Multi-sensor Gaze-tracking

Joe Rice

March 2017

Abstract

In this paper, we propose a method using multiple gaze-tracking-capable sensors along with fuzzy data-fusion techniques to improve gaze-estimation. When compared to existing gaze-tracking systems, using multiple sensors should improve the effective field of view by covering more area as well as improve estimation accuracy by utilizing multiple perspectives of the scene.

1 Introduction

Gaze tracking provides indispensable utility to the field of human-computer interaction; it presents a useful observation for usability testing and provides a unique accessibility vector for computer applications. This utility is not lost on the consumer market; gaze tracking is found on most modern cellphones and the price of entry-level gaze-tracking hardware continues to decline. However, high-accuracy equipment is still cost prohibitive and lacks the flexibility necessary for certain advanced computer interfaces. In this paper, we propose a method to improve gaze-tracking performance by utilizing multiple gaze-predictors in a system.

Most modern gaze-tracking systems use a single sensor or a fixed binocular system. For most basic usability testing purposes, this is sufficient, until larger field of view requirements are considered. With the prevalence of multi-monitor displays and the current trend toward ultra wide curved displays, the field of view on most gaze-tracking equipment seems inadequate. By utilizing existing visual light cameras, we aim to increase system "flexibility", perhaps enabling the possibility of accurate remote gaze-tracking.

The remainder of this paper is laid out as follows: First, we will recall some basic gaze-tracking concepts including the common types of sensors and calibration techniques. Then, we will describe both decision-level and feature-level fusion and how we will use them in our experiment. Lastly, we will describe our experiment's evaluation process and how we will quantify the results of our experiment.

2 Background

In the following section, we provide some background information on gaze tracking systems and sensor fusion techniques. More specifically, we describe various types of sensors commonly used in remote-sensor gaze-tracking applications as well as general techniques for decision-level fusion.

2.1 Visible Light Camera

Gaze-tracking using visible light sensors is particularly appealing due to their ubiquity and affordability. One prolific example are webcams. Omnipresent in modern computing environments, webcams provide gaze-tracking utility on hardware that may already be a sunk cost. In the case of multi-monitor desktop usage, webcams built into the monitors may provide adequate observation from multiple viewpoints. In video surveillance systems, the cost of implementing gaze-tracking may be reduced to software processing. This, however, is a trait shared with the increasingly common infrared video surveillance systems.

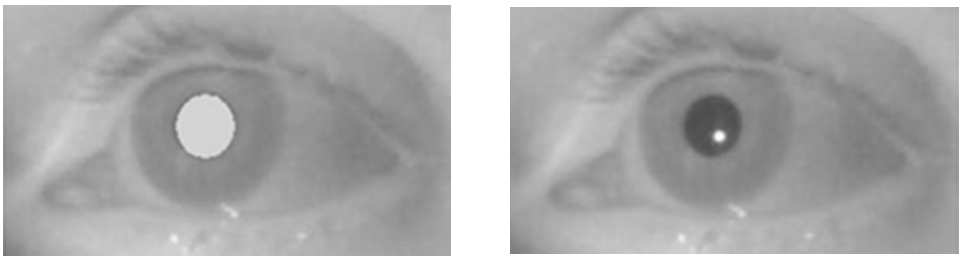
While visible light sensors are inexpensive and common, they pose additional constraints that often affect accuracy. Commonly, webcams are used as "passive" sensors. In other words, they rely on external light sources to reflect information into the sensor. A camera's "flash" is an example of an "active" visible light sensor. A passive visible light sensor in a dimly lit room may severely hinder the accuracy of a visible light gaze-tracking system.

In some cases, providing the system enough light is not an option; for example, when a using the system in dimly-lit room is desired or when the required intensity would be irritating to the human eye. A consequence of multi-sensor systems is the increased utilization environments' light.

2.2 Infrared Camera

Systems utilizing active infrared (IR) or near-infrared (NIR) sensors aim to resolve the problems posed by visible light sensors by emitting electromagnetic radiation at a wavelength undetectable by the human eye. This allows active infrared sensors to emit "infrared light" at a much higher intensity without irritating users' eyes. While some gaze-trackers utilizing infrared sensors are passive systems [1], high-accuracy systems often include (NIR) emitters [9, 4].

The intensity attainable from active infrared sensors allows the use of corneal reflection given the positioning of the emitters is centered relative to the sensor. This pupil tracking technique is considered a "bright pupil" technique [12] in contrast to "dark pupil" techniques commonly used in visible light systems. An additional benefit of "bright pupil" techniques is the disambiguation of pupils and arbitrary dark regions in the scene as well as the ease of accurately measuring the diameter of the pupil. Both techniques are shown in Figure 1.



(a) Bright pupil (b) Dark pupil

Figure 1: Pupillary tracking techniques.[1]

2.3 Depth Sensor

Modeling the human face in 3 dimensions provides a useful frame of reference for gaze tracking. Some systems [13] utilize eye corners and other facial features serve as consistent reference points for normalizing scale and rotations. Other systems [10] utilize depth sensors such as those found on RGBD cameras, like the Microsoft Kinect, to provide depth cues that aid in the 3D modeling of the scene.

[1] Images thanks to WikiMedia Commons.

3 Calibration

In this section, we recall basic calibration techniques as well as describe some of the more modern dynamic approaches.

3.1 Gaze Calibration

In many modern calibration schemes, users are simply instructed to follow a known point with their eyes; this often takes the form of following a dot on a computer display. The system assumes that users are correctly staring at the center of the dot. Predictions provide bias (mean) and error (variance) which may then be used to update the system parameters.

Adaptive gaze calibration is an appealing concept because it further simplifies setup and maintenance for the user. One technique that has been around for some time is building system configuration profiles for each user and updating them when the user calibrates [2]. Modern techniques aim to be even more dynamic by updating the user profile when the user fixates [4] on some item; this is taken as an opportunity for some minor automatic calibration.

3.2 Sensor Localization

To begin modeling a scene, sensors must know the relative positions of the other sensors. This positioning can be configured manually by the user, but this adds significant complexity to setup and is not robust to change. One method used frequently by high-accuracy binocular gaze-trackers [12] is to fix the relative position of the sensors via hardware (i.e. two sensors mounted on the same device).

Pose estimation techniques like Simultaneous Localization and Mapping [3] (SLAM) can be used to estimate the relative positions of the sensors; this simplifies setup for the user at the cost of some software complexity and possible cost of position accuracy.

4 Decision-level Fusion

In the case of systems where each sensor (or subsystem) produces a prediction (as described in Equation 1), each prediction can be aggregated to into a

single estimate representing the group decision; a linear example of this is shown in Equation 2. In the following sections, we describe a few aggregation functions to evaluate in our proposed experiment.

$$\hat{\boldsymbol{y}} = \begin{cases} \hat{y}_{1...3} : \text{prediction origin} \\ \hat{y}_{4...6} : \text{prediction direction} \\ \hat{y}_{7...9} : \text{sensor direction} \end{cases} \tag{1}$$

where $\hat{\boldsymbol{y}} =$ is an individual prediction

$$\tilde{Y} = \sum_{i=1}^{n} \hat{Y}_i \cdot \lambda_i \tag{2}$$

where $\hat{Y}_i =$ the individual prediction i

$\lambda_i =$ the aggregation weight of prediction i

$\tilde{Y} =$ the group decision

4.1 Rule-based Systems

When a problem requires the flexibility of multi-sensor systems (i.e. video surveillance), modern gaze-trackers often utilize simple rule-based control systems. One example of such a rule-base is for the sensor with the best view of the target to make the prediction. This one prediction is treated as the group prediction instead of aggregating decisions from each of the subsystems.

4.2 Fuzzy Aggregation

Fuzzy integrals with respect to fuzzy measures are a powerful tool for aggregation of arbitrary information sources. Fuzzy measures are used to describe the marginal benefit provided by each additional sensor. Fuzzy integrals are then used to compute the aggregation. By trading the additive property of linear aggregations for the weaker property of monotonicity, fuzzy integrals are able to perform arbitrary non-linear aggregations of data sources (i.e. sensor features or predictions).

In previous work [11], we applied fuzzy sensor fusion techniques to machine learning methods such as multiple kernel support vector machines. In this work, we aim to extend those techniques to aggregate gaze predictions from multiple sources. More specifically, offline processing can be used to learn an ideal aggregation function using a genetic algorithm to choose the parameters.

One assumption made in this work is that each predictor is confident in it's own predictions. A more appropriate assumption might be that each prediction is a distribution about a point and a direction. This work could be extended to support the aggregation of arbitrary distributions of sensors [6, 5].

5 Feature-level Fusion

Before the gaze estimate is calculated, features are collected from a frame of the video feed and a model of the scene is developed. Feature-level fusion is the aggregation of these individual models before a decision is made. If the entire model is aggregated, then a single prediction, representing the group decision, will be produced. If only some proper subset of the model is aggregated (i.e. the head pose estimation), then an ensemble of predictions will be produced, requiring decision-level fusion.

Features sharing similar structures may be aggregated using techniques such as those described in Section 4. However, complex systems may comprise significantly different predictors, whose features require non-trivial aggregations. One simple example of feature fusion used in modern gaze-tracking systems [7], is the use of a separate subsystem of depth sensors that build a 3D model of a user's head and share this model with the predictor. In our proposed work, we generalize this concept to any combination of predictors.

6 Evaluation

To provide interesting comparisons on known gaze tracking systems, we will use the iTracker [8] and xLabs [14] webcam gaze-tracking systems. The use of existing gaze tracking systems should lend some credibility to our results.

To measure system performance, we will collect user gaze data alongside

a separate, high-accuracy "trusted" system and process each experiment "offline". This will allow us to assess the effectiveness of the system without the constraints of running the system in real time. Results will include analyses of the differences between the trusted system's predictions and those of the evaluated systems.

To provide interesting analysis, we will compare fuzzy decision-level fused systems to conventional rule-based systems and to simple well-known prediction aggregations. Additional analyses may be performed on the real-time performance of the systems, however, this should be considered more an example of computational feasibility than a significant benchmark.

7 Conclusion

While current market trends have the price of entry-level gaze-tracking hardware declining, tracking with high precision is a task currently reserved for specialized hardware which can quickly become very expensive. The use of several sensors in the gaze-tracking system has the potential to improve the accuracy of the predictions and provide a degree of flexibility uncommon in modern gaze-tracking systems. In this paper, we proposed several experiments to evaluate the effectiveness of multi-sensor gaze-tracking systems. The proposed work will identify the following: (1) if more accuracy can be extracted from a system using well-known sensor fusion techniques or (2) if we can create a more flexible system while maintaining the accuracy of more rigid systems.

References

[1] J. Chen and Q. Ji. "3D gaze estimation with a single camera without IR illumination". In: *2008 19th International Conference on Pattern Recognition*. Dec. 2008, pp. 1–4. DOI: 10.1109/ICPR.2008.4761343.

[2] Z. R. Cherif et al. "An adaptive calibration of an infrared light device used for gaze tracking". In: *IMTC/2002. Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No.00CH37276)*. Vol. 2. 2002, 1029–1033 vol.2. DOI: 10.1109/IMTC.2002.1007096.

[3] H. Durrant-Whyte and T. Bailey. "Simultaneous localization and mapping: part I". In: *IEEE Robotics Automation Magazine* 13.2 (June 2006), pp. 99–110. ISSN: 1070-9932. DOI: 10.1109/MRA.2006.1638022.

[4] K. Han et al. "A novel remote eye gaze tracking approach with dynamic calibration". In: *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*. Sept. 2013, pp. 111–116. DOI: 10.1109/MMSP.2013.6659273.

[5] T. C. Havens, D. T. Anderson, and C. Wagner. "Data-Informed Fuzzy Measures for Fuzzy Integration of Intervals and Fuzzy Numbers". In: *IEEE Transactions on Fuzzy Systems* 23.5 (Oct. 2015), pp. 1861–1875. ISSN: 1063-6706. DOI: 10.1109/TFUZZ.2014.2382133.

[6] T. C. Havens et al. "Fuzzy integrals of crowd-sourced intervals using a measure of generalized accord". In: *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. July 2013, pp. 1–8. DOI: 10.1109/FUZZ-IEEE.2013.6622343.

[7] T. Kocejko, A. Bujnowski, and J. Wtorek. "Eye mouse for disabled". In: *2008 Conference on Human System Interactions*. May 2008, pp. 199–202. DOI: 10.1109/HSI.2008.4581433.

[8] Kyle Krafka et al. "Eye Tracking for Everyone". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. URL: http://gazecapture.csail.mit.edu/.

[9] H. C. Lee et al. "Gaze tracking system at a distance for controlling IPTV". In: *IEEE Transactions on Consumer Electronics* 56.4 (Nov. 2010), pp. 2577–2583. ISSN: 0098-3063. DOI: 10.1109/TCE.2010.5681143.

[10] K. A. Funes Mora and J. M. Odobez. "Geometric Generative Gaze Estimation (G3E) for Remote RGB-D Cameras". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. June 2014, pp. 1773–1780. DOI: 10.1109/CVPR.2014.229.

[11] A. Pinar et al. "Feature and decision level fusion using multiple kernel learning and fuzzy integrals". In: *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Aug. 2015, pp. 1–7. DOI: 10.1109/FUZZ-IEEE.2015.7337934.

[12] Tobii Eye Tracking. *An Introduction to eye tracking and Tobii eye-trackers, White Paper*. 2010.

[13] R. Valenti and T. Gevers. "Accurate Eye Center Location through Invariant Isocentric Patterns". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.9 (Sept. 2012), pp. 1785–1798. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2011.251.

[14] xLabs. "xLabs Gaze-tracking System". In: (). URL: https://xlabsgaze.com/.