Gaze Estimation for Assisted Living Environments

Philipe A. Dias[1] Damiano Malafronte[2] Henry Medeiros[1] Francesca Odone[3]
[1]Marquette University (EECE), USA [2]Italian Institute of Technology [3]University of Genova (MaLGa-DIBRIS), Italy
{philipe.ambroziodias,henry.medeiros}@marquette.edu damiano.malafronte@iit.it francesca.odone@unige.it

Abstract

Effective assisted living environments must be able to perform inferences on how their occupants interact with one another as well as with surrounding objects. To accomplish this goal using a vision-based automated approach, multiple tasks such as pose estimation, object segmentation and gaze estimation must be addressed. Gaze direction provides some of the strongest indications of how a person interacts with the environment. In this paper, we propose a simple neural network regressor that estimates the gaze direction of individuals in a multi-camera assisted living scenario, relying only on the relative positions of facial keypoints collected from a single pose estimation model. To handle cases of keypoint occlusion, our model exploits a novel confidence gated unit in its input layer. In addition to the gaze direction, our model also outputs an estimation of its own prediction uncertainty. Experimental results on a public benchmark demonstrate that our approach performs on par with a complex, dataset-specific baseline, while its uncertainty predictions are highly correlated to the actual angular error of corresponding estimations. Finally, experiments on images from a real assisted living environment demonstrate that our model has a higher suitability for its final application.

1. Introduction

The number of people aged 60 years or older is expected to nearly double by 2050 [27]. The future viability of medical care systems depends upon the adoption of new strategies to minimize the need for costly medical interventions, such as the development of technologies that maximize health status and quality of life in aging populations. Currently, clinicians use evaluation scales that incorporate mobility and Instrumented Activities of Daily Living (IADL) assessments (i.e., a person's ability to use a tool such as a telephone without assistance) [28] to determine the health status of elderly patients and to recommend habit changes.

Despite the potential of recent advances in many areas of computer vision, no current technology allows automatic and unobtrusive assessment of mobility and IADL over extended periods of time in long-term care facilities or patients' homes. Patient activity analysis to date has been limited to simplistic scenarios [9], which do not cover a wide range of relatively unconstrained and unpredictable situations.

Vision-based analysis of mobility and characterization of ADLs is challenging. As the examples in Figs. 1 and 2 illustrate, images acquired from assisted living environments cover a wide scene where multiple people can be performing different activities in a varied range of scenarios. Moreover, it encompasses multiple underlying complex tasks including: detection of subjects and objects of interest, identification of body joints for pose estimation, and estimation of the gaze of the subjects in the scene.

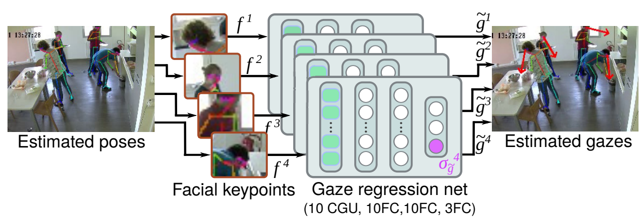

Figure 1. Overview of our apparent gaze estimation approach. The anatomical keypoints of all the persons present in the scene are detected using a pose estimation model [4]. The facial keypoints of each person are then provided as inputs to a neural network regressor that outputs estimations of their apparent gaze and its confidence on each prediction.

In this paper we focus on gaze estimation, which is a critical element to determine how humans interact with the surrounding environment. It has been applied to design human-computer interaction methods [23] and to analyze social interactions among multiple individuals [30]. For our application, in conjunction with object detection [10], gaze direction could define mutual relationships between objects and their users (e.g. the user is sitting on a chair with a book on his/her lap vs. sitting on a chair reading the book) and classify sim-

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

ple actions (e.g. mopping the floor, getting dressed, cooking food, eating/drinking).

The contributions of the present work can be summarized in three main points:

- *we propose an approach that relies solely on the relative positions of facial keypoints to estimate gaze direction.* As shown in Fig. 1, we extract these features using the off-the-shelf OpenPose model [4]. From the coordinates and confidence levels of the detected facial keypoints, our regression network estimates the apparent gaze of the corresponding subjects. From the perspective of the overall framework for ADL analysis, leveraging the facial keypoints is beneficial because a single feature extractor module can be used for two required tasks: pose estimation and gaze estimation. Code is available at coviss.org/codes
- the complexity of gaze estimation varies according to the scenario, such that the quality of predictions provided by a gaze regressor is expected to vary case-by-case. For this reason, *our model is designed and trained to provide an estimation of its uncertainty for each prediction of gaze direction.* To that end, we leverage concepts used by Bayesian neural networks for estimation of aleatoric uncertainty.
- in cases such as self-occlusion, one or more facial keypoints might not be detected, and OpenPose assigns a confidence of zero to the corresponding feature. To handle the absence of detections, *we introduce the concept of Confidence Gated Units (CGU)* to induce our model to disregard detections for which a low confidence level is provided.

2. Related Work

Ambient assisted living applications may benefit from computer vision methods in a variety of scenarios, including safety, well-being assessment, and human-machine interaction [5, 21]. Our aim is to monitor the overall health status of a patient by observing his/her behavior, or the way he/she interacts with the environment or with others. Summarized in Section 4.2 and detailed in [25, 6], the assisted living environment where our research takes place has been used for studies on automatic assessment of mobility information and frailty [24]. Related to our system are the methods presented in [4, 2, 36], which propose different smart systems designed to monitor human behavior and way of life incorporating computer vision elements.

Estimating the relative pose of subjects is crucial to perform high level tasks such as whole body action recognition and understanding the relationship between a person and the environment. Appearance-based pose estimation systems attempt to infer the positions of the body joints of the subjects present in a scene. Traditional methods relied on models fit to each of the individual subjects found in a given image frame [33, 3]. More recent approaches employ convolutional architectures [31, 4] to extract features from the entire scene, therefore making the whole process relatively independent of the number of subjects in the scene.

At a finer level, the analysis of human facial features may provide additional information [1] about well-being. For example, facial expression recognition [22, 32] can be used in sentiment analysis [15]. Facial analysis can also provide information on gaze direction, which is useful to better understand the interaction between a person and his/her surrounding environment [30]. Recent contributions in this area attempt to infer the orientation of a person's head by fitting a 3D face model to estimate both 2D [34] and 3D gaze information [35]. Other contemporary methods resort to different types of information, which include head detection, head orientation estimation, or contextual information about the surrounding environment [26]. In the context of human-computer interaction, the work in [20] employs an end-to-end architecture to track the eyes of a user in real-time using hand-held devices.

However, most works and datasets on inference of head orientation and gaze focus on specific scenarios, such as images containing close-up views of the subjects' heads [11, 34], with restricted background size and complexity. More similar to our scenario of interest, the GazeFollow dataset introduced in [29] contains more than 120k images of one or more individuals performing a variety of actions in relatively unconstrained scenarios. Together with the dataset, the authors introduce a two-pathway architecture that combines contextual cues with information about the position and appearance of the head of a subject to infer his/her gaze direction. A similar model is introduced in [8], with applicability extended to scenarios where the subject's gaze is directed somewhere outside the image.

Gaze estimation is a task with multiple possible levels of difficulty, which vary according to the scenario of observation. Even for humans, it is much easier to tell where someone is looking if a full-view of the subject's face is possible, while the task becomes much harder when the subject is facing backwards with respect to the observer's point of view. In modeling terms, this corresponds to heteroscedastic uncertainty, i.e., uncertainty that depends on the inputs to the model, such that some inputs are associated to more noisy outputs than others.

As explained in [17], conventional deep learning models do not provide estimations of uncertainties for their outputs. Classification models typically employ softmax in their last layer, such that prediction scores are normalized and do not necessarily represent uncertainty. For regression models, usually no information on prediction confidence is provided by the model. Bayesian deep learning approaches are becoming increasingly more popular as a way to understand and estimate uncertainty with deep learning models [12, 16, 18]. Under this paradigm, uncertainties are formalized as probability distributions over model parameters and/or outputs. For the estimation of heteroscedastic uncertainty in regressor models, the outputs can be modeled as corrupted with Gaussian random noise. Then, as we detail in Section 3.2, a customized loss function is sufficient for learning a regressor model that also predicts the variance of this noise as a function of the input [17], without need for uncertainty labels.

3. Proposed Approach

Our method estimates a person's apparent gaze direction according to the relative locations of his/her facial keypoints. As Fig. 1 indicates, we use OpenPose [4] 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.

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.

Aiming at a scale-invariant representation, for each person j in the scene we centralize all detected keypoints with respect to the head-centroid $h^j = [x^j_h, y^j_h]$, which is computed as the mean coordinates of all head keypoints detected in the scene. Then, the obtained relative coordinates are normalized based on the distance of the farthest keypoint to the centroid. In this way, for each detected person we form a feature vector $f \in \mathbb{R}^{15}$ by concatenating the relative vectors $\hat{p}^j_{k,s} = [\hat{x}^j_{k,s}, \hat{y}^j_{k,s}, c^j_{k,s}]$

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

3.1. Network architecture using gated units

Images acquired from assisted living environments can contain multiple people performing different activities, such that their apparent pose may vary significantly and self-occlusions frequently occur. For example, in lateral-views at least an ear is often occluded, while in back-views nose and eyes tend to be occluded. As consequence, an additional challenge intrinsic to this task is the representation of missing keypoints. In such cases, OpenPose outputs 0 for both the spatial coordinates $(x, y)^j_{k,s}$ and also the detection confidence value $c^j_{k,s}$. Since the spatial coordinates are centralized with respect to the head-centroid h^j as the $(0, 0)$ reference of the input space, a confidence score $c^j_{k,s} = 0$ plays a crucial role in indicating both the reliability and also the absence of a keypoint.

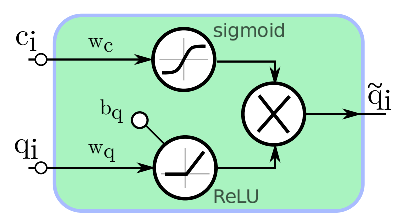

Figure 3. The proposed Confidence Gated Unit (CGU).

Inspired on the Gated Recurrent Units (GRUs) employed in recurrent neural networks [7], we propose a Confidence Gated Unit (CGU) 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. As depicted in Figure 3, we opt for a sigmoid unit without a bias parameter, to avoid potential biases towards models that disregard c_i when trained with unbalanced datasets where the majority of samples are detected with high confidence. Finally, the outputs of both units are then multiplied into an adjusted CGU output $\tilde{q}_i$.

For our application, a CGU is applied to each pair coordinate-confidence $(\hat{x}^j_{k,s}, c^j_{k,s})$ and $(\hat{y}^j_{k,s}, c^j_{k,s})$. To properly exploit the full range of the sigmoid function and thus reach output values near 0 for $c^j_{k,s} = 0$, we centralize and standardize the input confidence scores according to the corresponding dataset statistics. In this way, our proposed network for gaze regression has a combination of 10 CGUs as input layer.

Moreover, the variety of view-points from which a subject might be visible in the scene, occlusions and unusual poses lead to a vast range of scenarios where the difficulty of the gaze estimation varies significantly. Hence, we design a model that incorporates an uncertainty estimation method, which indicates its level of confidence for each prediction of gaze direction. From an application perspective, this additional information would allow us to refine the predictions by choosing between different cameras, models, or time instants.

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. In terms of architecture design, 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 $\sigma_{\tilde{g}}$.

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).

3.2. Training strategy

While all the weights composing the fully-connected layers are initialized as in [14], we empirically observed better results when initializing the parameters composing CGU units with *ones*. Since these compose only the input layer, initializing the weights as such does not represent a risk of gradient explosion as no further backpropagation has to be performed. Intuitively, our rationale is that the input coordinate features should not be strongly transformed in this first layer, as at this initial point no information from additional keypoints is accessible. Regarding regularization, we empirically observed better results without regularization in the input and output layers, while a L2 penalty of 10^{-4} is applied to parameters of both FC hidden layers.

Regardless of the dataset, we trained our network only on images where at least two facial keypoints are detected. Since we are interested on estimating direction of gaze to verify whether any object of interest is within a person's field of view, we opt for optimization and evaluations based on angular error. Thus, training was performed using a cosine similarity loss function that is adjusted based on [17] 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 $\sigma_{\tilde{g}}$ represents the model's uncertainty prediction. Our cost function is then given by

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

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

In terms of model optimization, all experiments were performed using the Adam [19] optimizer with early stopping based on angular error on the corresponding validation sets. Additional parameters such as batch size and learning rate varied according to the dataset. Hence, we describe them in detail in Section 4.

4. Experiments and Results

We evaluate our approach on two different datasets. The first is the GazeFollow dataset [29], on which we compare our method against two different baselines. The second dataset, which we refer to as the *MoDiPro* dataset, comprises images acquired from an actual discharge facility as detailed in Section 4.2.

4.1. Evaluation on the GazeFollow dataset

Dataset split and training details. The publicly available GazeFollow dataset contains more than 120k images, with corresponding annotations of the eye locations and the focus of attention point of specific subjects in the scene. We use the direction vectors connecting these two points to train and evaluate our regressors. In terms of angular distribution, about 53% of the samples composing the GazeFollow training set correspond to subjects whose gaze direction lies within the quadrant $[-90°, 0°]$ with respect to the horizontal axis. On the other hand, in only 29% of the cases their gaze direction is within the $[-180°, -90°]$ quadrant. To compensate such bias, we augment the number of samples in the later quadrant by mirroring with respect to the vertical-axis a subset of randomly selected samples from the most frequent quadrant. Finally, for training our model we split the training set into two subsets: 90% for *train*, and 10% for validation *val* subset. Training is performed using a learning rate 5×10^{-3}, batches of 1024 samples and early-stopping based on angular error on the *val* subset. The *test* set comprises 4782 images, with ten different annotations per image. For evaluation, we follow [29] and assess each model by computing the angular error between their predictions and the average annotation vector.

The GazeFollow dataset is structured such that for each image only the gaze from a specific subject must be assessed. For images containing multiple people, this requires identifying which detection provided by OpenPose corresponds to the subject of interest. To that end, we identify which detected subject has an estimated head-centroid that is the closest to the annotated eye-coordinates E_{GT} provided as ground-truth. To avoid mismatches when the correct subject is not detected but detections for other subjects on the scene are available, we impose that gaze is estimated only if E_{GT} falls within a radius of $1.5 \times \delta$ around the head-centroid, where δ corresponds to distance between the centroid and its farthest detected facial keypoint.

We compare our method against two baselines. The first, which we refer to as GEOM, relies solely on linear geometry to estimate gaze from the relative facial keypoints positions. Comparison against this baseline aims at evaluating if training a network is needed to approximate the regression $f \to g$, instead of directly approximating it by a set of simple equations. The second baseline is the model introduced together with the GazeFollow dataset in [29], which consists of a deep neural-network that combines a gaze pathway and a saliency pathway that are jointly trained for gaze estimation. We refer to this baseline as GF-MODEL.

Comparison against geometry-based baseline. We refer the reader to our Supplementary Material for a more detailed description of GEOM. This baseline is a simplification of the model introduced in [13] for face orientation estimation, which makes minimal assumptions about the facial structure [13] but additionally requires mouth keypoints and pre-defined model ratios. In short, let $\vec{s}$ represent the facial symmetry axis that is computed as the normal of the eye-axis. We estimate the facial normal $\vec{n}$ as a vector that is normal to $\vec{s}$ while intersecting $\vec{s}$ at the detected nose position. Then, the head pitch ω is estimated as the angle between the ear-centroid and the eye-centroid, i.e., the average coordinates of eyes and ears detections, respectively. Finally, gaze direction is estimated by rotating $\vec{n}$ with the estimated pitch ω.

The GEOM baseline requires the detection of the nose and at least one eye. Out of the 4782 images composing the GazeFollow *test* set, GEOM is thus restricted to a subset *Set*1 of 4258 images. As summarized on Tab. 1, results obtained on subset *Set*1 demonstrate that our model NET provide gaze estimations on average 23° more accurate than the ones obtained with the simpler baseline. Such a large improvement in performance suggests our network learns a more complex (possibly non-linear) relationship between keypoints and gaze direction. Examples available on Fig. 4 qualitatively illustrate how the predictions provided by our NET model (in green) are significantly better than the ones provided by the baseline GEOM (in red).

	Set1	*Set2*	*Full*
No. of images	4258	4671	4782
GEOM	42.63°	-	-
NET0	19.52°	25.70°	-
NET	19.41°	23.37°	-
GF-MODEL[29]	-	-	24°

Table 1. Comparison in terms of angular errors between our method and baselines on the GazeFollow test set.

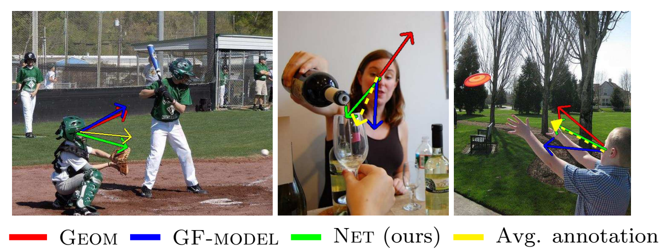

Figure 4. Examples of gaze direction estimations provided by the different models evaluated on GazeFollow.

Comparison against GazeFollow model. Since our network is trained on images where at least two facial keypoints are detected, we apply the same constraint for evaluation. In the test set, OpenPose detects at least two keypoints for a subset *Set*2 containing 97.7% of the 4782 images composing the full set.

The results of our evaluation are summarized in Tab. 1, while qualitative examples are provided in Fig. 4. As reported in [29], gaze predictions provided by the GF-MODEL present a mean angular error of 24° on the *test* set. Our NET model provides an mean angular error of 23.37° for 97.7% of these images, which strongly indicates that its performance is on par with GF-MODEL network despite relying solely on the relative position of 5 facial keypoints to predict gaze.

Impact of using Confidence Gated Units (CGU). To verify the benefits of applying our pro-

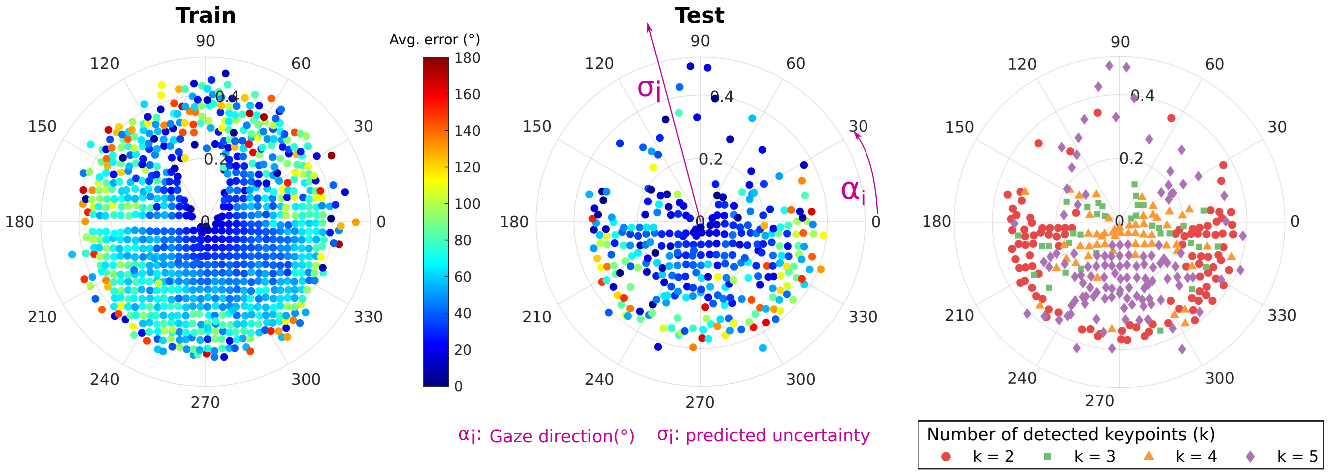

Figure 5. Distribution of gaze direction (α_i) and uncertainty predictions (σ_i) provided by our proposed model. *Left/center*: colormap depicts angular error of predictions. *Right*: colors represent the amount of keypoints detected by OpenPose for the corresponding samples. For better visualization, the samples are grouped into equally spaced bins.

posed CGU blocks to handle absent keypoint detections, i.e., keypoints with 0 confidence score, we evaluated the performance of our model with and without feeding the confidence scores as inputs. We refer to the latter case as the NET0, where the CGU blocks composing the input layer are replaced by simple ReLU units initialized in the same way as described in Section 3.2. Results summarized in Tab. 1 indicate an error decrease of 2.3° when providing confidence scores to an input layer composed of CGUs. In addition to experiments summarized in Tab. 1, we also evaluated a model where the CGU units are replaced by simple additional ReLU units to handle confidence scores. For the 1536 images where OpenPose detects less than 4 facial keypoints, a significant decrease on angular error is observed when using CGU units: 30.1° mean error, in comparison to 30.9° provided by the model with solely ReLU based input layer.

Quality of uncertainty estimations. In addition to the overall mean angular error, we also evaluate how accurate are the uncertainty estimations provided by our NET model for its gaze direction predictions. As depicted in Fig. 6, significantly lower angular errors are observed for gaze predictions accompanied by low uncertainty network predictions. Uncertainties lower than 0.1 are observed for 80% of the *test* set, a subset for which the gaze estimations provided by our NET model are on average off by only 16.5°.

Moreover, the high correlation between uncertainty predictions and angular error ($\rho = 0.56$) is clearly depicted by the plots provided in Fig. 5. For each sample in these plots, the radial distance corresponds to its predicted uncertainty σ_i, while the angle corresponds to predicted direction of gaze $\tilde{g}$, i.e $\alpha_i = tan^{-1}(-\tilde{g}_y/\tilde{g}_x)$. For both *train* and *test* sets, the associated colormap shows that lower errors (in dark blue) are observed for predictions with lower uncertainty, with increasingly higher errors (green to red) as the uncertainty increases (farther from the center).

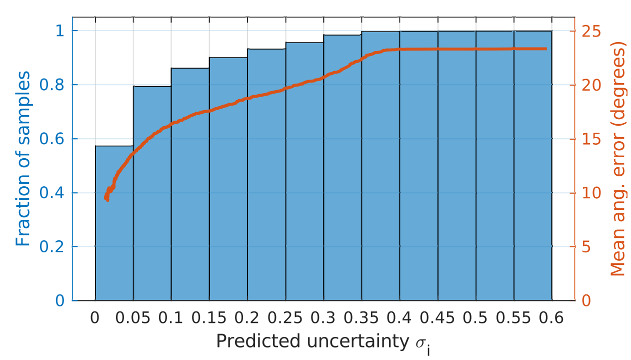

Figure 6. Cumulative mean angular error according to uncertainty predicted by our model for each sample.

Performance according to keypoint occlusions. Furthermore, the central and the right-most scatter plots in Fig. 5 also allow an analysis on how the performance of our model and its uncertainty predictions vary according to specific scenarios. For most cases, the number of detected keypoints (k) indicates specific scenarios: $k = 2$ is mostly related to back-views, where nose and two other keypoints (both eyes or a pair eye-ear) are missing; $k = 3$ and $k = 4$ are mostly lateral-views; $k = 5$ are frontal-views, where all keypoints are visible. Since images are $2D$ projections from the environment, back- and frontal-views are the ones more affected by the information loss implicit in the image formation process, while for lateral-views estimation of gaze direction tends to be easier.

An analysis of the scatter plots demonstrates that the predictions provided by our model reflect these expected behaviors. For samples with $k = 2$ (back-view), both uncertainty predictions and angular error tend to be higher, while for most cases of $k = 3$ and $k = 4$ the predictions are associated with lower uncertainty and higher angular accuracy. Predictions for $k = 5$ are spread, indicating that the model's uncertainty predictions are not just defined by the amount of available keypoints but also reflect the intrinsic uncertainty of determining the head orientation from frontal views.

4.2. Results on the assisted living dataset

This work is part of a project that focuses on elderly patients with partial autonomy but in need of moderate assistance, possibly in a post-hospitalization stage. Thus, it is critical to evaluate the performance of our gaze estimation model on data from real assisted living environments. To that end, we also evaluate our approach 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 for a few days. The facility is a fully-equipped apartment where patients may be monitored by various sensors, including localization systems, RGB-D, and two conventional video cameras, arranged as shown in Fig. 2.

Dataset split and training details. We compiled a dataset, which we call *MoDiPro*, consisting of 1,060 video frames collected from the two video cameras. For $CAM1$, 530 frames were sampled from 46 different video sequences; for $CAM2$, 530 frames were sampled from 27 different video sequences. To limit storage while discarding minimal temporal information, the resolution of the acquired frames was limited to 480×270 pixels, at 25 fps. In most frames multiple subjects are simultaneously visible, with a total of 22 subjects performing different activities.

As exemplified also in Fig. 7, cameras $CAM1$ and $CAM2$ cover different parts of the environment. Images acquired with $CAM2$ present significant distortion, which increases the complexity of the task. We randomly split the available sets of images into camera-specific training, validation and test subsets. Since frames composing the same video sequence can be highly correlated, we opt for a stratified strategy where video sequences are sampled. That is, all frames available from a certain video sequence are assigned to either *train*, *val* or *test* subsets. Aiming at an evaluation that covers a wide variety of scenes, the proportions

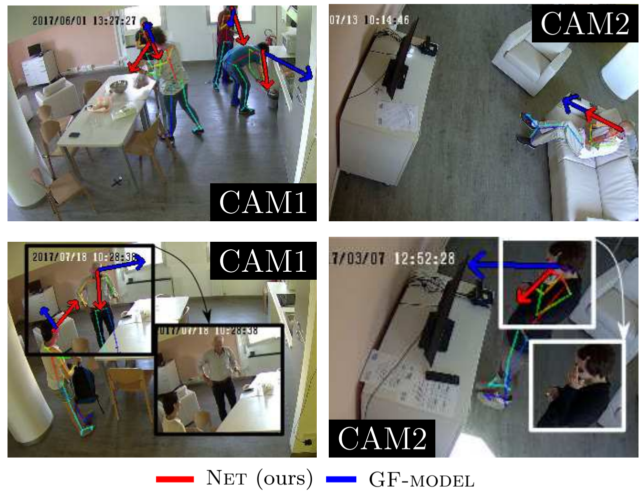

Figure 7. Examples of results for our gaze direction estimation approach in the *MoDiPro* dataset.

chosen in terms of total number of frames are: 50% for training, 20% for validation, 30% for testing. Fine-tuning experiments are performed using learning rates 1×10^{-5}, while 1×10^{-4} is adopted when training models only on *MoDiPro* images. Batches with 64 samples are used, with early-stopping based on angular error on the *val* subset. Moreover, all results reported on Tab. 2 and discussed below correspond to average values obtained after train/test on 3 different random splits.

To assess the cross-view performance of our method, we train our NET model with 7 different combinations of images from the *MoDiPro* and GazeFollow datasets. As summarized in Tab. 2, models NET#0-2 are trained in $CAM1$-only, $CAM2$-only, and both *MoDiPro* 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 three possible sets of *MoDiPro* frames.

	TRAIN			TEST		
Model	GF	*Cam1*	*Cam2*	*Cam1*	*Cam2*	*Mean*
NET#0		✓		16.16°	39.12°	-
NET#1			✓	29.56°	26.37°	-
NET#2		✓	✓	18.52°	23.02°	20.94°
NET#3	✓			27.64°	26.98°	27.31°
NET#4	✓	✓		16.17°	27.36°	-
NET#5	✓		✓	27.56°	24.01°	-
NET#6	✓	✓	✓	17.82°	20.15°	19.05°
GF-MODEL	✓			43.49°	60.82°	52.15°

Table 2. Performance of our method on the *MoDiPro* dataset for different combinations of training/testing sets.

Performance according to camera view. 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 composing unseen subsets. Trained on both $CAM1$ and $CAM2$, the model NET#2 demonstrates a more consistent performance across views. In comparison with the camera specific models, a 3° lower angular error on $CAM2$ is obtained at cost of only 1.4° error increase on $CAM1$.

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 17.82° for $CAM1$ and 20.15° for $CAM2$. This corresponds to an overall average error 1.9° lower than the model NET#2 not pre-trained on GF, while more than 7° better than the model NET#3 trained solely on GF. In terms of camera-specific performance, for $CAM1$ optimal performances with error below 17° are obtained when not training on $CAM2$. On the other hand, predictions for $CAM2$ are significantly better when training is performed using additional $CAM1$ and/or GazeFollow images. We hypothesize the distortions characteristic of $CAM2$ images easily lead to overfitting, thus the advantage of training on additional sets of images. As a final remark we may notice that overall NET#6 provides the best an most stable result across the two views.

Comparison against GF-model. Finally, we compare the predictions provided by our NET models to the ones obtained by the publicly available version of GF-MODEL[1]. As summarized in Tab. 2, gaze predictions provided by GF-MODEL on the *MoDiPro* dataset are remarkably worse in terms of angular error than the ones predicted by any of our NET#0-6 models, including the NET#3 also trained only on GF images.

Closer inspection of GF-MODEL predictions suggests two disadvantages of this model with respect to ours when predicting gaze on images from real assisted living environments: i) sensitivity to scale; ii) bias towards salient objects. Images composing the GazeFollow typically contain a close-view of the subject of interest, such that only a small surrounding area is covered by the camera-view. In contrast, images from assisted living facilities such as the ones in the *MoDiPro* dataset contain subjects covering a much smaller region of the scene, i.e., they are smaller in terms of pixel area. Our NET model profits from the adopted representation of keypoints, with coordinates centered at the head-centroid and normalized based on the largest distance between centroid and detected keypoints. Moreover, visual inspection of GF-MODEL predictions reveals examples such as the two bottom ones in Fig. 7: in the left, while our model correctly indicates that the subjects look at each other, GF-MODEL is misled by the saliency of the TV and possibly the window; in the right, the saliency of the TV again misguides the GF-MODEL, while our model properly indicates that the person is looking at the object she is holding.

[1]This version provides 25.8° mean angular error on the GazeFollow test set, in comparison to the 24° reported in [29]

4.3. Runtime Analysis

Our network requires on average $0.85ms$ per call on a NVIDIA GeForce 970M, with one feedforward execution per person. The overall runtime is thus dominated by OpenPose, which requires $77ms$ on COCO images with a NVIDIA GeForce 1080 Ti (as reported in [4]).

5. Conclusion

This paper presents a gaze estimation method that exploits solely facial keypoints detected by a pose estimation model. Our end goal is to assist clinicians assessing the health status of individuals in an assisted living environment, providing them with automatic reports of patients' mobility and IADL patterns. 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 the GazeFollow dataset demonstrate that our method estimates gaze with accuracy comparable to a complex task-specific baseline, without relying on any image features except the relative positions of facial keypoints. In contrast to conventional regression methods, our proposed model also provides estimations of uncertainty of its own predictions, with results demonstrating a high correlation between predicted uncertainties and actual gaze angular errors. Moreover, analysis of performance according to the number of detected keypoints indicates that the proposed Confidence Gate Units improve the model's performance for cases of partial absence of features.

Finally, evaluation on frames collected from a real assisted living facility demonstrate that our model has a higher suitability for IADL analysis in realistic scenarios, where images cover wider areas and subjects are visible at different scales and poses.

Acknowledgements Part of this work has been carried out at the Machine Learning Genoa (MaLGa) center, Università di Genova (IT) thanks to the students mobility supported by Erasmus+ K107. We acknowledge the NVIDIA Corporation for the donation of a GPU used for this research.

References

[1] T. Baltrušaitis, P. Robinson, and L. Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016. 2

[2] V. Bathrinarayanan, B. Fosty, A. Konig, R. Romdhane, M. Thonnat, F. Bremond, et al. Evaluation of a monitoring system for event recognition of older people. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 165–170, 2013. 2

[3] M. Brubaker, D. Fleet, and A. Hertzmann. Physics-based human pose tracking. In *NIPS Workshop on Evaluation of Articulated Human Motion and Pose Estimation*, 2006. 2

[4] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 8

[5] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012. 2

[6] M. Chessa, N. Noceti, C. Martini, F. Solari, and F. Odone. Design of assistive tools for the market. In M. Leo and G. Farinella, editors, *Assistive Computer Vision*. Elsevier, 2017. 2

[7] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 3

[8] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *European Conference on Computer Vision (ECCV)*, pages 383–398, 2018. 3

[9] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer. Monitoring activities of daily living in smart homes: Understanding human behavior. *IEEE Signal Processing Magazine*, 33(2):81–94, 2016. 1

[10] P. Dias, H. Medeiros, and F. Odone. 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. 1

[11] K. A. Funes Mora, F. Monay, and J.-M. Odobez. 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

[12] Y. Gal. *Uncertainty in deep learning*. PhD thesis. 3

[13] A. Gee and R. Cipolla. Determining the gaze of faces in images. *Image and Vision Computing*, 12(10):639–647, 1994. 5

[14] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034. IEEE Computer Society, 2015. 4

[15] J. Jayalekshmi and T. Mathew. Facial expression recognition and emotion classification system for sentiment analysis. In *2017 International Conference on Networks Advances in Computational Technologies (NetACT)*, pages 1–8, 2017. 2

[16] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 3

[17] A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems (NIPS)*, pages 5574–5584, 2017. 3, 4

[18] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 3

[19] D. Kinga and J. B. Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[20] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[21] M. Leo, G. Medioni, M. Trivedi, T. Kanade, and G. M. Farinella. Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154:1–15, 2017. 2

[22] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610 – 628, 2017. 2

[23] P. Majaranta and A. Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, pages 39–65. Springer, 2014. 1

[24] C. Martini, A. Barla, F. Odone, A. Verri, G. A. Rollandi, and A. Pilotto. Data-driven continuous assessment of frailty in older people. *Frontiers in Digital Humanities*, 5:6, 2018. 2

[25] C. Martini, N. Noceti, M. Chessa, A. Barla, A. Cella, G. A. Rollandi, A. Pilotto, A. Verri, and F. Odone. 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. 2

[26] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009. 2

[27] T. U. Nations. World population ageing. http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Report.pdf, 2017. Accessed: 2018-09-03. 1

[28] A. Pilotto, L. Ferrucci, M. Franceschi, L. P. D'Ambrosio, C. Scarcelli, L. Cascavilla, F. Paris, G. Placentino, D. Seripa, B. Dallapiccola, 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. 1

[29] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 3, 4, 5, 8

[30] J. Varadarajan, R. Subramanian, S. R. Bulò, N. Ahuja, O. Lanz, and E. Ricci. Joint estimation of human pose and conversational groups from social scenes. *International Journal of Computer Vision*, 126(2):410–429, Apr 2018. 1, 2

[31] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[32] K. Zhang, Y. Huang, Y. Du, and L. Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203, Sept 2017. 2

[33] X. Zhang, C. Li, X. Tong, W. Hu, S. Maybank, and Y. Zhang. Efficient human pose estimation via parsing a tree structure based human model. In *IEEE International Conference on Computer Vision (ICCV)*, 2009. 2

[34] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2, 3

[35] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308. IEEE, 2017. 2

[36] N. Zouba, F. Bremond, and M. Thonnat. An activity monitoring system for real elderly at home: Validation study. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 278–285. IEEE, 2010. 2