Uncertainty-Aware Gaze Tracking for Assisted Living Environments

Paris Her, Logan Manderle, Philipe A. Dias, *Member, IEEE*, Henry Medeiros, *Senior Member, IEEE*, and Francesca Odone

***Abstract*— Effective assisted living environments must be able to infer how their occupants interact in a variety of scenarios. Gaze direction provides strong indications of how a person engages with the environment and its occupants. In this paper, we investigate the problem of gaze tracking in multi-camera assisted living environments. We propose a gaze tracking method based on predictions generated by a neural network regressor that relies only on the relative positions of facial keypoints to estimate gaze. For each gaze prediction, our regressor also provides an estimate of its own uncertainty, which is used to weigh the contribution of previously estimated gazes within a tracking framework based on an angular Kalman filter. Our gaze estimation neural network uses confidence gated units to alleviate keypoint prediction uncertainties in scenarios involving partial occlusions or unfavorable views of the subjects. We evaluate our method using videos from the *MoDiPro* dataset, which we acquired in a real assisted living facility, and on the publicly available MPIIFaceGaze, GazeFollow, and Gaze360 datasets. Experimental results show that our gaze estimation network outperforms sophisticated state-of-the-art methods, while additionally providing uncertainty predictions that are highly correlated with the actual angular error of the corresponding estimates. Finally, an analysis of the temporal integration performance of our method demonstrates that it generates accurate and temporally stable gaze predictions.**

***Index Terms*— Machine learning, gaze tracking, neural network regressor, uncertainty, pose estimation, multi-camera assisted living scenario.**

Manuscript received 22 February 2022; revised 2 September 2022 and 17 December 2022; accepted 14 February 2023. Date of publication 7 April 2023; date of current version 21 April 2023. This work was supported in part by the National Science Foundation under Grant 1854158. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Soma Biswas. *(Corresponding author: Paris Her.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Comitato Etico Regionale—Regione Liguria (Regional Ethical Committee—Liguria Region, Italy) through the MoDiPro project.

Paris Her is with the Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI 53233 USA (e-mail: paris.her@marquette.edu).

Logan Manderle was with the Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI 53233 USA. He is now with Milwaukee Tool, Brookfield, WI 53005 USA (e-mail: loganmanderle@gmail.com).

Philipe A. Dias was with the Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI 53233 USA. He is now with the Geospatial Science and Human Security Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (e-mail: ambroziodiap@ornl.gov).

Henry Medeiros was with the Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI 53233 USA. He is now with the Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: hmedeiros@ufl.edu).

Francesca Odone is with the Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genoa, 16126 Genoa, Italy (e-mail: francesca.odone@unige.it).

Digital Object Identifier 10.1109/TIP.2023.3253253

I. Introduction

SCIENTIFIC and technological advances have led to a consistent increase in global life expectancy, which has now surpassed 73 years [1]. As a consequence, the percentage of the global population aged 65 or older will increase from 12% to nearly 24% by 2050 [2]. As the global population ages, there is an increasing demand for new and innovative healthcare practices [3]. Such practices include advances in cost-efficient, unobtrusive, and intelligent medical care systems. Current methods for monitoring the health status of older individuals include evaluation scales that assess mobility and Instrumented Activities of Daily Living (IADL) performance (i.e., a person's ability to use common household tools such as a TV remote or phone without assistance) [4]. These assessments are episodic and subjective, generally taking place at a healthcare facility and based on questionnaires or self-reported outcomes.

Existing technologies for automatically monitoring the health status of older individuals consist of methods that are obtrusive and require extensive expert supervision [5]. Moreover, these techniques provide limited information about the patient status and can only be employed in controlled environments. Modern computer vision techniques have the potential to play a significant role in the development of automated health evaluation methods [6], [7]. However, the intrinsic challenges of vision-based techniques, such as occluded views or illumination variations, call for the development of more sophisticated computer vision methods that can be reliably employed in uncontrolled environments.

To address the ongoing challenge of unobtrusively monitoring the health status of elderly individuals over extended periods of time, in a partnership with the Galliera Hospital in Genoa, Italy, we have developed an instrumented patient discharge facility for assisted living [8]. Fig. 1 shows the layout and different views of the common areas of the facility, which is designed to extract information on how subjects interact with other people and their surroundings. The facility provides a test-bed for the development of general multi-modal assisted living technologies [9], [10]. It has been used to carry out research on human mobility and frailty [11], multi-target segmentation and tracking [12], and gaze estimation [13], [14].

This paper focuses specifically on the important problem of human gaze estimation and tracking. Human gaze is directly related to how a person interacts with their surroundings and other people, which provides important information to determine the well-being of that person. Human gaze predictions have been applied to design human-computer interaction

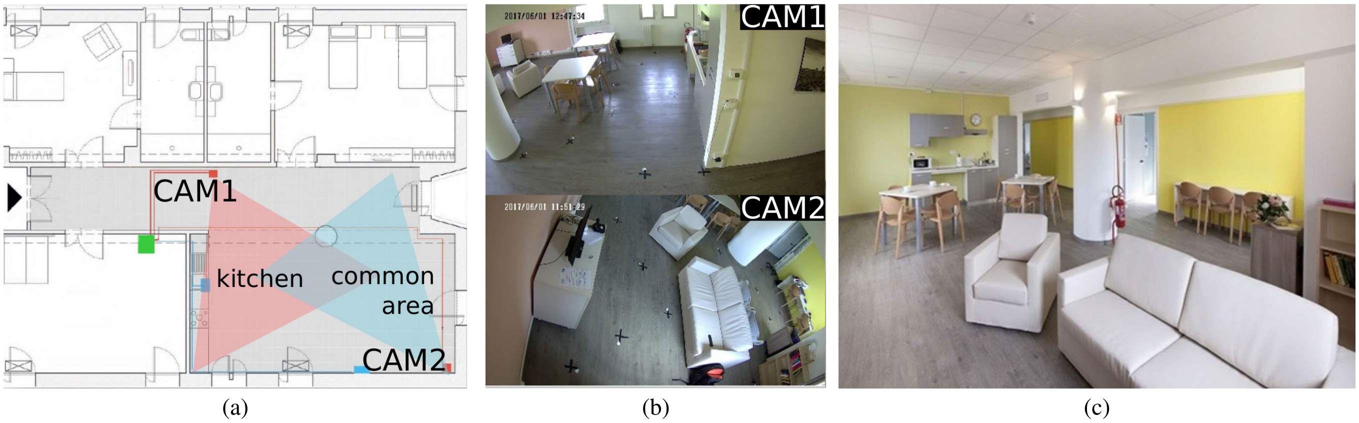


Fig. 1. Instrumented assisted living facility used for data collection. a) Apartment floorplan showing the placement of the two cameras and their fields of view; CAM1 in red and CAM2 in blue, b) Cameras views, c) Common area of the apartment.

methods [15], to analyze social interactions among multiple individuals [16], [17], and to estimate the cognitive load of vehicle operators [18]. For our application, gaze direction in conjunction with object detection [12] could define relationships among objects and their users (e.g. a person is sitting on a chair with a book on their lap vs. sitting on a chair reading the book), and classify simple actions (e.g. mopping the floor, getting dressed, cooking, eating/drinking, talking with other individuals [19], [20], [21]). More generally, gaze information can be used to determine changes in a person's social interactions or daily routine, preemptively detecting anomalies in their behavior.

We have first introduced the concept of estimating gaze direction and gaze prediction uncertainties based on facial keypoints using a regressor network in [13]. In [14], we extended that work into the temporal domain by using simple moving average techniques to integrate gaze predictions obtained in sequential video frames. In this paper, we build upon those findings to further improve the temporal consistency of the estimated gazes. To that end, we employ an angular Kalman filter to the gaze predictions to generate more accurate estimates and track the gaze direction. We utilize the uncertainties produced by the regressor network to adjust the predictions of the Kalman filter and further improve the robustness of our method. Thus, this work introduces five main contributions:

- *We propose an approach that relies solely on the relative positions of facial keypoints to estimate gaze direction.* As shown in Fig. 2, we extract these features using an off-the-shelf human pose estimation model [22]. From the coordinates and confidence levels of the detected facial keypoints, our regression network estimates the apparent gaze of the corresponding individuals. From the perspective of a general framework for IADL analysis, leveraging facial keypoints is beneficial because a single feature extractor module can be used for two important tasks: pose estimation and gaze tracking. Code is available at https://www.coviss.org/codes/.
- The complexity of gaze estimation varies according to the context, such that the quality of predictions provided by a gaze regressor is expected to vary on a case-by-case basis. For this reason, *our model is designed to provide an estimate of its uncertainty for each gaze prediction.* To that end, we leverage aleatoric uncertainty estimation techniques used in Bayesian neural networks.
- In cases involving unfavorable views or self-occlusion, one or more facial keypoints might not be detected with high confidence. To handle low-confidence detections, *we employ the concept of Confidence Gated Units (CGUs)* [13] to induce our model to reduce the impact of detections for which a low confidence level is provided. We further present an ablation study on the performance impact of the CGUs.
- We employ a *Kalman filter to track the angular trajectory of the gaze predictions* using predicted gaze uncertainties to adjust the estimations. We compare the performance of our method with different moving average schemes that utilize past gaze estimations to adjust current gaze predictions.
- We extend our assisted living activities dataset [13] to include two independent annotation sets of $\sim 24,000$ observable individual gazes and provide an analysis of the variability between the annotations. The dataset is available for research purposes upon request. Our extensive experimental evaluation demonstrates that our method substantially improves gaze estimation and tracking accuracy in assisted living scenarios and in publicly available benchmark datasets [23], [24].

II. Related Work

Due to the rapid advance of robust computer vision techniques, practical unobtrusive systems for well-being assessment and human-machine interaction that can operate under realistic conditions are now becoming a reality [25], [26]. Previous works present various methods for human behavior assessment using smart environments [22], [27], [28]. However, although recent developments in computer vision have the potential to automatically and unobtrusively quantify human mobility parameters [5], [29], patient activity analysis to date has been limited to simplistic scenarios. Our work in this paper utilizes data collected from an assisted living facility to precisely capture real-world living scenarios.

Fine-grained behavior analysis must take into consideration a person's focus of attention. Gaze direction provides important information regarding a person's intentions as they interact with their surroundings [30], [31], [32]. Most works

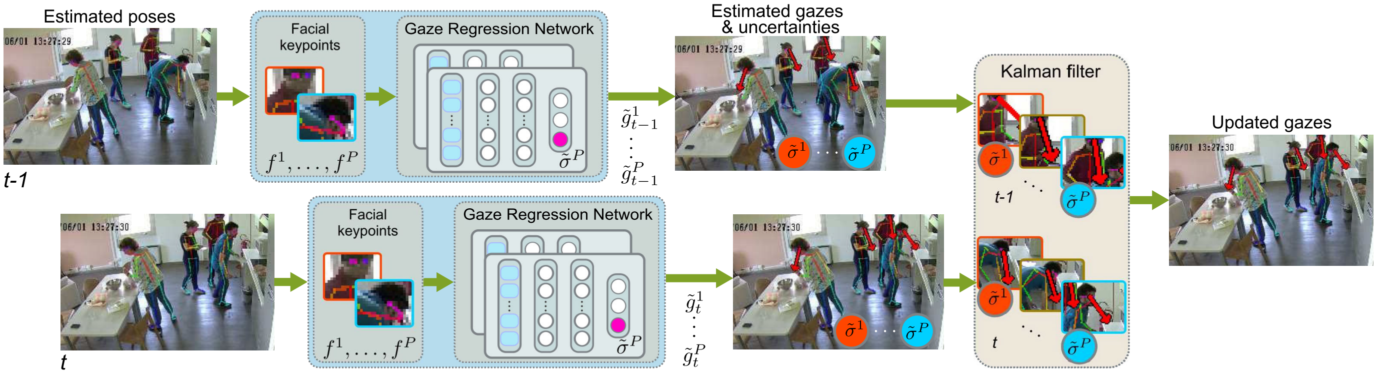


Fig. 2. Proposed gaze tracking approach. The facial keypoints of each person in the scene are collected using a pose estimation model [22] and provided to a neural network regressor that estimates their apparent gazes and corresponding confidence levels $\tilde{\sigma}^j$. An uncertainty-weighed Kalman filter combines the estimates collected in the current and previous frames, generating temporally consistent gaze estimations at each time instant.

on gaze estimation focus on monitoring eye movement [33], [34], [35], [36], [37] or perform head pose regression using facial appearance models [38], [39], [40], [41], [42]. These methods use close-up head features, generally acquired from a single static snapshot of the subject's face in scenarios with limited complexity that do not reflect real-world applications. Although head pose and gaze estimation are closely related, head pose estimation only provides a coarse estimate of a subject's visual attention [43], [44], [45], [46]. In a sense, gaze estimation is a more refined version of the head pose estimation problem. Whereas our method relies on head pose, more specifically relative facial keypoint locations, to determine gaze direction, the ground truth gaze directions are not determined solely based on the head position. That is, annotating and learning gaze direction encompasses the subject's apparent head position as well as contextual cues in the image. Therefore, head pose and gaze direction may not always align. For our method, we avoid employing estimation techniques using contextual cues altogether and shift the problem to estimating gaze based on facial keypoints, but we inherently utilize contextual information in the gaze direction annotation process instead of relying only on head pose.

Most popular gaze estimation datasets are collected in laboratory settings that do not contain real-world interactions and may comprise only eye patches or cropped facial images [47]. The dataset introduced in [20] includes gaze tracking data corresponding to multiple individuals performing common daily activities, but it is limited to first person views. The Gaze-Follow [23] and Gaze360 [24] datasets attempt to encourage the development of more robust gaze estimation methods by providing rigorously annotated gazes in a third-person view of natural scenes in images and videos. Together with the dataset, the authors of [23] introduce a two-pathway CNN architecture that combines saliency maps with the position of the subjects' head direction to generate gaze predictions. Gaze360 is a recent 3D gaze tracking dataset that contains a variety of indoor and outdoor activities performed by various individuals. For temporal gaze prediction, the authors introduce a model based on bidirectional LSTM modules that incorporate future and past gazes to determine a single central gaze. Both of these datasets allow the evaluation of gaze estimation in natural settings, and the Gaze360 dataset specifically provides gazes that are collected sequentially to be used in temporal methods.

Gaze estimation techniques are broadly divided into two domains: 3D and 2D approaches. Methods in the 3D domain require complex and expensive data acquisition systems to capture ground truth annotations [24], [48], whereas in the 2D domain annotators are able to simply use videos obtained from off-the-shelf cameras and annotate them utilizing readily available annotation tools. To address a specific gaze estimation problem in real-world conditions, such as gaze estimation for assisted living environments, it is not practical nor efficient to obtain deployment-specific 3D gaze annotations. Furthermore, although 3D projected gaze directions provide more information than 2D gazes, this additional information is not necessarily useful for the overarching task of inferring subject-environment interactions in 2D images and videos. When using gaze estimation to establish relationships between subjects and their environment, a 2D line of sight should be sufficient to establish connections between the subject and nearby objects.

Thus, in this work we focus on developing a 2D approach for gaze estimation, which is already a challenging task, especially when only a static snapshot of a person is taken into consideration. However, with the inclusion of previous frames that show the movement of a subject's head, we can better estimate and track their gaze. Methods that leverage sequential information have shown promising results, especially in low resolution images obtained using off-the-shelf cameras [49]. The impressive gaze estimation performance shown in [24] highlights the importance of utilizing temporal information to estimate gazes. In addition, gaze prediction is intrinsically more difficult for certain points of views. For instance, when subjects are facing vertically (i.e., frontward/backward) relative to the camera, this uncertainty creates large discrepancies in accuracy. In Bayesian neural networks [50], this corresponds to heteroscedastic uncertainty, i.e., uncertainty that depends on the model inputs, such that each input is associated with a different level of noise.

As explained in [50], conventional deep learning models are unable to estimate the uncertainty of their outputs. Classification models typically employ a softmax function in their last layer, such that prediction scores are normalized but do not necessarily represent uncertainty. For regression problems, usually no information on prediction confidence is provided by the model. Bayesian deep learning approaches are becoming an increasingly popular strategy to understand and estimate uncertainty in deep learning models [51], [52], [53]. Under this paradigm, uncertainties are formalized as probability distributions over model parameters and/or outputs. For the estimation of heteroscedastic uncertainty in regression models, the outputs can be modeled as corrupted by random noise. Then, as we discuss in Section III-B, a customized loss function is sufficient for training a regression model that also predicts the variance of the noise as a function of the input [50], without the need for uncertainty labels.

III. Proposed Approach

Human gazes are highly correlated with their body poses, particularly their facial orientation. This paper builds upon our previous work described in [14], where a human pose estimation algorithm [22] is used to obtain the facial keypoints of interest shown in Fig. 2. The information extracted from the pose estimation model includes the coordinates as well as the confidence scores of the eyes, ears, and nose.

Let $p^j_{k,s} = [x^j_{k,s}, y^j_{k,s}, c^j_{k,s}]$ represent the horizontal and vertical coordinates of a keypoint and its corresponding detection confidence value, respectively. The subscript $k \in \{n, e, a\}$ represents the nose, eyes, and ears features, while the subscript $s \in \{l, r, \emptyset\}$ encodes the side of the feature points. For each person j in the scene, we centralize the detected keypoints with respect to that person's head centroid $h^j = [x^j_h, y^j_h]$, which is computed as the mean coordinates of that person's head keypoints detected in the scene. These relative coordinates are then normalized based on the distance $m^j = [x^j_m, y^j_m]$ of the farthest keypoint from the centroid. For each detected person, we form a feature vector $f^j \in \mathbb{R}^{15}$ by concatenating the relative vectors $\hat{p}^j_{k,s} = [\hat{x}^j_{k,s}, \hat{y}^j_{k,s}, c^j_{k,s}]$, where $\hat{x}^j_{k,s} = {}^{(x^j_{k,s}-x^j_h)}/_{x^j_m}$ and $\hat{y}^j_{k,s} = {}^{(y^j_{k,s}-y^j_h)}/_{y^j_m}$, to obtain

$$f^j = \left[\hat{p}^j_{n,\emptyset}, \hat{p}^j_{e,r}, \hat{p}^j_{e,l}, \hat{p}^j_{a,r}, \hat{p}^j_{a,l}\right]. \tag{1}$$

A. Network Architecture Using Confidence Gated Units

Gaze direction is approximated by the vector $\tilde{g}^j = \left[\tilde{g}^j_x, \tilde{g}^j_y\right]$, which consists of the projection onto the image plane of the unit vector centered at the centroid h^j. That is, let $\tilde{\rho}^j$ be the apparent gaze angle with respect to the horizontal image axis. Then, $\tilde{g}^j_x = \sin(\tilde{\rho}^j)$, $\tilde{g}^j_y = \cos(\tilde{\rho}^j)$, and $\|\tilde{g}^j\| = 1$. Our model has an output layer with three units: two that regress the $\left[\tilde{g}^j_x, \tilde{g}^j_y\right]$ vector of gaze direction, and an additional unit that outputs the regression uncertainty $\tilde{\sigma}^j$. For that purpose, we train a fully-connected regression neural network that learns the function $\left[\tilde{g}^j, \tilde{\sigma}^j\right] = g(f^j)$.

Oftentimes, real world scenarios contain complex scenes involving various subjects in different positions. These scenarios may include subjects in poses that have occluded or missing keypoints, and keypoints that are estimated with low confidence. For example when a person is facing away from the camera, the detection of the eyes and nose are missing, or when the illumination of the room is uneven, the confidence of the facial keypoints may be low. We use our Confidence Gated Units (CGU), first introduced in [13], to handle these situations. As illustrated in Fig. 3, the CGU comprises two internal units: i) a ReLU unit acting on an input feature q_i, in our case the x and y coordinates of facial keypoints; and ii) a sigmoid unit without bias to emulate the behavior of a gate according to a confidence value c_i. The outputs of both units are multiplied into an adjusted CGU output $\tilde{q}_i$.

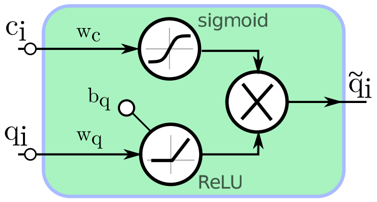


Fig. 3. The proposed Confidence Gated Unit (CGU) adjusts the contribution to gaze estimation of the i-th keypoint q_i according to its confidence level c_i.

Following our work in [13], our network architecture comprises a CGU-based input layer followed by two fully-connected (FC) hidden layers with 10 units each, and the output layer with three units. The architecture has a total of 283 learnable parameters and can be summarized as: (10 CGU, 10 FC, 10 FC, 3 FC).

B. Network Loss Function

Since our goal is to estimate gaze direction, our optimization and evaluation metrics are based on the angular error between the predictions and the ground truth gaze vectors. That is, training is performed using an uncertainty-weighed loss function adapted from [50] modified to use cosine similarity. Let $\mathcal{T}$ be the set of annotated orientation vectors g, while $\tilde{g}$ corresponds to the estimated orientation produced by the network and $\tilde{\sigma}$ represents the model's uncertainty prediction.[1] Our cost function is then given by

$$\mathcal{L}_{\cos}(g, \tilde{g}) = \frac{1}{|\mathcal{T}|} \sum_{g \in \mathcal{T}} \frac{e^{-\tilde{\sigma}}}{2} \frac{-g \cdot \tilde{g}}{||g|| \cdot ||\tilde{g}||} + \frac{\log \tilde{\sigma}}{2}. \tag{2}$$

This loss function requires no additional labels for the model to learn to predict its uncertainty. During training, the predicted uncertainty is used to weigh the contribution of the corresponding sample. As the prediction uncertainties increase, so does the overall loss. As a consequence, high uncertainty samples have little impact on the update of the model parameters during back-propagation. Hence, the model intrinsically learns the prediction uncertainty in relation to the angular error to help reduce the overall loss. More specifically, the $e^{-\tilde{\sigma}}$ component is a numerically stable representation of ${}^{1}/_{\tilde{\sigma}}$, which encourages the model to output a higher $\tilde{\sigma}$ when the error is higher. On the other hand, the regularizing component $\log(\tilde{\sigma})$ helps to avoid unbounded uncertainty predictions. From a Bayesian perspective, this loss function corresponds to a

[1] To simplify the notation, we omit the person-specific superscript j in this section.

von Mises distribution [54] where the Bessel function on the normalization term is approximated using a second-order series [55].

C. Temporal Integration

The gazes estimated by the regressor network are integrated over time using a Kalman filter. At time t, the network estimates the gaze of each person in the scene using the detected keypoints for that person and their corresponding confidence values. The Kalman filter uses a motion model to predict the individual gazes based on the estimated gazes at time $t-1$ and combines these predictions with the observed gazes at time t to produce a refined estimate. The Kalman filter state vector is given by

$$s_t = \begin{bmatrix} \tilde{\rho}_t \\ \omega_t \end{bmatrix}, \tag{3}$$

where $\tilde{\rho}_t = \arctan(\tilde{g}_y/\tilde{g}_x)$ is the apparent gaze orientation in polar coordinates at time instant t and ω_t is its corresponding angular velocity. We model the dynamic behavior of our system as a constant angular velocity motion corrupted with normally distributed noise, i.e.,

$$s_{t+1} = F \cdot s_t + w_t, \tag{4}$$

where $w_t \sim \mathcal{N}(0, \sigma_w)$ is the zero-mean, normally distributed process noise with variance σ_w and F is the system transition matrix, which is given by

$$F = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \tag{5}$$

Since we can directly observe the value of $\tilde{\rho}_t$ based on the output of our regressor, the observation model is given by

$$z_t = H \cdot s_t + v_t, \tag{6}$$

where

$$H = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{7}$$

is the observation matrix and $v_t \sim \mathcal{N}(0, \sigma_v)$ is the observation noise. The observation noise variance σ_v is a function of the gaze prediction uncertainty $\tilde{\sigma}$. As explained in detail in our experimental results, we evaluate two strategies to compute the observation variance: $\sigma_v = {}^{1}/_{\tilde{\sigma}}$ and $\sigma_v = e^{-\tilde{\sigma}}$. The value of σ_w relative to σ_v is empirically estimated using the Expectation-Maximization algorithm. The estimated state $\hat{s}_t = \left[\hat{\rho}_t, \hat{\omega}_t\right]$ is then obtained using the Kalman filter equations, and the estimated gaze in Cartesian coordinates is obtained from the estimated state according to

$$\hat{g}_t = \begin{bmatrix} \sin(\hat{\rho}_t) \\ \cos(\hat{\rho}_t) \end{bmatrix}, \tag{8}$$

where $\hat{\rho}_t$ is the estimated apparent gaze in polar coordinates. Fig. 4 illustrates the estimated apparent gaze in the 2D image plane.

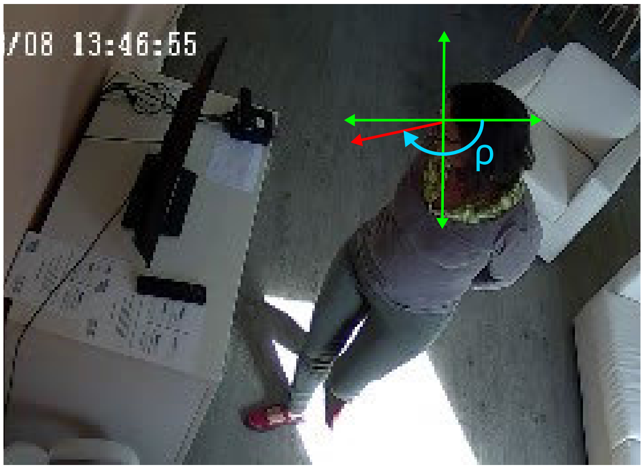


Fig. 4. Illustration of the apparent gaze direction, ρ, on the *MoDiPro* dataset (red).

IV. Experiments and Results

In our previous work [13], we have established the performance improvement of our regression network on an earlier version of the *MoDiPro* dataset. In this section we verify that our gaze regression network still outperforms other static gaze estimation methods on an extended version of the *MoDiPro* dataset. We also show the correlation between our network prediction uncertainties and angular error. Following that, we compare our temporal integration approach with other temporal methods such as the moving average method we introduced in [14] and the temporal method proposed in [24] on the extended *MoDiPro* and the Gaze360 datasets [24]. Lastly, we discuss the performance under keypoint occlusions, the impact of the CGUs, and an analysis on the uncertainty variance.

A. Datasets

We perform experiments on three publicly available gaze estimation datasets: the MPIIFaceGaze [38] and GazeFollow static gazes datasets [23], and the Gaze360 dataset [24], which contains temporal gaze sequences. While the MPIIFaceGaze is composed of relatively high-resolution, close-up facial images, the GazeFollow dataset contains subjects and gazes in natural settings including indoor and outdoor environments and is not constrained to a laboratory setup. We follow the training/testing split of the datasets suggested in their original papers for a fair comparison. We also conduct evaluations on our own assisted living environment gaze dataset, *MoDiPro*.

1) MPIIFaceGaze: The MPIIFaceGaze dataset, which is an extension of the MPIIGaze dataset [56], contains facial patch images of 15 subjects with annotations on eye gaze direction. The dataset consists of 213, 659 images from a laptop camera view. This ensures variations in background, lighting, and appearance. Eye gaze annotations are captured by a software on the computer screen, with participants looking at focal points on screen. Fig. 5 summarizes the 2D gaze distribution in the image plane for the MPIIFaceGaze dataset.

2) GazeFollow: The GazeFollow dataset [23] is a publicly available static gaze estimation dataset that contains more than

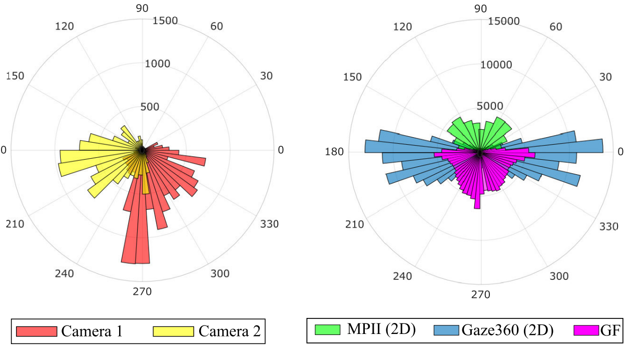


Fig. 5. Distributions of gaze annotations on the datasets used in this work. Left) Merged *MoDiPro* dataset with Cam1 (orange) and Cam2 (yellow). Right) Gaze360 gazes projected onto a 2D plane (blue), MPIIFaceGaze in 2D (green) and GazeFollow (purple).

130k people in 122k images. These images are collected from various well-known image datasets, such as MS COCO [57], PASCAL [58], and ImageNet [59]. GazeFollow is one of the first gaze estimation datasets that captures unconstrained and real-world gazes in natural scenes and activities. The images contain subjects performing natural activities in various environments, and the gaze annotations are defined with two points in the image plane: the center of the subject's head and the corresponding focus of attention. There are 10 annotated gaze vectors per subject in a given image, and we use the average endpoint location as the ground truth gaze vector, as suggested in [23]. Since our network is trained on images where at least two facial keypoints are detected, we retain $153,828$ training samples and $4,677$ testing samples, which corresponds to 98% of the dataset. Fig. 5 shows the angular gaze distribution for the gazes used in our experiments.

3) Gaze360: The Gaze360 [24] dataset is composed of images captured with a Ladybug 360^o panoramic camera mounted on a tripod with the subjects looking at a moving target. The dataset contains gazes of 238 subjects in 80 different recordings with a varying number of subjects in each video. The recordings take place in a variety of natural environments in indoor and outdoor locations, with varying illumination, subject visibility (i.e. frontal and back views), and background. All these features contribute to the creation of a natural dataset of people in real-world environments and situations. For each subject there is a ground truth gaze vector with the origin given by the midpoint between the eyes. This dataset is composed of 3D gaze annotations. Since our method focuses on 2D gaze estimations, we convert these three-dimensional vectors to two dimensions by taking the projection of the 3D gaze vector onto the camera image plane. These 2D vectors become our new annotations. Again, we exclude frames where the pose estimation algorithm cannot detect at least two facial keypoints, such that a total of $126,812$ frames are used for training, $17,011$ frames for validation, and $25,949$ frames for testing. Fig. 5 also shows the angular distribution of the Gaze360 dataset.

4) MoDiPro: Our *MoDiPro* dataset focuses on patients in a discharge facility whose health status must be monitored. It allows us to evaluate our method on a real assisted living environment. We obtain videos of volunteers in a patient discharge facility in the Galliera Hospital. This discharge environment contains various sensors such as localization systems, RGB-D, and two conventional cameras. The cameras are arranged as shown in Fig. 1 and capture 480 × 720 pixel resolution frames at 25 frames per second. We collected videos of 22 individuals over a period of 40 days. Recordings took place throughout the day and the videos contain a variety of illuminations based on the hour of the day.

We generate the ground truth gaze annotations using the MATLAB$^{(R2018)}$ VideoLabeler annotation tool.[2] Two annotators manually label the gaze directions independently. Cam1 has 47 videos with a total of $15,750$ frames while Cam2 has 30 videos with $10,750$ frames, totaling $26,500$ frames. The annotation process took on average 9 minutes for each video with an average of 20 frames manually annotated and the rest interpolated by the annotation tool and manually verified by the annotators. Annotation set 1 has a total of $24,509$ annotated gazes with at least two keypoints, and annotation set 2 has $24,494$. The discrepancy in the number of gazes comes from partially occluded facial keypoints and/or subjects near the edge of the frame where annotators may deem there to be a gaze or not. We also average the annotated gaze vectors in the two annotation sets to obtain a merged annotation set for our evaluations.

One of the significant challenges in the gaze tracking problem is its inherent uncertainty. The level of uncertainty in estimating human gazes is a challenge even for human annotators. Gazes are largely determined by body language, head, and eye locations and sometimes cues from the subjects' hands. Occluded gazes are oftentimes determined using the best judgment with inference from previous and upcoming frames. With that said, two annotators looking at the same frame may estimate each gaze differently. In the *MoDiPro* dataset, although the average difference in annotation was 0.08°, the variance was significantly higher with a standard deviation of 23.30°. The angular differences follow a Laplace distribution with 75% of the differences within 11.5°. This variability in annotations demonstrates the intrinsic uncertainty in the problem of gaze estimation. This uncertainty is an indicator of the expected performance levels that can be achieved by a gaze estimation model.

Fig. 5 illustrates the gaze distributions of the merged annotation set. The angle distributions are displayed from the perspective of the camera image frame (see Fig. 4). Specifically, gazes within the Cam1 dataset comprise mostly vertical gazes while Cam2 consists of more lateral gazes. As the camera frames shown in Fig. 1 indicate, subjects in Cam1 tend to move vertically along the path between the tables and the kitchen area. Subjects in Cam2 are more likely to look toward objects of interest such as the TV on the left, or sit on the sofa directing their gazes horizontally.

Frames used for training and evaluation are grouped by video since frames from the same video sequence are highly correlated. We randomly select 50% of the videos from each camera for training, 20% for validation, and 30% for testing.

[2] https://www.mathworks.com/help/vision/ref/videolabeler-app.html

Final mean angular errors are the average values obtained after training and testing on three different random splits of the data.

B. Static Gaze Estimation Performance

In this section, we compare our regression network against the GF model and the method introduced in [60]. This publicly available approach represents the state of the art for the GazeFollow dataset that is most similar to our method to create a fair comparison. That is, in [60] predictions are generated without the need for eye patch images [48], [61], [62]. Methods that effectively estimate gaze without this limitation provide practical results in tackling gaze estimation in real-world scenarios. Additionally, we show results on the MPIIFaceGaze eye gaze dataset with the method introduced in the original paper. Moreover, we show that existing methods present substantial performance differences when applied to our *MoDiPro* dataset.

1) Network Implementation and Training Details: Our network is implemented using TensorFlow. We train the network using only samples where at least two facial keypoints are detected. We use the Adam [63] optimizer with early stopping based on the validation angular error. The initialization and optimization of our model follow the strategy described in [13]. Specifically, the fully-connected layers are initialized as in [64] and the CGU units are initialized with *ones*. Regarding regularization, we empirically observed better results without regularization in the input and output layers, while a L2 penalty of 1e−4 is applied to both FC hidden layers.

2) Impact of the Training Dataset: First, we show the performance of our network trained on various combinations of the datasets described above and evaluated on the *MoDiPro* dataset. The corresponding results are shown in Table I. Training our model only on the MPIIFaceGaze dataset (NET #0) results in the worst performance on the *MoDiPro* dataset with a mean angular error of 60.64°. When trained only on the Gaze360 dataset (NET#1), our network has a mean angular error of 35.79°. However, when only the GazeFollow dataset is used for training (NET#2), the average error is reduced by more than 11°. As Fig. 5 shows, the GazeFollow data more closely represents the *MoDiPro* dataset than the Gaze360 or the MPIIFaceGaze datasets, which explains this significant difference in performance. For additional training, we do not include MPIIFaceGaze data as it consists of close-up facial images, unlike the natural gazes in the other datasets. Furthermore, training our model using both the GazeFollow and Gaze360 datasets (NET#3) and then fine-tuning it on one of the *MoDiPro* camera views (NET#4 and NET#5) leads to noticeable performance improvements in both cameras. However, the NET#6 and NET#7 models, which are trained only on GazeFollow and fine-tuned on the CAM1 and CAM2 views, respectively show a more pronounced improvement in CAM2. Training our model using both the GazeFollow and Gaze360 datasets and then fine-tuning it on the two *MoDiPro* camera views (NET#8) leads to a slight performance improvement for CAM1, but this gain is outweighed by the degradation observed in CAM2. The NET#9 model, which is trained only on the GazeFollow dataset and fine-tuned on both views of the *MoDiPro* dataset has the lowest mean angular error (21.70°). Hence, subsequent evaluations are based on this model.

TABLE I

MEAN ANGULAR ERROR FOR EACH CAMERA ON THE *MoDiPro* MERGED ANNOTATION SET USING DIFFERENT TRAINING SETS

	TRAIN					TEST		
Model	MPII	GF	Gaze360	CAM1	CAM2	CAM1	CAM2	*Mean*
NET#0	✓					79.93°	41.34°	60.64°
NET#1			✓			38.06°	33.52°	35.79°
NET#2		✓				23.29°	25.90°	24.60°
NET#3		✓	✓			21.77°	25.31°	23.54°
NET#4		✓	✓	✓		20.20°	24.11°	22.16°
NET#5		✓	✓		✓	20.05°	24.77°	22.41°
NET#6		✓		✓		20.60°	23.36°	21.98°
NET#7		✓			✓	21.58°	**22.70**°	22.14°
NET#8		✓	✓	✓	✓	**19.72**°	24.11°	21.91°
NET#9		✓		✓	✓	20.47°	22.92°	**21.70**°

TABLE II

MEAN ANGULAR ERROR COMPARISONS WITH STATIC AND TEMPORAL METHODS. FOR A FAIR COMPARISON, ALL THE MODELS ARE TRAINED ONLY ON THE MPIIFACEGAZE, GF, OR GAZE360 DATASET, RESPECTIVELY

Method	MPII	GF	Gaze360	CAM1	CAM2	*Mean*
FullFace [38]	9.49°	-	-	66.98°	110.07°	88.53°
NET#0	38.89°	-	-	79.93°	41.34°	60.64°
GF-Model [23]		24.00°	-	45.82°	76.55°	61.19°
Lian *et al.* [60]		**17.60°**	-	43.91°	117.88°	80.90°
NET#2		23.37°	-	23.29°	25.90°	**24.60**°
Gaze360 [24]		-	**24.48**°	49.10°	84.05°	66.58°
NET#1		-	31.38°	38.06°	33.52°	35.79°
NET#1 + WMA & $e^{(-\sigma)}$		-	31.28°	37.49°	33.16°	35.33°
NET#1 + KF & $\frac{1}{\sigma}$		-	27.15°	28.79°	29.16°	28.98°
NET#1 + KF & $e^{(-\sigma)}$		-	25.96°	**28.10**°	**28.32**°	**28.21**°

3) Comparison With Static Gaze Estimation Methods: The first two rows of Table II show the performance of our method and [38] on the MPIIFaceGaze Dataset and our merged *MoDiPro* dataset. The two methods are trained only on the MPIIFaceGaze dataset for a fair evaluation. To compare with our model, the results of the method in [38] are converted to the 2D domain as the intended model outputs 3D gaze predictions. In MPIIFaceGaze, [38] achieves 9.49° of error compared to our method at 38.89°. This large error in our method comes from the nature of the dataset. The dataset contains gazes that are more related to eye ball orientation than natural gazes which our method is intended for. As shown in Fig. 6, most samples are of subjects' frontal view which makes it difficult to pinpoint a 2D gaze direction using facial keypoint locations. The illumination in some examples also introduces cases where OpenPose fails to identify any keypoints. When evaluated on the *MoDiPro* dataset, our model achieves a mean angular error of 60.64° while [38] obtains 88.53°. The higher error in [38] is because of the lower quality of the facial image features along with the fact that our dataset is geared towards natural gaze direction rather than strictly eye gaze.

The next three rows of Table II compare the performance of our method on the GazeFollow dataset and our merged *MoDiPro* dataset with the state-of-the-art static gaze

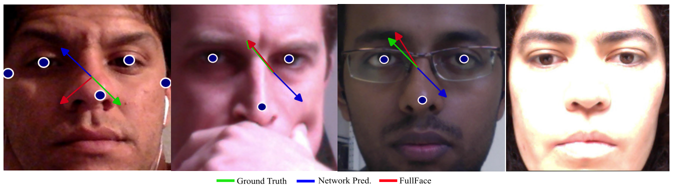


Fig. 6. Examples of gaze direction estimation by our method and [38] on the MPIIFaceGaze dataset. The blue markers denote keypoints detected by OpenPose.

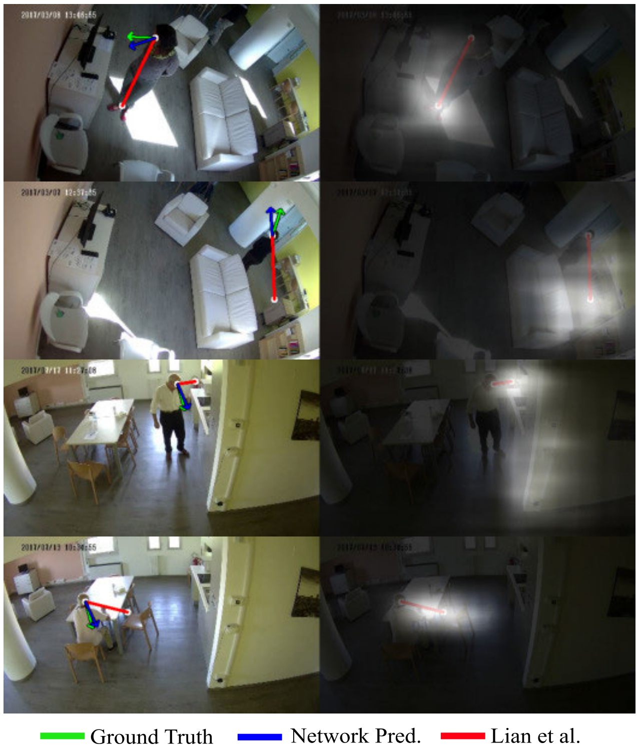


Fig. 7. Left) Examples of gaze direction estimation by our method and [60] on the *MoDiPro* dataset. Right) The generated prediction heatmap from [60].

estimation techniques proposed in [23] and [60]. For a fair comparison, all the models are trained only on the GazeFollow dataset (i.e., our model corresponds to NET#2 in Table I) since training the methods described in [23] and [60] requires additional ground truth information that is not available in the *MoDiPro* dataset. As the results indicate, [60] is the best-performing method on the GazeFollow data with an average angular error of 17.60°. However, when evaluated on the *MoDiPro* data, our method achieves an average angular error of 24.60°, which is approximately 36° lower than [23] and 56° than [60]. Although [60] performs well on images that contain saliency information, its performance decreases drastically in scenarios where the subjects are not necessarily looking at salient image features, such as in the *MoDiPro* dataset. As our method does not rely on regions of interest of the subjects' gaze, it is able to maintain a consistent and reliable performance across both datasets. This is illustrated in the examples shown in Fig. 7 where [60] maps the subjects' gaze to objects of interest but clearly the gaze is not in that direction.

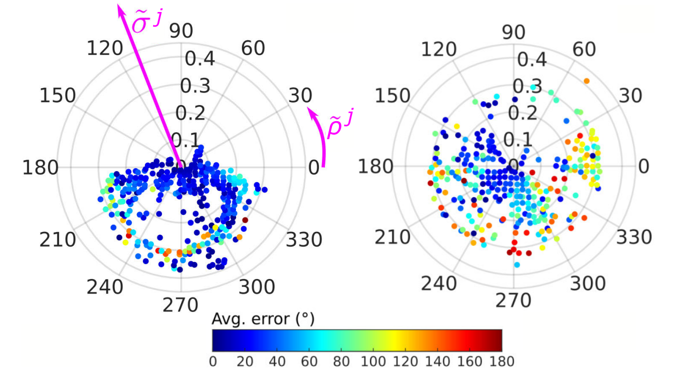


Fig. 8. Distribution of gaze direction ($\tilde{\rho}^j$) and uncertainty predictions ($\tilde{\sigma}^j$) for cameras 1 (left) and 2 (right) on the *MoDiPro* dataset. The color map represents the angular error of the predictions.

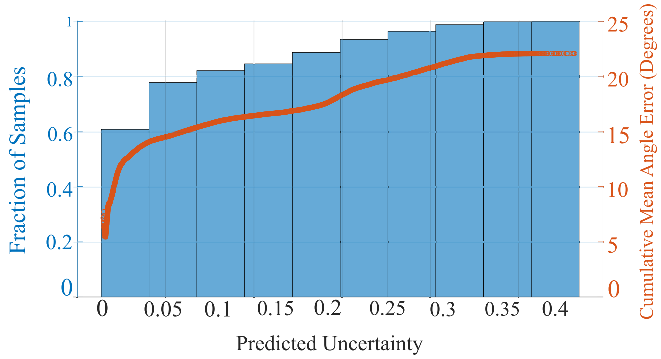


Fig. 9. Cumulative mean angular error according to the uncertainty predicted by our model for each sample in the test set.

4) Quality of Uncertainty Estimates: Fig. 8 shows the correlation between uncertainty predictions generated by our network (NET#9) and the corresponding angular errors on the *MoDiPro* dataset. For each sample in these plots, the angle corresponds to the predicted gaze direction, i.e., $\tilde{\rho}^j = \arctan(\tilde{g}_y^j/\tilde{g}_x^j)$ and the radial distance corresponds to its predicted uncertainty $\tilde{\sigma}^j$. The color map shows that lower errors (dark blue) are observed for predictions with lower uncertainty (small radial distance), with higher errors observed as the uncertainty increases.

To determine whether our model produces well calibrated uncertainties, we compare the uncertainties estimated by our model with their corresponding angular errors. Fig. 9 demonstrates the high correlation between the uncertainty predictions and the mean angular error. The figure demonstrates that the higher the mean angular error the higher the uncertainty prediction. For 80% of the *test* set the uncertainties are lower than 0.1, and the gaze estimations provided by our model for this subset are on average off by only $\sim 15°$.

5) Performance Under Keypoint Occlusions: Fig. 10 illustrates the performance of our model according to the number of visible keypoints. The left plot represents predictions for CAM1 whereas the right corresponds to CAM2. For samples with $k = 2$ (back-view), both uncertainty predictions and angular errors tend to be higher, while for most cases of $k = 3$ and $k = 4$ (lateral views) the predictions are associated with lower uncertainty and higher angular accuracy. Predictions for

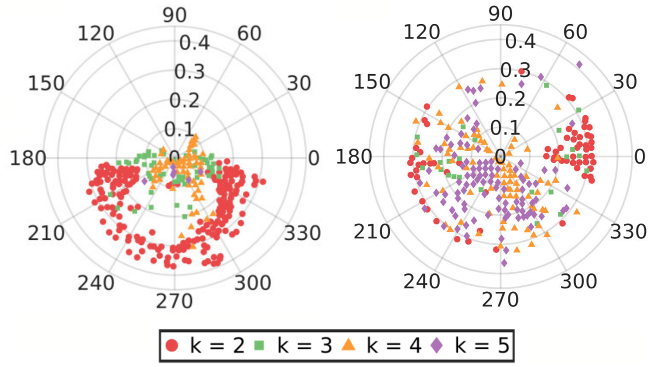


Fig. 10. Distribution of gaze direction for cameras 1 (left) and 2 (right) with colors representing the number of keypoints, k, detected by OpenPose [22] for the corresponding sample.

$k = 5$ (frontal views) are diverse, indicating that the model's uncertainty predictions are not just defined by the number of available keypoints but also reflect the intrinsic uncertainty of determining the head orientation from frontal views.

6) Impact of the Confidence Gated Units: We evaluate the extent to which the CGUs reduce the median angular error of our model for different average keypoint confidence levels. We perform our analysis only for samples containing no missing keypoints because low average confidence scores and missing keypoints are relatively orthogonal sources of uncertainty and must hence be assessed separately. We train a new model that replaces the 10 CGU units in the first layer of the network with ReLU units initialized the same way as described in Section III-B. We then evaluate the performance of both models as a function of the average keypoint confidence scores. We divide the average keypoint confidences into eight ranges and compute the median for each bin. The box plots in Fig. 11 show that the performance benefits of the model with the CGUs increase as the average keypoint confidence decreases. The difference in performance increases from 1.74° in the highest confidence bin to 3.12° in the bin with the lowest average keypoint confidences.

C. Gaze Tracking Performance

We compare the performance of our uncertainty-weighed angular Kalman filter described in Section III-C on the Gaze360 and *MoDiPro* datasets with the temporal method introduced in [24]. To our knowledge, that is the only temporal gaze tracking method proposed to date (see also [65]). We also demonstrate the superior performance of the proposed approach with respect to the best moving averaging scheme we originally presented in [14]. All temporal methods are trained only on the Gaze360 dataset to achieve a fair comparison because the *MoDiPro* dataset does not contain the 3D annotations required by the method proposed in [24]. Specifically, we train and evaluate our network using the 2D projections of the Gaze360 gazes since our model is designed for 2D inputs.

1) Comparison With Gaze Tracking Methods: As shown in Table II, the Gaze360 method performs slightly better than our approach in its respective dataset. However, when evaluated on the *MoDiPro* dataset, our method achieves a mean angular error of 28.21° compared to 66.58° for Gaze360. This large error in the Gaze360 method is a limitation with appearance-based gaze estimation techniques. These methods require high resolution images of the individuals' facial features. When we downsample the Gaze360 test set to the same dimensions as our *MoDiPro* facial crops, the method introduced in [24] shows a significant drop in performance, with a mean angular error of 40.35° compared to the 24.48° error obtained in the original images. Despite the simple architecture of our method, it performs on par with the sophisticated model proposed in [24] on the Gaze360 dataset. Furthermore, we achieve more stable results across the Gaze360 and *MoDiPro* dataset, showing that our method better generalizes to unseen data and is not constrained by the quality of the facial features.

TABLE III

Mean Angular Error Comparisons of Our Optimal Model Using Different Temporal Integration Strategies

Method	Cam1	Cam2	*Mean*
Net#9	20.47°	22.92°	21.70°
Net#9 + WMA & $e^{(-\sigma)}$	20.33°	22.82°	21.58°
Net#9 + KF & $\frac{1}{\sigma}$	19.16°	21.14°	20.15°
Net#9 + KF & $e^{(-\sigma)}$	**18.92°**	**20.77°**	**19.85°**

Experimental results for the moving average scheme and the Kalman filter using the optimal model Net#9 are summarized in Table III. The moving average scheme leads to a modest improvement over the Net#9 model. This small improvement can be partially attributed to the fact that the moving average strategy does not take into consideration the dynamic nature of the gazes. As the results in Table III indicate, incorporating a motion model through the use of an angular Kalman filter leads to substantial performance gains. When using the inverse of the uncertainties as the observation covariance, we observe a reduction in angular error across both cameras with a mean angular error improvement of 1.38°. Using $e^{-\tilde{\sigma}}$ as the observation covariance further reduces the mean angular error by 1.51°. We suspect that using $e^{-\tilde{\sigma}}$ to update the observation covariance promotes a more significant increase in performance because it better reflects the actual covariances learned by the network. Fig. 12 illustrates the performance of the Kalman filter on a sample video from the *MoDiPro* dataset. The smoothing nature of the filter moves sporadic noisy predictions closer to the ground truth.

2) Uncertainty Variance Analysis: In this section, we explore the impact of the uncertainties on the temporal integration techniques. We hypothesize that the temporal integration methods have a higher impact on videos with higher uncertainty variances. Over extended periods with low uncertainty variance, meaning that the uncertainties are relatively constant, incorporating the uncertainties is essentially equivalent to weighing the raw predictions using a constant factor thus providing no actual impact on the adjusted predictions. Fig. 13 shows the distribution of uncertainty variance over the videos in the *MoDiPro* dataset. As the figure indicates, over 70% of the videos have an uncertainty variance of 0.005 or less.

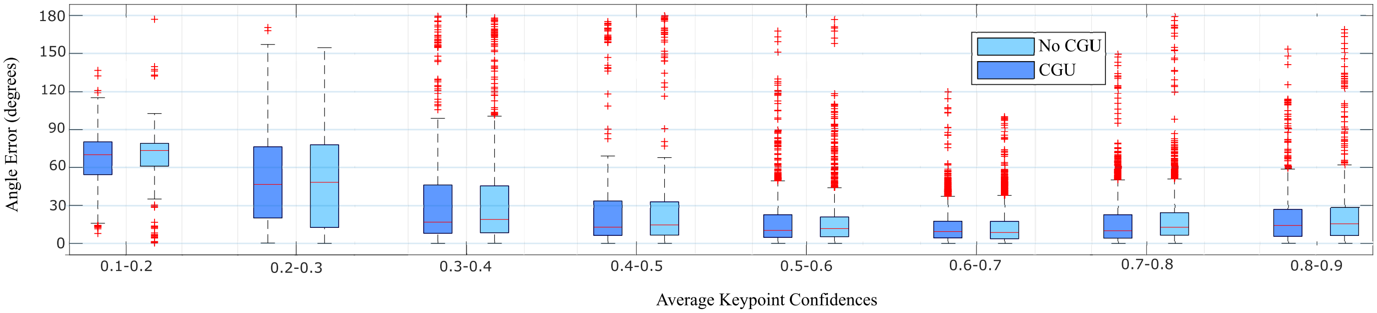
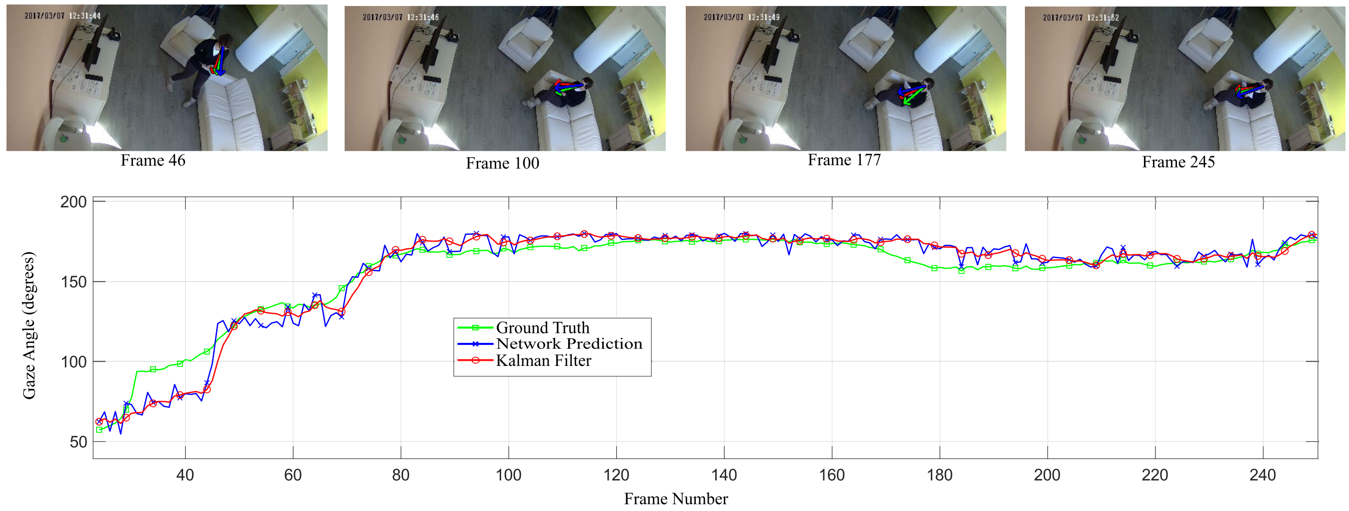


Fig. 11. Median angular errors as a function of average keypoint confidences on data with no missing keypoints for the models with CGUs and without CGU. The boxes represent the 75^{th} and 25^{th} percentiles of the error distribution, the horizontal red lines correspond to the median values, the whiskers represent the maximum and minimum values, and the red crosses indicate outliers. The performance improvement of the CGUs is more pronounced at lower confidences.



Frame 46

Frame 100

Frame 177

Frame 245

Fig. 12. Sample video from the *MoDiPro* dataset (CAM2) showing the smoothing effect of the Kalman filter compared to the raw predictions obtained from the network.

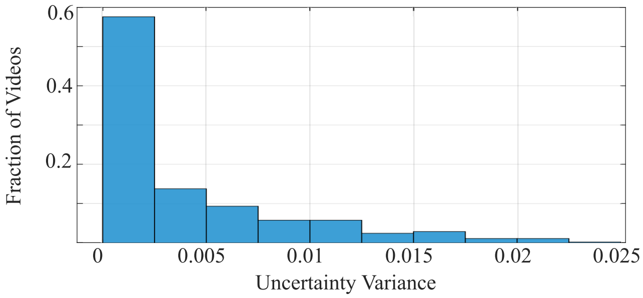


Fig. 13. Frequency of uncertainty variance ranges of videos as percentages.

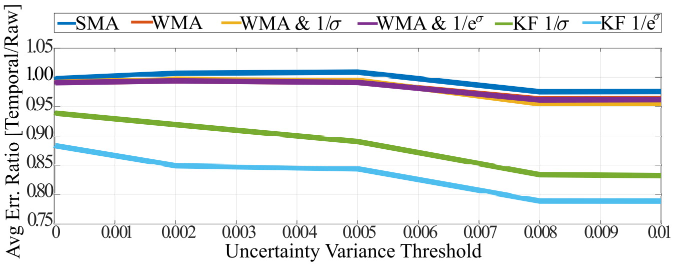


Fig. 14. Ratio of average angular error of temporal integration methods to network raw predictions for different uncertainty variance ranges.

Fig. 14 shows the ratio of average angular error to the raw network predictions as a function of video uncertainty variance for each of the proposed temporal integration methods. That is, values lower than one indicate relative performance gains, whereas values higher than one indicate performance degradation. The moving average schemes show no performance improvements for uncertainty variances lower than 0.005. In fact, the simple moving average (SMA) method shows a small degradation in that range. For higher variances, the moving average schemes show improvements of up to 5% over the raw predictions, whereas the Kalman filter methods show up to 16% and 19% improvements.

To further elucidate the impact of high uncertainty videos on the performance of our methods, we conduct an error analysis on a separate set of high uncertainty variance videos. These high uncertainty videos often correspond to unfavorable scenarios such as those involving significant illumination variations and occlusions among individuals. The average uncertainty variance in these videos is 0.012. These results are shown in Table IV. Although the performance gains provided by the moving average schemes are higher, they remain relatively modest. On the other hand, the strategies based on the Kalman filter show average improvements of 9.67° and 11.35° for the two annotation sets and improvements

TABLE IV

COMPARISON OF MEAN ANGULAR ERRORS OF THE ANGULAR KALMAN FILTER PREDICTIONS WITH THE MOVING AVERAGES AND NETWORK PREDICTIONS FOR HIGH UNCERTAINTY VARIANCE VIDEOS

	ANNOTATION SET 1			ANNOTATION SET 2					MERGED ANNOTATIONS			
	CAM1	CAM2	*Both*	CAM1	CAM2	*Both*	*Mean*	*Std. Dev*	CAM1	CAM2	*Both*	*Mean*
NET#9	49.51°	46.91°	47.40°	49.43°	49.13°	48.62°	48.50°	1.00°	50.00°	48.34°	48.60°	48.98°
Simple MA	48.94°	46.32°	46.74°	49.19°	48.70°	48.36°	48.04°	1.10°	49.58°	47.80°	48.24°	48.54°
Weighted MA	48.93°	46.22°	46.78°	49.07°	48.63°	48.08°	47.95°	1.08°	49.55°	47.77°	48.18°	48.50°
WMA & $\frac{1}{\sigma}$	49.14°	46.51°	46.94°	49.34°	48.93°	48.32°	48.20°	1.09°	49.78°	47.96°	48.31°	48.68°
WMA & $e^{(-\sigma)}$	48.29°	45.56°	46.28°	48.35°	47.89°	47.54°	47.32°	1.05°	49.03°	47.21°	47.70°	47.98°
KF & $\frac{1}{\sigma}$	38.95°	37.45°	37.02°	39.29°	40.83°	39.45°	38.83°	1.28°	40.81°	39.97°	39.99°	40.26°
KF & $e^{(-\sigma)}$	37.35°	35.58°	35.44°	37.64°	39.04°	37.86°	**37.15°**	1.27°	39.00°	38.32°	38.38°	**38.57°**

of 8.72° and 10.41° on the merged annotations for these high uncertainty variance videos. This large improvement in performance strongly supports the claim that high uncertainty variance videos greatly benefit from our uncertainty-weighed temporal integration methods.

3) Runtime Analysis: The most significant bottleneck in our proposed method is the pose estimation module. OpenPose processes one frame in 77 ms as reported in [22]. Our neural network regressor and Kalman filtering together average 0.013 ms per frame in the *MoDiPro* dataset. All values are obtained on a NVIDIA® GeFore® GTX-1080 Ti GPU.

V. CONCLUSION

The overarching goal of assisted living environments is to extract information regarding the behavior of its occupants to make inferences about their health status. Gaze direction is a key element in the evaluation of human behavior. In this paper, we present a gaze estimation and tracking approach that incorporates a temporal integration strategy to track the gaze direction of multiple individuals. Our gaze estimation regressor relies solely on the relative positions of the subject's facial keypoints and also provides an estimate of its uncertainty for each gaze prediction. We also introduce Confidence Gated Units, which we incorporate into our network architecture to mitigate the impact of low-confidence or occluded keypoints. The uncertainties from our model's predictions are used to generate the observation covariance of an angular Kalman filter for more robust and accurate gaze predictions. More importantly, our method relies exclusively on information extracted from the occupants of the environment. That is, it does not depend on salient features in the scene, which shift the gaze estimation problem from understanding human intentions to analyzing the environment itself and is thus prone to dataset biases.

Our experimental results demonstrate the importance of taking into consideration estimation uncertainties in the gaze tracking problem. Both our uncertainty-aware regressor and our CGUs play significant roles in reducing gaze estimation errors, particularly in unfavorable conditions. The high correlation between the uncertainties and the network prediction errors make it possible to use these uncertainties within an angular Kalman filter tracking framework that further improves the accuracy of our method. Experimental results on the *MoDiPro* dataset demonstrate the effectiveness of our method in a real assisted living environment. Furthermore, results on publicly available gaze datasets illustrate the generalization capability of our approach. These results indicate that our method generates prediction errors that are comparable to the variability observed in gaze directions manually estimated by human annotators.

In the future, we plan to investigate strategies for effectively utilizing additional keypoints for gaze estimation to better leverage relevant body language cues in the determination of gaze. Furthermore, we intend to continue to explore strategies to increase the robustness and the accuracy of our gaze tracking method. In particular, we plan to incorporate additional facial image features generated by a CNN to complement the facial keypoint information. We are also investigating strategies to incorporate temporal integration methods designed specifically for problems corrupted by noise that follows a von Mises distribution [66], [67], which would better reflect the gaze prediction errors observed in our system. Finally, we intend to investigate the impact of illumination on gaze estimation performance [68], [69].

ACKNOWLEDGMENT

The authors would like to thank NVIDIA Corporation for the donation of the GPU used for this research.

REFERENCES

[1] H. Wang et al., "Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: A comprehensive demographic analysis for the global burden of disease study 2019," *Lancet*, vol. 396, no. 10258, pp. 1160–1203, 2020.

[2] The United Nations. (2017). *World Population Ageing*. Accessed: Mar. 31, 2021. [Online]. Available: http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Report.pdf

[3] F. Odone, G. Grossi, R. Lanzarotti, H. Medeiros, and N. Noceti, "Guest editorial assistive computing technologies for human well-being," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 3, pp. 1231–1233, Jul. 2021.

[4] A. Pilotto et al., "Development and validation of a multidimensional prognostic index for one-year mortality from comprehensive geriatric assessment in hospitalized older patients," *Rejuvenation Res.*, vol. 11, no. 1, pp. 151–161, Feb. 2008.

[5] G. Grossi, R. Lanzarotti, P. Napoletano, N. Noceti, and F. Odone, "Positive technology for elderly well-being: A review," *Pattern Recognit. Lett.*, vol. 137, pp. 61–70, Sep. 2020.

[6] S. Mihradi, T. Dirgantara, and A. I. Mahyuddin, "Development of an optical motion-capture system for 3D gait analysis," in *Proc. 2nd Int. Conf. Instrum., Commun., Inf. Technol., Biomed. Eng.*, Nov. 2011, pp. 391–394.

[7] T. Zult, J. Allsop, J. Tabernero, and S. Pardhan, "A low-cost 2-D video system can accurately and reliably assess adaptive gait kinematics in healthy and low vision subjects," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, Dec. 2019.

[8] C. Martini et al., "Visual computing methods for assessing the well-being of older people," in *Proc. Int. Joint Conf. Comput. Vis., Imag. Comput. Graph.*, 2018, pp. 195–211.

[9] C. Martini et al., "A visual computing approach for estimating the motility index in the frail elder," in *Proc. 13th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2018, pp. 439–445.

[10] M. Chessa et al., "Designing assistive tools for the market," *Comput. Vis. Assistive Healthcare*, vol. 2, pp. 337–362, Jan. 2018.

[11] C. Martini et al., "Data-driven continuous assessment of frailty in older people," *Frontiers Digit. Humanities*, vol. 5, p. 6, Apr. 2018.

[12] P. Dias, H. Medeiros, and F. Odone, "Fine segmentation for activity of daily living analysis in a wide-angle multi-camera set-up," in *Proc. 5th Activity Monit. Multiple Distrib. Sens. Workshop (AMMDS) Conjunct. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[13] P. A. Dias, D. Malafronte, H. Medeiros, and F. Odone, "Gaze estimation for assisted living environments," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 290–299.

[14] P. Her, L. Manderle, P. Dias, H. Medeiros, and F. Odone, "Keypoint-based gaze tracking," in *Proc. ICPR Int. Workshops Challenge*, 2020, pp. 144–155.

[15] P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction," in *Advances in Physiological Computing*. Berlin, Germany: Springer, 2014, pp. 39–65.

[16] J. Varadarajan, R. Subramanian, S. R. Bulo, N. Ahuja, O. Lanz, and E. Ricci, "Joint estimation of human pose and conversational groups from social scenes," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 410–429, Apr. 2018.

[17] L. Fan, W. Wang, S.-C. Zhu, X. Tang, and S. Huang, "Understanding human gaze communication by spatio-temporal graph reasoning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5723–5732.

[18] G. Prabhakar, A. Mukhopadhyay, L. Murthy, M. Modiksha, D. Sachin, and P. Biswas, "Cognitive load estimation using ocular parameters in automotive," *Transp. Eng.*, vol. 2, Dec. 2020, Art. no. 100008.

[19] Z. Zuo, L. Yang, Y. Peng, F. Chao, and Y. Qu, "Gaze-informed egocentric action recognition for memory aid systems," *IEEE Access*, vol. 6, pp. 12894–12904, 2018.

[20] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 619–635.

[21] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman, "LAEO-Net: Revisiting people looking at each other in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3477–3485.

[22] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[23] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 1–9.

[24] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6912–6921.

[25] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Exp. Syst. Appl.*, vol. 39, pp. 10873–10888, Sep. 2012.

[26] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. M. Farinella, "Computer vision for assistive technologies," *Comput. Vis. Image Understand.*, vol. 154, pp. 1–15, Jan. 2017.

[27] C. F. Crispim et al., "Evaluation of a monitoring system for event recognition of older people," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2013, pp. 165–170.

[28] N. Zouba, F. Bremond, and M. Thonnat, "An activity monitoring system for real elderly at home: Validation study," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2010, pp. 278–285.

[29] M. Leo, A. Furnari, G. G. Medioni, M. Trivedi, and G. M. Farinella, "Deep learning for assistive computer vision," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2018, pp. 1–12.

[30] K. Min and J. J. Corso, "Integrating human gaze into attention for egocentric activity recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1069–1078.

[31] R. Singh, T. Miller, J. Newn, E. Velloso, F. Vetere, and L. Sonenberg, "Combining gaze and AI planning for online human intention recognition," *Artif. Intell.*, vol. 284, Jul. 2020, Art. no. 103275.

[32] J. Schwarz, C. C. Marais, T. Leyvand, S. E. Hudson, and J. Mankoff, "Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Apr. 2014, pp. 3443–3452.

[33] Y. Yu and J.-M. Odobez, "Unsupervised representation learning for gaze estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7314–7324.

[34] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Trans. Image Process.*, vol. 29, pp. 5259–5272, 2020.

[35] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu, "Where and why are they looking? Jointly inferring human attention and intentions in complex tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6801–6809.

[36] F. Lu, X. Chen, and Y. Sato, "Appearance-based gaze estimation via uncalibrated gaze pattern recovery," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1543–1553, Apr. 2017.

[37] Y. Cheng, F. Lu, and X. Zhang, "Appearance-based gaze estimation via evaluation-guided asymmetric regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 100–115.

[38] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2299–2308.

[39] S. Lathuiliere, R. Juge, P. Mesejo, R. Munoz-Salinas, and R. Horaud, "Deep mixture of linear inverse regressions applied to head-pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4817–4825.

[40] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1087–1096.

[41] R. Berral-Soler, F. J. Madrid-Cuevas, R. Munoz-Salinas, and M. J. Marin-Jimenez, "RealHePoNet: A robust single-stage ConvNet for head pose estimation in the wild," *Neural Comput. Appl.*, vol. 33, pp. 7673–7689, Jul. 2021.

[42] Y. Wang, W. Liang, J. Shen, Y. Jia, and L.-F. Yu, "A deep coarse-to-fine network for head pose estimation from synthetic data," *Pattern Recognit.*, vol. 94, pp. 196–206, Oct. 2019.

[43] S. Jha and C. Busso, "Analyzing the relationship between head pose and gaze to model driver visual attention," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 2157–2162.

[44] R. Ranjan, S. De Mello, and J. Kautz, "Light-weight head pose invariant gaze tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2156–2164.

[45] Z. Wang et al., "Learning to detect head movement in unconstrained remote gaze estimation in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3443–3452.

[46] G. Cantarini, F. F. Tomenotti, N. Noceti, and F. Odone, "HHP-Net: A light heteroscedastic neural network for head pose estimation with uncertainty," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3521–3530.

[47] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," 2021, *arXiv:2104.12668*.

[48] T. Fischer, H. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 334–352.

[49] C. P. Cantarino, O. V. Komogortsev, and S. S. Talathi, "Benefits of temporal information for appearance-based gaze estimation," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, Jun. 2020, pp. 1–5.

[50] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5574–5584.

[51] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, Dept. Eng., University of Cambridge, Cambridge, U.K., 2016.

[52] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder–decoder architectures for scene understanding," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–11.

[53] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.

[54] S. Prokudin, P. Gehler, and S. Nowozin, "Deep directional statistics: Pose estimation with uncertainty quantification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 534–551.

[55] A. Giusti and F. Mainardi, "On infinite series concerning zeros of Bessel functions of the first kind," *Eur. Phys. J. Plus*, vol. 131, no. 6, pp. 1–7, Jun. 2016.

[56] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.

[57] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[58] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and W. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2010.

[59] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[60] D. Lian, Z. Yu, and S. Gao, "Believe it or not, we know what you are looking at!" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, May 2019, pp. 35–50.

[61] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10623–10630.

[62] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2018, pp. 309–324.

[63] D. Kinga and J. B. Adam, "A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[65] S. Ghosh, A. Dhall, M. Hayat, J. Knibbe, and Q. Ji, "Automatic gaze analysis: A survey of deep learning based approaches," 2021, *arXiv:2108.05479*.

[66] G. Kurz, I. Gilitschensk, and U. D. Hanebeck, "Recursive Bayesian filtering in circular state spaces," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 31, no. 3, pp. 70–87, Mar. 2016.

[67] G. Kurz, I. Gilitschenski, and U. D. Hanebeck, "Unscented von Mises–Fisher filtering," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 463–467, Apr. 2016.

[68] L. R. D. Murthy and P. Biswas, "Deep learning-based eye gaze estimation for military aviation," in *Proc. IEEE Aerosp. Conf. (AERO)*, Mar. 2022, pp. 1–8.

[69] L. R. D. Murthy, A. Mukhopadhyay, K. Anand, S. Aggarwal, and P. Biswas, "PARKS-Gaze—A precision-focused gaze estimation dataset in the wild under extreme head poses," in *Proc. 27th Int. Conf. Intell. User Interfaces*, Mar. 2022, pp. 81–84.

Paris Her received the B.S. degree in electrical engineering from the Milwaukee School of Engineering (MSOE) in 2019 and the M.S. degree in electrical and computer engineering from Marquette University in 2021, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research consists of utilizing machine learning techniques in various computer vision applications. These applications include violence detection in surveillance videos, generating synthetic lung and colon cancer images, and gaze estimation for the elderly in assisted living environments in collaboration with the University of Genoa, Italy.

Logan Manderle received the Bachelor of Science degree in electrical engineering from Marquette University in Fall 2020. His advanced course work focused on signals, systems, and controls. His research interests include machine learning. As a recipient of an International Engineering Research Fellowship, he collaborated with the University of Genoa, Italy, for his summer research. Since graduating, he has been with Milwaukee Tool on Embedded Systems and Power Electronics.

Philipe A. Dias (Member, IEEE) received the M.Sc. degree in information technology from Hochschule Mannheim, Germany, the M.S. degree in electrical and computer engineering from the Federal University of Technology Paraná (UTFPR), Brazil, and the Ph.D. degree in electrical and computer engineering from Marquette University. He is currently a Research and Development Associate with the Oak Ridge National Laboratory (ORNL). His research interests include supervised and unsupervised learning, combined with probability theory, and stochastic simulation applied to computer vision models. His work has found application in agricultural automation, semantic segmentation, and image annotation tools, with current emphasis on analysis of remote sensing imagery. His Ph.D. studies included a period with the University of Genoa, Italy, as part of a collaboration on applying such techniques for healthcare-related scenarios.

Henry Medeiros (Senior Member, IEEE) received the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University. He is currently an Associate Professor in agricultural and biological engineering with the University of Florida. He has published over 50 papers and peer-reviewed conference papers and holds several patents in the United States and other countries. Before joining the University of Florida, he was an Associate Professor of electrical and computer engineering with Marquette University and the Chief Technology Officer of Spensa Technologies, a technology start-up company. His research interests include computer vision and robotics, and his work focuses on the application of machine learning and signal processing techniques to solve problems of practical relevance in areas ranging from manufacturing to agricultural automation and assisted living environments. He was a recipient of the National Science Foundation Faculty Early Career Development Program (CAREER) Award. He has been an Associate Editor for the IEEE International Conference on Robotics and Automation, the IEEE International Conference on Intelligent Robots, and the IEEE Winter Conference on Applications of Computer Vision.

Francesca Odone received the Laurea degree in information sciences and the Ph.D. degree in computer science from the University of Genova. From 1999 to 2000, she was a Visiting Student with Heriot-Watt University, Edinburgh, U.K., with an EU Marie Curie Research Grant. She is currently a Professor of computer science with the University of Genova and a Founding Member of the Machine Learning Genoa (MaLGa) Center. She has authored over 100 papers on international conferences and journals. Her research interests include computer vision and machine learning, including multi-resolution signal processing, feature extraction, feature selection, and data-driven representations for visual data. Often times her research is linked to applied tasks in the fields of robotics, ambient-assisted living, rehabilitation, and video-surveillance. She has been involved in various research projects and acted as a scientific coordinator of technology transfer contracts with SMEs, large companies, and hospitals. For more information visit the link (https://person.dibris.unige.it/odone-francesca).