Keypoint-Based Gaze Tracking

Paris Her[1][0000−0002−7704−5587], Logan Manderle[1], Philipe A. Dias[1][0000−0001−9427−7112], Henry Medeiros[1][0000−0002−7704−5587], and Francesca Odone[2][0000−0002−3463−2263]

[1] Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI, USA
paris.her@marquette.edu, logan.manderle@marquette.edu
philipe.ambroziodias@marquette.edu, henry.medeiros@marquette.edu
[2] Department of Informatics, Bioengineering, Robotics, and Systems Engineering at the University of Genoa, Italy
francesca.odone@unige.edu

Abstract. Effective assisted living environments must be able to perform inferences on how their occupants interact with their environment. Gaze direction provides strong indications of how people interact with their surroundings. In this paper, we propose a gaze tracking method that uses a neural network regressor to estimate gazes from keypoints and integrates them over time using a moving average mechanism. Our gaze regression model uses confidence gated units to handle cases of keypoint occlusion and estimate its own prediction uncertainty. Our temporal approach for gaze tracking incorporates these prediction uncertainties as weights in the moving average scheme. Experimental results on a dataset collected in an assisted living facility demonstrate that our gaze regression network performs on par with a complex, dataset-specific baseline, while its uncertainty predictions are highly correlated with the actual angular error of corresponding estimations. Finally, experiments on videos sequences show that our temporal approach generates more accurate and stable gaze predictions.

Keywords: Gaze tracking · Neural networks · Assisted living environments.

1 Introduction

Official prospects from the United Nations (UN) indicate an expected 15% of the world's population to be over age 65 by 2050 [18]. As the older population grows, advances in intelligent medical care systems will prove essential for providing these individuals with improved quality of life, consequently avoiding costly medical interventions. In contrast to conventional methods based on sporadic questionnaires and self-reported outcomes, there is great interest in developing health-assessment techniques that are cost-effective, unobtrusive, objective, and informative over longer periods.

Thus, many studies have been attempting to leverage recent advances in robotics and artificial intelligence for assessment of patterns related to health

status, such as mobility and Instrumented Activities of Daily Living (IADL) assessments [19]. Ambient assisted living applications can particularly benefit from modern computer vision algorithms, as applications on safety, well-being assessment, and human-machine interaction demonstrate [3, 14].

Yet, to date systems exploiting computer vision for patient activity analysis have been limited to simplistic scenarios [6]. In contrast, as Figs. 1 and 2 illustrate, images acquired from assisted living environments cover a wide scene where different activities involving multiple people can take place.

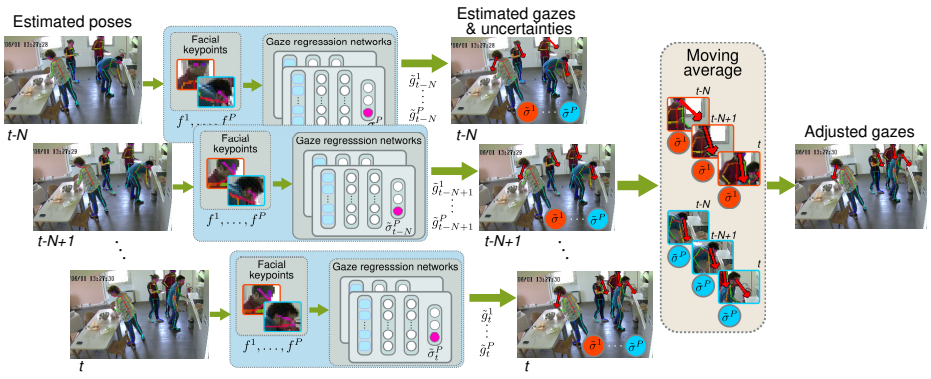

Fig. 1. Overview of our apparent gaze tracking approach. The facial keypoints of each person in the scene are collected using a pose estimation model [2] and provided as inputs to a neural network regressor that outputs estimations of their apparent gaze and its confidence $\tilde{\sigma}$ on each prediction. An uncertainty-weighted moving average scheme then combines the estimations collected from the last N frames, generating temporally consistent gaze estimations at each time instant.

Our long-term goal is to exploit video analytics to monitor the overall health status of patients by observing their behavior in terms of human-human and human-object interactions. To that end, multiple underlying complex tasks must be addressed, including: i) human and object detection; ii) human pose estimation; and iii) subjects' gaze estimation and tracking.

We have introduced in [7] an approach for precise segmentation of individuals and objects of interest in video-streams acquired from assisted living environments. In conjunction with object detection, gaze direction is crucial to differentiate relationships between objects and their users (e.g. person with a book on his/her lap vs. actually reading a book) and classify simple actions (e.g. watching television, cooking food, socializing).

In [8], we introduced a novel strategy for gaze estimation that relies solely on the relative positions of facial keypoints to estimate gaze direction. These features can be extracted using off-the-shelf human pose estimation models such as [2], with the advantage that a single feature extractor module can be used for both pose estimation and gaze estimation.

Fig. 2. Images and layout of the instrumented assisted living facility; in color, the fields of view of the video cameras.

Operating on a frame-by-frame basis, our network introduced in [8] also outputs an estimate of its own uncertainty for each prediction of gaze direction. In this paper, we build upon that model to provide the following contributions:

- we design an effective framework that temporally integrates gaze estimates collected at different frames, while leveraging the uncertainty associated with each estimation;
- for model optimization and evaluation, we augment our previous MoDiPro dataset [8] with annotations of full video sequences;
- we evaluate different *moving average schemes that utilizes past gaze estimations to adjust current gaze predictions*. We evaluate different weighing strategies to correct the current gaze using the gaze uncertainties determined from the regressor network.

2 Related Work

Many studies exploit human facial features for the estimation of well-being status [1]. In addition to examples including facial expression recognition [15, 22] for sentiment analysis [10], facial analysis is also commonly used for gaze estimation, since gaze direction provides valuable information on the interaction between a person and his/her surrounding environment [21]. Recent studies in this area included approaches based on the estimation of head orientation by fitting a 3D face model, to estimate both 2D [23] and 3D gaze information [24]. In the context of human-computer interaction, the work in [13] employs an end-to-end architecture to track the eyes of a user in real-time using hand-held devices.

Most works and datasets on gaze estimation focus, however, on images with close-up views of a single subject's face, acquired through webcams or smartphones [9, 23]. The GazeFollow dataset introduced in [20] is an exception, containing images of individuals performing actions in relatively unconstrained scenarios. In addition to the dataset, the authors introduced a CNN-based architecture for gaze estimation that combines image saliency with head appearance analysis. A similar model is introduced in [5], with applicability extended to scenarios where the subject's gaze is directed somewhere outside the image.

Even for humans, it is much easier to estimate the gaze direction of a subject when a full-view of the subject's face is available, with the task becoming significantly harder if only a back view from the subject is available. Yet, the aforementioned CNN-based approaches for gaze estimation lack the ability of estimating the uncertainty associated with their estimations, a limitation inherited from conventional deep learning models in general [12].

As detailed in Section 3, our gaze estimation model introduced in [8] overcomes this limitation by exploiting a customized loss function that, based on the modelling of outputs as corrupted with Gaussian random noise, allows learning a regression model that also predicts the variance of this noise as a function of the input [12], without the need for any extra labels. However, both model design and evaluation using such datasets focus on frame-by-frame scenarios that disregard any available temporal information. From an application perspective, gaze tracking across longer time periods is crucial for the identification of activities taking place in an environment of interest.

In contrast to most datasets, the publicly available Gaze360 dataset [11] contains images and annotations for full video sequences, with 238 subjects in total and 80 different recordings. Images are acquired in a variety of natural environments, including indoor and outdoor locations, with variations in lighting and background.

3 Proposed Approach

Our algorithm uses the method we have previously proposed in [8] to simultaneously estimate the gazes of all the people observed at each frame. As Fig. 1 indicates, the gaze estimation method uses a pose estimation model [2] to detect the anatomical keypoints of all the persons present in the scene. Of the detected keypoints, we consider only those located in the head (i.e., the nose, eyes, and ears) of each individual to estimate their corresponding gazes.

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 k and its corresponding detection confidence value, respectively. The subscript $k \in \{n, e, a\}$ represents the nose, eyes, and ears features, with the subscript $s \in \{l, r, \emptyset\}$ encoding the side of the feature points. For each person j in the scene, we centralize the detected keypoints with respect to the head centroid $h^j = [x^j_h, y^j_h]$, which is computed as the mean coordinates of the head keypoints for each individual. Then, the obtained relative coordinates are normalized based on the distance of the farthest keypoint to the centroid. Hence, for each person we form a feature vector $f \in \mathbb{R}^{15}$ by concatenating the relative vectors $\hat{p}^j_{k,s} = [\hat{x}^j_{k,s}, \hat{y}^j_{k,s}, c^j_{k,s}]$

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

To account for low-confidence or missing keypoints, for each feature $\hat{p}^j_{k,s}$, the corresponding coordinate-confidence pairs $(\hat{x}^j_{k,s}, c^j_{k,s})$ and $(\hat{y}^j_{k,s}, c^j_{k,s})$ are used as input to a Confidence Gated Unit (CGU). As described in [8], each CGU is

composed of two internal units: i) a ReLU unit acting on an input feature q_i, and ii) a sigmoid unit to emulate the behavior of a gate according to a confidence value c_i. The outputs of both units are multiplied to generate an adjusted CGU output $\tilde{q}_i$.

The gaze direction is approximated by the vector $\tilde{g}^j = [\tilde{g}_x, \tilde{g}_y]$, which consists of the projection onto the image plane of the unit vector centered at the centroid h^j. Our model further incorporates an uncertainty estimation method, which indicates its level of confidence for each prediction of gaze direction. In terms of network architecture, this corresponds to an output layer with 3 units: two that regress the $(\tilde{g}_x, \tilde{g}_y)$ vector of gaze direction, and an additional unit that outputs the regression uncertainty σ.

To train the network to learn gaze direction, we use a cosine similarity loss function modified according to [12] to allow uncertainty estimation. Let $\mathcal{T}$ be the set of annotated orientation vectors g, while $\tilde{g}$ corresponds to the estimated orientation produced by the network and σ represents the model's uncertainty prediction. Our cost function is then given by

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

With this loss function, no additional labels are needed for the model to learn to predict its own uncertainty. The $\exp(-\sigma)$ component is a more numerically stable representation of $\frac{1}{\sigma}$, which encourages the model to output a higher σ when the cosine error is higher. On the other hand, the regularizing component $\log(\sigma)$ helps avoiding an exploding uncertainty prediction.

Following ablative experiments and weight visualization to identify dead units, we opt for an architecture where the CGU-based input layer is followed by 2 fully-connected (FC) hidden layers with 10 units each, and the output layer with 3 units. Thus, the architecture has a total of 283 learnable parameters and can be summarized as: (10 CGU, 10 FC, 10 FC, 3 FC).

3.1 Temporal Integration

After generating the raw predictions using the regressor network, we employ a moving average strategy to integrate gaze predictions over multiple frames. Let $\tilde{g}_t$ represent the gaze direction vector estimated by the neural network described above at time t. The refined gaze estimate that incorporates information from the previous N frames is given by

$$\hat{g}_t = \frac{\sum_{n=0}^{N} \alpha_{t-n} \tilde{\sigma}_{t-n} \omega_{t-n} \tilde{g}_{t-n}}{\sum_{n=0}^{N} \tilde{\sigma}_{t-n}}, \tag{3}$$

where α_{t-n} are empirically defined weights, $\tilde{\sigma}_{t-n}$ is a function of the estimated gaze uncertainty at time $t-n$, and the forgetting factor ω_{t-n} is given by

$$\omega_{t-n} = \frac{N - n + 1}{[N(N+1)]/2}. \tag{4}$$

We consider two forms for the uncertainty weights $\tilde{\sigma}_{t-n}$:

$$\tilde{\sigma}_{t-n} = \frac{1}{\sigma_{t-n}} \text{ and } \tilde{\sigma}_{t-n} = \frac{1}{e^{\sigma_{t-n}}}. \tag{5}$$

We evaluate four combinations of the parameters above. In our first approach, which we call *Simple MA* (or *SMA* for conciseness), we consider $\alpha_{t-n} = 1$ and $\tilde{\sigma}_{t-n} = 1$ for $n = 0, 1, \ldots, N$. Our second strategy, *Weighed MA* (or *WMA*), uses an empirically defined weight for the current frame $\alpha_t = \alpha$ and identical weights for previous frames, i.e., $\alpha_{t-n} = 1 - \alpha$ for $n = 1, 2, \ldots, N$. Finally, the strategies in which the value of $\tilde{\sigma}_{t-n}$ is given by the functions in Eq. (5) are deemed *WMA* & $\frac{1}{\sigma}$ and *WMA* & $e^{-\sigma}$.

4 Experiments and Results

We evaluate our method on videos acquired in an assisted living facility situated in the Galliera Hospital (Genova, Italy), in which the patient, after being discharged from the hospital, is hosted and monitored for a few days. This furnished apartment contains various sensing systems. Specifically, we utilize the two video cameras illustrated in Fig. 2, which acquire videos at a resolution of 480×270 pixels at 25 frames per second. For more details, we refer the reader to [4, 16, 17].

MoDiPro Dataset. Our dataset contains 47 videos captured by Camera 1 with a total of 15,750 frames and 30 videos from Camera 2 with 10,750 frames, totaling to 26,500 frames in which 22 individuals are observed. Two annotators manually labelled the gaze directions in each video frame. Annotation Set 1 contains a total of 24,509 observable gazes with at least 2-keypoints, and Annotation Set 2 has 24,494 observable gazes with at least 2 key-points. Fig. 3 illustrates the gaze distribution of the two annotation sets for the observable gazes only. The angle distributions are consistent with how a human viewer would see the gazes in the video frame. In Camera 1, subjects tended to look vertically, typically towards objects on the table. In Camera 2, subjects tended to look east, typically in the direction of the television.

One of the significant challenges in the gaze tracking problem is its inherent uncertainty. Different viewers looking at the same image or video of a person may estimate significantly different values for the subject's gaze. This is evident if we observe the statistics of the two sets of annotations for the *MoDiPro* dataset. Although we observed little bias among the annotations (average difference of 0.08°), the variability was substantial with a standard deviation of 23.30°.

Network training. We use 50% of the videos from each camera for training, 20% for validation, and 30% for testing. Since frames from the same video are highly correlated, all frames from a given video sequence are assigned to the same training, validation, or test set. To combine the two annotation sets described above into one cohesive set, we calculate the mean gaze vector between Annotation Sets 1 and 2 as our ground truth. To analyze the effect of

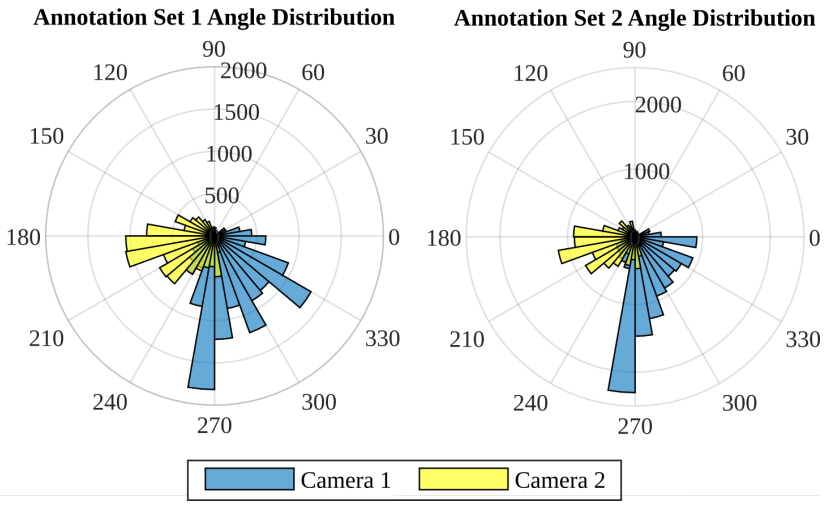

Fig. 3. Distribution of the annotation datasets for Camera 1 (blue) and Camera 2 (yellow). Left) Annotation Set 1. Right) Annotation Set 2.

human annotation error, we also evaluate our methods on the two annotations sets separately.

We train our model with 7 different combinations of images from the *MoDiPro* and GazeFollow datasets. As summarized in Table 1, models NET#0-2 are trained using only images from Camera 1 (*Cam1*), Camera 2 (*Cam2*), and both cameras. NET#3 corresponds to the model trained only on GazeFollow frames (GF for shortness) while NET#4-6 are obtained by fine-tuning the pre-trained NET#3 on the three subsets of *MoDiPro* frames. All models are trained using a learning rate of 3×10^{-7}, batches of 64 samples, and early stopping based on the validation loss. The results reported in Table 1 correspond to the average values obtained after train/test on 3 different random splits. The mean column below is the average of the result from experiments using both camera videos as training.

4.1 Gaze Regression Performance

Cross-view results obtained by NET#0 on $Cam2$ and NET#1 on $Cam1$ demonstrate how models trained only on a camera-specific set of images are less robust to image distortions, with significantly higher angular errors for images captured by a different camera. Trained on both $Cam1$ and $Cam2$, NET#2 demonstrates a more consistent performance across views. We can observe a slight decrease in performance for the camera-specific tests in $Cam1$ and $Cam2$ of 0.67° and 2.57° respectively, but the performance for the non camera-specific tests improve dramatically by 41.67° and 23.72° for $Cam1$ and $Cam2$, respectively.

Table 1. Mean angular error obtained on three different random splits of train/val/test sets for each camera across from the merged annotation set.

	Train			Test		
Model	GF	*Cam1*	*Cam2*	*Cam1*	*Cam2*	*Mean*
GF-Model				45.82°	76.55°	61.18°
Net#0		✓		21.85°	49.75°	-
Net#1			✓	64.19°	23.46°	-
Net#2		✓	✓	22.52°	26.03°	24.28°
Net#3	✓			23.29°	25.90°	24.60°
Net#4	✓	✓		19.71°	22.94°	-
Net#5	✓		✓	22.40°	23.92°	-
Net#6	✓	✓	✓	21.17°	23.56°	22.37°

In addition, error comparisons between models Net#0-2 and Net#4-6 demonstrate that pre-training the model on the GF dataset before fine-tuning on *MoDiPro* images leads to consistently lower mean angular errors, with an optimal performance of 21.17° for $Cam1$ and 23.56° for $Cam2$. This corresponds to an overall average error 1.91° lower than the model Net#2, which is not pre-trained on GF, and more than 2.23° improvement over the model Net#3, which is trained solely on GF. In terms of camera-specific performance, for $Cam1$ optimal performances with errors below 20° are obtained when not training on $Cam2$. On the other hand, predictions for $Cam2$ are significantly better when training is performed using $Cam1$ and/or GazeFollow images. We hypothesize that the distortions characteristic of $Cam2$ images easily lead to overfitting, thus confirms the advantage of training on additional sets of images. As a final remark we note that overall Net#6, which is pre-trained on GF and further trained using images from both camera views, provides the best and most stable result across the two cameras. NET#6 has a mean angle difference across the two views being 22.37°, compared to NET#2 and #3 being 24.28° and 24.60° respectively.

Furthermore, Fig. 4 illustrates the high-correlation between uncertainty predictions and angular error. The figure shows the cumulative mean angular error versus gaze uncertainty estimations. When looking at gazes with lower predicted uncertainties, the overall mean angular error is significantly lower. Gaze prediction uncertainties below 0.1 correspond to 80% of the *MoDiPro* data. Hence, for 80% of our predictions, the mean angular error is only $\approx 16°$ compared to over 20° for the entire set.

4.2 Temporal Integration Performance

Experimental results corresponding to different moving average strategies are summarized in Table 2. The table shows separate results sections for the two annotation sets and the average and standard deviation of the error over the two datasets. We compute the average for the two coordinates of the gaze direction vector separately for each of the methods. The values of the parameters N,

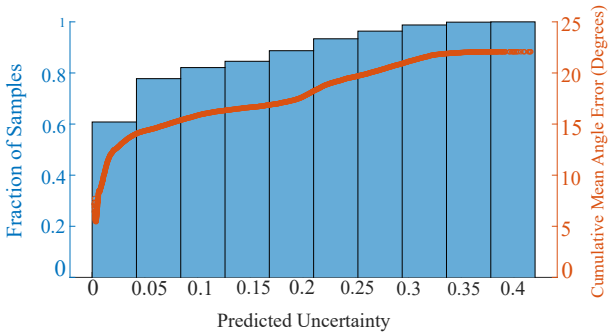

Fig. 4. Cumulative mean angular error versus predicted uncertainty.

α_{t-n}, and $\tilde{\sigma}_{t-n}$ were determined experimentally. For the *Simple MA* method, we observed that values greater than $N = 3$ led to diminishing returns on the average angular error across both cameras. For *Weighed MA*, a value of $N = 6$ leads to similar behavior, whereas for *WMA* & $\frac{1}{\sigma}$ and *WMA* & *WMA* & $e^{(-\sigma)}$, $N = 5$. An iteration over the values of α from 0.05 with a step size of 0.05 to 0.95 was performed to find the optimal α. For *Simple MA*, we found the optimal value of $\alpha = 0.85$ and for the remaining methods we use $\alpha = 0.60$. As the table indicates, improvements can be seen from using just a simple moving average, and then again using weighted moving averages. The best improvements are observed with the *Weighted MA*, *WMA* & $\frac{1}{\sigma}$, and *WMA* & $e^{(-\sigma)}$. The relatively small improvements obtained using the uncertainty estimates can be explained by the variances of the uncertainties, which are discussed in the next section.

Table 2. Comparison of mean angular errors of the network predictions with the moving averages for both sets

	Annotation Set 1			Annotation Set 2				
	Cam1	**Cam2**	**Both**	**Cam1**	**Cam2**	**Both**	**Mean**	**Std. Dev**
NET#6	21.54°	22.71°	21.98°	23.64°	25.42°	24.16°	23.24°	1.23°
Simple MA	21.50°	22.59°	21.87°	23.54°	25.26°	24.11°	23.15°	1.21°
Weighted MA	21.40°	22.45°	21.73°	23.47°	25.22°	23.98°	23.04°	1.23°
WMA & $\frac{1}{\sigma}$	21.47°	22.43°	21.72°	23.52°	25.19°	24.00°	23.06°	1.22°
WMA & $e^{(-\sigma)}$	21.42°	22.46°	21.72°	23.47°	25.21°	23.98°	23.04°	1.22°

Uncertainty Variance Analysis. In this section, we explore the impact of the uncertainties on the temporal integration method. We hypothesize that videos with higher uncertainty variances would show larger performance improvements. Over extended periods of low uncertainty variance, meaning that the predicted uncertainties are relatively constant, when we introduce the uncertainties, we are essentially multiplying the raw predictions by a constant factor

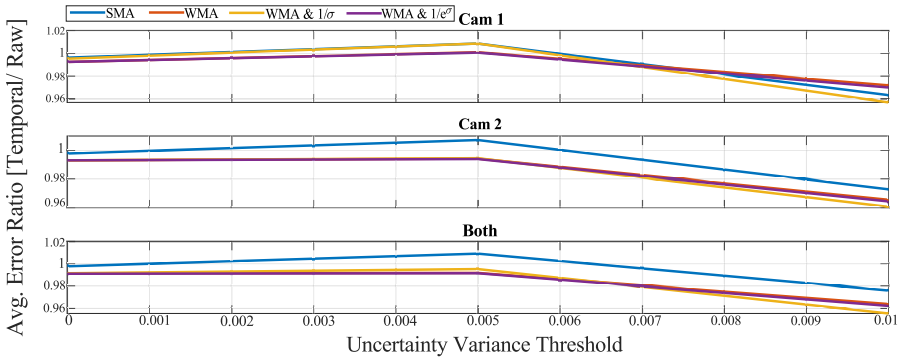

Fig. 5. Average angle error ratio trend across *Cam1*, *Cam2*, and *Both* for the given uncertainty variance thresholds .

thus providing no actual impact on the adjusted prediction. Since in our dataset nearly 80% of the videos show an uncertainty variance of 0.01 or less, our hypothesis suggests that this would lead to a marginal impact of the uncertainties on overall performance. This can be observed by comparing the performance obtained by the methods that consider the uncertainty with those that do not in Table 2. As the table indicates, the differences in angular mean error from our raw predictions and moving average methods are relatively small.

The greatest impact of incorporating the uncertainties would occur in situations involving significant fluctuations of the uncertainties. Hence, we conduct an analysis of the mean angular error as a function of the variance of the uncertainty. We partition the test video sets according to the variance of the uncertainty and measure the corresponding angular error for each subset. Fig. 5 illustrates the results when the variance threshold varies between 0, which is equivalent to the scenario evaluated in Table 2, to 0.01, which is the highest uncertainty variance we considered. In the figure, we plot the ratio between the raw mean angular error and each of the moving average methods. That is, a value less than one indicates performance gain whereas values above one indicate performance degradation.

As Fig. 5 indicates, the benefits of the uncertainty-weighed methods increase at higher variance threshold values. Although the *WMA* & $e^{-(\sigma)}$ method performs on par or slighlty better than *WMA* on both cameras for uncertainty variances under 0.07, the *WMA* & $\frac{1}{\sigma}$ method outperforms both methods by more than 1° at higher variances. This indicates that more sophisticated mechanisms to incorporate the uncertainties are a promising future research direction.

5 Conclusion

This paper presents a gaze tracking method based solely on facial keypoints detected by a pose estimation model. Our end goal is to assist clinicians in the assessment of the health status of individuals in assisted living environments, providing them with automatic reports of patients' mobility and IADL pat-

terns. Thus, we plan to combine gaze estimations with a semantic segmentation model to identify human-human and human-object interactions. Exploring a single feature extraction backbone for both pose and gaze estimation also reduces the complexity of the overall model.

Results obtained on datasets acquired at a real assisted living facility demonstrate that our method estimates gaze with higher accuracy than a complex task-specific baseline, without relying on any image features except the relative positions of facial keypoints. Our proposed model also provides estimations of uncertainty of its own predictions, and our results demonstrate a high correlation between predicted uncertainties and actual gaze angular errors.

We then showed that a simple moving average mechanism can be used to improve the temporal consistency and slightly reduce the estimation error of the gazes throughout a video sequence. In particular, our experimental results demonstrate that in scenarios where high gaze estimation uncertainty is present, moving average methods that leverage the estimated uncertainty can lead to more significant improvements.

References

1. Baltrušaitis, T., Robinson, P., Morency, L.: Openface: an open source facial behavior analysis toolkit. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–10. IEEE (2016) 3
2. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2, 4
3. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Expert Systems with Applications **39**(12), 10873–10888 (2012) 2
4. Chessa, M., Noceti, N., Martini, C., Solari, F., Odone, F.: Design of assistive tools for the market. In: Leo, M., Farinella, G. (eds.) Assistive Computer Vision. Elsevier (2017) 6
5. Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., Rehg, J.M.: Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In: European Conference on Computer Vision (ECCV). pp. 383–398 (2018) 3
6. Debes, C., Merentitis, A., Sukhanov, S., Niessen, M., Frangiadakis, N., Bauer, A.: Monitoring activities of daily living in smart homes: Understanding human behavior. IEEE Signal Processing Magazine **33**(2), 81–94 (2016) 2
7. Dias, P., Medeiros, H., Odone, F.: Fine segmentation for activity of daily living analysis in a wide-angle multi-camera set-up. In: 5th Activity Monitoring by Multiple Distributed Sensing Workshop (AMMDS) in conjunction with British Machine Vision Conference (2017) 2
8. Dias, P.A., Malafronte, D., Medeiros, H., Odone, F.: Gaze estimation for assisted living environments. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 290–299 (2020) 2, 3, 4
9. Funes Mora, K.A., Monay, F., Odobez, J.M.: EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D

Cameras. In: ACM Symposium on Eye Tracking Research and Applications. ACM (Mar 2014) 3
10. Jayalekshmi, J., Mathew, T.: Facial expression recognition and emotion classification system for sentiment analysis. In: 2017 International Conference on Networks Advances in Computational Technologies (NetACT). pp. 1–8 (2017) 3
11. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: IEEE International Conference on Computer Vision (ICCV) (October 2019) 4
12. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Advances in Neural Information Processing Systems (NIPS). pp. 5574–5584 (2017) 4, 5
13. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 3
14. Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.M.: Computer vision for assistive technologies. Computer Vision and Image Understanding **154**, 1–15 (2017) 2
15. Lopes, A.T., de Aguiar, E., Souza, A.F.D., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. Pattern Recognition **61**, 610 – 628 (2017) 3
16. Martini, C., Barla, A., Odone, F., Verri, A., Rollandi, G.A., Pilotto, A.: Data-driven continuous assessment of frailty in older people. Frontiers in Digital Humanities **5**, 6 (2018) 6
17. Martini, C., Noceti, N., Chessa, M., Barla, A., Cella, A., Rollandi, G.A., Pilotto, A., Verri, A., Odone, F.: La visual computing approach for estimating the motility index in the frail elder. 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2018) 6
18. Nations, T.U.: World population prospects: the 2019 revision. https://population.un.org/wpp/ (2019), accessed: 2020-10-22 1
19. Pilotto, A., Ferrucci, L., Franceschi, M., D'Ambrosio, L.P., Scarcelli, C., Cascavilla, L., Paris, F., Placentino, G., Seripa, D., Dallapiccola, B., et al.: Development and validation of a multidimensional prognostic index for one-year mortality from comprehensive geriatric assessment in hospitalized older patients. Rejuvenation research **11**(1), 151–161 (2008) 2
20. Recasens, A., Khosla, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems (NIPS) (2015) 3
21. Varadarajan, J., Subramanian, R., Bulò, S.R., Ahuja, N., Lanz, O., Ricci, E.: Joint estimation of human pose and conversational groups from social scenes. International Journal of Computer Vision **126**(2), 410–429 (Apr 2018) 3
22. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Transactions on Image Processing **26**(9), 4193–4203 (Sept 2017) 3
23. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015) 3
24. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2299–2308. IEEE (2017) 3