




Development of family level assessment of screen use in the home for television (FLASH-TV)

Anil Kumar Vadathya¹ · Tom Baranowski² · Teresia M. O'Connor² · Alicia Beltran² · Salma M. Musaad² · Oriana Perez² · Jason A. Mendoza^{3,4} · Sheryl O. Hughes² · Ashok Veeraraghavan¹ 

Received: 5 October 2022 / Revised: 5 October 2022 / Accepted: 12 December 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Screen use, including TV viewing, among children is associated with their physical and mental development. The most common assessment of TV viewing are self-reports and these introduce significant error. Objective measures are needed to improve research approaches to inform screen use guidelines. We present an objective approach to assess TV viewing as participant's gaze on the screen. **Family Level Assessment of Screen use in the Home (FLASH-TV)** uses state-of-the-art computer vision methods for face detection, recognition, and gaze estimation to process images and estimate the amount of time a child in the family spends watching TV. We recruited 21 triads of participants for the development of the FLASH-TV system who took part in 1.5 h observation studies with 5 in participants' homes. We evaluated each step of FLASH-TV by comparing to human-labeled ground truth data. Face detection and recognition methods achieved more than 90% sensitivity in detecting the target child under the challenging conditions of low lighting and poor resolution on a subset of test frames. Our final step of gaze estimation achieved more than 70% sensitivity and 85% specificity when evaluated on all of 3 million gaze/no-gaze labeled frames from 21 triads. Finally, our combined three-step system achieved 4.68 min mean absolute error of the TV watching time with a mean ground-truth TV watching time of 21.72 min. This method offers an objective approach to measure a child's TV viewing, with validation studies underway.

Keywords Objective · Passive · Assessment · Screen media use · TV viewing

✉ Ashok Veeraraghavan
vashok@rice.edu

¹ Department of Electrical & Computer Engineering, Rice University, 6100 Main St. – MS 366, Houston, TX 77005-1892, USA

² USDA/ARS Children's Nutrition Research Center, Baylor College of Medicine, Houston, TX, USA

³ Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

⁴ General Pediatrics, Department of Pediatrics, University of Washington, Seattle, WA, USA

1 Introduction

Precise measurement provides the most confidence in scientific findings and the strongest foundation for related policy. Most measures of TV viewing rely on self- or parent-proxy reports with known limitations on validity [1]. Institutions interested in the health of children have proposed time restrictions on younger children's TV watching [2]. However, these times may be based on faulty data and thereby may be incorrect which could lead to undue anxiety for parents if too restrictive or even harmful for children if too liberal.

To redress these problems, some investigators developed objective methods, including attachment of electronic sensors to measure the time TV is powered on, or use of wearable sensors like wrist based optical sensors and head mounted cameras. Robinson et al. [3] used the TV Allowance™ (Mindmaster Inc., Miami, FL) device which measures duration the screen media platform is ON along with a user specific code to authenticate. Fletcher et al. [4] developed a wearable wrist band with an optical color sensor to measure the light emitted by screen media devices like TV and computers. Zhang and Rehg [5] introduced a head mounted camera, whose video feed is analyzed using machine learning to identify and categorize various screen media use including TV, mobiles and tablets. Kerr et al. [6] proposed SenseCam, a wearable camera that takes images at 10 to 15 s intervals throughout the day which must be analyzed by hand. A limitation of these methods is that they do not assess whether the targeted research participant was actually watching the TV for the time estimated and are often costly because of human labelling. Moreover, introduction of wearable devices might alter the behavior of the target child towards watching TV thus affecting the actual screen time.

We developed the Family Level Assessment of Screen use in the Home-Television (FLASH-TV), a camera along with a processor, placed next to a TV, monitoring the area in front of the screen at a 30 frames/second or less fine granularity [7], and reported elsewhere that FLASH provides reasonably reliable and valid assessments of TV use of a targeted child participant, under diverse conditions [7]. The current paper presents a more detailed report of the methods in the development of FLASH-TV, which may be useful to those interested in modifying the current version, taking next steps in further developing it, or exploring new approaches to assessing screen media use.

2 Methods

FLASH-TV processes the images making three sequential decisions [7]: 1) Is anyone in the room, as determined by the presence of a face (face detector)? 2) Is that person, the targeted research participant (face verification)? 3) For the targeted research participant, is s/he watching the TV as determined by their gaze at the screen (gaze estimation)? We developed methods for FLASH-TV based on state-of-the-art approaches for each of these steps and modifying them appropriately to specifically fit our task. The remaining part of this section describes the data collection protocol for our design studies, and the details of each of the three steps of FLASH-TV.

2.1 Design studies

The target group to participate in these studies was a family of a parent with two children, one of whom was 6 to 11 years old (the child of primary interest) and a sibling 6 to 14 years old. Recruiting family triads ensured the FLASH-TV system was developed to differentiate the primary child of interest from others with whom they are likely to watch TV and from siblings with whom they share facial characteristics which may make face verification more challenging. In addition, the family needed to be fluent in English, and the parent should be willing to have their children engage with age-appropriate TV, movies or videogames. Children with a developmental, medical, mental or physical problem that might prevent them from adhering to the research protocol were excluded. The Institutional Review Board of the Baylor College of Medicine reviewed and approved the protocol, and reciprocity approval was obtained from the Rice University Institutional Review Board. The parent provided informed consent and both children provided assent. In addition, the parent was asked if s/he would be willing to opt-in to having the family's images used in publications or presentations to depict study findings, and if so, asked to sign the Baylor College of Medicine media release form. One family's data were corrupt, leaving data from 21 of 22 families analyzed for this paper.

We designed four studies to generate images for training and testing the FLASH-TV methods. Study 1 had ten participating family triads. Family members were asked to remain in a lab for a duration of 1.5 h while being filmed in a simulated living room with a TV and FLASH-TV unit. The protocol required participants to watch TV, engage with a mobile tablet, or play with physical toys while being video recorded by the observation room cameras as well as the prototype FLASH-TV system. Participants were asked to change positions in the room (e.g. from the couch to the floor) while performing each task for a few minutes at a time. For certain protocol segments, participants were asked to leave the room for a short period to ensure FLASH-TV would detect their absence and return. The lighting of the room was varied for some tasks during several of the design tests to assess the robustness of FLASH-TV during bright, dim and dark conditions, and included a 20–30 min free-play portion to capture more naturalistic viewing of a TV-screen by children when toys and a mobile device were also available. The room set-up varied for each family including different locations of the TV and chairs in the room and different room decorations, including portraits, to differentiate real and unreal faces.

To assess test–retest reliability, Study 3 had five family-triads who came to the lab for 30 min and returned a week later for 30 min to assess FLASH-TV's ability to recognize faces over time. To assess external validity, Study 4 had five family triads who were studied in their home. For in-home data collection, the equipment was setup in the participant's home and the protocol was implemented to capture children's natural TV viewing behavior. Together, we had 16 triads in our in-lab design studies and 5 triads in our in-home design studies.

FLASH-TV data were captured using Logitech webcams at 1080p resolution and 15–30fps (frames per second) with a large field of view (90 degrees). For each triad's video, a randomly selected 4% of the frames were selected with uniform probability as test frames. These test frames were labelled by trained staff for validation of the first two steps of FLASH-TV: face detection and verification. The publicly available video annotation toolbox, vatic [8], was used for labeling the bounding boxes along with identities as to who the target child was. This test set consisted of about 37,000 test frames

from 16 in-lab triads averaging about 2,000 frames per triad. For 5 in-home triads, we had about 20,000 frames averaging about 4000 frames per triad.

All the frames from the video data of about 3 million images at 1080p resolution from 21 participants, 162,000 frames per participant, were labeled for the target child's gaze on the TV by trained staff with the vatic labeling tool. The staff were trained for labeling the participant's gaze or no-gaze on TV. They were allowed to label once they achieved a minimum accuracy of 95% among three experts on the benchmark/criterion training set. Inter-coder reliability was high with a Kappa score of 0.88 for 10% of video clips that were double coded by separate staff.

2.2 Face detection

For face detection, the first step of the FLASH-TV algorithm, YoLo [9], previously trained to detect a set of objects present in an image, was selected considering its efficiency and excellent performance. The YOLO object detection approach treats object detection as a single regression problem of mapping image pixels to bounding box coordinates and class probabilities without the need for generating regional proposals. Thus, YOLO can run much faster compared with other region proposal methods [10]. Since YOLO targets general object detection, we limit the object detection to detect only one object of interest i.e., face, instead of 10 classes predicted by YOLO. We employed the default network structure named Darknet-19, which has 19 convolutional layers and 5 max-pooling layers with the largest input size of 608×608 to train the network. The default training parameters (e.g., learning rate, momentum, and weight decay) and the trade-off parameters in the objective function were used. In the training stage, we trained YOLO on the mixture of three popular face datasets WIDER Face [11], FDDB [12], and AFLW [13] for 100,000 steps with batch size set as 6. In the prediction stage, we obtained locations of the detected bounding boxes after performing non-maximum suppression (NMS) and simultaneously output their confidence score (0,1).

FLASH-TV achieved 92.5% sensitivity on the test FLASH-TV dataset [7]. ROC analysis was employed to select the threshold on the detected regions (bounding box) above which would be considered faces. We used 10 k test frames from our initial design tests (11 triads) for this from FLASH-TV datasets. The bounding boxes for these frames were labeled by our staff using the vatic toolbox. The face detector's threshold was set at 0.18, which resulted in 92.5% sensitivity with 0.79 false positives per second (i.e. 4 false positives frames every 5 s). FLASH-TV can tolerate higher false positives in favor of higher sensitivity since false positives from this step are ruled out at the next step of face verification. Higher sensitivity enables detecting faces in difficult scenarios (e.g. low-lighting or unexpected poses).

2.3 Face verification

Face verification identifies the target-child's face among all the faces that the detector outputs in the first step. Given a pair of facial images face verification determines if they belong to the same person or not. For FLASH-TV face verification, DeepFace [11] was chosen as it is efficient and easier to train compared with other then state-of-the-art approaches like FaceNet [12]. DeepFace learns to recognize faces by posing the task as an image classification problem. Each facial identity is assigned a specific label which the algorithm learns to predict. In this way, DeepFace encodes discriminative features

ensuring that the two facial images of the same person are closer in the feature space compared to the encoding of two dissimilar facial images.

We used the publicly available VGG Face database [13] consisting of 3.3 million images of over 9 k identities for training DeepFace. While testing, features extracted by the DeepFace from the penultimate layer, a 512-dimensional feature vector, was used for facial image representation. We concatenated both the facial image and its flipped image features to compute a feature representation which is used for computing a cosine similarity score [0,1]. Facial images above the 0.93 similarity score were considered to be the same person. The threshold was selected to have highest true-positive rate while minimizing the false-positive rate. The algorithm achieved state-of-the-art performance on the well-established LFW test benchmark [14] resulting in 99.65% sensitivity. To improve low-light performance on the FLASH-TV dataset, we retrained the DeepFace with data augmentation by simulating low-light conditions with gamma-correction and gaussian noise addition. This helped improve sensitivity by 5% on our initial design tests with 11 triads.

For FLASH-TV target-child face recognition, a gallery of template face images was built for each individual in our participating triad (child, sibling and parent), as shown in Fig. 1. During the test time, the test facial image from the face detector was matched against the template facial images for each identity using the face verification feature representation. The identity with maximum matches was assigned to the test facial image. Since this is temporal data at regular time intervals, a high confidence match was added to the gallery for every identity. For example, after every 3 min, if a facial image was identified as the target child with a similarity score of more than 0.975 it was added to the gallery. Similarly for the other identities, this resulted in a dynamic gallery with up-to-date templates of the identities. This helped improve face recognition sensitivity by almost 10% (Table 1). With all these modifications FLASH-TV face verification was able to achieve 94% sensitivity in identifying the target child.

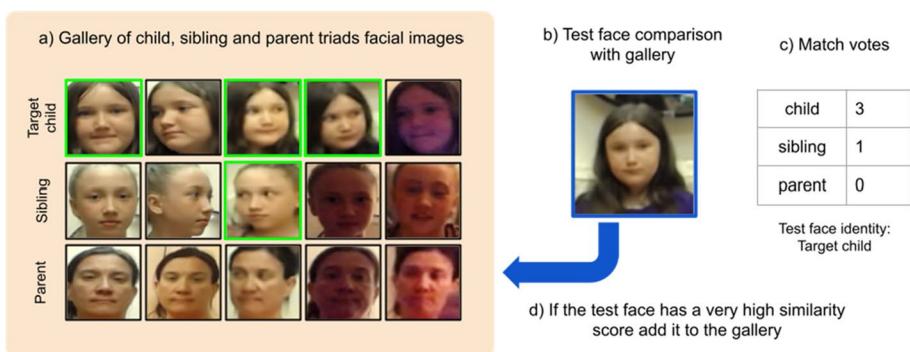


Fig. 1 FLASH-TV face recognition. **A** A gallery of facial images is built for each identity of the triad (Parent, Sibling and Child) in the participating family under different lighting conditions and poses. **B** During the test time, the input test facial image from the face detector is matched using the DeepFace features with each template image in the gallery. FLASH face verification compares each pair of facial images and outputs a similarity score [0,1] for which above 0.93 was considered the same person. **C** The match votes are summed across the templates for each identity and the test facial image is assigned an identity with the maximum votes. **D** Note that, the gallery is updated over time where the faces that have very high similarity (> 0.97) with the identities are added every 3 min to increase our recognition accuracy

Table 1 FLASH-TV face verification results with low-light training and tracking

Design tests 1 and 2 ($n = 11$ triads)	Original DeepFace Mean % (Min–Max)	Modified DeepFace (Low-light training) Mean % (Min–Max)	Modified DeepFace + tracking Mean % (Min–Max)
Sensitivity			
With gaze	79.42 (55.66–99.67)	84.75 (59.34–99.67)	93.10 (77.11–99.33)
With no-gaze	58.30 (37.63–87.57)	61.26 (42.56–89.42)	71.93 (51.88–95.92)
PPV			
With gaze	97.57 (90.51–100.0)	97.81 (90.26–100.0)	98.19 (90.50–100.0)
With no-gaze	89.79 (70.34–98.87)	83.81 (49.77–98.1)	85.87 (54.80–98.83)

2.4 Gaze estimation

Once the target child's face was detected, we computed a 3D direction vector along which the child was looking to determine if the child was watching the TV or not (see Fig. 11). For FLASH-TV, the TV viewing distances were usually a few meters from the screen resulting in facial images with poor resolution in contrast with popular gaze estimation approaches, such as MPIIGaze [17] or TabletGaze [18], where the camera to subject distance is less than a meter from the screen (see Fig. 2). The poorer resolution of the FLASH-TV images to be processed resulted in imprecise details around the eye region making it difficult to judge the gaze direction. Recent approaches, Gaze360 and RT-GENE, propose gaze estimation in a real-world outdoor setting with the camera to subject distances ranging a few meters. Notice the similarity of Gaze360 training facial images [20] to FLASH-TV data, unlike earlier datasets shown in the Fig. 2. The Gaze360 approach takes as input the facial images from 7 consecutive frames and computes feature representation for each image using resnet18 CNN architecture [21]. These features are fed to a bidirectional recurrent neural network which exploits the temporal aspects to finally predict gaze direction in terms of azimuth angle and elevation angle. These are compared with the ground truth as feedback to train the network parameters. Gaze360 yielded more accurate results on the FLASH-TV dataset due to similar data characteristics in comparison to RT-GENE.


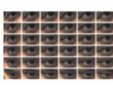





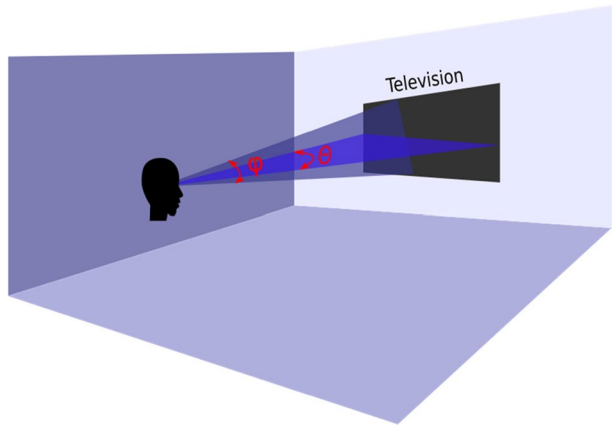
	Columbia gaze	UTMultiview	MPIIGaze	TabletGaze	RTGENE	Gaze360	FLASH
Input	High res. face image	Eye area + eye patches	Face + eye patches	Eye patches	Face + eye patches	Face frames	Face frames
Distance to Camera	2 mts, fixed	1 mts, fixed	< 0.6 mts	< 0.6 mts	0.5-2.9 mts	1 - 3 mts	1 - 4 mts
Location	indoors	indoors	indoors / outdoors	indoors	indoors	outdoors	indoors
Samples							

Fig. 2 Gaze estimation approach of different datasets—Columbia gaze [15], UTMultiview [16], MPIIGaze [17], TabletGaze [18], RTGENE [19], Gaze360 [20]. Our FLASH-TV has poor resolution due to large camera to person distance, up to 4 mts, making it difficult to infer the gaze direction directly from the eye-region details unlike MPIIGaze [17], TabletGaze [18] and other earlier approaches which only have a maximum distance of a meter. The Gaze360 dataset most closely resembles our FLASH dataset

Fig. 3 Gaze angle limits to identify the direction as gaze/no-gaze on TV. Notice that, from the position the person is watching TV the directions indicated with in the conical region formed by (θ, ϕ) correspond to watching TV, that is, gaze on TV. The directions outside this conical region would be considered as no-gaze on TV



Gaze estimation outputs a gaze direction in terms of azimuth and elevation angles (θ, ϕ) . FLASH-TV, however, requires a binary output indicating whether the target child is watching TV or not. As shown in Fig. 3, to convert the gaze angles to the binary gaze/no-gaze on TV we need to find the limits on the angles $(\theta_l, \theta_u, \phi_l, \phi_u)$. The directions indicated within the limits, that is, when $\theta_l < \theta < \theta_u; \phi_l < \phi < \phi_u$ are classified as gaze on TV and outside these limits are classified as no-gaze on TV. This FLASH-TV binary estimate is compared with human labeled ground truth gaze/no-gaze for every frame. Note that, the gaze direction by which one faces to watch TV changes depending on the head position (x,y) in the room. For example, if the TV is positioned in the center (see Fig. 4), the gaze direction to watch TV is different from sitting in the chair on the left, position (1), vs the chair on right, position (3). The direction also varies relative to the position of the TV, e.g. if the TV is in the center or to the left corner. Since the gaze direction is location dependent, the limits $(\theta_l, \theta_u, \phi_l, \phi_u)$ vary with the position (x,y) in the image. To account for this, the image was divided into regular grids. For each of these regions individually, the limits for azimuth (θ_l, θ_u) and elevation angles (ϕ_l, ϕ_u) were optimized. To account for the TV position, the data was stratified according to the TV position, and the limits were determined based on the TV position.

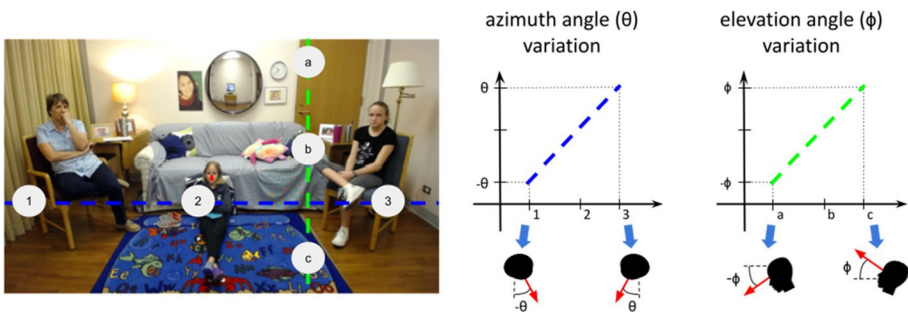


Fig. 4 The TV gaze direction is dependent on the location inside the room. For each region, limits on the azimuth and elevation angles were determined to indicate gaze direction for the target child's gaze on TV. Moreover, notice that along a row shown in blue line the azimuth angle monotonically increases from position (1) through (3). Similarly, along the column shown in green line, elevation angle monotonically increases from positions (a) through (c). This information is used to additionally regularize the gaze limits

2.4.1 Geometry based regularization for the gaze angular limits

For video frame resolutions with 1080 pixels, the images were divided into 20×25 for the TV position in the center case due to higher data density, for TV positioned in the left we divided the region into 10×12 pixel grids. For in-home case, we use the limits obtained from the TV positioned in center as most of the participants' TV position matched this setting. Regularization on the optimized limits was enforced to ensure that they agreed with the geometric constraints. For example, for gaze directions corresponding to watching the TV located in the center (shown in Fig. 4), the azimuth angles θ_{ij} monotonically increase along any i^{th} row of the image, that is, $\theta_{i,j-1} < \theta_{ij} < \theta_{i,j+1}; \forall j$. Moreover, θ_{ij} remains constant along any j^{th} column of the image regions, that is, $\theta_{ij} = \theta_{i+1,j}; \forall i$. The angles θ_{ij} are further optimized to satisfy the above constraints. The ratio of number of gaze samples over no gaze samples in each region are normalized along the rows and used as weights to optimize θ_{ij} . As the limits estimated in regions containing more gaze samples are more reliable compared with the regions having less gaze samples. Similarly, elevation angles ϕ_{ij} (as shown in Fig. 4) monotonically increase along the columns while remaining constant along the rows of the regions.

Algorithm pipeline for region-based gaze angular limits

Set azimuth angle limits $\theta_S \sim \{\theta_1, \dots, \theta_K\}$;

Set elevation angle limits $\phi_S \sim \{\phi_1, \dots, \phi_K\}$;

Set training gaze data $X_{train} \sim \{(\theta, \phi)_{x,y}^k\}$, Y_{train} ; where $k \in \{1, \dots, N_{train}\}$, x, y are the location coordinates of the data sample in the image, Y_{train} are binary gaze/no-gaze ground-truth labels.

Divide the image into $N \times M$ regions with N rows and M columns; R_{ij} where $i \in \{1, \dots, N\}$, $j \in \{1, \dots, M\}$.

For each region R_{ij}

- Get train data, $X_{ij}, Y_{ij} \subset X_{train}$ i.e., $\{(\theta, \phi)_{x,y}^k\}; \forall k$ where $(x, y) \in R_{ij}$;
- Find optimal gaze limits (θ_{ij}, ϕ_{ij})
 - Set best accuracy $\alpha_{ij} = 0$;
 - For each pair $(\theta, \phi) \in \theta_S \times \phi_S$
 - Apply angle limits (θ, ϕ) to X_{ij} to obtain Y'_{ij} , predicted labels;
 - Calculate accuracy α comparing (Y_{ij}, Y'_{ij}) ;
 - If $\alpha > \alpha_{ij}$; set $\alpha_{ij} = \alpha$
 - Set (θ, ϕ) as optimal limits $\theta_{ij} = \theta, \phi_{ij} = \phi$;

$\theta_{ij} = \theta, \phi_{ij} = \phi$;

Here the weights are normalized along the columns. This applies to the other positions of TV as well. Sensitivity improved by 10% on in-lab test frames with this additional regularization.

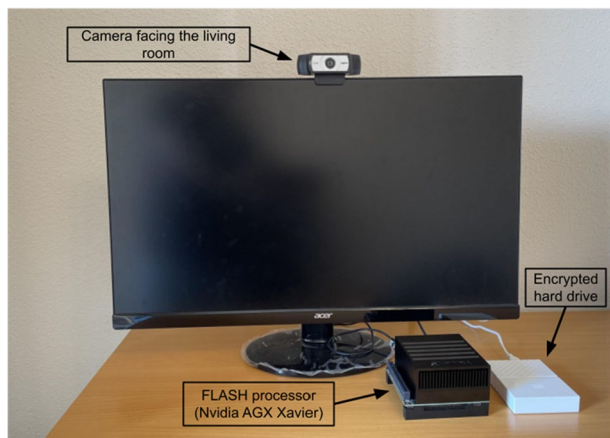
Given the small size of the number of participants (< 10), for each TV position, we adopted a leave-one-out strategy to test the gaze estimation algorithm, wherein the data of one of the participant's was held out for testing and the rest for training. This was repeated for each participant in the dataset. During the training step, the limits for each region were optimized by maximizing the accuracy of dichotomized gaze directions with staff coded GT. We performed a combinatorial optimization through a search space of azimuth and elevation angle limits choosing the optimal limits that resulted in highest accuracy with the training data. The obtained limits were applied on the test set in each region to compute test set accuracy. More detailed description of this algorithm is described below.

2.5 FLASH-TV system

The FLASH-TV system combines the above three steps of face detection, verification and gaze estimation sequentially processing each video frame to a binary decision as gaze/no-gaze on TV. These decisions, corresponding to small finite durations, were accumulated over time giving the overall TV viewing time by the child. We evaluated the system by comparing the FLASH-TV estimate with that of gold standard TV viewing time coded by the staff. For real-time deployment in the home, the three steps of the FLASH-TV algorithm were embedded in the platform, NVIDIA Jetson AGX Xavier. This device processed the data at 3fps and computed the real time TV watching behavior of the target-child.

Figure 5 below shows the FLASH-TV system in operation. The FLASH processor analyzes the video frames at 3fps in real-time and produces a time-stamped log indicating when the target child watched TV. Note that the video/images not stored anywhere on the system and are deleted permanently after processing in real-time to preserve privacy of the study participants. Only the time-stamped log is shared with research staff for further analysis.

Fig. 5 The FLASH-TV system set-up in practice. The camera mounted on the TV streams the video frames to the processor, NVIDIA AGX Xavier kit, which analyzes them and produces a time-stamped log indicating when the target child watched TV. The images/video are not stored anywhere and are deleted permanently to preserve the privacy of the participants



3 Results

3.1 Sample characteristics

The sample (reported in more detail elsewhere) [7] included 21 parent–child–sibling triads of whom the children were an average 10.2 ± 2.1 years old, slightly more female (56.8%), with race/ethnicity of 38.1% non-Hispanic White, 19% Hispanic White, 23.8% Non-Hispanic Black, 4.8% Hispanic Black, 4.8% Asian and 9.6% Other. The mean parent age was 43.9 ± 8.7 years, 90.5% female with similar race/ethnicity distribution and with education of 19.0% graduate school, 42.9% college educated, 28.6% some college, and 9.5% completed high school.

3.2 Face detector results

FLASH-TV face detector achieved an overall sensitivity of 93.9% in detecting the faces of the target child, sibling and parent in our in-lab design test. Our face detector's bounding boxes were compared against the human labeled ground truth bounding boxes. True positives were counted when the detector's box had a minimum of 30% IOU with the ground truth box. The results were evaluated using sensitivity and positive predictive value (PPV). Since our primary interest was measuring the TV viewing time, missing the child's face when they were not watching TV did not result in any errors. Even if these faces were detected and identified, they would be categorized as no-gaze and would not contribute to TV viewing time. Thus, errors in the no-gaze case were not as important. The detector for target child's face conditioned on gaze achieved sensitivity of 95.9% for in-lab tests and 97.9% for in-home tests (see Fig. 6). The PPVs were 68.7% and 52.5% respectively. Note that our higher sensitivity values came at the cost of lower PPVs indicating high false positive rates. This is not a problem, as most false positives, 97.5%, identified by the face detector get rejected at the next step of face verification.

Figure 7 shows the FLASH-TV face detector results from the test set frames. The detector picks faces across various poses, lighting conditions and relative position of the camera. The most undetected faces had no-gaze on the TV (third row) and thus do not affect the primary goal of measuring TV viewing time. The detector was not robust to faces oriented horizontally (third from left) where the child was lying on the sofa and watching TV.

Fig. 6 FLASH-TV Face detector results on our dataset. For more details see Online Resource 1

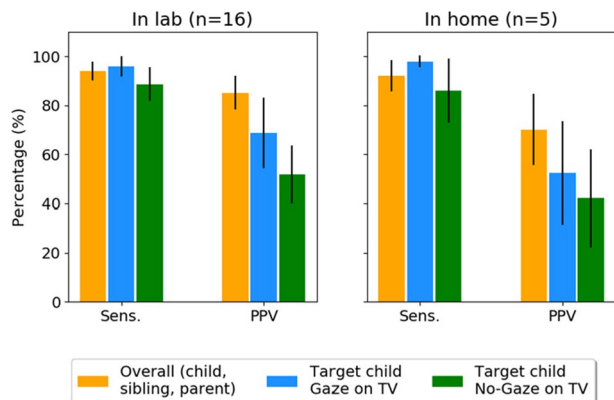




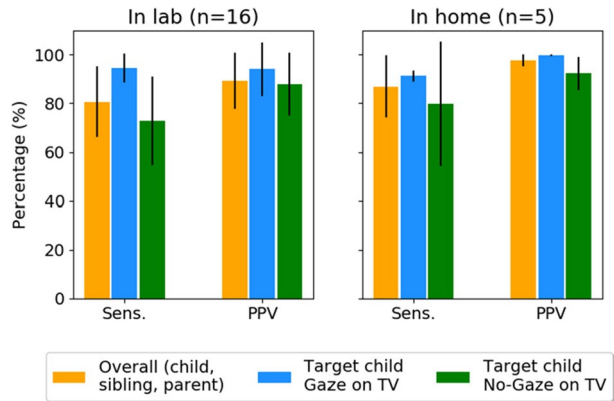
Fig. 7 FLASH-TV face detection results. Our detector is robust to various poses, low-lighting and low resolution

3.3 Face verification results

Face verification labels whether each detected face was a target child or not. FLASH-TV face verification achieved sensitivity of 94.5% when the child had gaze on the TV in our in-lab design tests. Similar to the face detector, conditional results are reported when the target child was gazing on TV, since this accounts towards our primary interest of measuring the TV viewing time. The original method of face verification, DeepFace, was modified with low-light training and tracking. The improvements in sensitivity and PPV achieved with each of these steps is in Table 1. The results are shown from our initial 11 triads of our in-lab design tests. Our sensitivity with gaze improved by 20% to 93.1% with proposed changes of tracking and low-light training. These improvements came at the cost of a small decrease in PPV with no-gaze of about 4%.

Face verification achieved conditional sensitivity of 94.5% for in-lab tests and 91.3% for in-home tests (see Fig. 8). This was quite successful in recognizing the child in the family triad when the child was watching TV. The relatively lower sensitivity for

Fig. 8 FLASH-TV face verification results. For more details see Online Resource 2



in-home cases was primarily due to very similar looking siblings with one of the triads. The conditional PPVs were 94.1% and 99.7% respectively for in-lab and in-home tests. The high PPVs indicate that our method very rarely falsely identified someone else in the family as target child. The PPVs needed to be high even for the no-gaze condition, as false positives account someone else's TV viewing time towards target child's.

In order to test the face verification algorithm to identify the child across days, one of our design tests involved five families with two visits spread a week apart. The facial images gallery for verification was built during their first visit and used for their second visit. The sensitivity and PPVs in Table 2, show the performance did not suffer. The participants had some changes in their visual appearance, eyewear, etc. to which face verification showed robustness in recognizing the child.

Figure 9 shows the images submitted to the face verification method on the FLASH-TV test set from our in-lab design tests. True positives showed that the target child was identified across various poses and lighting conditions. False negatives, where TC was not identified even though they were present in the view of the camera happened mostly when the child was not gazing on the TV (middle and rightmost image) and when the TC's face was partially obstructed with hands (left most image). False positives, the relatively high PPV values (> 85%) indicate these errors occurred less frequently, primarily during low-lighting (middle and right image) or when the facial features were not clearly visible (left image).

Table 2 FLASH-TV face verification robustness across days

Design tests 3 ($n=5$ triads)	Visit-1 Mean % (min-max)	Visit-2 Mean % (min-max)
Sensitivity		
With gaze	95.76% (85.88–98.59)	96.51% (94.87–97.56)
With no-gaze	74.58% (55.63–93.55)	73.26% (24.35–98.18)
PPV		
With gaze	89.99% (55.73–100.0)	89.07% (73.53–99.33)
With no-gaze	90.81% (72.14–97.22)	89.35% (65.52–100.0)

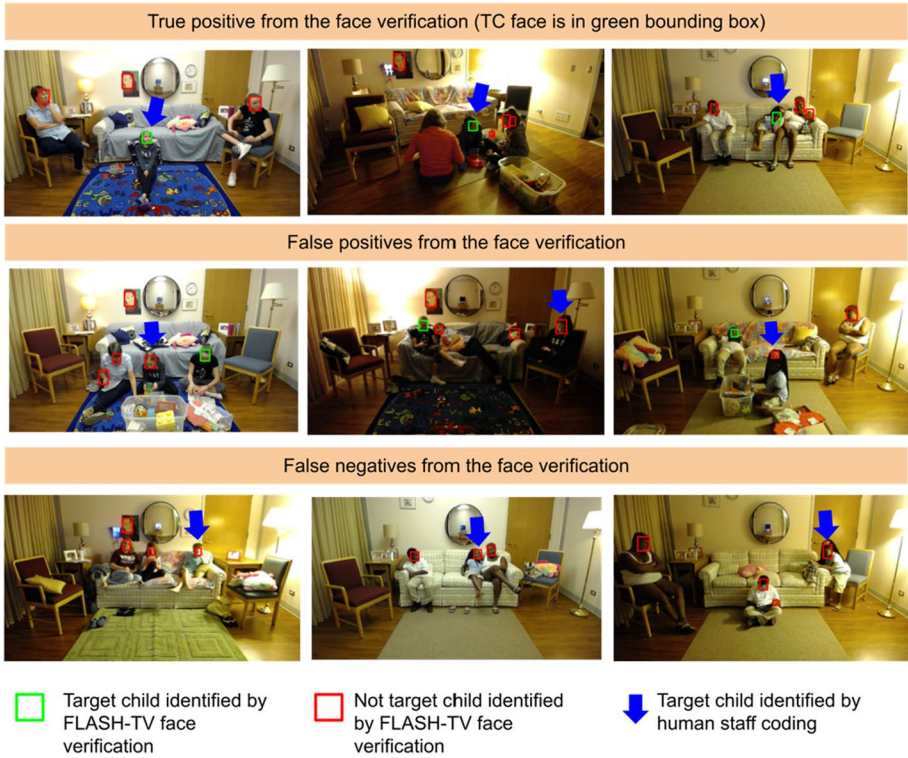
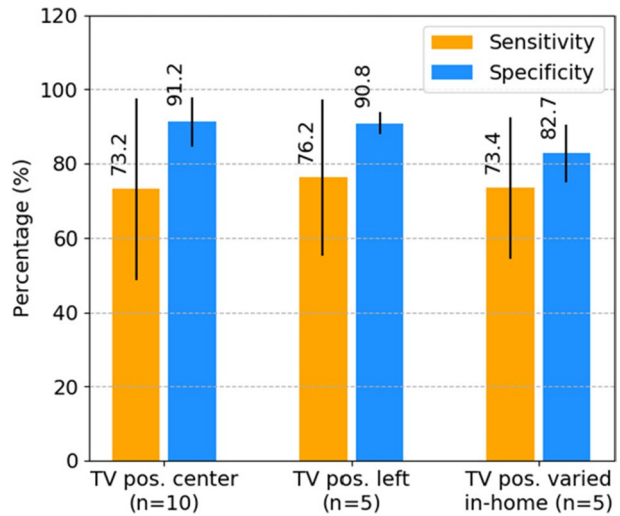


Fig. 9 FLASH-TV face verification results. Our method is able to identify target child across various poses, low-lighting and low resolution

Fig. 10 FLASH-TV gaze estimation results. For more details see Online Resource 3



3.4 Gaze estimation results

FLASH-TV gaze estimation predicts whether the child is watching TV or not, based on the facial image provided from the first two steps of face detection and verification. The estimator achieved sensitivity of more than 70% and specificity of more than 80% across our data with different TV positions (Fig. 10). The FLASH-TV labels of gaze/no-gaze were compared with human labeled data. Too many false negatives (no-gaze) result in FLASH-TV underestimating the TV viewing time as gaze samples get classified as no-gaze. On the other hand, too many false positives result in overestimation of the TV viewing time. The relatively higher specificity over sensitivity indicates that FLASH made more errors in misclassifying the gaze samples than no-gaze samples.

Table 3 shows the results of gaze estimation over our initial design tests ($n = 10$), with and without the geometry-based regularization. Notice the improvements in our sensitivity with our initial design tests for different TV positions in the center and to the left corner. Also, notice how Gaze360 performed much better than the RT-GENE approach. This is due to the fact that Gaze360 dataset captures several aspects similar to the FLASH-TV gaze estimation requirement.

Figure 11 shows the qualitative results of the gaze estimation algorithm on the FLASH-TV video frames. True positives and negatives shown suggest it is able to estimate the gaze direction and classify it as gaze/no-gaze across different poses, lighting conditions and relative position of the TV (center and left corner). False positives, where FLASH-TV incorrectly classified as gaze, mostly occurred while the child was simultaneously using a handheld device (first and second images). False negatives, where FLASH-TV incorrectly classified as no-gaze were mostly when the facial image had low-lighting (middle) or poor resolution (left most).

3.5 TV viewing time estimation using FLASH-TV system

FLASH-TV labels of gaze/no-gaze were summed at 30fps over the 90-min protocol giving the final estimate of TV viewing time for the target child in minutes. The summed TV viewing times were compared with gold-standard viewing time using mean absolute error (MAE). This achieved a value of 4.68 min (SD 4.58) over a mean gold-standard TV viewing time of 21.65 min (SD 11.11). MAE for TV positioned in the center ($n = 10$) was 3.19 (SD 3.32) whereas GT was 20.05 (SD 9.85); for TV

Table 3 FLASH-TV gaze estimation comparison with RT-GENE and improvements with additional geometry-based regularization

Metric	RT-GENE approach	Gaze360 without regularization	Gaze360 with regularization
TV position in the center ($n = 5$)			
Sensitivity	46.32% (34.68–58.14)	52.37% (25.64–74.09)	65.29% (27.78–85.30)
Specificity	66.20% (54.66–72.43)	96.18% (94.87–97.23)	95.56% (93.41–96.90)
TV position in the left corner ($n = 5$)			
Sensitivity	81.54% (70.43–89.37)	68.03% (43.83–85.56)	76.21% (53.98–96.13)
Specificity	56.69% (44.82–69.81)	91.57% (86.53–95.28)	90.81% (87.29–94.38)



Fig. 11 FLASH-TV gaze estimation results. Our approach is able to estimate gaze and no-gaze directions robustly. The insets show the zoomed facial images of target child along with a red arrow pointing along the gaze direction

positioned to the left ($n=5$) MAE was 2.54 (SD 1.80) and GT was 13.24 (SD 5.45). For in-home images ($n=5$), they were 9.80 (SD 5.24) and 33.25 (SD 8.69) respectively (Table 4).

Figure 12 shows the ratio of estimated viewing time vs ground truth for all the 20 triads. A perfect agreement between FLASH-TV and ground truth would be a ratio of 1, lower than one indicates underestimation and higher than one indicates overestimation of TV viewing. Half of the data lies in our acceptable error rates of 20% within the ground truth TV time estimation, indicating the range 0.8 to 1.2 for the ratio.

Table 4 FLASH-TV estimation of target child's TV viewing time

TV position	Mean Absolute Error (MAE) minutes (Min–Max)	Gold-standard TV viewing time, minutes (Min–Max)
Overall	4.68 (0.30–11.66)	21.72 (8.93–43.0)
Center of wall ($n=10$) ^a	3.19 (0.30–9.13)	20.2 (12.2–43.0)
Left corner of room ($n=5$) ^a	2.54 (0.43–4.50)	13.24 (8.93–22.5)
In Home, TV position varied ($n=5$) ^b	9.80 (4.04–11.66)	33.3 (23.3–42.7)

^aDesign tests 1–3 (in observation lab data collection)

^bDesign test 4 (in home data collection)

^cPrevalence and bias adjusted Kappa statistic

Total TV viewing during a 90-min task-based, observation period. One family's data from design test 1–3 were obtained from a unique position (below TV) and could not be used in gaze estimation training or testing

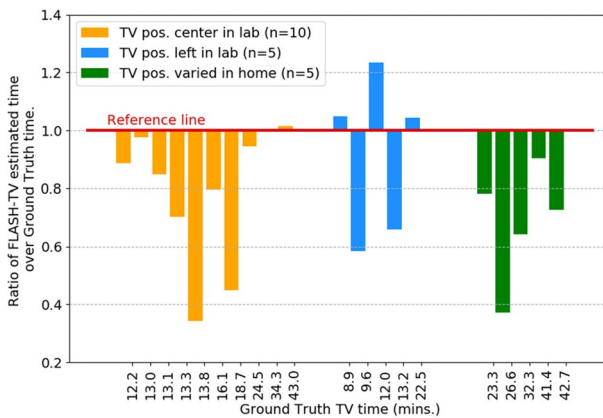


Fig. 12 TV viewing time estimated by FLASH-TV. The plot shows the ratio of FLASH-TV estimated time over the ground truth time labeled by our trained staff across our 20 participants. At the reference line, when the ratio is 1.0 both the estimates are equal. The bars show the deviation from the reference, ratio above reference lines implies that FLASH-TV over-estimates whereas below the reference line implies under-estimation. The ground truth time in minutes for each family is indicated on the x-axis below the bar

4 Discussion

We report the steps in the development of FLASH-TV. In each step state-of-the-art methods were selected and adapted to the requirements of our dataset. YoLo based object detector was modified for face detection. We modified the DeepFace face verification approach for low-light enhancement along with regularly updating our gallery of facial images with confidence matches. For the last step of gaze estimation, we performed optimization to find optimal spatially varying angular limits and enforce additional geometry-based regularization to convert the 3-D gaze vectors into binary gaze/no-gaze labels. All these changes resulted in satisfactory performance on the FLASH-TV dataset.

The estimates of TV viewing time from FLASH-TV are not comparable to estimates available in the literature [3–6] as the approaches are not directly comparable. These approaches involve TV allowance measuring device [3] and wearable devices including

wristbands and head-mounted cameras [4–6] which passively measure TV viewing time based on child's proximity to TV or if the TV is in the field of view of the child. Unlike these, FLASH-TV processes the direct video feed from the camera mounted on the TV actively measuring if the child is looking at the TV. This manuscript is an endeavor towards a multidisciplinary (behavioral, engineering and medical sciences) research effort to objectively assess child TV viewing behavior using a passive stationary camera which assures objective measurement; and developing the algorithms under varying conditions (lighting, position of the camera, ethnicity of the participants with varying skin colors, presence of possibly confounding faces, e.g. a portrait of girl, stuffed animals) to enhance utility and validity. We believe use of FLASH-TV can lead to more precise estimates of childrens' screentime than possible with self or proxy reports. Further development could involve use of FLASH-TV to set limits on child TV use.

In the future, we want to address the issues with the current version of the FLASH-TV methods. This includes improving the performance of face detector in picking up the faces that are not oriented upright; and improving the gaze estimator's performance in the low-lighting condition and extreme head poses e.g. when the child is watching TV lying on the sofa or when their facial features are partially obstructed with their hands.

The limitations of this research include not addressing the use of other forms of screen media (which we report using a different algorithm in another manuscript [Perez et al. submitted]); the small non-representative sample (but providing large numbers of frames for processing); use of family triads in preference to more representative samples of people watching TV; the possibility of other more effective AI approaches; collecting most data in a laboratory rather than in the homes of families with children; not addressing the assessment of TV viewing by adolescents or adults; the inability to identify what programming was seen by the child; and the inability to assess whether having the child's face being exposed to or gazing at the TV screen actually means attentiveness to the programming. Much important work remains to be done.

5 Conclusion

FLASH-TV, an instrument to take images and process them with algorithms to sequentially assess the presence of faces in an image, determine if the face is the target child's face, determine if the target child's face is gazing at the TV screen, and accumulate the time the target child watched the TV screen, has been developed and preliminarily tested. Future research must refine this approach, and assess the validity of the newer versions with larger samples in more settings and circumstances. This method offers promise of changing the way in which child TV screen use is measured, and lead to more precise data upon which to make judgments about the appropriate length of daily TV screen media exposure. The code for FLASH-TV is publicly available at <https://github.com/anilgukt/flash-tv-scripts> for research purposes.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11042-023-17852-y>.

Funding This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) of the National Institutes of Health [grant number R01DK113269]. This work was also supported by the United States Department of Agriculture/Agricultural Research Service (USDA/ARS) [cooperative agreement 58–3092-0–001]. This work was additionally supported by National Institute of Child Health and Human Development (NICHD) of the NIH [grant number P01HD109876], National Science Foundation

(NSF) Expeditions in Computing [grant number IIS-1730574], and NSF PATHS-UP ERC [grant number EEC-1648451]. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIH, NSF or USDA.

Data availability The datasets generated during and/or analysed during the current study are not publicly available due to HIPPA and Human Subjects privacy constraints on the video data. De-identified processed data (not image data itself) are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors have no competing interests to declare.

References

1. Byrne R, Terranova CO, Trost SG (2021) Measurement of screen time among young children aged 0–6 years: a systematic review. *Obes Rev* 22(8):e13260. <https://doi.org/10.1111/obr.13260>
2. Council on Communications and Media (2016) Media use in school-aged children and adolescents. *Pediatrics* 138(5):e20162592. <https://doi.org/10.1542/peds.2016-2592>
3. Robinson JL, Winiewicz DD, Fuerch JH, Roemmich JN, Epstein LH (2006) Relationship between parental estimate and an objective measure of child television watching. *Int J Behav Nutr Phys Act* 3:43. <https://doi.org/10.1186/1479-5868-3-43>
4. Fletcher RR, Chamberlain D, Richman D, Oreskovic N, Taveras E (2016) Wearable sensor and algorithm for automated measurement of screen time. 2016 IEEE Wireless Health (WH). IEEE, Bethesda, pp 109–116. <https://doi.org/10.1109/WH.2016.7764564>
5. Zhang YC, Rehg JM (2018) Watching the TV watchers. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2(2):88. <https://doi.org/10.1145/3214291>
6. Kerr J, Marshall SJ, Godbole S, Chen J, Legge A, Doherty AR, Kelly P, Oliver M, Badland HM, Foster C (2013) Using the SenseCam to improve classifications of sedentary behavior in free-living settings. *Am J Prev Med* 44(3):290–296. <https://doi.org/10.1016/j.amepre.2012.11.004>
7. Kumar Vadathya A, MUSAAD S, Beltran A, Perez O, Meister L, Baranowski T, Hughes SO, Mendoza JA, Sabharwal A, Veeraraghavan A, O'Connor TM (2022) An objective system for quantitative assessment of TV viewing among children: FLASH-TV. *JMIR Pediatr Parent* 5(1):e33569. <https://doi.org/10.2196/33569>
8. Vondrick C, Patterson D, Ramanan D (2013) Efficiently scaling up crowdsourced video annotation. *Int J Comput Vis* 101(1):184–204. <https://doi.org/10.1007/s11263-012-0564-1>
9. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Las Vegas, pp 779–788. <https://doi.org/10.1109/CVPR.2016.91>
10. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Columbus, pp 580–587. <https://doi.org/10.1109/CVPR.2014.81>
11. Taigman Y, Yang M, Ranzato MA, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Columbus, pp 1701–1708. <https://doi.org/10.1109/CVPR.2014.220>
12. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: a unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on computer vision and pattern recognition. IEEE, Boston, pp 815–823. <https://doi.org/10.1109/CVPR.2015.7298682>
13. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: a dataset for recognising faces across pose and age. Proceedings of the 13th IEEE international conference on automatic face & gesture recognition. IEEE, Xi'an, pp 67–74. <https://doi.org/10.1109/FG.2018.00020>
14. Huang GB, Learned-Miller E (2014) Labeled faces in the wild: updates and new reporting procedures. Technical Report UM-CS-2014-003. University of Massachusetts Amherst. http://www.cs.umass.edu/~elm/papers/lfw_update.pdf. Accessed 4 Aug 2022
15. Smith BA, Yin Q, Feiner SK, Nayyar SK (2013) Gaze locking: passive eye contact detection for human-object interaction. Proceedings of the 26th annual ACM symposium on user interface software and technology. ACM, New York, pp 271–280. <https://doi.org/10.1145/2501988.2501994>

16. Sugano Y, Matsushita Y, Sato Y (2014) Learning-by-synthesis for appearance-based 3D gaze estimation. Proceedings of the IEEE Conference on computer vision and pattern recognition. IEEE, Columbus, pp 1821–1828. <https://doi.org/10.1109/CVPR.2014.235>
17. Zhang X, Sugano Y, Fritz M, Bulling A (2015) Appearance-based gaze estimation in the wild. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Boston, pp 4511–4520. <https://doi.org/10.1109/CVPR.2015.7299081>
18. Huang Q, Veeraraghavan A, Sabharwal A (2017) TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Mach Vis Appl* 28(5–6):445–461. <https://doi.org/10.1007/s00138-017-0852-4>
19. Fischer T, Chang HJ, Demiris Y (2018) Rt-gene: real-time eye gaze estimation in natural environments. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Proceedings of the European Conference on Computer Vision (ECCV). Springer, Cham, pp 334–352. https://doi.org/10.1007/978-3-030-01249-6_21
20. Kellnhofer P, Recasens A, Stent S, Matusik W, Torralba A (2019) Gaze360: physically unconstrained gaze estimation in the wild. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, pp 6912–6921. <https://doi.org/10.1109/ICCV.2019.00701>
21. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Las Vegas, pp 770–778. <https://doi.org/10.1109/CVPR.2016.90>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.