HHP-Net: A light Heteroscedastic neural network for Head Pose estimation with uncertainty

Giorgio Cantarini[1,2], Federico Figari Tomenotti[1], Nicoletta Noceti[1], and Francesca Odone[1]

[1]MaLGa-DIBRIS, Università degli Studi di Genova, via Dodecaneso 35, 16146-IT Genova,Italy
[2]IMAVIS srl, via Trento 5/2, 16145-IT Genova,Italy
giorgio.cantarini@imavis.com, federico.figaritomenotti@edu.unige.it, nicoletta.noceti, francesca.odone@unige.it

Abstract

In this paper we introduce a novel method to estimate the head pose of people in single images starting from a small set of head keypoints. To this purpose, we propose a regression model that exploits keypoints computed automatically by 2D pose estimation algorithms and outputs the head pose represented by yaw, pitch, and roll. Our model is simple to implement and more efficient with respect to the state of the art – faster in inference and smaller in terms of memory occupancy – with comparable accuracy.
Our method also provides a measure of the heteroscedastic uncertainties associated with the three angles, through an appropriately designed loss function; we show there is a correlation between error and uncertainty values, thus this extra source of information may be used in subsequent computational steps. As an example application, we address social interaction analysis in images: we propose an algorithm for a quantitative estimation of the level of interaction between people, starting from their head poses and reasoning on their mutual positions.

1. Introduction

Nowadays, 2D human pose estimators are growing in popularity, likewise the number of effective algorithms available in the literature [5, 25, 9, 29], since they provide a richer output than simple people detection. Their potential applications are countless, ranging from motion analysis and action recognition, to human-machine interaction, and social interaction analysis (se for instance [36, 20, 37]).

A key element in several of the aforementioned applications is the estimation of the head direction [21, 1, 7, 14], that aims at computing the pose of human heads with respect to a reference (frontal) pose. When observing human agents with conventional cameras, eyes may be difficult to detect, and head direction is often used as a proxy to gaze, knowing that the eyeball orientation can differ by $\pm 35^{\circ}$ degree from the head orientation [38].
In this paper we estimate the head direction only relying on the sparse and meaningful semantic features provided by pose estimators, with no need of additional input (Fig. 1).

We formulate the problem as a multi-task regression and design a Heteroscedastic Neural Network to estimate the head pose expressed in Euler angles. Thanks to the use of an appropriately designed loss function, the method also associates an uncertainty value – learned from the data – with each angle. The concept of uncertainty provides an additional cue that may help the interpretation of the output of the network. We estimate aleatoric heteroscedastic uncertainty, that is uncertainty due to data and varying on different inputs. Our head pose estimation method, with a negligible additional effort in terms of space and time resources, can be seen as a *plug-in to any given pose estimator*, and 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 a potential to run on mobile architectures, and performs in real time (at about 100 fps).

As an example application of our estimates, we consider social interaction analysis: in particular we propose a light method based on the estimated head poses, for detecting pair of people looking at each other in images [23, 22].

In summary, the main contributions of the paper are thus the following:

- We introduce a very light (in space and time) head pose estimation method to be used as a plug-in for 2D human pose estimators in RGB images and able to associate an uncertainty with each estimated (pose) angle;

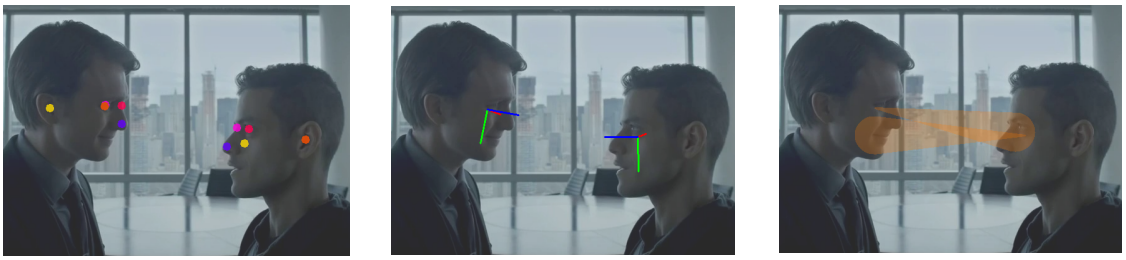

Figure 1. Pipeline of the proposed method: starting from the keypoints detected on the head by a human pose estimation method (left), our method provides the 3D head pose as a triplet of angles yaw, pitch and roll (center) and a measure of uncertainty associated with each angle in the triplet (right, for clarity we only derived a projection on the image plane from yaw and pitch using the Tait-Bryan angles, and represent the uncertainty as a cone around the estimated direction).

the estimated uncertainties represent a cue for model interpretation and strongly correlate with the estimation error.

- As a core element of the method, we propose a multi-task regression loss where the uncertainties act as weights of each sub-loss responsible for the estimation of each angle. To the best of our knowledge this is the first attempt to adopt the concept of uncertainty in multi-task regression for head pose estimation.
- We provide a thorough experimental assessment showing how our method is "ready" to use for mobile devices, has a space occupancy $\sim$12 times smaller than state-of-the-art while it provides comparable results, in the worst case with about 2 degrees of degradation.
- We present a proof-of-concept of the estimated head poses in detecting people looking at each other, as a first step for social interaction analysis; in this context the uncertainty contributes to obtain effective performances.

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 to mutual interaction, while Section 6 is left to a final discussion.

2. Related Work

Head pose estimation methods. Head pose estimation has been addressed by a number of relatively recent methodologies [27], with classical applications to Human-Machine Interaction or to social interaction analysis.
Some methods use additional information such as depth [11, 26] or time [15], but in this paper we will only refer to methods applicable to RGB images. Pose can be derived by an estimation of a 3D model [10], here we mention recent deep learning based methods, such as 3DDFA [43], a CNN able to fit a 3D model to an RGB image; FAN [3] is a state of the art facial landmark detection method, also estimating pose. These approaches propose complex computational pipelines and may obtain rather accurate results. One of the most recent challenges is in estimating pose directly from individual 2D images. On this respect, we start by mentioning a different but related task of estimating the 2D gaze. GazeFollow [32] is a two-pathway CNN architecture that estimates the apparent direction of 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 provide an estimate of the apparent direction of gaze is proposed in [7]: similarly to our work, the input is obtained by a 2D pose estimator.
3D head pose from images has been addressed in several works [8, 19], and nowadays it is often obtained by deep learning architectures that start from the output of face detectors: Shao et al [35] 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 [33]. FSA-net [39] is a two-stream multi-dimensional regression network able to provide accurate fine grained estimations.
We propose a 3D head pose estimation from RGB images, but instead of estimating pose directly from the image, we rely on the output of a human pose detector, similarly to the strategy proposed in [7] for 2D heading estimation. This allows us to design a very light architecture, still able to achieve accurate results.
In this section it is also worth mentioning multi-task approaches, exploiting the concept of training an architecture to achieve joint results, an approach proved to improve performances: KEPLER [18] predicts facial keypoints and pose with a modified GoogLeNet; a coarse pose is used to improve keypoints detection. Hyperface [31] simultaneously performs face and landmark detection, pose estimation and gender recognition. A very recent paper [4] pro-

poses a reformulation of the problem in terms of a rotation matrix to be used to formulate the output.

Uncertainty estimation. As recently observed in [42], where the WHENET architecture is proposed, head pose estimation is intrinsically harder on certain view-points. This paper extends [33], by changing the loss functions specifically addressing wide-range head pose, to perform well on lateral views. Instead our work follows the observation in [7]: certain view-points are associated with different levels of uncertainty, creating a large discrepancy in accuracy. This can be formalized with the concept of aleatoric heteroscedastic uncertainty [17], which depends on the inputs, and may be estimated from data. Conventional deep learning methods are unable to estimate the uncertainty of their inputs, for this Bayesian deep learning is becoming very popular on this respect [6, 2, 30]. In our method we propose a multi-task approach where a task is associated with one of the three pose angles, extending [17] to the multi-loss case. Indeed [6] 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.

3. The proposed method: HHP-Net

The starting point of our approach is the output of a pose detector (the literature today offers various alternatives, as OpenPose [5, 25], CenterNet [9], and PoseNet [29]) providing a set of keypoints roughly describing the pose of a human body in an image. These detectors commonly provide also a confidence on the keypoint location estimate, which represents an additional source of knowledge that can be injected in 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 i-th keypoint. Coordinates are centered with respect to their centroid and then normalized with respect to their corresponding 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 occur frequently 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

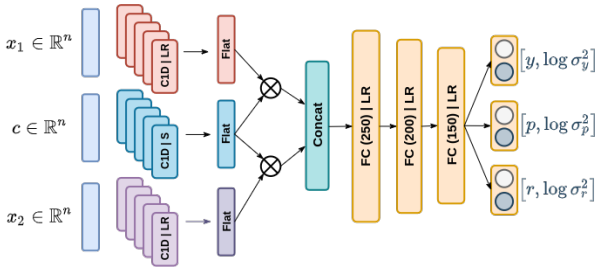

Figure 2. The light architecture of our approach. C1D = 1D Convolution, S = sigmoid, LR = leakyReLu, $\bigotimes$ = element-wise multiplication

$\mathbf{x}_1 = [x_1^1, \ldots, x_1^n]$, $\mathbf{x}_2 = [x_2^1, \ldots, x_2^n]$ and $\mathbf{c} = [c^1, \ldots, c^n]$. The input vectors are first processed in independent streams, with a 1D Convolutional layer followed by a non-linear activation: Leaky ReLU, to avoid vanishing gradient issues, for $\mathbf{x}_1$ and $\mathbf{x}_2$, and sigmoid activation for the confidence vector $\mathbf{c}$, to smoothly control the impact of different confidence values. After the convolution and the non-linear activation, from x_1 x_2 and c we obtain respectively x_1^*, x_2^* and c^*.

The latter are flattened and combined, using the element-wise multiplication to obtain two vectors $v_1 = x_1^* \otimes c^*$ and $v_2 = x_2^* \otimes c^*$, following the logic of the Confidence Gated Unit (CGU) proposed in [7]. The two gated outputs v_1 and 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.

3.2. Estimating the 3D head orientation

For training the network we adopt a loss function incorporating heteroscedastic aleatoric uncertainty. With respect to classical Neural Networks, an 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 with inherent keypoints detection which may be affected by difficult viewpoints or occlusions. Indeed, some poses are intrinsically noisier and more prone to self-occlusions.

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 $\mathbf{q} = [y, p, r]$, but also the uncertainty values associated with them $\mathbf{s} = [s_y, s_p, s_r]$

We use a multi-task loss function $\mathcal{L}_{HHP}$[1] for training our network, defined as follows:

[1]The derivation of the loss is in the supplementary material.

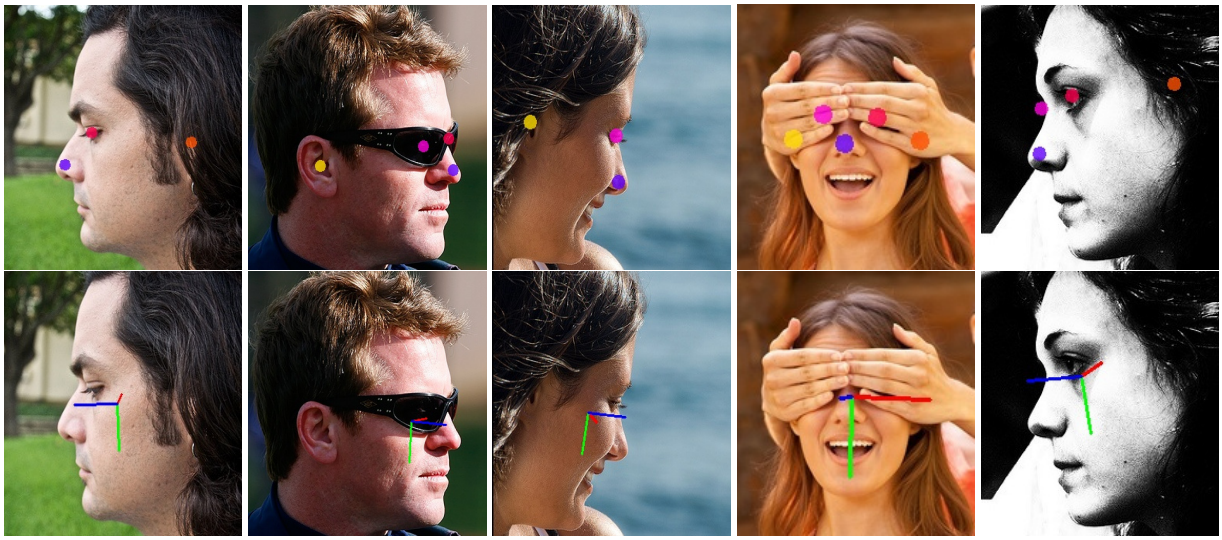

Figure 3. Examples on the AFLW2000 Dataset. Top: keypoints input of our estimate; bottom: estimated head pose.

$$\sum_{i \in \{y,p,r\}} \left(\frac{1}{2} \exp\left(-s_i\right) \left\| q_i - f_i\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}\right) \right\|^2 + \frac{1}{2} s_i \right)$$

where f_i is the $i-th$ component (associated with either yaw, pitch, or roll) of $\mathbf{f}$, the estimate obtained by the Heteroscedastic Neural Network, and $s_i = \log \sigma_i(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c})^2$ where σ_i^2 is the variance of a normal distribution we assume to have generated the noise [28]: learning the logarithm of the variance allows us to obtain a more stable solution, avoiding potential divisions by zero [17].

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

4. Experiments

4.1. Implementation details

In this work we adopt OpenPose [5] as a keypoints extractor, as it provides a good balance between efficiency (it performs in real-time) and effectiveness. Among the 25 body keypoints provided by OpenPose, 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 $\mathbf{x}_1 = [x_1^1, \ldots, x_1^5]$, $\mathbf{x}_2 = [x_2^1, \ldots, x_2^5]$ and $\mathbf{c} = [c^1, \ldots, c^5]$. 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 epochs that ranges from 100 to 1000 depending on the dataset using Adam as optimizer, with learning rate 0.001, and batch size of 64. The weights associated with the best validation loss have been selected as final model[2].

4.2. Datasets and protocols

We evaluate the effectiveness of our approach on three different datasets:

- BIWI [12] includes $\sim 15K$ images of 24 people acquired in a controlled scenario. The head pose orientation covers the range $\pm 75^o$ for the yaw angle, and $\pm 60^o$ for the pitch. The ground truth has been obtained by fitting a 3D face model.
- AFLW-2000 [40] contains the first 2000 images of the in-the-wild AFLW dataset [24], 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 [34] is a collection of different in-the-wild datasets, grouped and re-annotated to account for different types of variability, 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.

According to previous works [33, 39, 35], 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 [10].

[2]Code and pre-trained weights are available at https://github.com/cantarinigiorgio/HHP-Net

4.3. Method assessment

We start the discussion on the experimental analysis showing samples of qualitative results in Fig. 3, where we derived the directions on the image plane according to the Tait-Bryan angles. In spite of the sparseness of the input representation (keypoints overlaid on the original image, top row), the estimated pose is very accurate on a variety of challenging conditions (bottom row).

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

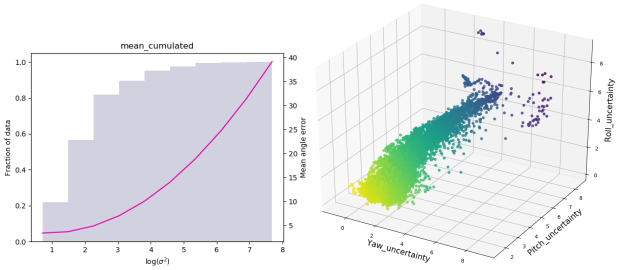

Figure 4. Left: Cumulative angular error as a function of the average uncertainty. Right: Linear correlation among the three uncertainties (Pearson correlation between (uncertainties of) *yaw* and *pitch* is 0.72, the one between *yaw* and *roll* is 0.78, the one between *pitch* and *roll* is 0.92)

On the quality of the uncertainty estimations. We report some qualitative experiments to highlight the properties of the uncertainty measures incorporated in our model. Fig. 4 (left) reports a cumulative analysis on the amount of data associated with a given uncertainty, highlighting how the average error grows with the uncertainty – in agreement with [7]. Similarly to what observed in [13], we also notice a strong correlation between the uncertainty values associated with the three predicted angles: in Fig. 4 (right) we provide a visual representation where a linear correlation between the uncertainties associated with the three predictions can be easily appreciated.

Uncertainty estimation and model interpretation. in Fig. 5 we report the trend of the uncertainty associated with the prediction obtained from a video sequence where a subject rotates the head offering different test poses to the method. Representative frames – that provide 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 keypoints on the face – the uncertainty tends to be higher. The lowest uncertainty values are associated with frontal views, the ones providing the most visible and non-ambiguous keypoints.
Inspired by these observations, we now evaluate the dependence of the uncertainty and the error on the quality

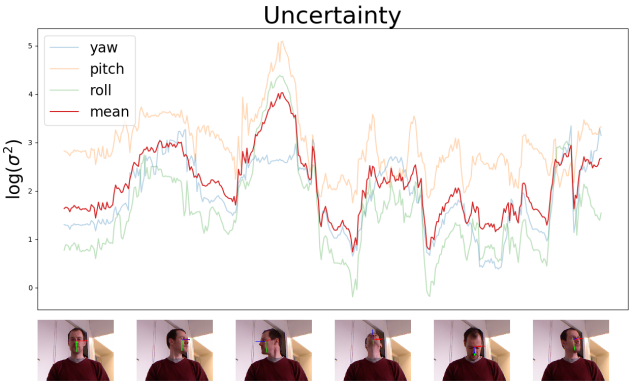

Figure 5. Uncertainty values are influenced by the head pose; the red line is the mean between the three uncertainties associated to each angle.

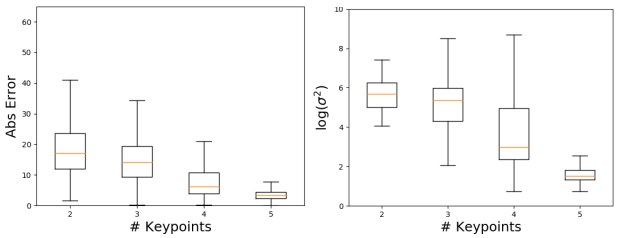

Figure 6. Performance of our method (left: mean angular error, right: uncertainty) with respect to the number of input points provided in input. Training: 300W-LP, Test: BIWI (plots refer to the latter).

and quantity of the input keypoints.

On the number of keypoints In Fig. 6 we analyse the performance of our method in terms of uncertainty values (right) and absolute angular error (left) as we group the input data according to the number of keypoints provided in input. Since in the worst case OpenPose provides at least three keypoints for the dataset used in this experiment, we randomly dropped points from the input to evaluate the behavior of the method in more challenging scenarios.

When all the 5 keypoints are available, the uncertainty is compactly lower 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 keypoints available in input – the uncertainty progressively decreases, but we also have a higher degree of variability, as some keypoints configurations are more significant than others and thus the amount of information they provide to the model may be uneven, also reflecting the concept that the noise could be different for each input sample.

Removing the uncertainty: an ablation study In this final experiment, we perform an ablation study by removing the uncertainty from our model. To this purpose, we consider two variations: in the first (MSE) we directly regress the

Table 1. Comparison among different loss functions (see text). **All errors are expressed in degrees** (°): MAE = Mean Absolute Error (the subscript refers to the loss)

Train	Val	MAE_{MSE}	MAE_{COMB}	MAE_{UNC}
BIWI	BIWI	3.70	3.80	**3.68**
300WLP	BIWI	5.28	5.88	**5.18**
300WLP	AFLW2000	8.07	8.04	**7.70**
AFLW	AFLW2000	6.26	6.18	**6.16**

three angles adopting a loss computed as the sum of the Mean Squared Error on each angle. In the second (COMB) we employ an alternative loss function that jointly solves a regression and a classification task, which has been successfully applied to the same estimation task [33].
In Tab. 1 we report the angular errors we obtain with the three different loss functions. As it can be observed, learning the angles associated with the uncertainty provides the best average performance[3].

4.4. 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. We always report the performance provided by our method in its largest variant (0.4MB) and we apply the same rule for alternative approches.
The analysis is reported in Tab. 2, 3, and 4. 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 $\sim$12 times smaller than the closest model in literature. According to the protocol followed by other works – all requiring a face detector but not including its size in their analysis[4] – the size of our model does not include the pose estimator. In terms of performances, Tab. 2 reports a comparison with respect to Protocol P2 (BIWI dataset for training and test): the results we obtain are superior to [26, 8, 10] and slightly below [15, 19, 39, 41] (less than 0.1 degree of difference for the first three, less than 0.4 for the latter).

Tab. 3 refers to Protocol P1 (training carried out on 300W-LP, BIWI for 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 [43, 18, 16, 3] and methods based on analysing RGB image portions obtained by face detectors, such as [35, 33, 39].

We share with the latter group the main motivation of 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. Tab. 3 reports results that are more accurate than all methods with the exception of FSA-Caps, although the difference is on average only slighlty 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 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, Tab. 4 follows again Protocol P1, on a more complex test set (as images are acquired in a less controlled environment). In this case our methodology is reporting slightly worse results, but with a loss always less that 3 degrees on average. We noticed this is due in particular to keypoint detection errors, as the the synthetic data manipulation (see examples in Fig. 7) introduced artifacts. 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, among the very recent approaches, our comparative analysis does not include WHENET [42] since it addresses specifically wide-angle cases and thus it adopts a different training set, not available at the moment, in its experimental analysis; its space occupancy is not reported in the paper, although the implementation they provide is 18 MB. We also do not report a comparison with [4], since their problem definition and their evaluation is not immediately comparable with state of the art.

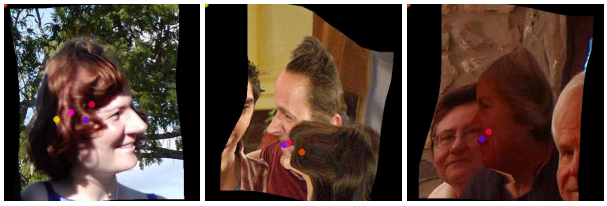

Figure 7. Examples of failures in the keypoints detection on the 300W-LP dataset caused by the artifacts due to the synthetic data manipulation; the points are clearly in the wrong positions.

5. An application to social interaction analysis

We finally discuss a task where our method finds a natural application, i.e. the analysis of social interaction, for which the gaze or the head directions represent a strong visual cue of non-verbal human-human communication [1]. 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).

[3] In the supplementary material details on the experiments are reported.
[4] as benchmark datasets provide the face bounding boxes.

Table 2. Comparison following Protocol P2: BIWI train-test. **All errors are expressed in degrees** (°): err_y= yaw error, err_p=pitch error, err_r= roll error. MB = Models size in megabytes.

Method	MB	err_y	err_p	err_r	MAE
D-HeadPose [26]	-	-	5.67	5.28	-
Drounard et al [8]	-	4.9	5.9	4.7	5.16
Fanelli et al.[10]	-	3.8	3.5	5.4	4.23
DFA[15]	500	3.91	4.03	3.03	3.66
DMLIR [19]	500	3.12	4.68	3.07	3.62
FSA-Net [39]	5.1	2.89	4.29	3.60	3.60
FND [41]	5.8	3.0	3.98	2.88	3.29
Our approach	**0.4**	3.04	4.79	3.21	3.68

Table 3. Comparison following Protocol P1: 300W-LP train, BIWI test. **All errors are expressed in degrees** (°): err_y= yaw error, err_p=pitch error, err_r= roll error, MAE = Mean Absolute Error. MB = Models size in megabytes.

Method	MB	err_y	err_p	err_r	MAE
3DDFA [43]	-	36.2	12.3	8.78	19.1
KEPLER [18]	-	8.8	17.3	16.2	13.9
Dlib (68 points)[16]	-	16.8	13.8	6.19	12.2
FAN (12 points) [3]	183	8.53	7.48	7.63	7.89
Shao(K=0.5)[35]	93	4.59	7.25	6.15	6.00
Ruiz [33])(α=2)	95.9	5.17	6.98	3.39	5.18
FSA-Caps-Fusion [39]	5.1	4.27	4.96	2.76	4.00
Our approach	**0.4**	4.14	7.00	4.40	5.18

Table 4. Comparison following Protocol P1: 300W-LP train, AFLW 2000 test († = Trained on (AFLW - AFLW2000)) **All errors are expressed in degrees** (°): err_y= yaw error, err_p=pitch error, err_r= roll error, MAE = Mean Absolute Error. MB = Models size in megabytes.

Method	MB	err_y	err_p	err_r	MAE
Dlib (68 points)[16]	-	23.1	13.6	10.5	15.8
FAN (12 points)[3]	183	8.53	7.48	7.63	7.88
Ruiz[33])(α=2)	95.9	6.92	6.64	5.67	6.41
Shao(K=0.5)[35]	93	4.59	7.25	6.15	6.00
FSA-Caps-Fusion [39]	5.1	4.50	6.08	4.64	5.07
Our approach	**0.4**	5.26	10.12	7.73	7.70
Our approach†	**0.4**	7.40	6.63	4.47	6.16

LAEO algorithm. Fig. 8 provides a visual sketch with 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 pose estimated for each of them (3 Euler angles) and obtain a projection of the corresponding direction on the image plane. More in details for the subject A, given the triplet of angles (y_A, p_A, r_A), we derive the end-point of the head direction on the image plane (x'_A, y'_A) as $x'_A = \sin(y_A)$ and $y'_A = -\cos(y_A)\sin(p_A)$.
Then, we estimate a measure of interaction between each pair of people by proposing a simple but effective method.

Table 5. The performance of our method for LAEO detection on the UCO-LAEO dataset. AP is estimated as in [22], $\tau = 0.93$.

Method	PREC	REC	F	AP
LAEO-Net [23]	–	–	–	0.80
LAEO-Net++ [22]	–	–	–	0.87
Baseline	0.77	0.80	0.78	0.86
with uncertainty	0.80	0.72	0.76	0.88

We consider 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 a thresholding on such measure allows us to detect LAEO pairs.
We build on this baseline method exploiting the knowledge we derive from the uncertainty associated with the 3D angles. Given the triplets of uncertainties associated with the two heads poses, (s^y_A, s^p_A, s^r_A) and (s^y_B, s^p_B, s^r_B), we compute the averages, $\hat{s}_A = \frac{1}{2}(s^y_A + s^p_A)$ and $\hat{s}_B = \frac{1}{2}(s^y_B + s^p_B)$; 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 not reliable. 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 $\delta = 7$). The method is sketched in Algorithm 1.

LAEO assessment. We evaluate our method on the UCO-LAEO dataset [23], that includes sequences from four popular TV shows in the form or 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 labeled 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 finally used to detect the LAEO pairs.

We report in Tab 5 the performance provided by our baseline method and the one incorporating the uncertainty on the test set. The results suggest that using the prior knowledge derived from the uncertainty allows us to significantly reduce the number of false positive (-6%, with a slight increase of the precision) to the price of a small re-

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 $LAEO_{value}$

duction 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 [23, 22].

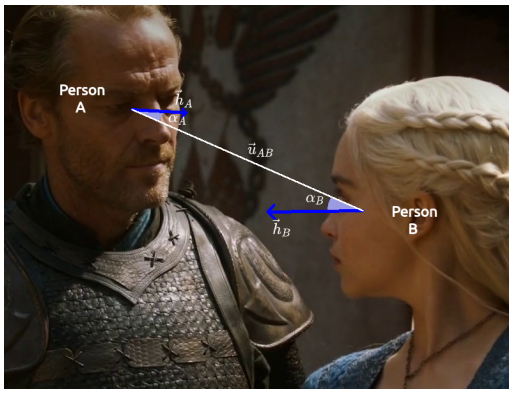

Figure 8. A visual sketch with our formulation of the LAEO detection task (for readability the vectors are denoted with arrows).

6. Discussion

In this work we introduced a method for head pose estimation from the head keypoints extracted from RGB images, that provides the head pose as a triplet of Euler angles, each one 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 of 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, each one related with the estimation of a certain angle and corresponding uncertainty.
We provided a thorough experimental assessment, where we discussed the connections between estimation error and uncertainty, that favor the interpretability of our model.

Figure 9. Examples of LAEO detections. The arrows represent the head direction estimated by HHP-Net and projected on the image plane, and is 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 case of LAEO, it specifies the identifier of the other interacting person. The identifiers are in red close the subjects.

We also compared our method with state-of-art approaches, showing comparable or slightly lower performance while providing the lightest method currently available. As an example applications, we discussed the application to social interaction analysis in images.
Our work will be extended in different directions. A straightforward one will be to enlarge the number of input points to fully exploit the information provided by the pose detector. This may favor on the one hand the robustness of the method in more challenging situations, and on the other the possibility of considering it as a first step for an activity recognition pipeline, in particular when considering social activities or actions involving the interaction with the environment or with other people. With reference to the latter task, we are currently working on the design of more refined measures of social interaction between small groups of people, incorporating more explicitly the uncertainty and the knowledge derived from the 3D head pose estimator, that may help to disambiguate challenging group configurations.

Acknowledgement

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

References

[1] Andrea Abele. Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior*, 10(2):83–101, 1986.

[2] Vijay Badrinarayanan Alex Kendall and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 57.1–57.12. BMVA Press, September 2017.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *WACV*, 2021.

[5] 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.

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

[7] Philipe A. Dias, Damiano Malafronte, Henry Medeiros, and Francesca Odone. Gaze estimation for assisted living environments. In *WACV*, 2020.

[8] Vincent Drouard, Silèye Ba, Georgios Evangelidis, Antoine Deleforge, and Radu Horaud. Head Pose Estimation via Probabilistic High-Dimensional Regression. In *IEEE International Conference on Image Processing, ICIP 2015*, Proceedings of the IEEE International Conference on Image Processing, pages 4624–4628, Quebec City, QC, Canada, Sept. 2015. IEEE.

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[10] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. van Gool. Random forests for real time 3D face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013.

[11] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In Rudolf Mester and Michael Felsberg, editors, *Pattern Recognition*, pages 101–110, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[12] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. Real time head pose estimation from consumer depth cameras. In *Joint pattern recognition symposium*, pages 101–110. Springer, 2011.

[13] Di Feng, Lars Rosenbaum, Fabian Timm, and Klaus Dietmayer. Leveraging heteroscedastic aleatoric uncertainties for robust real-time lidar 3d object detection. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1280–1287. IEEE, 2019.

[14] Giuliano Grossi, Raffaella Lanzarotti, Paolo Napoletano, Nicoletta Noceti, and Francesca Odone. Positive technology for elderly well-being: A review. *Pattern Recognition Letters*, 137:61–70, 2020.

[15] J. Gu, X. Yang, S. De Mello, and J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1531–1540, 2017.

[16] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[18] A. Kumar, A. Alavi, and R. Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 258–265, 2017.

[19] Stephane Lathuiliere, Remi Juge, Pablo Mesejo, Rafael Munoz-Salinas, and Radu Horaud. Deep mixture of linear inverse regressions applied to head-pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[20] Diogo C. Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2752–2764, 2021.

[21] Francisco Madrigal and Frederic Lerasle. Robust head pose estimation based on key frames for human-machine interaction. *EURASIP Journal on Image and Video Processing*, (1):13, 2020.

[22] Manuel Marín-Jiménez, Vicky Kalogeiton, Pablo Medina-Suárez, and Andrew Zisserman. Laeo-net++: revisiting people looking at each other in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2020.

[23] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019.

[24] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[25] Gines Hidalgo Martinez, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-network whole-body pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6981–6990. IEEE, 2019.

[26] Sankha S. Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video.

IEEE Transactions on Multimedia, 17(11):2094–2107, Nov. 2015. "This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014277/1, the MOD University Defence Research Collaboration in Signal Processing.".

[27] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):607–626, Apr. 2009.

[28] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, 1994.

[29] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.

[30] Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification, 2018.

[31] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019.

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

[33] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *CVPR Workshops*, 2018.

[34] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[35] Mingzhen Shao, Zhun Sun, Mete Ozay, and Takayuki Okatani. Improving head pose estimation with a combined loss and bounding box margin adjustment. In *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*, Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019. Institute of Electrical and Electronics Engineers Inc., May 2019.

[36] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[37] Ziyang Song, Ziyi Yin, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. Attention-oriented action recognition for real- time human-robot interaction. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7087–7094, 2021.

[38] J. S. Stahl. Amplitude of human head movements associated with horizontal saccades. *Experimental Brain Research*, (1):41–54, 1999.

[39] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *In Proceeding of International Conference on Computer Vision*, Venice, Italy, October 2017.

[41] Hao Zhang, Mengmeng Wang, Yong Liu, and Yi Yuan. FDN: feature decoupling network for head pose estimation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12789–12796. AAAI Press, 2020.

[42] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. 2020.

[43] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):78–92, Jan 2019.