has a mean error of 0.35°. Fig. 2 (a) shows example of captured retinal frames, and Fig. 2 (b) is the image mosaicked from a series of frames. We can see from the donut-shaped optic disc that the images have movement distortion. Note that the retina image has many regions with similar background, the imaging has low quality with local distortions on the still frame, which increase the difficulty of localization.

3. Proposed Method

Given a large FOV mosaicked image as reference, our task is the real-time localization of the captured SFE frames onto the search image as shown in Fig. 2 (b). Because of the challenges of data, we use deep learning method to extract representative deep features for analysis. However, the neural network has uncertainty and deep features are not always reliable, thus image registration method is used to compensate the performance of deep learning. On the other hand, the result of image registration is also noisy because of the data challenges. As described above, two process are combined with the Kalman filter, where the deep learning results are embedded in the transition model and registration results are taken as the measurement in Kalman filter. The Kalman filter requirements of linear Markov model and additive Gaussian noise are satisfied in our case. In this section, we introduce the form of our state transition model and measurement in the Kalman filter respectively.

3.1. State Transition Model with Deep Learning

The state transition model assumes the true state at time k is evolved from the state at $k-1$. In the proposed method, the transition model is formed as follows:

$$\begin{bmatrix} X_k \\ Y_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} X_{k-1} \\ Y_{k-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} \dot{x_{k-1}} \\ \dot{y_{k-1}} \end{bmatrix} + w_k. \quad (1)$$

X_k, Y_k represents the position state at time k in x and y directions. w_k is the process noise drawn from a zero mean multivariate normal distribution. $\dot{x_{k-1}}, \dot{y_{k-1}}$ forms the control vector of the first order state estimation model. It is the velocity within a time unit computed from the difference between the deep neural network results at time k and $k-1$. The proposed formation allows us to embed the deep learning into a classic Kalman filter model. Here one time step is the duration between continuous frames.

The deep learning framework we use is modified from the Siamese RPN [9]. Alexnet [8] is first used to extract the deep features of the frame and search image, then the frame feature is converted into two different features with convolution layers for classification and regression respectively. Two corresponding response maps are created by the convolution of the frame and search image features. One response map is used for the target region/non-target (positive/negative) region classification, another response map predicts the position refinement at each positive position. Different from learning robust representations of a specific object in Siamese RPN, we localized different templates on the same search image. The deep feature of the search image is saved and repeatedly used after the training process. Since the imaging scale will not change much in HMD, we focus on the target position in x and y instead of the bounding box with adjustable height and width.

3.2. Measurement with Outer Ring Registration

In Kalman filter, the measurement is obtained at the current time:

$$z_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} X_k \\ Y_k \end{bmatrix} + v_k, \quad (2)$$

where z_k is the measurement obtained by the image registration, and v_k is the measurement noise similar to w_k. Our image registration is based on SIFT method. The intensity-based registration methods (cross correlation and mutual information) are not as robust as SIFT on our data, while they can be used when the captured retinal image cannot provide detectable features, such as the near-infrared image.

As described above, the outer ring in a frame indicates more accurate retinal location, while directly matching the outer ring globally is difficult because of the very sparse features it contains. To reduce the interference of similar background and few features of outer ring registration, the image registration includes two steps: coarse registration of the whole frame and fine registration with the outer ring only. In the coarse localization, we detect feature points from two images and register the frame f to the corresponding regions $\tilde{f}$ on the search image. In the outer ring registration, the feature points within the enlarged region around $\tilde{f}$ on the search image are selected, and they are rematched with the feature points falling into the outer ring region on the frame. Using an enlarged region improves the robustness of the algorithm when the matched featured points in coarse registration are concentrated in the inside area. This method also avoids repeated computation of feature points. Because of the challenges of retina images, such measurement of Kalman filter occasionally drops out, the tracking system then relies on the deep neural work only until the next successful registration.

4. Experiments

Experiments are performed on two datasets: the synthetic retina movement videos and SFE videos introduced in Section. 2. We compare the performance of the proposed tracking method with Kalman filter using the deep learning only.

The synthetic data is generated from the public retina dataset STARE [2]. We generate overall 36000 frames of retinal movement videos from 300 retina images as the training set for deep learning, and 4000 frames from 15 retina images as the test set. We add four different levels of image degradations on test video frames to evaluate the robustness of our method: 1) Gaussian noise with mean 0 and variance selected from $0.001 \sim 0.005$; 2) Rotation and shear angle from $-10° \sim 10°$ and $-5° \sim 5°$ respectively; 3) Scale change from $0.8 \sim 1.2$. The degradation level increases uniformly within the parameter ranges. It is shown in Table 1 that our method has an acceptable accuracy $0.63°$ even under the largest degradation.

Table 1: Mean errors of the proposed tracking method and using deep learning only over the synthetic data.

Degradation level	0	1	2	3	4
Deep learning only	0.35°	0.43°	0.53°	0.60°	0.65°
Kalman filter	0.20°	0.33°	0.44°	0.53°	0.63°

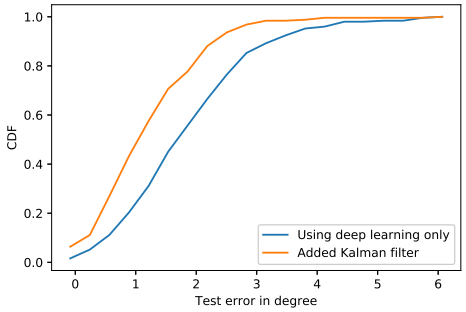

Figure 3: CDF of the retina tracking errors in degrees over 400 frames, and the annotation error is included in the CDF.

The experiment of the SFE video is implemented on one retina image [1]. We collected overall 7000 frames for training and 400 frames for test. The test errors are summarized as a cumulative distribution function (CDF) in Fig.3. We can see there are outliers over $5°$ in using deep learning only. The accuracy of the data annotation is around $0.35°$ as described before, and the mean error of our method is $0.68°$ excluding the influence of the annotation, whereas only using the neural network has a mean error of $1.01°$. The speed can reach 72fps with the GPU of Titan RTX.

5. Conclusion and Future Work

We present the application of retina-based eye tracking for HMD and a novel real-time localization method using Kalman filter to combine the performance of deep learning and image registration. To the best of our knowledge, this is the first systematic discussion of embedding retina tracking in AR/VR headset and providing algorithm solutions.

We expect to further improve the accuracy and generalize the model trained on SFE videos by enlarging the dataset from different user's retina, then evaluate our method on in vivo cases in the future. To improve the measurement in Kalman filter, deep learning can also be used to learn more robust feature points in the image registration part [15]. We predict the accuracy of the widely used pupil-glint methods may have reached its tracking resolution limit using current sensors, whereas using retina videos has the capacity for improving their degree of accuracy. Retina based eye tracking also provides a more precise gaze estimation model by computing the fovea position, and will enable more applications in the medical field such as the retinal diagnosis or surgery guidance [6].

References

[1] Widefield imaging module. *https://business-lounge. heidelbergengineering.com*.

[2] Structured analysis of the retina. *http://www.ces.clemson.edu /˜ahoover/stare/*, accessed on May 15, 2018.

[3] Michael Bajura, Henry Fuchs, and Ryutarou Ohbuchi. Merging virtual objects with the real world: Seeing ultrasound imagery within the patient. *ACM SIGGRAPH Computer Graphics*, 26(2):203–210, 1992.

[4] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. Wearable eog goggles: Seamless sensing and context-awareness in everyday environments. *Journal of Ambient Intelligence and Smart Environments*, 1(2):157–171, 2009.

[5] Thomas Adrian Furness III and Joel S Kollin. Virtual retinal display and method for tracking eye position, Nov. 13 2001. US Patent 6,317,103.

[6] Chen Gong, N Benjamin Erichson, John P Kelly, Laura Trutoiu, Brian T Schowengerdt, Steven L Brunton, and Eric J Seibel. Retinamatch: Efficient template matching of retina images for teleophthalmology. *IEEE Transactions on Medical Imaging*, 38(8):1993–2004, 2019.

[7] Hong Hua, Prasanna Krishnaswamy, and Jannick P Rolland. Video-based eyetracking methods and algorithms in head-mounted displays. *Optics Express*, 14(10):4328–4350, 2006.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018.

[10] Jannick P Rolland, Larry D Davis, and Yohan Baillot. A survey of tracking technologies for virtual environments. In *Fundamentals of wearable computers and augmented reality*, pages 83–128. CRC Press, 2001.

[11] Jannick P Rolland, Akitoshi Yoshida, Larry D Davis, and John H Reif. High-resolution inset head-mounted display. *Applied optics*, 37(19):4183–4193, 1998.

[12] Eric J Seibel, Richard S Johnston, and C David Melville. A full-color scanning fiber endoscope. In *Optical fibers and sensors for medical diagnostics and treatment applications VI*, volume 6083. International Society for Optics and Photonics, 2006.

[13] Christy K Sheehy, Qiang Yang, David W Arathorn, Pavan Tiruveedhula, Johannes F de Boer, and Austin Roorda. High-speed, image-based eye tracking with a scanning laser ophthalmoscope. *Biomedical optics express*, 3(10):2611–2622, 2012.

[14] PC Thomas and WM David. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *Hawaii International Conference on System Sciences*, pages 659–669, 1992.

[15] Prune Truong, Stefanos Apostolopoulos, Agata Mosinska, Samuel Stucky, Carlos Ciller, and Sandro De Zanet. Glampoints: Greedily learned accurate match points. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10732–10741, 2019.

[16] Greg Welch and Gary Bishop. An introduction to the kalman filter. 1995.

[17] Eric Whitmire, Laura Trutoiu, Robert Cavin, David Perek, Brian Scally, James Phillips, and Shwetak Patel. Eyecontact: scleral coil eye tracking for virtual reality. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, pages 184–191, 2016.