Computer Vision and Image Understanding 243 (2024) 103999

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Head pose estimation with uncertainty and an application to dyadic interaction detection

Federico Figari Tomenotti, Nicoletta Noceti *, Francesca Odone

MaLGa-DIBRIS, Università degli Studi di Genova, via Dodecaneso 35, Genova, 16146, Italy

ARTICLE INFO

Communicated by Liming Chen

MSC:
68T45
68T05

Keywords:
Head pose estimation
Multi-task regression
Neural networks
Heteroscedastic uncertainty
Dyadic interaction detection

ABSTRACT

Determining the visual focus of attention of people in a scene is a fundamental cue to understand social interactions from videos. Gaze direction is ideal for determining eye contact, a basic cue of non-verbal communication, but it is not always easy to recognize. Head direction is a well-known proxy of gaze direction, more robust to the variability of the scene, thus offering a valuable alternative. In this work, we consider HHP-net, a method for estimating the head direction from single frames based on a heteroscedastic neural network to estimate people's head pose from a minimal set of head key points. We formulate the problem as a multi-task regression, to predict the pose as a triplet of Euler angles from the output of a 2D pose estimator. HHP-net also provides a measure of the aleatoric heteroscedastic uncertainties associated with the angles, through an ad-hoc loss function we introduce. In a thorough experimental analysis, we show that our model is efficient and effective compared with the state of the art, with only ∼2 degrees of degradation in the worst case counterbalanced by a space occupation ∼12 times smaller. We also show the beneficial effects of uncertainty on interpretability. Finally, we discuss the robustness of our method to input variability, showing that it can be seen as a plug-in to different pose estimators. As a proof-of-concept, we address social interaction analysis, with an algorithm to detect dyadic interactions in images.

1. Introduction

Analysing the social attitude of humans is paramount for a number of application domains – ranging from assisted living (Grossi et al., 2020) to robotics (Saunderson and Nejat, 2019), health (Alghamdi et al., 2023), sports (Bagautdinov et al., 2017), education and arts (Schiavio et al., 2019) – where observing the interactions flow in a non-invasive way is fundamental to guarantee the naturalness of the experience, and thus the meaningfulness of the analysis. With these motivations, video-based approaches emerged in the last years as effective tools to address the task.

A trigger to social interaction analysis is the understanding of human attention, an important cue of non-verbal communication that can also deepen the understanding of human behaviours (see for instance (Cristani et al., 2013; Fan et al., 2018)). In video-based analysis, human attention is usually encoded by the gaze direction (Wang et al., 2021b), which unfortunately is very difficult to recognize if the subject is placed at a certain distance from the camera. It is known, however, that the eyeball orientation can differ only by $\pm 35^{\circ}$ degrees from the head orientation (Stahl, 1999). Hence, a suitable proxy for the gaze is the head direction (Madrigal and Lerasle, 2020; Dias et al., 2020; Grossi et al., 2020), encoded as the rotation of human heads with respect to a reference (frontal) pose.

Recent approaches to Head Pose Estimation (HPE) may vary on the input — the entire image or the face region provided by a face detector (Liu et al., 2023, 2021a; Hong et al., 2021; Liu et al., 2021b), or 3D models and point clouds (Zhu et al., 2019; Ruan et al., 2021; Xu et al., 2022) – and in their objectives – as some of them tackle multiple tasks concurrently (see e.g. Bulat and Tzimiropoulos (2017), Kumar et al. (2017), Ranjan et al. (2019), Xia et al. (2022)). Most of them provide very good estimates, at least in relatively uncluttered scenarios, to the price of significant complexity and limited modularity. For this, they may not be an ideal choice for real-time computation.

In this work, we discuss how to estimate the head pose of one or more subjects in images effectively and efficiently, proposing a very light and modular neural network that estimates the head direction as a triplet of yaw–pitch–roll angles. As an input, we consider key points obtained by 2D human pose estimators (Cao et al., 2019; Martinez et al., 2019; Duan et al., 2019; Lugaresi et al., 2019). Their widespread diffusion for a variety of tasks makes them an ideal input for our method so that it can be easily integrated into already existing pipelines and applications, without a significant further load in the computation. Specifically, we estimate the head direction only relying on the pose estimation key points, with no need for additional input (Fig. 1).

We formulate the problem as a multi-task regression and discuss the use of HHP-net (Heteroscedastic Head Pose network) to estimate

* Corresponding author.
E-mail address: nicoletta.noceti@unige.it (N. Noceti).

https://doi.org/10.1016/j.cviu.2024.103999
Received 8 September 2023; Received in revised form 15 January 2024; Accepted 18 March 2024
Available online 28 March 2024

Fig. 1. The pipeline of the proposed method: pose detection is used to extract the pose of each human in the image (left), and then the facial key points are injected into a neural network that retrieves the 3D head pose for each person. Our method provides the 3D head pose as a triplet of angles yaw, pitch and roll (centre) and a measure of uncertainty associated with each angle in the triplet (right). A higher uncertainty can be visually represented as a larger visual cone centred on the subject. Notice how uncertainty relates to the subject's visibility.

the head pose expressed in Euler angles (yaw-pitch-roll). Thanks to the adoption of an appropriately designed loss function, we also associate an uncertainty value – learned from the data – with each output angle. Specifically, we estimate aleatoric heteroscedastic uncertainty, that is uncertainty due to data and varying on different inputs (Kendall and Gal, 2017). The estimated uncertainty provides an additional cue that may help the interpretation of the network output.

Our method, with negligible additional effort in terms of space and time resources with respect to a 2D human pose estimator, allows us to extract precise (in line with more complex state-of-the-art algorithms) head poses. Indeed, as a positive side effect of operating on a very compact input, the architecture we propose is small in size (it occupies less than 0.5 MB) – with the potential to run on devices with limited space resources – and performs in real-time (at about 100 fps); for these reasons it may also be easily adapted to portable devices, considering appropriate pose estimators are now available (Choi et al., 2021; Lugaresi et al., 2019). Our approach can be seen as a *plug-in to any given pose estimator*, as we will experimentally assess associating it with different 2D pose estimators, specifically, OpenPose (Cao et al., 2019), CenterNet (Duan et al., 2019), MediaPipe (Lugaresi et al., 2019).

We report a thorough experimental analysis based on three reference benchmarks (Fanelli et al., 2011; Yin et al., 2017; Sagonas et al., 2013), with ablation studies and comparisons with existing approaches. We empirically show that our method is comparable or better than other approaches in terms of computational load and accuracy of the estimates, providing an ideal trade-off between them. Moreover, to show the potential use of our method, we consider its application to social interaction analysis, where the ability to understand the focus of attention of people is a core element. We focus in particular on dyadic interactions and show that the head poses provided by our method can be profitably employed to detect Looking-At-Each-Other (LAEO) events (Marin-Jimenez et al., 2019; Marín-Jiménez et al., 2020) in images and videos with a simple and lightweight approach. Here the uncertainty estimates contribute to obtaining improved performance.

This paper extends an early publication (Cantarini et al., 2022), with the following main contributions:

- From a methodological viewpoint, we thoroughly discuss a modular pipeline for head pose estimation based on a multi-task regression loss where the uncertainties act as weights of each sub-loss responsible for the estimation of each angle. We provide a theoretical derivation of loss and uncertainties and show that the latter is tightly related to the estimation errors, indicating its importance for the interpretability of the results.
- From the experimental viewpoint, we provide a thorough assessment of the pipeline including the impact of choosing different loss functions and ablation studies on different publicly available datasets. Moreover, we present an assessment with a variety of well-known 2D human pose estimators. In this way, we show the robustness of the pipeline against a variety of usages and challenging conditions.
- We show that our method has the potential to be used on devices with limited resource availability and has a space occupancy ∼12 times smaller than state-of-the-art while providing comparable results (in the worst case with about 2 degrees of degradation). This may widen the range of its applicability.
- We apply the pipeline to a problem of dyadic interaction analysis, considering the LAEO events detection. We discuss the use of a simple LAEO measure, showing how it smoothly changes over time, allowing the detection of the event and possibly enabling its anticipation.

We provide code for the network[1] and a demo on Hugginface Spaces.[2]

The remainder of the paper is organized as follows: Section 2 covers related works on head pose estimation in images and uncertainty evaluation; Section 3 presents the proposed HHP-Net architecture, Section 4 focuses on the method assessment and the comparison with state-of-the-art. Section 5 is about the application of our method to mutual interaction, while Section 6 is left to a final discussion.

2. Related works

2.1. Human pose estimation

Human Pose Estimation, aiming to extract the semantics and topology of the human body from images, finds applications in several domains, including human motion analysis in sports (Colyer et al., 2018) and medicine (Moro et al., 2022), action recognition (Luvizon et al., 2020; Shi et al., 2019), human–machine and social interaction analysis (Luvizon et al., 2020; Song et al., 2021), biometric recognition (Barra et al., 2020) and driver attention detection (Campbell, 2012; Wang et al., 2021a).

Comprehensive studies (Gong et al., 2016; Zheng et al., 2023; Wang et al., 2021a) analyse differences in approaches like 2D (Cao et al., 2019; Duan et al., 2019; Bazarevsky et al., 2020) versus 3D (Yu et al., 2017; Zhou et al., 2021), handcrafted features versus deep learning, and single-person versus multi-person scenarios. The focus here is on

[1] https://github.com/Malga-Vision/HHP-Net

[2] https://huggingface.co/spaces/FedeFT/Head_Pose_Estimation_and_LAEO_computation

2D pose estimation from monocular images, where we may identify bottom-up algorithms like Openpose (Cao et al., 2019) and top-down approaches like AlphaPose (Fang et al., 2022). Regarding computational performances, one of the fastest and most recent methods is in the MediaPipe framework (Lugaresi et al., 2019), employing a combined heatmap, offset, and regression approach (Bazarevsky et al., 2020). Distinctions in algorithms include the number of key points and topology, with 'standards' like COCO format (Lin et al., 2014) with 17 key points, OpenPose (Cao et al., 2019) with 25 key points, and MediaPipe encompassing 33 key points. More complex approaches aim to incorporate spatial and appearance consistency (Yang et al., 2016) and video-based methods (Luo et al., 2018), but are out of the scope of this work.

2.2. Head pose estimation

Head pose estimation has been addressed by a number of relatively recent methodologies (Cao et al., 2021a; Zhou and Gregson, 2020; Madrigal and Lerasle, 2020; Zhang et al., 2020; Liu et al., 2021a; Dhingra, 2022; Liu et al., 2022), with classical applications to Human–Machine Interaction or to social interaction analysis. The more recent and comprehensive survey on the topic is probably (Abate et al., 2022).

Some methods use additional information such as depth (Fanelli et al., 2011; Mukherjee and Robertson, 2015; Hong et al., 2018) or time (Gu et al., 2017), but also points clouds as in Xu et al. (2022) or infrared as in Liu et al. (2021b). In the field of driver-assistive technology and safety, infrared cameras are used to estimate the head pose (Ju et al., 2022), but with ad-hoc solutions due to the camera setup (the camera is commonly located in the centre of the rear-view mirror of the car). Differently from these approaches which use different sensors, in our work, we only employ RGB images, which guarantee less expensive and more general applications. In alternatives methodologies, the head pose is derived by fitting an image onto a 3D face model, or on some approximation of it. An estimation of a 3D model is first presented in Fanelli et al. (2013), while more recently deep learning-based methods have been presented: such as 3DDFA (Zhu et al., 2019), a CNN able to fit a 3D model to an RGB image, or SADRNet (Ruan et al., 2021) which specifically tackles the problem of face occlusions. Furthermore, FAN (Bulat and Tzimiropoulos, 2017) is a state-of-the-art facial landmark detection method, that performs also face alignment. These approaches propose complex computational pipelines and have demonstrated the potential to yield notably accurate results.

One of the most recent challenges is in estimating pose directly from individual 2D images. In this respect, we start by mentioning a different but related task of estimating the 2D gaze: GazeFollow (Recasens et al., 2015) is a two-pathway CNN architecture that estimates the apparent direction of the human gaze and the object being observed; it combines saliency maps of the whole image with the position of subjects' head to obtain a pose prediction. A very efficient strategy to estimate the apparent direction of gaze is proposed in Dias et al. (2020).

Moving to 3D head pose estimation, nowadays it is mostly obtained by deep learning architectures that start from the output of face detectors, often implemented as a Convolutional Neural Network (CNN). Besides recent few exceptions such as Bisogni et al. (2021), the literature presents several works starting from images: Shao et al. (2019) propose an adjustment of the ROI obtained by face detection (it incorporates an offset around the face) and a combined regression and classification loss. HopeNet is a regression method with ResNet and a joint MSE and cross-entropy loss (Ruiz et al., 2018). LwPosr (Dhingra, 2022) introduces a lightweight architecture based on a mixture of depthwise separable convolutional and transformer encoder layers, structured in two streams and three stages to provide fine-grained regression. Transformers have also been used in Liu et al. (2023) which specifically addresses challenges related to occlusions, illuminations, and extreme orientations. FSA-net (Yang et al., 2019) is a two-stream multi-dimensional regression network able to provide accurate fine-grained estimations. Rahmaniar et al. (2022) presents an approach based on a combination of coarse and fine feature map classification to train a multi-loss CNN architecture. CNNs are also used in Hsu et al. (2018), where an L2 regression loss and an ordinal regression loss are jointly employed, and in Albiero et al. (2021), which regresses 6DoF pose in a Faster R-CNN–based framework.

In contrast to these methods, our 3D head pose estimation pipeline relies exclusively on the output of a human pose detector, similar to the strategy proposed in Dias et al. (2020). This allows us to design a very lightweight architecture, capable of attaining accurate results, acting sequentially to a 2D pose estimator. Our approach also differs from multi-task approaches such as KEPLER (Kumar et al., 2017) – predicting facial key points and pose–, Hyperface (Ranjan et al., 2019) – simultaneously performing face and landmark detection, pose estimation and gender recognition –, or the method proposed in Xia et al. (2022) – jointly learning Head Pose Estimation, face alignment and face tracking. Concerning these methodologies, our approach prioritizes modularity, offering the flexibility to seamlessly integrate with various pose estimators.

As recently observed in Zhou and Gregson (2020), head pose estimation is intrinsically harder on certain viewpoints. Starting from this observation, in Ruiz et al. (2018) an approach for improving on lateral views is proposed, to obtain wide-range head pose estimation. Instead, our work follows the observation in Dias et al. (2020): certain viewpoints are associated with different levels of uncertainty, creating a large discrepancy in accuracy. This can be formalized with the concept of aleatoric heteroscedastic uncertainty (Kendall and Gal, 2017), which depends on the inputs and may be estimated from data. Conventional deep learning methods cannot estimate the uncertainty of their inputs, consequently, Bayesian deep learning is becoming very popular as an effective approach to address this limitation. In our method we propose a multi-task approach where a task is associated with one of the three pose angles, extending (Kendall and Gal, 2017) to the multi-loss case. Indeed (Cipolla et al., 2018) reports a loss with homoscedastic uncertainty, also called task-dependent uncertainty, that is constant across different inputs. In this way, their model can learn the weight of each task.

2.3. Social interaction analysis

A specific case within the realm of social interaction analysis is that of dyadic interactions involving two individuals. In this context, the initial phase involves identifying pairs of individuals engaged in interaction, commonly achieved through the detection of mutual gaze or "Looking At Each Other" (LAEO) as referenced in the literature (Marín-Jiménez et al., 2014; Fan et al., 2019). Early work on LAEO detection in images utilized a Gaussian process predicting yaw and pitch, generating a LAEO score per frame (Marín-Jiménez et al., 2014). More recently, LAEO-Net and LAEO-Net++, introduced by the same authors, proposed a CNN-based extension to estimate LAEO over temporal windows (Marin-Jimenez et al., 2019; Marín-Jiménez et al., 2020). Recent advancements include an end-to-end pipeline based on Transformers for mutual gaze detection (Guo et al., 2022) and a late fusion approach combining head and scene features encoded with variants of a ResNet (Chang et al., 2023).

Multimodal approaches have also been explored for mutual gaze detection. In Trabelsi et al. (2017), authors leverage RGB data with depth information, while Kukleva et al. (2020) presents an approach integrating vision and text to jointly address interaction detection and long-term relationship prediction. Joint learning of LAEO and 3D gaze estimation is discussed in Doosti et al. (2021).

In comparison to existing methods, our LAEO method relies solely on head orientation as a proxy for mutual gaze, which introduces limitations concerning the usage of the proper gaze. However, it offers notable advantages, such as effectiveness even when subjects are

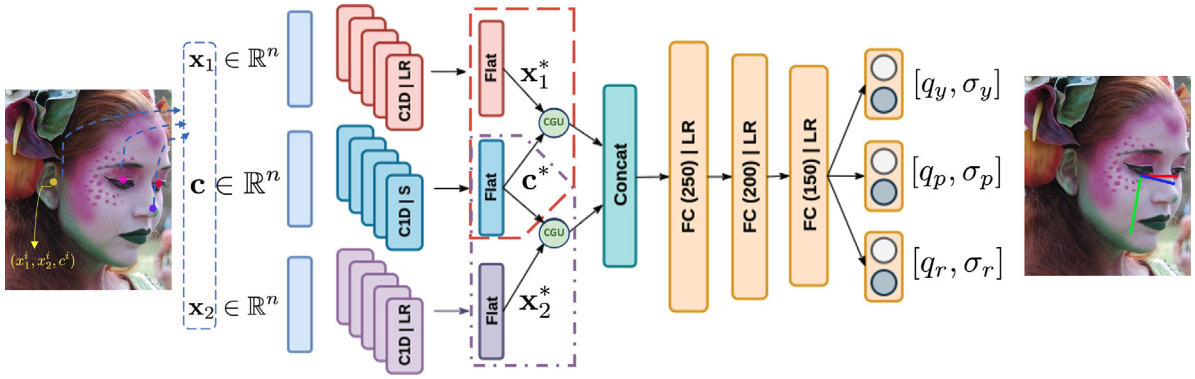

Fig. 2. A visual representation of our architecture. A set of key point locations with associated confidences $\{x_1^i, x_2^i, c^i\}_{i=1}^n$ is provided in input to the network, and processed with 1D convolutional layers. With a CGU we combine the intermediate outputs, that is then provided to the second part of the networks, composed of 3 FC layers to produce the final output, i.e. yaw, pitch and roll estimates with associated uncertainties.

positioned at a distance from the camera, expanding its applicability to diverse real-world scenarios. Moreover, our presented approach is engineered to be effective but simple enough to be placed in cascade to our HHP-net without compromising its speed and lightweight nature. Our HHP-net is fast and very lightweight, so the usage of a deep network as in Guo et al. (2022) was not an option for us.

3. Head pose estimation method: HHP-net

The starting point of our approach is the output of a 2D pose estimator providing a set of key points describing the pose of a human body in an image. These detectors commonly provide also a confidence measure on the keypoint location estimate, which represents an additional source of knowledge that can be injected into our approach.

We model the estimation of the head orientation as a multi-task regression problem, where a Neural Network predicts the 3D vector of the head orientation with angles in Euler notation (*yaw, pitch* and *roll*). The input is formed by a set of n semantic keypoints located on the image plane: $\{(x_1^i, x_2^i, c^i)\}_{i=1}^n$, with x_1^i the horizontal and x_2^i vertical coordinates and c_i the confidence of the ith keypoint. Coordinates are centred and normalized according to, respectively, their centroid and maximum value; c_i is provided in the range $[0, 1]$. The value of confidence is particularly important, as it encodes missing points ($c = 0$) and low confidence points. These situations may frequently occur in real-world applications, in particular in human–human interaction, because of occlusions, self-occlusions or lateral poses.

3.1. The architecture

Fig. 2 provides a sketch of our architecture. We formalize the input of the network as a triplet of vectors $\mathbf{x}_1 = [x_1^1, \dots, x_1^n]$, $\mathbf{x}_2 = [x_2^1, \dots, x_2^n]$ and $\mathbf{c} = [c^1, \dots, c^n]$ incorporating positions and confidence of n key points describing a face. The input vectors are first processed in independent streams, with 5-channels 1D convolutions, followed by a Leaky ReLU for $\mathbf{x}_1$ and $\mathbf{x}_2$ – to avoid vanishing gradient issues – and sigmoid activation for the confidence vector $\mathbf{c}$ – to smoothly control the impact of different confidence values.

The outputs of the 1D convolutional layers are flattened to obtain $\mathbf{x}_1^*$, $\mathbf{x}_2^*$ and $\mathbf{c}^*$ from the independent streams. They are then combined, using an element-wise multiplication to obtain two vectors $\mathbf{v}_1 = \mathbf{x}_1^* \otimes \mathbf{c}^*$ and $\mathbf{v}_2 = \mathbf{x}_2^* \otimes \mathbf{c}^*$, following the logic of the Confidence Gated Unit (CGU) proposed in Dias et al. (2020). The CGU is composed as ReLU+sigmoid activation functions. As visualized in Fig. 3, ReLU and sigmoid are applied, respectively, to coordinates (x_1^i or x_2^i) and confidence (c_i); their outputs are finally multiplied. The CGU emulates the behaviour of a gate, controlled by the confidence, as it returns values near 0 in the case of low confidence.

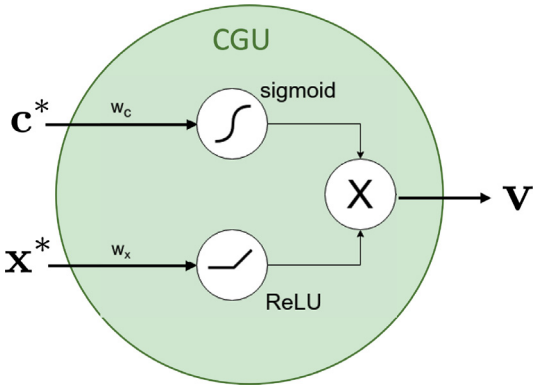

Fig. 3. Confidence Gated Unit (CGU).

The two gated outputs $\mathbf{v}_1$ and $\mathbf{v}_2$ are concatenated to obtain a single vector, which is provided to the intermediate part of the architecture, where a sequence of three fully connected layers consisting of 250, 200 and 150 neurons respectively is employed. Each layer includes a LeakyReLU, again to avoid vanishing gradients, as a non-linear activation function. Three output layers return the estimated angles, each of which is associated with its uncertainty value — details are reported in the next section.

3.2. HHP-net multi-task loss

To train the network we design a multi-task loss function incorporating heteroscedastic aleatoric uncertainty. With respect to classical Neural Networks, a Heteroscedastic Neural Network provides an estimate of the uncertainty of each prediction. This is particularly useful to capture noise within input observations: noise in our case is related to inherent noise in key point localization which may be affected by difficult viewpoints or occlusions. Indeed, some poses are intrinsically noisier and more prone to self-occlusions (see for instance examples in Fig. 4). This type of uncertainty may be learned as a function of the data, thus the output will include not only the three angles (yaw, pitch, roll), stored in a vector $\mathbf{q} = [q_y, q_p, q_r]$, but also the uncertainty values associated with them $\boldsymbol{\sigma} = [\sigma_y, \sigma_p, \sigma_r]$. We now discuss how we derive the multi-task loss function starting from a simple heteroscedastic loss formulation

3.2.1. A heteroscedastic loss function

Without loss of generality, we reason on a simple regression problem where we want to estimate a function $f_\omega : \mathbb{R}^n \to \mathbb{R}$ so that

$$y = f_\omega(\mathbf{x}) + \epsilon(\mathbf{x}). \tag{1}$$

The output is thus the sum between the function $f_\omega(\mathbf{x})$ – that depends on some parameters ω and the input $\mathbf{x}$ – and $\epsilon(\mathbf{x})$, that is the noise only depending on the input $\mathbf{x}$ (Nix and Weigend, 1994).

To quantify the uncertainty, we train a model to learn from a training set $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ a function that estimates both the mean and the variance of a target distribution using a maximum-likelihood formulation of a neural network (MacKay, 1992). To this purpose, we need to assume that the errors are normally distributed $\epsilon(\mathbf{x}_i) \sim \mathcal{N}\left(0, \sigma(\mathbf{x}_i)^2\right)$ hence the likelihood for each point $\mathbf{x}_i$ is:

$$p(y_i|\mathbf{x}_i;\omega) = \mathcal{N}(f_\omega(\mathbf{x}_i), \sigma(\mathbf{x}_i)^2) = \frac{1}{\sqrt{2\pi\sigma(\mathbf{x}_i)^2}} e^{\left(-\frac{(y-f_\omega(\mathbf{x}_i))^2}{2\sigma(\mathbf{x}_i)^2}\right)} \tag{2}$$

where y_i is the mean of this distribution and $\sigma(\mathbf{x}_i)^2$ is the variance. Hence, from a structural point of you, in addition to the estimation of the y_i, the heteroscedastic neural network architecture must be modified to also output a prediction of the variance: the latter quantifies the uncertainty associated with the prediction based on the noise in the training samples. Notice that the uncertainty is a function of the input e.g. if the noise is uniform over all the input values, the uncertainty should be constant.

Applying the logarithm to both sides of Eq. (2), we obtain a log-likelihood to maximize over the training set, i.e.

$$\max_{\omega} \quad \frac{1}{n}\sum_{i=1}^{\ell} -\frac{1}{2\hat{\sigma}\left(\mathbf{x}_i\right)^2}(y_i - \hat{f}_\omega\left(\mathbf{x}_i\right))^2 - \frac{1}{2}\log\hat{\sigma}\left(\mathbf{x}_i\right)^2 - \frac{1}{2}\log(2\pi) \tag{3}$$

where $\hat{f}_\omega$ and $\hat{\sigma}$ are, respectively, the prediction function and uncertainty estimated by the heteroscedastic neural network.[3] Equivalently:

$$\min_{\omega} \quad \frac{1}{n}\sum_{i=1}^{\ell} \frac{1}{2\hat{\sigma}\left(\mathbf{x}_i\right)^2}(\mathbf{y}_i - \hat{f}_\omega\left(\mathbf{x}_i\right))^2 + \frac{1}{2}\log\hat{\sigma}\left(\mathbf{x}_i\right)^2 \tag{4}$$

An alternative formulation based on the change of variable $\hat{s}_i = \log\hat{\sigma}(\mathbf{x}_i)^2$ can be adopted to avoid exploding uncertainties during training (Kendall and Gal, 2017), leading to the final problem formulation:

$$\min_{\omega} \quad \frac{1}{n}\sum_{i=1}^{\ell} \frac{1}{2}e^{(-\hat{s}_i)}(y_i - \hat{f}_\omega\left(\mathbf{x}_i\right))^2 + \frac{1}{2}\hat{s}_i \tag{5}$$

from which we derive the heteroscedastic loss function in Eq. (6) similarly to Kendall and Gal (2017)

$$\mathcal{L}_H(y, \hat{f}_\omega(\mathbf{x}), \hat{s}) = \frac{1}{2}e^{(-\hat{s})}(y - \hat{f}_\omega(\mathbf{x}))^2 + \frac{1}{2}\hat{s} \tag{6}$$

Notice finally that

$$\mathcal{L}_H(y, \hat{f}_\omega(\mathbf{x}), \hat{s}) = \frac{1}{2}e^{(-\hat{s})}\mathcal{L}_{MSE}(y, \hat{f}_{\omega(\mathbf{x})}) + \frac{1}{2}\hat{s} \tag{7}$$

where $\mathcal{L}_{MSE}$ is the classical square loss.

3.2.2. The heteroscedastic multi-task (HMT) loss function

We now specify to our problem the general formulation of the heteroscedastic loss function derived in the previous section.

We extend the model in Eq. (1) to represent a multi-task problem where the three components of the output are $\mathbf{q} = [q_y, q_p, q_r]$ and the associated uncertainties are $\sigma = [\sigma_y, \sigma_p, \sigma_r]$ (as usual we refer to the three angles yaw (y), pitch (p) and roll (r)). Hence, the single tasks within the multi-task formulation refer to the estimation of the three angles separately. In our solution, we estimate them by optimizing a unique function and exploiting their synergies. The input is composed of $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{c}$, that are respectively the coordinates of the key points detected on the face and the confidence in their detection.

We can now derive the multi-task heteroscedastic loss function we employ in our method:

$$\mathcal{L}_{HMT}(\mathbf{q}, \hat{\mathbf{q}}, \hat{\sigma}) = \sum_{k\in\{y,p,r\}} \mathcal{L}_H(q_k, \hat{f}_k(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}), \hat{s}_k)$$

$$= \sum_{k\in\{y,p,r\}} \left(\frac{1}{2}e^{(-\hat{s}_k)}(q_k - \hat{f}_k\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}\right))^2 + \frac{1}{2}\hat{s}_k\right). \tag{8}$$

where

$$\hat{\mathbf{q}} = \hat{\mathbf{f}}_\omega(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}) = [\hat{f}_y(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}), \hat{f}_p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}), \hat{f}_r(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c})] \tag{9}$$

and

$$\hat{\sigma}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}) = [\hat{\sigma}_y(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}), \hat{\sigma}_p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}), \hat{\sigma}_r(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c})] \tag{10}$$

are, respectively, function and uncertainty estimated by the heteroscedastic neural network, and for $k \in \{y, p, r\}$

$$\hat{s}_k = \log\hat{\sigma}_k(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c})^2. \tag{11}$$

With this formulation, we obtain a data-driven uncertainty estimation for each angle, used as a weight of each sub-loss. The uncertainty can increase the robustness of the network when dealing with noisy input data, we will empirically show a correlation between uncertainty and estimation error

4. Experiments

In this section we report the experimental analysis we performed to assess our approach. We first discuss in detail the implementation, the datasets and the experimental protocols we adopt, and then provide qualitative and quantitative results. In particular, we perform ablation studies to show the benefit of each element in the method, discuss the role of the uncertainty and the relation with the estimated error, and evaluate the transfer capability of the model across datasets.

It is worth observing that there are no free parameters to be tuned in our method.

4.1. Implementation details

Unless otherwise stated, we adopt OpenPose (Cao et al., 2019) as a key points extractor, as it provides a good balance between efficiency and accuracy. Among the 25 body key points it provides, in this work we focus on the five located on the face – left and right eye, left and right ear, nose – thus obtaining a triplet of input vectors $\mathbf{x}_1 = [x_1^1, \dots, x_1^5]$, $\mathbf{x}_2 = [x_2^1, \dots, x_2^5]$ and $\mathbf{c} = [c^1, \dots, c^5]$.

For the initialization, the weights of each layer are randomly sampled from a normal distribution with $\mu = 0$ and $\sigma^2 = 0.05$. The network has been trained for a number of epochs that ranges from 100 to 1000 depending on the dataset. We used Adam as an optimizer, with a learning rate 0.001, and a batch size of 64. The weights associated with the best validation loss have been selected as the final model.[4]

4.2. Datasets and protocols

We evaluate the effectiveness of our approach on three different datasets (see sample frames in Fig. 4):

- BIWI (Fanelli et al., 2011) includes $\sim 15K$ images of 24 people acquired in a controlled scenario. The head pose orientation covers the range $\pm 75°$ for the yaw angle and $\pm 60°$ for the pitch. The ground truth has been obtained by fitting a 3D face model.
- AFLW-2000 (Yin et al., 2017) contains the first 2000 images of the in-the-wild AFLW dataset (Koestinger et al., 2011), a large-scale collection of face images with a large variety in appearance and environmental conditions. The annotation has been obtained by fitting a 3D face model.
- 300W-LP (Sagonas et al., 2013) is a collection of different in-the-wild datasets, grouped and re-annotated to account for different types of variability, such as pose, expression, illumination, background, occlusion, and image quality. A face model is fit on each image, distorted to vary the yaw of the face.

[3] In the following the last term is ignored as it is a constant.

[4] Code and pre-trained weights are available at https://github.com/Malga-Vision/HHP-Net.

Fig. 4. Sample frames from the public datasets we adopted in our experimental analysis: BIWI (top row), AFLW-2000 (middle row), and 300W-LP (bottom). For readability of the figures, we report their greyscale version with an arrow in red which is the 2D projection of the head direction. Being the projection of a 3D vector, it can also be a point, e.g. like in the top-left image where the direction of view is 'outside' the page. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For all the datasets, the ground truth takes the form of a triplet of angles in Euler notation expressed with respect to a reference frontal pose

According to previous works (e.g. Yang et al., 2019), in the comparative analysis we adopt two main protocols:

- P1 Training is performed on a single dataset (300W-LP), while BIWI and AFLW-2000 are used as test.
- P2 Training and test set are derived from the BIWI dataset using the split 16–9 sequences, for training and test respectively, following the procedure proposed in Fanelli et al. (2013).

4.3. Method assessment

In this section, we present an experimental assessment to discuss the core properties of our approach.

The output is visualized by projecting the angles on the image plane according to the Tait-Bryan angles. The projections are computed as

$$
\begin{aligned}
x_r &= \cos q_y \cdot \cos q_r + \Delta_x \\
y_r &= \cos q_p \cdot \sin q_r + \cos q_r \cdot \sin q_y \cdot \sin q_p + \Delta_y \\
x_g &= -\cos q_y \cdot \sin q_r + \Delta_x \\
y_g &= \cos q_p \cdot \cos q_r - \sin q_y \cdot \sin q_p \cdot \sin q_r + \Delta_y \\
x_b &= \sin q_y + \Delta_x \\
y_b &= -\cos q_y \cdot \sin q_p + \Delta_y
\end{aligned}
\tag{12}
$$

where (x_r, y_r), (x_g, y_g) and (x_b, y_b) are the image coordinates of the endpoints of red, green and blue vectors, while (Δ_x, Δ_y) is the application point they have in common.

In the following we provide an assessment of the properties and meaningfulness of the uncertainty measures.

Uncertainty estimations quality. Fig. 5 (above) reports a cumulative analysis of the amount of data associated with a given uncertainty, highlighting how the average error grows with the uncertainty — in agreement with what has been reported in Dias et al. (2020). Given the assumptions in Section 3.2.1 of a normal distribution for the errors, the $\sigma(\mathbf{x_i})^2$ is the variance of this distribution and the parameter we are going to estimate for each angle in the regression process. So, fixed one angle (e.g. yaw) it can be seen as the variance of the retrieved angle (yaw). Under this perspective, it is straightforward to read it in degrees. However, having implemented the algorithm and defined the uncertainty as $log(\hat{\sigma}(\mathbf{x_i}))^2$ or better $log(\hat{\sigma}(\mathbf{x_1}, \mathbf{x_2}, \mathbf{c}))^2$, we retrieved the degree information as

$$
log(\hat{\sigma}(\mathbf{x_1}, \mathbf{x_2}, \mathbf{c}))^2 = s_i \Leftrightarrow \sigma = \sqrt{e^{s_i}} \tag{13}
$$

Hence, the interpretability of our uncertainty measure is strengthened by the fact it can be expressed in degrees, as the estimated angles. In this way, the two outcomes of our model can be directly compared. In Fig. 5 (bottom) we report the histogram of the absolute value of the difference between the angle and corresponding uncertainty. It can be observed that it is predominantly very low, in 71% of the cases below 3 degrees, 88% below 5 degrees and 98% below 10 degrees. This shows that the uncertainty measures can be adopted as an indicator of the reliability of our estimated angles.

Similarly to what was observed in Feng et al. (2019), we also notice a strong correlation between the uncertainty values associated with the three predicted angles. To quantify the correlation we computed the Pearson correlation between the uncertainties of all pairs of angles, obtaining 0.72 for (*yaw*, *pitch*), 0.78 for (*yaw*, *roll*), and 0.92 for (*pitch*, *roll*).

Uncertainty estimation and model interpretation. We now analyse the factors that may influence the uncertainty estimation, with a focus on the characteristics of the head pose to be predicted. In Fig. 6 we report the trend of the uncertainty associated with the prediction obtained from video sequences where a subject rotates the head offering

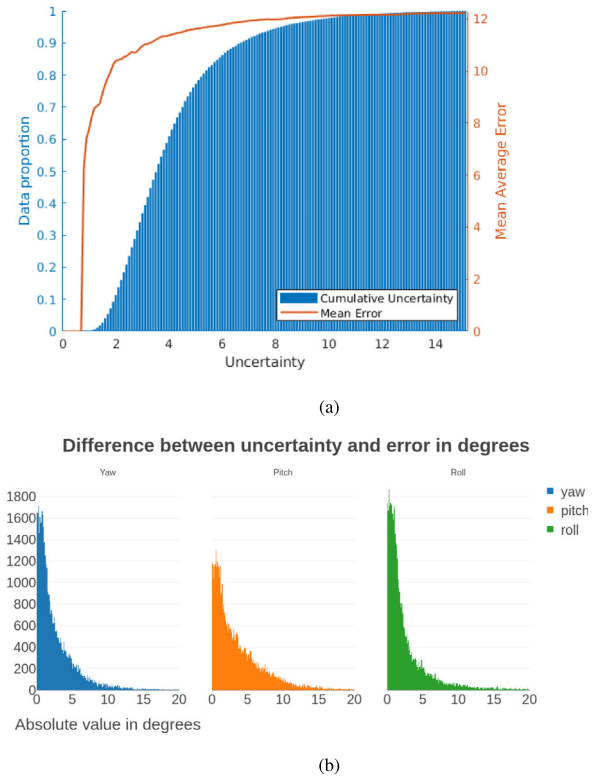

Fig. 5. (a) Cumulative angular error as a function of the average uncertainty (red), and data proportion with at least the uncertainty written in the x-axis (blue). (b) Occurrences of test data divided in bins of the absolute difference between uncertainty (in degrees) and error; the zeroth bin on the x-axis is when error and uncertainties coincide. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

different test poses to the method. Representative frames, providing an intuition about the transitions between poses in the sequence, are reported below the plot. It is easy to observe that for some poses (the ones associated with ambiguous views or partial occlusions that hide some key points on the face) the uncertainty is higher. The lowest uncertainty values are associated with frontal views, the ones providing the most visible and non-ambiguous key points.

Inspired by these observations, we now evaluate the dependence of the uncertainty and the error on the quality and quantity of the input key points.

Number of key points. We observe the influence of the quality and quantity of input semantic features on the final head pose estimate. In Fig. 7, we analyse the performance of our method in terms of uncertainty values (bottom) and absolute angular error (top) as the number of available key points changes. On the left, we cluster faces according to the number of key points detected by OpenPose. When only 3 key points are available the uncertainty is rather high on average. Increasing the number of points uncertainty is progressively reduced, with a similar trend shown by the error. This confirms the intuition that the more input points the method has, the higher its confidence in the prediction, which is more reliable and accurate.

On the right, we randomly drop points from the available input to simulate an even more challenging scenario for our method. When points are randomly dropped, we only consider samples with more than two points. When all 5 key points are available, the uncertainty is compactly lower (confirming what was already observed in the previous experiment) as the method can rely on a more comprehensive representation of the input. In the intermediate cases – where we may have 2, 3, or 4 key points available in input – the uncertainty

Table 1
Training and testing HHP-Net with inputs from different 2D pose estimators on the BIWI dataset. In the table, we report the MAE (Mean Absolute Error in degree) averaged over the three angles and the standard deviation.

Train \ Test	Centernet	Mediapipe	Openpose
Centernet	3.33 ± 0.91	7.57 ± 2.48	6.73 ± 5.88
Mediapipe	13.31 ± 7.96	5.99 ± 1.56	14.55 ± 7.59
Openpose	4.64 ± 1.99	7.08 ± 2.37	4.51 ± 1.27

progressively decreases, but we also have a higher degree of variability, as some key-point configurations are more significant than others and thus the amount of information they provide to the model may be unevenly reflecting the concept that the noise could be different for each input sample. With respect to the plots in the left column of Fig. 7, the box plots at right show a higher standard deviation since randomly dropping points from the input we simulate a higher variability in the input configurations with respect to the ones usually provided by OpenPose and from the datasets we used.

Plugging in different Pose detectors. Here we assess the robustness of our approach to different choices of 2D pose estimators. More specifically, we employ OpenPose (Cao et al., 2019), CenterNet (Duan et al., 2019), and MediaPipe (Lugaresi et al., 2019) and consider all the pairs for train–test. The results are reported in Table 1. If we read the table row-wise we may analyse the behaviour of models obtained from the different pose detectors on the output of different nature. It shows that CenterNet and Openpose are rather interchangeable (OpenPose in particular provides very similar results when tested on itself or Centernet), while Mediapipe is not. The reason is that its output is rather different in terms of localization of the key points and behaviour in the presence of occlusions (MediaPipe never provides a zero confidence for occluded key points), reducing the benefit of the Confidence Gated Unit.

4.4. Removing the uncertainty: An ablation study

We perform an ablation study by removing the uncertainty from our model. To this purpose, we consider two variations of the method (with y = yaw, p = pitch and r = roll):

MSE: we directly regress the three angles adopting a loss computed as the sum of the Mean Squared Error (MSE) on each angle:

$$\mathcal{L}_{MSE-MT}(\mathbf{q},\hat{\mathbf{q}}) = \sum_{k\in\{y,p,r\}} (q_k - \hat{f}_k\left(\mathbf{x}_1,\mathbf{x}_2,\mathbf{c}\right))^2. \tag{14}$$

where $\mathbf{q} = [q_y, q_p, q_r]$, and $\hat{\mathbf{q}} = [\hat{f}_y(\mathbf{x}_1,\mathbf{x}_2,\mathbf{c}), \hat{f}_p(\mathbf{x}_1,\mathbf{x}_2,\mathbf{c}), \hat{f}_r(\mathbf{x}_1,\mathbf{x}_2,\mathbf{c})]$.

COMB: we employ an alternative loss function $\mathcal{L}_{COMB}$ proposed in Ruiz et al. (2018) which has been proved to be very successful on the same estimation task. The loss allows for jointly solving a multi-class classification (with N classes corresponding to binned angles) and a regression task, and it can be formalized as follows :

$$\mathcal{L}_{COMB}(\mathbf{q},\hat{\mathbf{q}}) = \mathcal{L}_{CE-MT}(\mathbf{q},\hat{\mathbf{q}}) + \alpha * \mathcal{L}_{MSE-MT}(\mathbf{q},\hat{\mathbf{q}}) \tag{15}$$

where

$$\mathcal{L}_{CE-MT}(\mathbf{q},\hat{\mathbf{q}}) = \sum_{k\in\{y,p,r\}} \left[-\sum_{j=1}^{N} q_k^j \log\left(\hat{f}_k^j(\mathbf{x}_1,\mathbf{x}_2,\mathbf{c})\right)\right] \tag{16}$$

is the cross-entropy loss adapted to our multi-task problem, while $\mathcal{L}_{MSE-MT}$ is the multi-task square loss of Eq. (14). Hence, the loss combines the cross entropy, computed between the binned angles, and the MSE loss, computed between the scalar angles; α is a hyperparameter that controls the weight of the regression loss. According to the original work, in the experiment we set $\alpha = 1$.

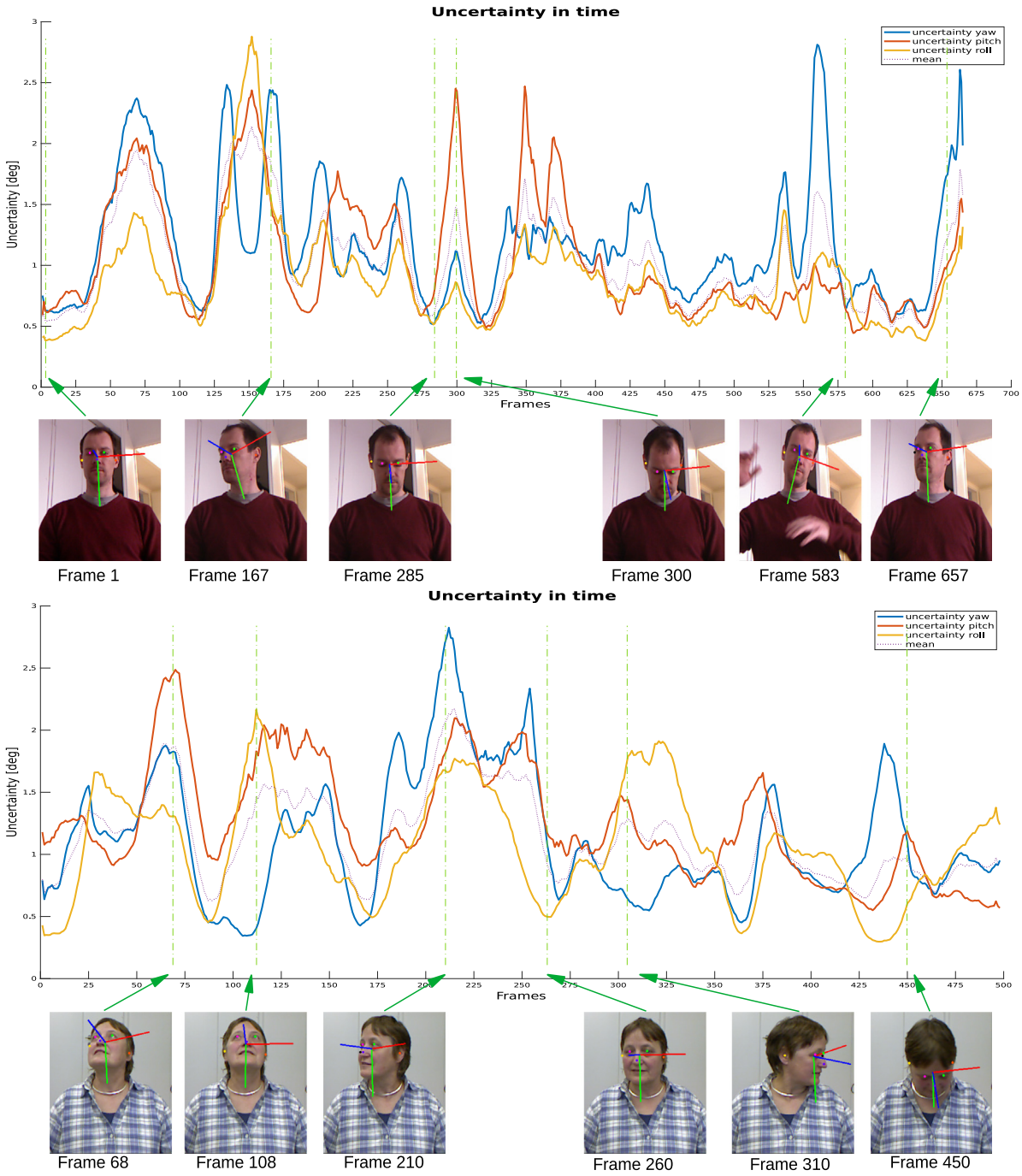

Fig. 6. Examples of how the uncertainties (in degrees) are influenced by the instantaneous head pose of a subject moving in front of a camera over time. We report in blue the yaw uncertainty, in orange the pitch uncertainty and in yellow the roll uncertainty. In dotted-purple we mark the mean uncertainty. The scale is in degrees of uncertainty. It can be observed that the uncertainties are very close to zero for the neutral head pose (frame 1 of the first sequence) and start to increase when the head rotates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In Table 2 we report the angular errors we obtain with the three different losses. As it can be observed, learning the angles associated with the uncertainty provides the best average performance, showing the benefit of the uncertainty not only in terms of the interpretability of the model but also as a way to improve its effectiveness.

4.5. Comparisons with other approaches

We now perform a comparative analysis with state-of-the-art head pose estimators. For a fair comparison, we consider methods that use RGB images as inputs or features extracted from them.

The analysis is reported in Tables 3, 4, and 5, where all errors are expressed in degrees (err_y = yaw error, err_p = pitch error, err_r = roll error), the model size is reported in MegaBytes (MB), and the MAE is the Mean Absolute Error.

As a first important observation, notice that our approach produces a significantly smaller model (0.4 MB). This was the main purpose of our work and it has been clearly achieved, as our method is about ∼12 times smaller than the closest model in the literature. According to the protocol followed by other works – all requiring a face detector but not including its size in their analysis – the size of our model does not include the pose estimator.

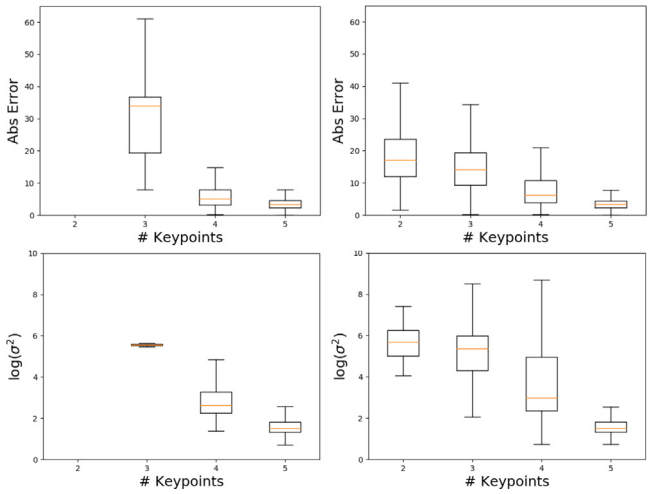

Fig. 7. Performance of our method (top row: mean angular error in angles, bottom row: uncertainty) with respect to the number of input points, considering the outputs of OpenPose (left) and randomly dropping points (right). Training: 300W-LP Test: BIWI. Uncertainty is presented in a log scale visualization for a clearer view.

Table 2
Comparison among different loss functions (see text). **All errors are expressed in degrees** (°): err_y = yaw error, err_p = pitch error, err_r = roll error, MAE = Mean Absolute Error.

Train	Val	Loss	err_y	err_p	err_r	MAE
BIWI	BIWI	$\mathcal{L}_{MSE}$	2.90	4.80	3.34	3.70
BIWI	BIWI	$\mathcal{L}_{COMB}$	3.15	4.85	3.40	3.80
BIWI	BIWI	$\mathcal{L}_{HMT}$	3.04	4.79	3.21	**3.68**
300WLP	BIWI	$\mathcal{L}_{MSE}$	4.75	6.65	4.45	5.28
300WLP	BIWI	$\mathcal{L}_{COMB}$	4.67	8.08	4.87	5.88
300WLP	BIWI	$\mathcal{L}_{HMT}$	4.14	7.00	4.40	**5.18**
300WLP	AFLW2000	$\mathcal{L}_{MSE}$	5.72	10.41	8.08	8.07
300WLP	AFLW2000	$\mathcal{L}_{COMB}$	5.55	10.39	8.18	8.04
300WLP	AFLW2000	$\mathcal{L}_{HMT}$	5.26	10.12	7.73	**7.70**
AFLW	AFLW2000	$\mathcal{L}_{MSE}$	7.60	6.43	4.76	6.26
AFLW	AFLW2000	$\mathcal{L}_{COMB}$	7.31	6.55	4.68	6.18
AFLW	AFLW2000	$\mathcal{L}_{HMT}$	7.40	6.63	4.47	**6.16**

In terms of performances, Table 3 reports a comparison with respect to Protocol P2 (BIWI dataset for training and test): the results we obtain are superior to Mukherjee and Robertson (2015), Drouard et al. (2015), Fanelli et al. (2013) and slightly below (Gu et al., 2017; Lathuiliere et al., 2017; Yang et al., 2019; Zhang et al., 2020) (less than 0.1 degrees of difference for the first three, less than 0.4 for the latter).

Table 4 refers to Protocol P1 (training carried out on 300W-LP, BIWI for the test): the experiment mainly evaluates the transfer potential to a different dataset with different properties. The table reports results obtained with methods relying on the estimation of 3D face models (Zhu et al., 2019; Kumar et al., 2017; Kazemi and Sullivan, 2014; Bulat and Tzimiropoulos, 2017) and methods based on analysing RGB image portions obtained by face detectors, such as Shao et al. (2019), Ruiz et al. (2018), Yang et al. (2019).

We share with the latter group the main motivation for designing simple and more efficient procedures while keeping competitive performances. In this sense, our approach does not require complex pre-processing steps or highly resource-demanding training, but at the same time, it wisely leverages structural information on the face. Table 4 reports results that are more accurate than all methods with the exception of FSA-Caps, although the difference is on average only slightly above 1 degree. This small accuracy loss is counterbalanced by the benefits in terms of a smaller size, and it may be explained by the

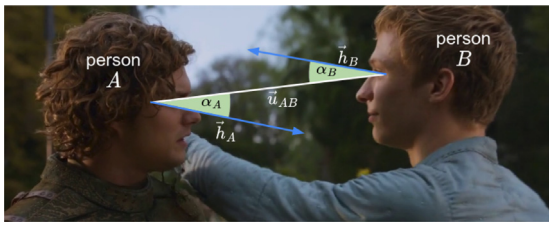

Fig. 8. A visual sketch with our formulation of the LAEO detection task (for readability the vectors are denoted with arrows).

simplicity and compactness of our input: while nicely behaving in the majority of non-ambiguous situations, our sparse input is more severely influenced by occlusions, and missed or noisy detections.

Finally, Table 5 follows again Protocol P1, on a more complex test set, AFLW2000, where images are acquired in a less controlled environment. In this case, our methodology is reporting slightly worse results, but with a loss always less than 3 degrees on average. We noticed this is due in particular to keypoint detection errors, as the synthetic data manipulation introduced artefacts.

To further evaluate the transfer potential of our approach we also report the result we obtained on the same test set when training the network on a related dataset (AFLW without the AFLW2000 section): the results are in this case comparable to the previous experiments.

We conclude by mentioning that we do not include in our comparison the approach in Fanelli et al. (2013) since it uses the depth as input, and the methods Dlib (Kazemi and Sullivan, 2014) and FAN (Bulat and Tzimiropoulos, 2017) that solve a different problem (face alignment). Also, among the very recently proposed approaches, our analysis does not mention EVA-GCN (Xin et al., 2021) and KEPLER (Kumar et al., 2017) as they solve a different problem (jointly solving different tasks, one of them being head pose estimation), and 3DDFA (Zhu et al., 2019) that uses a richer input (image and 3D model).

4.6. Model size and inference time

We now show the robustness of our method with respect to reductions of size, which may be needed when the available computational resources are very limited. More specifically, we analyse how the performance changes as we reduce the size of the model. We choose 300W-LP training and BIWI test (protocol P1) for their larger training and test sets and decrease the number of neurons in the fully connected layers so the backbone remains the same as proposed in the paper, while its size decreases. Given a reduction factor $\beta \in (0, 1)$, we obtain a "reduced" version of our architecture by multiplying the original number of neurons in each layer (250, 200 and 150 in, respectively, the first, second and third layer) by β.

By varying β in the range $(0, 1)$ we reduce the model size (the number of parameters) and thus also the number of sum and multiplication operations. Table 6 compares our baseline ($\beta = 1$) with two reduced models (overall size in MB up to 10× smaller) causing a very limited degradation in the MAE (below 1 degree). This experiment highlights the possibility of further reducing the size of the architecture, with a very limited performance loss, if required by the system.

We finally briefly mention the computational performance of our method, which is an average of 142 fps (approximately an inference time of 7×10^{-3} s per frame). In the full inference pipeline, we should also consider the cost of running the key points detection, which depends on the specific approach. Empirical estimation of inference times can be found in Lugaresi et al. (2019) for Openpose and Mediapipe, and in Duan et al. (2019) for Centernet.

Table 3
Comparison following Protocol P2: BIWI is both training and test. Our model is the smallest (∼0.4 MB) while providing only a small degradation with respect to the best result (∼ 0.4°).

Method	MB	Par. × 10^6	err_y	err_p	err_r	MAE
D-HeadPose (Mukherjee and Robertson, 2015)	–	–	–	5.67	5.18	–
Drouard et al. (2015)	–	–	4.9	5.9	4.7	5.16
DFA (Gu et al., 2017)	500†	138‡	3.91	4.03	3.03	3.66
DMLIR (Lathuiliere et al., 2017)	500	–	3.12	4.68	3.07	3.62
FSA-Caps-Fusion (Yang et al., 2019)	5.1	1.2	2.89	4.29	3.60	3.60
FND (Zhang et al., 2020)	5.8	–	3.0	3.98	2.88	3.29
img2pose (Albiero et al., 2021)	–	–	4.57	3.55	3.24	3.79
LwPosr (Dhingra, 2022)	–	0.15	3.62	4.65	3.78	4.01
QTNet (Hsu et al., 2018)	–	–	4.01	5.49	2.94	4.15
Ruiz (Ruiz et al., 2018) (α = 2)	–	–	3.29	3.39	3.00	3.23
TriNet (Cao et al., 2021b)	–	26‡	2.44	3.04	2.93	2.80
Our approach	**∼0.4**	**∼0.09**	3.04	4.79	3.21	3.68

†, ‡ *data respectively from* Yang et al. (2019),Dhingra (2022).

Table 4
Comparison following Protocol P1, where 300W-LP is the training, while BIWI is the test. Our method is still the smallest and performs better than all other approaches but Yang et al. (2019). The latter is however associated with a model significantly larger than ours.

Method	MB	Par. ×10^6	err_y	err_p	err_r	MAE
Shao (K = 0.5) (Shao et al., 2019)	93	24.6††	4.59	7.25	6.15	6.00
Ruiz (Ruiz et al., 2018) (α = 2)	95.9†	23.9	5.17	6.98	3.39	5.18
Ruiz (Ruiz et al., 2018) (α = 1)	95.9†	23.9	4.81	6.61	3.27	4.90
LwPosr α (Dhingra, 2022)	–	0.15	4.41	5.11	3.24	4.25
LwPosr (Dhingra, 2022)	–	0.15	4.11	4.87	3.19	4.05
FSA-Caps-Fusion (Yang et al., 2019)	5.1	1.2	4.27	4.96	2.76	4.00
TriNet (Cao et al., 2021b)	–	26‡	3.05	4.76	4.11	3.97
FND (Zhang et al., 2020)	5.8	–	4.52	4.70	2.56	3.93
WHENet-V (Zhou and Gregson, 2020)	–	4.4	–	–	–	3.48
Our approach	**∼0.4**	**∼0.09**	4.14	7.00	4.40	5.18

†, ‡, ††, ∗ *data respectively from* Yang et al. (2019), Dhingra (2022), Zhou and Gregson (2020), Ruiz et al. (2018).

5. An application to dyadic interaction detection

We finally discuss a task where our method finds a natural application, i.e. the analysis of social interactions, for which the head directions represent a strong visual cue of non-verbal human–human communication (Abele, 1986). We consider scenarios where a small group of people is involved in a social experience, and we pay particular attention to people *looking at each other* (LAEO).

LAEO algorithm. Fig. 8 provides a visual sketch of our formulation of the task. Let us consider the two people present in the scene, A and B in our example, whose positions can be compactly described with the head centroids (x_A, y_A) and (x_B, y_B). We start from the head pose estimated for each of them, $\mathbf{q}^A$ and $\mathbf{q}^B$ respectively, and obtain a projection of the corresponding direction on the image plane: for the subject A, given the triplet of angles (q_y^A, q_p^A, q_r^A), we derive the end-point of the head direction on the image plane (x'_A, y'_A) as $x'_A = \sin(q_y^A)$ and $y'_A = -\cos(q_y^A)\sin(q_p^A)$. The projection is computed as $\mathbf{h}_A = (x'_A - x_A, y'_A - y_A)$. Similarly we obtain $\mathbf{h}_B$ for the other subject.

Then, we estimate a measure of interaction between each pair of people considering the vector $\mathbf{u}_{AB}$ connecting the two head centroids, the vector $\mathbf{h}_A$ and the angle α_A between the two: the measure of the interaction is given by the cosine of the angle α_A. The same applies to person B with $\mathbf{u}_{BA} = -\mathbf{u}_{AB}$ and α_B. The average between the two measures gives the LAEO value and thresholding on such measure allows us to detect LAEO pairs.

We build our approach on this baseline method, incorporating knowledge from the uncertainty associated with the 3D angles (the method is sketched in Algorithm 1). Given the triplets of uncertainties associated with the two heads poses, (s_A^y, s_A^p, s_A^r) and (s_B^y, s_B^p, s_B^r), we compute the averages, $\hat{s}_A = \frac{1}{2}(s_A^y + s_A^p)$ and $\hat{s}_B = \frac{1}{2}(s_B^y + s_B^p)$; the roll component is discarded because it does not affect the gaze vector projection on the image plane. Following the intuition that estimates with high uncertainty should be less reliable, we compute a weight to adjust the contribution of each subject to the interaction measure depending on the confidence we have in it, essentially deciding a threshold above which the estimate is considered unreliable. For the subject A this can be formulated as $w_A = \mathbb{1}_X(\hat{s}_A)$ where $X = [0, \delta]$ with δ an appropriate threshold on the uncertainty, and $\mathbb{1}_X : \mathbb{R} \to \{0, 1\}$ the indicator function on the interval X. δ is computed as the average uncertainty plus the standard deviation, both of them computed on the entire training set (in the experiments $\delta = 7$).

Table 5
Comparison following Protocol P1, where 300W-LP is the training and AFLW 2000 is the test (note: ✠ = Trained on AFLW - AFLW2000). The performances show a slightly higher worsening with respect to alternative approaches, but the difference is still very limited (always less than 3°).

Method	MB	Par. $\times 10^6$	err_y	err_p	err_r	MAE
3DDFA (Zhu et al., 2019)	–	–	5.40	8.53	8.25	7.39
Ruiz (Ruiz et al., 2018)) ($\alpha = 1$)	95.9†	23.9	6.92	6.64	5.67	6.41
Ruiz (Ruiz et al., 2018) ($\alpha = 2$)	95.9†	23.9	6.47	6.56	5.44	6.16
Shao (K = 0.5) (Shao et al., 2019)	93	24.6††	4.59	7.25	6.15	6.00
FSA-Caps-Fusion (Yang et al., 2019)	5.1	1.2	4.50	6.08	4.64	5.07
Shao (K = 0.5) (Shao et al., 2019)	93	24.6	5.07	6.37	4.99	5.48
WHENet-V (Zhou and Gregson, 2020)	–	4.4	–	–	–	4.83
LwPosr (Dhingra, 2022)	–	0.15	4.80	6.38	4.88	5.35
LwPosr α (Dhingra, 2022)	–	0.15	4.44	6.06	4.35	4.95
TriNet (Cao et al., 2021b)	–	26‡	4.20	5.77	4.04	4.67
FND (Zhang et al., 2020)	5.8	–	3.78	5.61	3.88	4.42
Our approach	**~0.4**	**~0.09**	5.26	10.12	7.73	7.70
Our approach✠	**~0.4**	**~0.09**	7.40	6.63	4.47	6.16

†, ‡, †† *data respectively from* Yang et al. (2019), Dhingra (2022), Zhou and Gregson (2020).

Table 6
Comparison among models with different sizes (Protocol P1: 300W-LP train, BIWI test). β = neurons reduction factor (see text), MAE = Mean Absolute Error.

β	MAE	Parameters	MB
1	5.18	94 031	0.385
0.6	5.43	37 206	0.158
0.2	5.54	6006	0.032

Table 7
The performance of our method for LAEO detection on the UCO-LAEO dataset. AP is estimated as in Marín-Jiménez et al. (2020), $\tau = 0.93$.

Method	PREC	REC	F	AP
LAEO-Net (Marin-Jimenez et al., 2019)	–	–	–	0.80
LAEO-Net++ (Marín-Jiménez et al., 2020)	–	–	–	0.87
Gaze Pattern Rec. (Chang et al., 2023)	–	–	–	0.80
Baseline (Ours)	0.77	0.80	0.78	0.86
With uncertainty (Ours)	0.80	0.72	0.76	0.88

LAEO estimation assessment. We evaluate our method on the UCO-LAEO dataset (Marin-Jimenez et al., 2019), which includes sequences from four popular TV shows in the form of 129 shots of variable length. The annotation is provided at a frame level – *is there a pair of LAEO people in the frame?* – and at a pair level – i.e. each head pair is labelled as LAEO or not. The task we solve is a binary classification task: for each frame in the sequence we consider all pairs of people detected in the frame and label them as LAEO or not using the method in Algorithm 1. Finally, a threshold τ, selected on the ROC curve of the training set, is used to detect the LAEO pairs.

We report in Fig. 9 examples to show how our LAEO measure smoothly changes during the interaction event.

We report in Table 7 the performance provided by our baseline method and the one incorporating the uncertainty on the test set.

Algorithm 1 Fast LAEO Detection

Input: Head centroids (x_A, y_A) and (x_B, y_B); projections of head directions (x'_A, y'_A) and (x'_B, y'_B); uncertainty weights w_A and w_B
$\mathbf{u}_{AB} \leftarrow (x_B - x_A, y_B - y_A)$
$\mathbf{h}_A \leftarrow (x'_A - x_A, y'_A - y_A)$
$\mathbf{h}_B \leftarrow (x'_B - x_B, y'_B - y_B)$
$cos(\alpha_A) \leftarrow \frac{\mathbf{u}_{AB} \cdot \mathbf{h}_A}{|\mathbf{u}_{AB}| \cdot |\mathbf{h}_A|}$
$cos(\alpha_B) \leftarrow \frac{-\mathbf{u}_{AB} \cdot \mathbf{h}_B}{|\mathbf{u}_{AB}| \cdot |\mathbf{h}_B|}$
Compute the level of mutual interaction $LAEO_{value} = w_A cos(\alpha_A) + w_B cos(\alpha_B)$
Return $LAEO_{value}$

The results suggest that using the prior knowledge derived from the uncertainty allows us to significantly reduce the number of false positives (−6%, with a slight increase of the precision) to the price of a small reduction of true positive (−7%, with a small reduction of the recall). Overall, the uncertainty brings improvements as the AP increases (+0.02). As a reference, we also show in the table the results provided by Marin-Jimenez et al. (2019), Marín-Jiménez et al. (2020), Chang et al. (2023).

Examples of the obtained results are reported in Fig. 10, where we show that our method is tolerant to the presence of more than 2 people, and to the scene variability.

6. Conclusions

In this work, we discussed a method for head pose estimation from the head key points extracted on RGB images, that provides the head pose as a triplet of Euler angles. Each angle is also associated with a measure of the aleatoric heteroscedastic uncertainty. We approached the problem as a multi-task regression and designed a neural network which is very efficient both in terms of space occupancy (less than 0.5 MB) and inference time (it runs at 100 fps), thus providing the potential to run on mobile devices. A core element of the architecture is the multi-task loss we employed, in which the data-driven uncertainties act as a weight of the sub-losses.

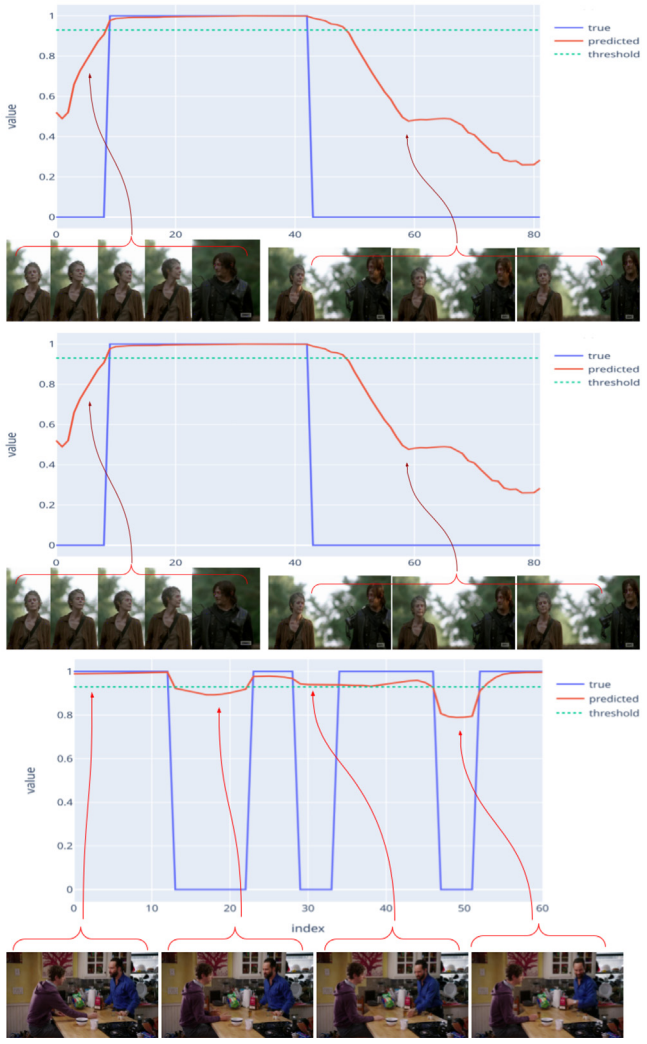

Fig. 9. Examples of LAEO measures over time. In the plots we report in blue the ground truth, and in red our LAEO measure. In green, we mark the threshold we adopt. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We provided a thorough experimental assessment, showing our method couples a very light computation with precision in the estimates superior or very close to state-of-the-art methods. When our performance does not reach such precision, the loss is always very limited (in the order of a few degrees) and counterbalanced by a significant saving in terms of space–time computational demand. Indeed, this may happen mainly when the quality of the input is not sufficient, i.e. when some of the keypoints are missing (e.g. due to occlusions or more in general detection failures) or when the localization of the key points (that may be uneven, depending on the specific pose detector that has been employed) is . Nevertheless, for its inherent characteristics, our method is particularly suitable for settings with a limited computational budget: with respect to alternative approaches relying on images of the detected face or on 3D face models, our very sparse input makes the cost of our method negligible with respect to a full end-to-end pipeline.

In this respect, its use in the robotics setting is currently under investigation.

We also discussed the connections between estimation error and uncertainty, which improves the interpretability of our model. In particular, we observed tolerance to variability of the input points (for instance the feasibility of using different pose detectors). On the negative side, as mentioned before, we experienced the negative impact

Fig. 10. Examples of LAEO detections. The arrows represent the head direction estimated by HHP-Net and projected on the image plane and are green if the corresponding person has been found involved in a LAEO. The prediction of our method for LAEO detection is reported in yellow and, in the case of LAEO, it specifies the identifier of the other interacting person. The identifiers are in red close to the subjects. In the last row, we report examples of failures, due to the ambiguities of the information on the image plane. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of key point detection failures that could be attenuated by adopting a video-based instead of an image-based analysis. This extension will be addressed in future research, along the line of Her et al. (2023). A further future direction of improvement, that may help addressing more challenging scenarios, is related with exploiting orientation relationships, as observed in Liu et al. (2023).

As an application, we discussed a proof-of-concept application for the detection of Looking-At-Each-Other events. In consideration of the encouraging results we obtained, we are currently performing a more comprehensive investigation of the use of the proposed methodology in analysing the activity of small groups of people.

CRediT authorship contribution statement

Federico Figari Tomenotti: Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Nicoletta Noceti:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Francesca Odone:** Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nicoletta Nocet reports financial support was provided by Cariplo Foundation. Nicoletta Noceti reports financial support was provided by Air Force Office of Scientific Research.

Data availability

Publicly available datasets have been used for this work. The code is publicly available, the link can be found in the manuscript.

Acknowledgements

This work has been carried out at the Machine Learning Genoa (MaLGa) center, Università di Genova (IT), supported by Fondazione Cariplo, Italy with grant no. 2018-0858, and by AFOSR (Air Force Office of Scientific Research), grant n. FA8655-20-1-7035.

References

Abate, A.F., Bisogni, C., Castiglione, A., Nappi, M., 2022. Head pose estimation: An extensive survey on recent techniques and applications. Pattern Recognit. 127.

Abele, A., 1986. Functions of gaze in social interaction: Communication and monitoring. J. Nonverbal Behav. 10 (2), 83–101.

Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T., 2021. img2pose: Face alignment and detection via 6dof, face pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7617–7627.

Alghamdi, M., Alhakbani, N., Al-Nafjan, A., 2023. Assessing the potential of robotics technology for enhancing educational for children with autism spectrum disorder. Behav. Sci. 13 (7), 598.

Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S., 2017. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4315–4324.

Barra, P., Barra, S., Bisogni, C., De Marsico, M., Nappi, M., 2020. Web-shaped model for head pose estimation: An approach for best exemplar selection. TIP 29, 5457–5468.

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M., 2020. Blazepose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204.

Bisogni, C., Nappi, M., Pero, C., Ricciardi, S., 2021. FASHE: A fractal based strategy for head pose estimation. IEEE Trans. Image Process. 30, 3192–3203.

Bulat, A., Tzimiropoulos, G., 2017. How far are we from solving the 2D and 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: ICCV.

Campbell, K.L., 2012. The SHRP 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety. Tr News 282.

Cantarini, G., Tomenotti, F.F., Noceti, N., Odone, F., 2022. HHP-net: A light heteroscedastic neural network for head pose estimation with uncertainty. In: WACV. pp. 3521–3530.

Cao, Z., Chu, Z., Liu, D., Chen, Y., 2021a. A vector-based representation to enhance head pose estimation. In: WACV.

Cao, Z., Chu, Z., Liu, D., Chen, Y., 2021b. A vector-based representation to enhance head pose estimation. In: CVPR. pp. 1188–1197.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A., 2019. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. PAMI.

Chang, F., Zeng, J., Liu, Q., Shan, S., 2023. Gaze pattern recognition in dyadic communication. In: Proceedings of the 2023 Symposium on Eye Tracking Research and Applications. pp. 1–7.

Choi, S., Choi, S., Kim, C., 2021. MobileHumanPose: Toward real-time 3D human pose estimation in mobile devices. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2328–2338.

Cipolla, R., Gal, Y., Kendall, A., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR. pp. 7482–7491.

Colyer, S.L., Evans, M., Cosker, D.P., Salo, A.I., 2018. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. Sports Med.-Open 4 (1), 1–15.

Cristani, M., Raghavendra, R., Del Bue, A., Murino, V., 2013. Human behavior analysis in video surveillance: A social signal processing perspective. Neurocomputing 100, 86–97.

Dhingra, N., 2022. Lwposr: Lightweight efficient fine grained head pose estimation. In: WACV. pp. 1495–1505.

Dias, P.A., Malafronte, D., Medeiros, H., Odone, F., 2020. Gaze estimation for assisted living environments. In: WACV.

Doosti, B., Chen, C.-H., Vemulapalli, R., Jia, X., Zhu, Y., Green, B., 2021. Boosting image-based mutual gaze detection using pseudo 3D gaze. In: AAAI. pp. 1273–1281.

Drouard, V., Ba, S., Evangelidis, G., Deleforge, A., Horaud, R., 2015. Head pose estimation via probabilistic high-dimensional regression. In: ICIP. pp. 4624–4628.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q., 2019. CenterNet: Keypoint triplets for object detection. In: ICCV.

Fan, L., Chen, Y., Wei, P., Wang, W., Zhu, S.-C., 2018. Inferring shared attention in social scene videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6460–6468.

Fan, L., Wang, W., Huang, S., Tang, X., Zhu, S.-C., 2019. Understanding human gaze communication by spatio-temporal graph reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5724–5733.

Fanelli, G., Dantone, M., Gall, J., Fossati, A., van Gool, L., 2013. Random forests for real time 3D face analysis. IJCV 101 (3), 437–458.

Fanelli, G., Weise, T., Gall, J., van Gool, L., 2011. Real time head pose estimation from consumer depth cameras. In: Joint PR Symp.. pp. 101–110.

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., Lu, C., 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Trans. Pattern Anal. Mach. Intell..

Feng, D., Rosenbaum, L., Timm, F., Dietmayer, K., 2019. Leveraging heteroscedastic aleatoric uncertainties for robust real-time lidar 3d object detection. In: Intelligent Vehicles Symp.. pp. 1280–1287.

Gong, W., Zhang, X., Gonzàlez, J., Sobral, A., Bouwmans, T., Tu, C., Zahzah, E.-h., 2016. Human pose estimation from monocular images: A comprehensive survey. Sensors 16 (12), 1966.

Grossi, G., Lanzarotti, R., Napoletano, P., Noceti, N., Odone, F., 2020. Positive technology for elderly well-being: A review. Pattern Recognit. Lett. 137, 61–70.

Gu, J., Yang, X., De Mello, S., Kautz, J., 2017. Dynamic facial analysis: From Bayesian filtering to recurrent neural network. In: CVPR. pp. 1531–1540.

Guo, H., Hu, Z., Liu, J., 2022. MGTR: End-to-end mutual gaze detection with transformer. In: Proceedings of the Asian Conference on Computer Vision. pp. 1590–1605.

Her, P., Manderle, L., Dias, P.A., Medeiros, H., Odone, F., 2023. Uncertainty-aware gaze tracking for assisted living environments. IEEE Trans. Image Process. 32, 2335–2347.

Hong, C., Chen, L., Liang, Y., Zeng, Z., 2021. Stacked capsule graph autoencoders for geometry-aware 3D head pose estimation. Comput. Vis. Image Underst. 208, 103224.

Hong, C., Yu, J., Zhang, J., Jin, X., Lee, K.-H., 2018. Multimodal face-pose estimation with multitask manifold deep learning. IEEE Trans. Ind. Inform. 15 (7), 3952–3961.

Hsu, H.-W., Wu, T.-Y., Wan, S., Wong, W.H., Lee, C.-Y., 2018. Quatnet: Quaternion-based head pose estimation with multiregression loss. IEEE Trans. Multimed. 21 (4), 1035–1046.

Ju, J., Zheng, H., Li, C., Li, X., Liu, H., Liu, T., 2022. AGCNNs: Attention-guided convolutional neural networks for infrared head pose estimation in assisted driving system. Infrared Phys. Technol. 123.

Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In: Adv. in Neural Information Processing Systems, vol. 30.

Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H., 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: Int. Work. on Benchmarking Facial Image Analysis Technologies.

Kukleva, A., Tapaswi, M., Laptev, I., 2020. Learning interactions and relationships between movie characters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9849–9858.

Kumar, A., Alavi, A., Chellappa, R., 2017. KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors. In: Int. Conf. on Automatic Face Gesture Recognition. pp. 258–265.

Lathuiliere, S., Juge, R., Mesejo, P., Munoz-Salinas, R., Horaud, R., 2017. Deep mixture of linear inverse regressions applied to head-pose estimation. In: CVPR.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, H., Fang, S., Zhang, Z., Li, D., Lin, K., Wang, J., 2021a. MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation. IEEE Trans. Multimed. 24, 2449–2460.

Liu, H., Liu, T., Zhang, Z., Sangaiah, A.K., Yang, B., Li, Y., 2022. Arhpe: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction. IEEE Trans. Ind. Inform. 18 (10), 7107–7117.

Liu, T., Wang, J., Yang, B., Wang, X., 2021b. NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom. Neurocomputing 436, 210–220.

Liu, H., Zhang, C., Deng, Y., Liu, T., Zhang, Z., Li, Y.-F., 2023. Orientation cues-aware facial relationship representation for head pose estimation via transformer. IEEE Trans. Image Process. 32, 6289–6302.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., et al., 2019. Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.

Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J., Lin, L., 2018. Lstm pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5207–5215.

Luvizon, D.C., Picard, D., Tabia, H., 2020. Multi-task deep learning for real-time 3D human pose estimation and action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43 (8), 2752–2764.

MacKay, D.J.C., 1992. A practical Bayesian framework for backpropagation networks. Neural Comput. 4 (3), 448–472.

Madrigal, F., Lerasle, F., 2020. Robust head pose estimation based on key frames for human-machine interaction. EURASIP J. Image Video Process. 2020, 1–19.

Marin-Jimenez, M.J., Kalogeiton, V., Medina-Suarez, P., Zisserman, A., 2019. Laeo-net: revisiting people looking at each other in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3477–3485.

Marín-Jiménez, M., Kalogeiton, V., Medina-Suárez, P., Zisserman, A., 2020. LAEO-net++: revisiting people looking at each other in videos. PAMI 1–16.

Marín-Jiménez, M.J., Zisserman, A., Eichner, M., Ferrari, V., 2014. Detecting people looking at each other in videos. Int. J. Comput. Vis. 106 (3), 282–296.

Martinez, G.H., Raaj, Y., Idrees, H., Xiang, D., Joo, H., Simon, T., Sheikh, Y., 2019. Single-network whole-body pose estimation. In: ICCV. pp. 6981–6990.

Moro, M., Marchesi, G., Hesse, F., Odone, F., Casadio, M., 2022. Markerless vs. Marker-based gait analysis: A proof of concept study. Sensors 22 (5).

Mukherjee, S., Robertson, N., 2015. Deep head pose: Gaze-direction estimation in multimodal video. IEEE Trans. Multimed. 17 (11), 2094–2107.

Nix, D., Weigend, A., 1994. Estimating the mean and variance of the target probability distribution. In: ICNN.

Rahmaniar, W., ul Haq, Q.M., Lin, T.-L., 2022. Wide range head pose estimation using a single RGB camera for intelligent surveillance. Sensors.

Ranjan, R., Patel, V.M., Chellappa, R., 2019. HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. PAMI 41, 121–135.

Recasens, A., Khosla, A., Vondrick, C., Torralba, A., 2015. Where are they looking? In: NIPS.

Ruan, Z., Zou, C., Wu, L., Wu, G., Wang, L., 2021. Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. IEEE Trans. Image Process. 30, 5793–5806.

Ruiz, N., Chong, E., Rehg, J.M., 2018. Fine-grained head pose estimation without keypoints. In: CVPR-W.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013. 300 Faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCV-W. pp. 397–403.

Saunderson, S., Nejat, G., 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. Int. J. Soc. Robotics 11, 575–608.

Schiavio, A., Gesbert, V., Reybrouck, M., Hauw, D., Parncutt, R., 2019. Optimizing performative skills in social interaction: Insights from embodied cognition, music education, and sport psychology. Front. Psychol. 10, 1542.

Shao, M., Sun, Z., Ozay, M., Okatani, T., 2019. Improving head pose estimation with a combined loss and bounding box margin adjustment. In: Int. Conf. on Automatic Face Gesture Recognition.

Shi, L., Zhang, Y., Cheng, J., Lu, H., 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: CVPR.

Song, Z., Yin, Z., Yuan, Z., Zhang, C., Chi, W., Ling, Y., Zhang, S., 2021. Attention-oriented action recognition for real- time human-robot interaction. In: ICPR. pp. 7087–7094.

Stahl, J.S., 1999. Amplitude of human head movements associated with horizontal saccades. Exp. Brain Res. 41–54.

Trabelsi, R., Varadarajan, J., Pei, Y., Zhang, L., Jabri, I., Bouallegue, A., Moulin, P., 2017. Robust multi-modal cues for dyadic human interaction recognition. In: Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes. ACM, pp. 47–53.

Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., Shao, L., 2021a. Deep 3D human pose estimation: A review. Comput. Vis. Image Underst. 210, 103225.

Wang, X., Zhang, J., Zhang, H., Zhao, S., Liu, H., 2021b. Vision-based gaze estimation: A review. IEEE Trans. Cogn. Dev. Syst. 14 (2), 316–332.

Xia, J., Zhang, H., Wen, S., Yang, S., Xu, M., 2022. An efficient multitask neural network for face alignment, head pose estimation and face tracking. Expert Syst. Appl..

Xin, M., Mo, S., Lin, Y., 2021. Eva-gcn: Head pose estimation based on graph convolutional networks. In: CVPR. pp. 1462–1471.

Xu, Y., Jung, C., Chang, Y., 2022. Head pose estimation using deep neural networks and 3D point clouds. Pattern Recognit. 121.

Yang, T.-Y., Chen, Y.-T., Lin, Y.-Y., Chuang, Y.-Y., 2019. FSA-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: CVPR.

Yang, W., Ouyang, W., Li, H., Wang, X., 2016. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3073–3082.

Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M., 2017. Towards large-pose face frontalization in the wild. In: ICCV.

Yu, J., Hong, C., Rui, Y., Tao, D., 2017. Multitask autoencoder model for recovering human poses. IEEE Trans. Ind. Electron. 65 (6), 5060–5068.

Zhang, H., Wang, M., Liu, Y., Yuan, Y., 2020. FDN: feature decoupling network for head pose estimation. In: AAAI. pp. 12789–12796.

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M., 2023. Deep learning-based human pose estimation: A survey. ACM Comput. Surv. 56 (1), 1–37.

Zhou, Y., Gregson, J., 2020. WHENet: Real-time fine-grained estimation for wide range head pose. In: BMVC.

Zhou, H., Hong, C., Han, Y., Huang, P., Zhuang, Y., 2021. MH pose: 3D human pose estimation based on high-quality heatmap. In: 2021 IEEE International Conference on Big Data (Big Data). IEEE, pp. 3215–3222.

Zhu, X., Liu, X., Lei, Z., Li, S.Z., 2019. Face alignment in full pose range: A 3D total solution. PAMI 41, 78–92.