



# On the representation and methodology for wide and short range head pose estimation

Alejandro Cobo <sup>a</sup>, Roberto Valle <sup>a</sup>, José M. Buenaposada <sup>b,\*</sup>, Luis Baumela <sup>a</sup>

<sup>a</sup> Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte, 28660, Madrid, Spain

<sup>b</sup> Universidad Rey Juan Carlos, Calle Tulipán s/n, Móstoles, 28933, Madrid, Spain

## ARTICLE INFO

### Keywords:

Short- and wide-range head pose estimation  
Orientation representation  
Error metrics  
Cross-data set evaluation methodology

## ABSTRACT

Head pose estimation (HPE) is a problem of interest in computer vision to improve the performance of face processing tasks in semi-frontal or profile settings. Recent applications require the analysis of faces in the full 360° rotation range. Traditional approaches to solve the semi-frontal and profile cases are not directly amenable for the full rotation case. In this paper we analyze the methodology for short- and wide-range HPE and discuss which representations and metrics are adequate for each case. We show that the popular Euler angles representation is a good choice for short-range HPE, but not at extreme rotations. However, the Euler angles' gimbal lock problem prevents them from being used as a valid metric in any setting. We also revisit the current cross-data set evaluation methodology and note that the lack of alignment between the reference systems of the training and test data sets negatively biases the results of all articles in the literature. We introduce a procedure to quantify this misalignment and a new methodology for cross-data set HPE that establishes new, more accurate, SOTA for the 300W-LP/Biwi benchmark. We also propose a generalization of the geodesic angular distance metric that enables the construction of a loss that controls the contribution of each training sample to the optimization of the model. Finally, we introduce a wide range HPE benchmark based on the CMU Panoptic data set. code:[https://github.com/pr-upm/opal23\\_headpose](https://github.com/pr-upm/opal23_headpose)

## 1. Introduction

Head Pose Estimation (HPE) aims to compute the three-dimensional orientation of human heads in images or videos. This is a problem that has been widely studied in computer vision (CV) because it is a key element in many facial analysis workflows [1]. Present state-of-the-art (SOTA) solutions achieve a Mean Absolute Euler angle Error (MAE) below 4° [2–4] in realistic cross-data set experiments with Short Range Head Poses (SRHP) involving semi-frontal and profile faces with yaw angles roughly between ±90°. This is appropriate for boosting the performance of facial analysis algorithms that recognize faces [5,6], estimate facial attributes [7] or facial expressions [8] (see Fig. 1). However, they fail in the context of, e.g., driver monitoring [9], group interactions [10] and surveillance [11] applications, involving Wide Range Head Poses (WRHP) with yaw angles of up to ±180° (see Fig. 1). For this reason, the WRHP estimation problem is a topic of renewed interest [12,13].

The immediate application of the traditional SRHP training and evaluation methodology to the WRHP setting produces HPE models with poor performance. The most popular approach to SRHP estimation involves the use of Euler angles to represent the orientation of the

head and the Mean Absolute Error (MAE) to measure the estimation error [1]. Although Euler angles have been widely adopted in the robotics community, when used in a WRHP setting, their representation is discontinuous [14] and ambiguous, in presence of the so-called *gimbal lock*. In this case, the MAE of two orientations is not a good measure of their distance because two nearby rotations may have a large MAE (see Section 3.1). Further, a discontinuous orientation representation poses a difficult learning problem and in practice models with such representation perform worse than those using a continuous one [14]. So, to train a WRHP estimation model, we need a continuous orientation representation and a proper metric [15] to define the training algorithm loss function and performance evaluation measure.

In this paper we study the methodology for training and evaluating HPE algorithms to propose a suitable representation, loss function and evaluation metric for both SRHP and WRHP problems (see Fig. 2). The gimbal lock is the reason given for discarding the use of the Euler angles representation [4,19–21] in HPE problems. In Sections 5.3 and 5.4 we analyze this issue and show that the gimbal lock is not a problem for HPE representation. In fact, Euler angles are an excellent choice in SRHP estimation problems. It is the issue of discontinuity arising

\* Corresponding author.

E-mail address: [josemiguel.buenaposada@urjc.es](mailto:josemiguel.buenaposada@urjc.es) (J.M. Buenaposada).

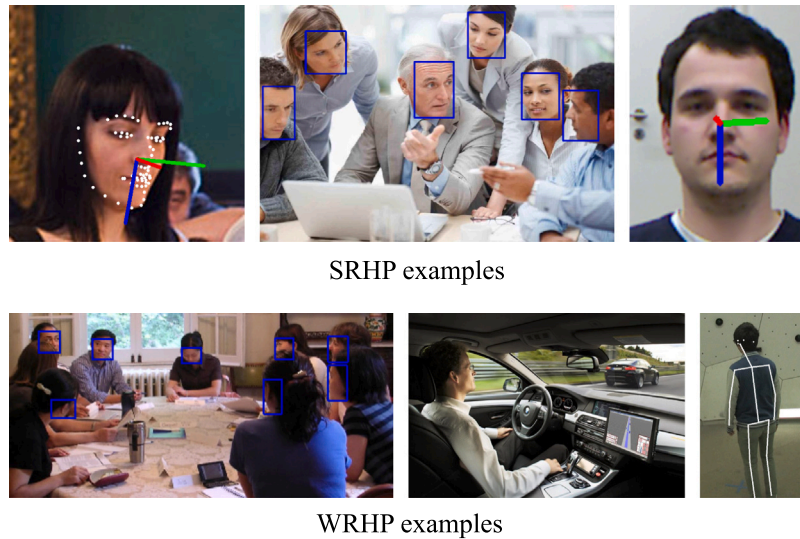


Fig. 1. Applications involving SRHP and WRHP configurations. Images from 300W-LP [16], WIDER Face [17], Biwi [18] and CMU Panoptic [10] data sets.

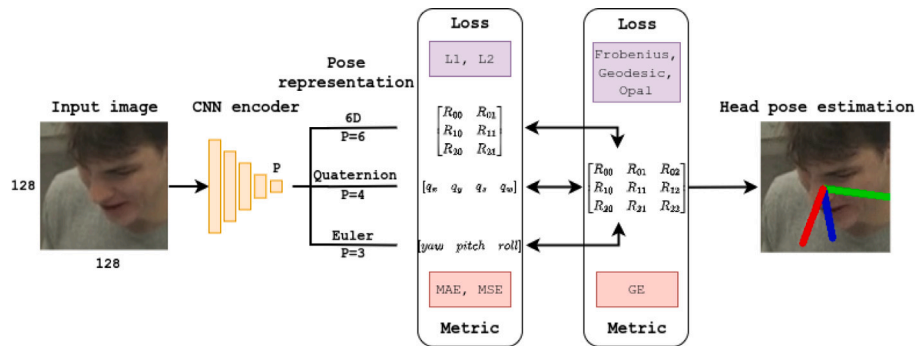


Fig. 2. Concept diagram of our analysis. Given an RGB image containing a cropped face, the HPE estimation algorithm produces a pose representation,  $p$ . Independently of the internal representation used by the model, predictions can be converted to Euler/quaternion angles or rotation matrices and measure the estimation error using different metrics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

at extreme rotations that prevents Euler angles, and the traditional quaternion-based alternative, from being used in the WRHP setting.

Concerning the evaluation metric, we show that it is the gimbal lock issue that prevents MAE, the most popular error evaluation measure, *e.g.*, [1,13,19,20], from being used for HPE in SRHP and WRHP settings. We propose the use of the angular geodesic distance [15], valid in any angular range. Based on the interpretability of this metric, we propose Opal, a loss function that allows to control the contribution of each sample to the model training as a function of its geodesic distance to the ground truth.

The most realistic methodology for evaluating HPE is a cross-data set procedure, in which the model is trained with one data set and evaluated with a different one [2–4,20,22]. In this context, a critical issue that has been ignored so far in the literature is the lack of alignment between training and test data set reference systems. In Section 3.3 we introduce a procedure to quantify this misalignment and establish a new SOTA in the popular cross-data set HPE evaluation methodology that uses 300W-LP [16] for training and Biwi [18] for testing.

In summary, our contributions are:

- We analyze the HPE pipeline in terms of the orientation representation (Euler angles, Quaternions or rotation matrix), and distance used, and discuss when each are amenable either for SRHP or WRHP.
- We introduce a new loss function, Opal, based on a generalization of the geodesic distance between two rotation matrices, which

allows us to weigh the contribution of each image in the data set to the minimization depending on the distance to the ground truth.

- We propose a method to estimate the misalignment between train and test data set reference systems providing a more accurate evaluation protocol in cross-data set settings. We establish a new, more accurate, SOTA when training in 300W-LP and evaluating in Biwi.

## 2. Related work

Object pose estimation is a topic of interest not only in the face processing area [2,3,12,13] but in CV in general [23,24]. When we talk about HPE, we refer to the estimation of the 3 DoF representing the orientation of the head w.r.t. the camera (see Fig. 3). We can organize the HPE literature in two groups. The traditional SRHP approach, when the yaw angle is in the range  $[-90^\circ, 90^\circ]$  (frontal to profile faces) [2–4,19–21,25,26] and the WRHP [12,13] when the range of the yaw angle is in  $[-180^\circ, 180^\circ]$ .

To train a HPE model we need to define the orientation representation, the loss function and the evaluation metric between ground truth and estimated poses. In this section we review different approaches in the literature for each of these elements.

### 2.1. Head pose representation

The usual orientation representation in HPE is based on Euler angles [1,2,13,25,26]. The rotation matrix is broken down into three

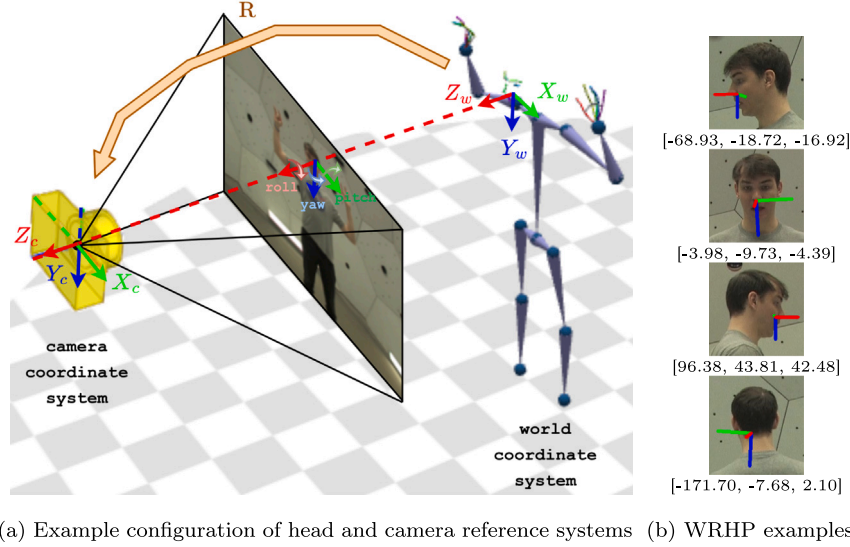


Fig. 3. WRHP means estimating the rotation matrix  $R$  to align the reference frame of the head with that of the camera. We show some WRHP results projecting a 3D axis onto the image plane coordinates. The text below represent the [yaw, pitch, roll] angles.

rotations around each camera axis, corresponding to the pitch, yaw and roll pose angles (see Fig. 3). In this configuration the gimbal lock problem (see Section 3.1) occurs when the face is in profile orientation, so the yaw angle is  $90^\circ$  or  $-90^\circ$ . Quaternions are also a popular alternative in object pose [23] and in HPE [19] because they have no gimbal lock. The most significant problem in WRHP estimation is that both Euler angles and quaternion representations are discontinuous, which makes learning with deep networks difficult [12,14]. In fact, any 3D orientation representation with less than 5 parameters is discontinuous [14].

Different representations for HPE have been introduced in the literature with the aim of solving the discontinuity problem [12,14,21]. Beyer et al. [12] propose a continuous representation for the head yaw angle, termed *bitemion*, which is modeled with the vector  $\mathbf{y} = [\cos(\phi), \sin(\phi)]$ , the first column of a 2D rotation matrix. This was later generalized to a continuous 3D rotation model with a 6D representation composed of the first two columns of a 3D rotation matrix [14], that has been used in the general object pose estimation [24] as well as in SRHP [4,20]. We will denote this representation in our paper as 6D. Cao et al. [21] propose a representation based on the full rotation matrix in the context of HPE.

In Section 5 we show that, although the 6D representation is immune to gimbal lock and discontinuities and therefore preferable in a WRHP setting, plain Euler angles are the best choice for SRHP estimation.

## 2.2. Loss functions and metrics for HPE

3D rotations arise in many contexts in the scientific literature and there are different functions to measure the distance between two rotations [15]. We need them to define the loss functions used to train the models and to evaluate their performance.

There are a few metrics to compare rotations in terms of Euler angles representation. Let the vector  $\mathbf{p} = (\alpha_p, \alpha_y, \alpha_r)$  represent a pose configuration with the three Euler angles and  $\hat{\mathbf{p}}$  the pose estimated by a model. The MAE and Mean Squared Error (MSE) of an estimation are given by the summation extended to the data set of respectively  $g_{MAE}(\hat{\mathbf{p}}, \mathbf{p}) = \|\hat{\mathbf{p}} - \mathbf{p}\|_1$  and  $g_{MSE}(\hat{\mathbf{p}}, \mathbf{p}) = \|\hat{\mathbf{p}} - \mathbf{p}\|_2^2$ , where  $\|\cdot\|_p$  denotes the  $L_p$  norm. There are other metrics that take into account the periodicity of angular representation. In this case the  $i$ -normed difference between two angles  $a, b$ , is given by  $d_i(a, b) = \min\{\|a - b\|_i, \|360^\circ - |a - b|\|_i\}$ , and define alternative metrics such as the Euclidean Distance between

Table 1  
Comparison of the most relevant HPE works with our methodology.

Method	Continuous representation	Gimbal lock-free metric	Wide range evaluated
HopeNet [25]	×	×	×
FSA-Net [26]	×	×	×
WHENet [13]	×	×	✓
TriNet [21]	✓	✓	×
6DRepNet [20]	✓	×	×
<b>Ours</b>	✓	✓	✓

Euler angles [15],  $g_{EUC}(\hat{\mathbf{p}}, \mathbf{p}) = \|\mathbf{d}_1(\hat{\mathbf{p}}, \mathbf{p})\|_1$ , and the, so-called, wrapped yaw distance [13]  $g_w(\hat{\alpha}_y, \alpha_y) = d_2(\hat{\alpha}_y, \alpha_y)$ .

As we discuss in Section 3.1, all distances based on Euler angles can lead to erroneous results in presence of gimbal lock. In the usual pitch–yaw–roll representation used in HPE, this happens when the face is close to a profile configuration,  $\alpha_y \approx \pm 90^\circ$ .

Other metrics directly compare rotation matrices and have also been used in HPE [4,21]. Some are based on the Frobenius norm, such as the *chordal distance* [27],  $g_F(\mathbf{R}, \hat{\mathbf{R}}) = \|\mathbf{R} - \hat{\mathbf{R}}\|_F$ , its squared form [21], the deviation from identity matrix,  $g_{DI}(\hat{\mathbf{R}}, \mathbf{R}) = \|\mathbf{I} - \hat{\mathbf{R}}\mathbf{R}^T\|_F$  [4,15], or the *geodesic distance* [14,15,27],

$$g_{GE}(\hat{\mathbf{R}}, \mathbf{R}) = \cos^{-1} \left( \frac{\text{tr}(\hat{\mathbf{R}}\mathbf{R}^T) - 1}{2} \right). \quad (1)$$

Both  $g_{GE}$  and  $g_{DI}$  are proper metrics in  $\text{SO}(3)$  and quantify the rotation required to bring  $\hat{\mathbf{R}}$  in coincidence with  $\mathbf{R}$  [15]. The former, however, has a direct interpretation as an angle. Thus, we will use the geodesic distance between the predicted rotation matrix for the  $i$ th image,  $\hat{\mathbf{R}}_i$ , and its ground truth,  $\mathbf{R}_i$ , for all samples in a data set to evaluate the quality of a model. We denote this measure as *Geodesic Error* (GE):

$$f_{GE} = \frac{1}{N} \sum_{i=1}^N g_{GE}(\hat{\mathbf{R}}_i, \mathbf{R}_i), \quad (2)$$

where  $N$  is the number of images in the data set.

In summary, to build a model adequate for WRHP estimation we have to select a representation that is continuous, use a correct metric and evaluate it in a wide range data set. To the best of our knowledge, as we can see in Table 1, we are the first to use a continuous representation with an accurate metric free from gimbal lock, and evaluate results using a WRHP data set.

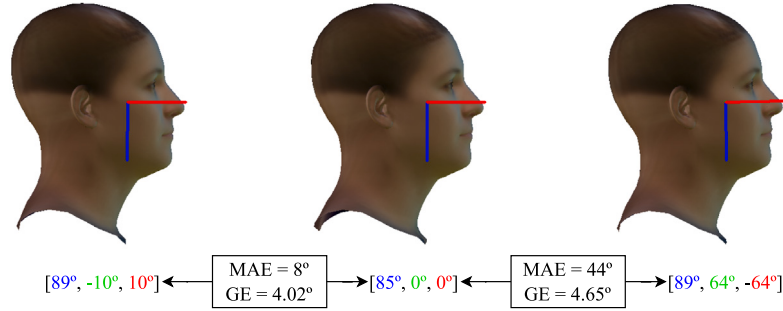


Fig. 4. All faces have visually very similar configuration but MAE is very large due to gimbal lock. However, the geodesic distance is coherent. Color code: [yaw, pitch, roll]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

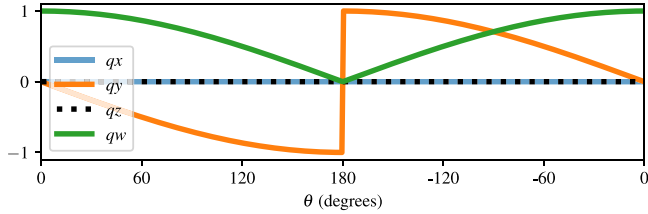


Fig. 5. Discontinuity in quaternions under a rotation around the yaw axis (from  $0^\circ$  to  $360^\circ$ ). Component  $q_y$  shows an abrupt change from  $-1$  to  $+1$  when the yaw reaches  $180^\circ$ .

### 3. Representation and methodology for HPE

The accuracy of a HPE algorithm depends on the range of possible head orientations and the right choice of rotation representation, error metric and evaluation methodology. In this section we first revisit the gimbal lock problem and discuss its impact in the Euler angle representation and the MAE. Finally, we discuss and solve the problem of miss-alignment in cross-data set HPE evaluation.

#### 3.1. The gimbal lock

There are different ways of decomposing a rotation matrix in three Euler angles representations. In HPE the representation is built by rotating first around the  $X$  axis of the camera coordinate system, pitch angle, then around the  $Y$  axis, yaw angle, and then around the  $Z$  axis, roll angle (see Fig. 3) and the head reference system is configured in such a way that the rotation matrix,  $R$ , of a head facing frontally to the camera is the identity.

So, a rotation matrix is the product of three elementary rotations matrices that follow the pitch, yaw, roll order, producing

$$R(\alpha_p, \alpha_y, \alpha_r) = \begin{pmatrix} \cos \alpha_r & \sin \alpha_r & 0 \\ -\sin \alpha_r & \cos \alpha_r & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \alpha_y & 0 & -\sin \alpha_y \\ 0 & 1 & 0 \\ \sin \alpha_y & 0 & \cos \alpha_y \end{pmatrix} \\ \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha_p & \sin \alpha_p \\ 0 & -\sin \alpha_p & \cos \alpha_p \end{pmatrix},$$

where  $\alpha_y$ ,  $\alpha_p$  and  $\alpha_r$  refer to yaw, pitch and roll angles respectively.

The use of Euler angles in a WRHP setting introduces two problems: the gimbal lock and the lack of continuity in the representation.

The first issue occurs when the yaw angle reaches  $\pm 90^\circ$ . This causes the other two axes to align and the representation collapses because one degree of freedom is lost. Algebraically we get

$$R\left(p, \frac{\pi}{2}, r\right) = \begin{pmatrix} 0 & \sin(\alpha_p + \alpha_r) & -\cos(\alpha_p + \alpha_r) \\ 0 & \cos(\alpha_p + \alpha_r) & \sin(\alpha_p + \alpha_r) \\ 1 & 0 & 0 \end{pmatrix}. \quad (3)$$

As we can see in Eq. (3), pitch and roll angles become indistinguishable and a given orientation may have infinite representations in terms of Euler angles that satisfy the linear relation

$$\alpha_p + \alpha_r = \alpha. \quad (4)$$

So, given any pitch and roll configuration satisfying Eq. (4), we can immediately recover  $R$  from Eq. (3). The gimbal lock is not a significant problem with Euler angles representation, if our model is able to learn the dependency between the angles of the two aligned axes. Previous works [19,21] changed the representation in SRHP because of gimbal lock. In Section 5.3 we show experimentally that a CNN using Euler angles representation is able to produce accurate estimations in a gimbal lock configuration.

However, Euler angles cannot be used to measure the distance between two rotations. First because in the gimbal lock configuration one pose may be represented by an infinite number of different pitch and roll values. Second, in configurations close to the gimbal lock, *i.e.*,  $\alpha_y = \frac{\pi}{2} + \delta$ , for a small  $\delta$ ,

$$R\left(p, \frac{\pi}{2} + \delta, r\right) = \begin{pmatrix} -\delta \cos \alpha_r & \sin(\alpha_p + \alpha_r) & -\cos(\alpha_p + \alpha_r) \\ \delta \sin \alpha_r & \cos(\alpha_p + \alpha_r) & \sin(\alpha_p + \alpha_r) \\ 1 & -\delta \sin \alpha_p & \delta \cos \alpha_p \end{pmatrix}, \quad (5)$$

two close poses may be represented by very different Euler angles (see Fig. 4). This completely invalidates the use of any distance measure using Euler angles, such as  $g_{MAE}$ ,  $g_{MSE}$ ,  $g_{EUC}$  and wrapped yaw distance,  $g_w$ , to compare two orientations either in SRHP or WRHP configurations. Conversely, the GE in Eq. (2), provides a coherent metric (see Fig. 4). Thus, the direct comparison of Euler angles should be abandoned in any HPE setting in favor of other well behaved metrics such as the geodesic distance.

#### 3.2. Discontinuity

The gimbal lock prompted some HPE researchers to use alternative representations without that problem, such as quaternions [19]. Both Euler angles and quaternions present discontinuity in their representation [14]. In Euler angles it appears when any angle reaches  $\pm 180^\circ$ . In HPE this happens when the head rotates in the yaw angle half a circle. One component of quaternions also shows a discontinuity in the most extreme yaw configuration (see Fig. 5).

The existence of a discontinuity affects learning because, near the discontinuity, similar facial appearances may have yaw angles (or quaternion components) with very different magnitude, therefore making learning more difficult [14]. Thus, discontinuity is a problem in a WRHP configuration but not in SRHP since, in the latter, faces are far from that problematic pose.

In Section 5.3 we show experimentally that for SRHP, Euler angles, quaternions and 6D representations achieve similar results. However, in a WRHP configuration, the continuous 6D representation provides significantly better results.

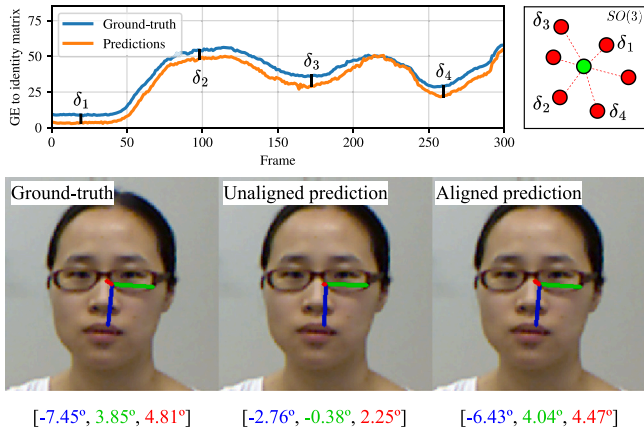


Fig. 6. Reference systems alignment problem. Top: (left) Unaligned video sequence. Time and GE from identity matrix shown in horizontal and vertical axes respectively; (right) Different alignment errors around the fixed alignment matrix. Bottom: Ground-truth, unaligned and aligned predictions in a specific frame with their corresponding Euler angles.

### 3.3. Reference systems alignment for cross-data set evaluation

Data sets present built-in biases that negatively affect the accuracy of the usual intra-data set evaluation protocols [28]. This is especially pronounced when evaluating a HPE algorithm, and it is mostly caused by the biases produced by the annotation procedure [2]. So, the most realistic approach to evaluate HPE involves a cross-data set approach, training the model in one data set and evaluating it in a different one [2–4,20,22]. This poses the additional problem of aligning the train and test coordinate reference systems. In Fig. 6 we show this problem for a video sequence from the Biwi data set. In the plot we can see that the angular distances between the ground truth and the predictions are composed of a fixed gap, constant for all the sequence, and a random estimation noise. The fixed term is caused by the misalignment between the train and test data set reference systems. The random estimation noise is the HPE error we want to measure. Although cross-data set evaluation is a common practice in the community, this issue has been ignored in the literature. Here we propose a procedure to compensate this misalignment.

Let  $\mathcal{R} = \{\mathbf{R}_1, \dots, \mathbf{R}_N\}$  be the set of ground truth rotation matrices of a data set with  $N$  images, or video sequence with  $N$  frames, and  $\hat{\mathcal{R}} = \{\hat{\mathbf{R}}_1, \dots, \hat{\mathbf{R}}_N\}$  be the set of predictions. If the estimations were perfect and the reference systems of train and test data sets were aligned, then  $\mathbf{R}_i^T \hat{\mathbf{R}}_i = \mathbf{I} \forall i$ , where  $\mathbf{I}$  is the identity matrix. However,  $\hat{\mathbf{R}}_i$ 's are affected by an estimation error matrix different for each image,  $\delta \mathbf{R}_i$ , and an alignment matrix,  $\Delta$ , common to all estimations in the set  $\hat{\mathcal{R}}$ , so  $\hat{\mathbf{R}}_i \delta \mathbf{R}_i \Delta = \mathbf{R}_i \forall i$ . To estimate  $\Delta$  we assume the  $\delta \mathbf{R}_i$ 's are random perturbations around the identity,  $\mathbf{I}$ , caused by a combination of different HPE errors. Hence,  $\delta_i = \delta \mathbf{R}_i \Delta = \hat{\mathbf{R}}_i^T \mathbf{R}_i$  are randomly distributed around  $\Delta$  (see Fig. 6 top right). So, we can estimate the alignment matrix  $\Delta$  as the mean of all  $\delta_i$ 's. This is a *single rotation averaging* problem that can be computed with the Karcher mean [27]

$$\hat{\Delta} = \arg \min_{\mathbf{R} \in SO(3)} \sum_{i=1}^N f_{GE}(\mathbf{R}, \delta_i)^2,$$

for which there is a simple and convergent algorithm [29].

Finally, we compute a new set of aligned predictions  $\hat{\mathcal{R}}_a = \{\hat{\mathbf{R}}_1 \hat{\Delta}^T, \dots, \hat{\mathbf{R}}_N \hat{\Delta}^T\}$  that we use to evaluate the HPE.

## 4. Opal loss function for HPE

In this section we introduce a new OPTimAL loss function, Opal, based on a generalization of the geodesic loss in Eq. (2) to improve the

performance of HPE. It uses the geodesic error of each training sample to control its contribution to the learning, optimizing for example the accuracy for a specific subset of images.

The gradient of the distance used in the loss function, typically denoted as the influence function, drives the minimization process and determines the importance of each sample. As shown in Fig. 7 (left), the geodesic loss resembles an L1. Inspired by [30] we design Opal in such way that its influence function (Fig. 7 (right)) has a pseudo-Gaussian distribution that lets us specify which range of errors are more important, and by how much, while also linearly reducing the influence of small errors. Furthermore, we set the influence of large errors to 1, to avoid exploding gradients early in the training phase.

The Opal generalization to the geodesic distance between two rotation matrices is piece-wise function,

$$g_{opal}(\hat{\mathbf{R}}, \mathbf{R}) = \begin{cases} a \cdot G^2 + b & G < \epsilon \\ c \cdot (\tanh(\sigma \cdot G - \mu) + \tanh(\mu)) & \epsilon \geq G < \beta \\ G + d & G \geq \beta, \end{cases} \quad (6)$$

where  $G = g_{GE}(\hat{\mathbf{R}}, \mathbf{R})$ ;  $\epsilon$  and  $\beta$  represent the angular thresholds between L2 and L1 functions, respectively;  $\mu$  and  $\sigma$  control the mean and height of the pseudo-Gaussian influence function; and  $a, b, c$  and  $d$  are constants that ensure continuity and differentiability. Finally, the Opal loss is defined by extending  $g_{opal}$  to the training set,

$$f_{opal} = \frac{1}{N} \sum_{i=1}^N g_{opal}(\hat{\mathbf{R}}_i, \mathbf{R}_i). \quad (7)$$

As we can see in Figs. 7 and 10, the parameters of Opal loss can be adjusted to fit the distribution of the Geodesic Errors in the data set. This improves the generalization power of the network by balancing the contribution of different samples given their difficulty in training. In Section 5.5 we empirically demonstrate the usefulness of this loss function for wide-range HPE.

## 5. Experiments

In this section we test different combinations of head pose representations and loss functions. We use three configurations for HPE: (1) Euler angles and RMSE loss, defined as  $g_{RMSE}(\hat{\mathbf{p}}, \mathbf{p}) = \sqrt{g_{MSE}(\hat{\mathbf{p}}, \mathbf{p})}$ , denoted as *Euler* in our experiments, (2) quaternions with RMSE loss, denoted as *Quaternion*, and (3) 6D with geodesic error loss (Eq. (2)) denoted simply as *6D*.

### 5.1. Data sets

300W-LP and AFLW2000-3D are SRHP data sets introduced in [16] consisting of 61,225 and 2000 images acquired from 300W [31] and AFLW [32], respectively. They include automatically annotated 3D landmarks and head poses generated by fitting a 3DMM to each sample image. We follow the standard protocol in AFLW2000-3D [25], and discard 30 images with yaw, pitch or roll outside the range  $[-99^\circ, 99^\circ]$ .

Biwi [18] is also a SRHP data set that contains 15,677 frames from 24 videos of 20 subjects. It includes head pose annotations and facial depth masks obtained by a Kinect device without face bounding box annotations. In our experimentation we also follow common protocols presented in [26] using 300W-LP as train set and 13,219 frames of Biwi as test set, with bounding boxes provided by the MTCNN face detector [33].

CMU Panoptic [10] is a WRHP data set acquired in laboratory conditions that contains 84 video sequences organized into 8 groups, each one focusing on different scenarios, such as *range of motion, musical instruments, social games, haggling, dance, toddler* and *others*. Each video sequence consists of 31 HD quality videos with a resolution of  $1920 \times 1080$  pixels taken from a set of calibrated cameras located in a dome structure covering a full-hemisphere. Unfortunately, Panoptic does not include pose and bounding box annotations. Zhou

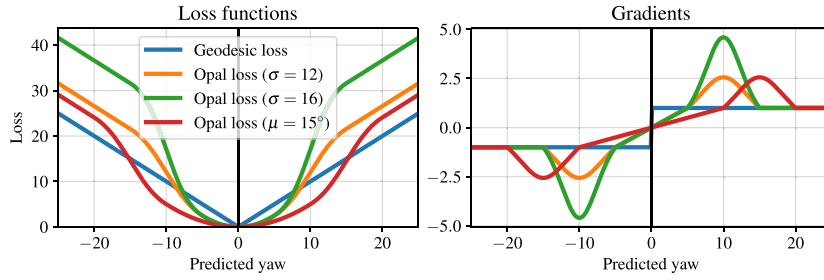


Fig. 7. Comparison of Geodesic and Opal losses, and the influence functions (gradients) obtained by both. The horizontal axis represent the predicted yaw (pitch and roll are fixed to  $0^\circ$ ), considering a ground-truth of  $[0^\circ, 0^\circ, 0^\circ]$ .

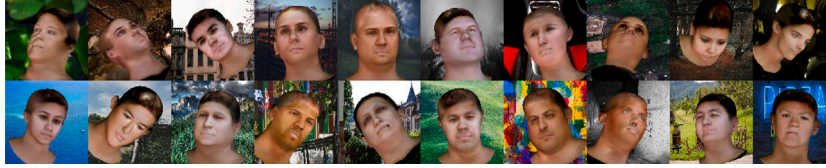


Fig. 8. Representative examples of the synthetic face images generated.

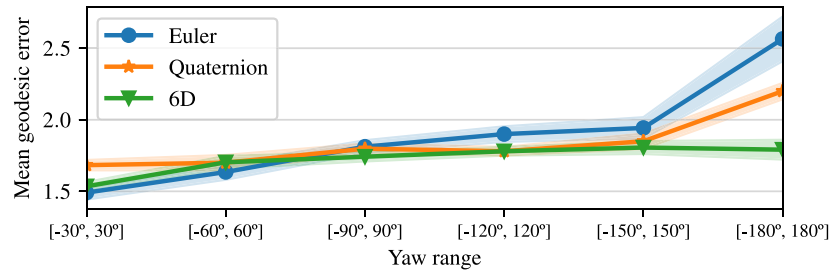


Fig. 9. Head pose representation experiment. Blue, orange and green colors compare the GE obtained at each interval using Euler, quaternion and 6D representations respectively. We also display with shading three standard deviations of the error. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

et al. [13] used the annotated 3D landmarks to generate them, but the training/testing protocols were not released. For this reason, we create a public benchmark containing 12 video sequences from *range of motion* and 5 video sequences from *haggling*. The number of images for training, validation and testing are, respectively, 61,008, 22,940 and 25,141. We discarded the frames where the landmarks could not be correctly projected onto all cameras. In the *haggling* sequences, where more than one subject is present in the same scene, we only considered the first person appearing in the annotation files.

## 5.2. Implementation details

During training, we perform data augmentation by applying the following random operations: horizontal flip, in plane rotations [34], scaling, synthetic occlusions, HSV color space manipulation and image blurring.

We build an encoder-like neural network which extracts relevant features from a  $128 \times 128$  RGB input image. The full architecture consists of 6 layers of stride 2 convolutions. At each one, we introduce “inverted residual bottleneck” modules from MobileNet-V2 [35]. As usual, each conv layer has batch normalization and a ReLU6 activation function. This is followed by a global average pooling layer to finally obtain a  $1 \times 1 \times 256$  tensor which is fed into a last fully connected layer responsible of estimating the  $P$  pose parameters.

Our networks are implemented in Pytorch and use Adam optimizer with an initial learning rate set to  $10^{-4}$ . We always select the model parameters with lowest validation error. For 300W-LP, we fine-tune the weights of a model pre-trained in the landmark detection task similar to [2]. We train the Euler model for 600 epochs, dropping the learning

rate to  $10^{-5}$  at epoch 300, and the 6D model for 900 epochs dropping the learning rate from  $5 \cdot 10^{-4}$  to  $5 \cdot 10^{-5}$  at epoch 450. For Panoptic, the learning rate is halved after 30 epochs without improvement in the validation. We stop training when the validation error stops improving for 60 epochs.

## 5.3. Synthetic experiments

In this section, we create a synthetic WRHP estimation benchmark to compare Euler, Quaternion and 6D representations. We use the Flame 3D model [36] with the albedo subspace of the Basel Face Model [37] to generate synthetic faces with head pose ranging between six intervals of yaw angles, *i.e.*,  $[-30^\circ, 30^\circ]$ ,  $[-60^\circ, 60^\circ]$ ,  $[-90^\circ, 90^\circ]$ ,  $[-120^\circ, 120^\circ]$ ,  $[-150^\circ, 150^\circ]$  and  $[-180^\circ, 180^\circ]$ , while pitch and roll angles have a fixed range of  $[-45^\circ, 45^\circ]$ , and neutral expressions. We add more realism to the generated images, using backgrounds taken from the BG-20K data set [38]. For each yaw interval, its corresponding data set contains 128,000 training images, 12,800 validation images and 38,400 test images. In Fig. 8 we display some representative examples of the synthetic face images generated.

In the first experiment (see Fig. 9), we train 3 different models for each interval. To make a fair comparison among the three representations, we use the GE as evaluation metric and the same loss function on all networks,  $\mathcal{L}_{RMSE}$ . As a global trend, we can see that the mean head pose errors increase as the yaw interval widens. This was expected since a wider yaw range requires the model to generalize across a broader spectrum of pose variations, making the learning task more challenging.

The gimbal lock has been traditionally perceived as a limitation of Euler angles for HPE [1,13,19–21]. Indeed, as discussed in Section 3,

**Table 2**  
MAE and GE in our synthetic data set with frontal faces and gimbal lock.

Parameterization	MAE				GE
	Yaw	Pitch	Roll	Mean	
Euler	<b>1.07</b>	<b>1.65</b>	<b>0.60</b>	<b>1.10</b>	<b>2.24</b>
6D	1.12	3.93	3.36	2.80	2.32

Euler angles cannot be used to measure the distance between two rotations, but in certain settings they are a good representation choice. Our results in Fig. 9 demonstrate that in a SRHP configuration, with yaw in the range  $[-90^\circ, 90^\circ]$ , the performance difference among all three competing representations is marginal. To confirm this result we perform a second experiment in which we decouple the inherent difficulty of predicting pose in profile faces from the gimbal lock. To this end we train our model to estimate the pose of semi-frontal faces with yaw angles ranging between  $[-10^\circ, 10^\circ]$ , but re-annotated as  $[80^\circ, 100^\circ]$ , so that all labels are close to the gimbal lock. In Table 2 we compare the results of this experiment for Euler angles and 6D representations. In terms of the GE error, both models have a very similar performance, with a marginal edge towards Euler. These results prove experimentally what we had discussed in Section 3, namely, in SRHP configuration Euler angles are a good representation, if the model is able to learn linear dependency between pitch and roll angles in the gimbal lock.

Differently to the gimbal lock case, we can also see in Fig. 9 that the prediction error for the Euler and quaternions models grow very rapidly when the range of yaw rotations is  $[-180^\circ, 180^\circ]$ . The discontinuity of Euler and quaternions representations near the yaw  $\pm 180^\circ$  limits their performance. So, the 6D representation, since it is continuous, provides the best performance in the WRHP setting.

#### 5.4. SRHP results

In this section we perform experiments on a SRHP configuration to confront the different representations with the existing literature. The most realistic SRHP evaluation is based on a cross-data set methodology. A widely used benchmark uses 300W-LP as training set and AFLW2000-3D and Biwi as test sets. To train the 300W-LP model, we shuffle the training set and split it into 90% train and 10% validation subset. We also crop faces using the bounding box of the landmarks annotations enlarged by 60%.

In Table 3 we compare our model using two representations with the SOTA in AFLW2000-3D. It is difficult to extract a conclusion about different representations since some methods use 3DMMs, e.g., DAD-3DNet [4], DSFNet [22], others additional training data, e.g., img2pose [3], DAD-3DNet [4], and, as expected, there are different CNN architectures. To address this issue we evaluate our model using the same network architecture, training images and data augmentation, only changing the representation. Our unaligned results in Table 3 confirm the synthetic experiments. Euler and 6D representations provide the same results using the GE metric, with a marginal edge in favor of the former. To compare with the literature we also provide the MAE.

The alignment procedure introduced Section 3.3 does not provide a reduction of the GE error in this experiment. This is an expected result since the training, 300W-LP, and test data sets, AFLW2000-3D, use the same annotation algorithm and, hence, their reference systems are aligned. Moreover, these results prove that our alignment procedure does not influence the result, if the reference systems of the training and test data sets are aligned.

In Table 4 we compare with the SOTA in Biwi. Like in the previous experiment, the results of models using Euler angles and 6D representation are very close. However, in this case and in contradiction with previous experiments, our model with 6D representation seems to be slightly better than Euler. The problem here, and with all previous results in the literature, is that we are ignoring the misalignment between

**Table 3**

SRHP results using AFLW2000-3D. GE is only available in methods that provide test code. Results ranked **first**, **second** and **third** are shown respectively in blue, green and red colors.

Method	Representation	MAE (°)				GE (°)
		Yaw	Pitch	Roll	Mean	
HopeNet [25]	Euler	6.47	6.56	5.44	6.15	9.93
FSA-Net [26]	Euler	4.50	6.08	4.64	5.07	8.16
WHENet [13]	Euler	4.44	5.75	4.31	4.83	–
TriNet [21]	Rot. matrix	4.19	5.76	4.04	4.66	–
TokenHPE [39]	Rot. matrix	4.36	5.54	4.08	4.66	–
QuatNet [19]	Quaternion	3.97	5.61	3.92	4.50	–
MFDNet [40]	Rot. matrix	4.30	5.16	3.69	4.38	–
img2pose [3]	Rot. vector	3.42	5.03	3.27	3.91	<b>6.41</b>
MNN [2]	Euler	3.34	4.69	3.48	3.83	–
DAD-3DNet [4]	6D	3.08	4.76	3.15	3.66	–
DSFNet [22]	Rot. matrix	<b>2.65</b>	<b>4.28</b>	<b>2.82</b>	<b>3.25</b>	–
<b>Ours (unaligned)</b>	Euler	<b>2.76</b>	<b>4.25</b>	<b>2.76</b>	<b>3.26</b>	<b>5.29</b>
<b>Ours (unaligned)</b>	6D	2.85	<b>4.59</b>	<b>3.04</b>	<b>3.49</b>	<b>5.37</b>
<b>Ours (aligned)</b>	Euler	2.75	4.23	2.76	3.25	5.28
<b>Ours (aligned)</b>	6D	2.83	4.55	3.04	3.47	5.34

**Table 4**

SRHP results using Biwi. GE only available in methods that provide test code. \* means methods where MTCNN detections are not used or is not stated. Results ranked **first**, **second** and **third** are shown respectively in blue, green and red colors.

Method	Representation	MAE (°)				GE (°)
		Yaw	Pitch	Roll	Mean	
HopeNet* [25]	Euler	4.81	6.61	3.27	4.89	9.53
QuatNet* [19]	Quaternion	4.01	5.49	2.93	4.14	–
FSA-Net [26]	Euler	4.27	4.96	2.76	4.00	7.64
DAD-3DNet [4]	6D	3.79	5.24	2.92	3.98	–
TriNet [21]	Rot. matrix	4.11	4.75	3.04	3.97	–
img2pose [3]	Rot. vector	4.56	<b>3.54</b>	3.24	<b>3.78</b>	<b>7.10</b>
TokenHPE [39]	Rot. matrix	3.95	<b>4.51</b>	<b>2.71</b>	3.72	–
MNN* [2]	Euler	3.98	<b>4.61</b>	<b>2.39</b>	<b>3.66</b>	–
MFDNet [40]	Rot. matrix	<b>3.40</b>	4.68	2.77	<b>3.62</b>	–
WHENet* [13]	Euler	<b>3.60</b>	<b>4.10</b>	<b>2.73</b>	<b>3.48</b>	–
Ours (unaligned)	Euler	4.54	5.05	2.80	4.13	<b>7.49</b>
Ours (unaligned)	6D	4.58	4.65	<b>2.71</b>	3.98	<b>7.30</b>
HopeNet* [25] (aligned)	Euler	4.53	<b>3.08</b>	2.83	3.48	6.60
FSA-Net [26] (aligned)	Euler	<b>3.59</b>	<b>2.90</b>	<b>2.27</b>	<b>2.92</b>	<b>5.36</b>
img2pose [3] (aligned)	Rot. vector	<b>4.04</b>	3.12	3.03	3.40	6.23
<b>Ours (aligned)</b>	Euler	<b>3.98</b>	<b>3.09</b>	<b>2.40</b>	<b>3.16</b>	<b>5.42</b>
<b>Ours (aligned)</b>	6D	<b>3.98</b>	3.24	<b>2.42</b>	<b>3.21</b>	<b>5.48</b>

the train and test data sets reference systems. Biwi was annotated by fitting a 3D model to the point cloud of a kinect depth map via ICP, whereas 300W-LP was labeled with the 3DDFA CNN [16]. In this case, the use of the alignment procedure reduces in 25% the GE error of our model using a 6D representation. Moreover, in the aligned results our Euler model is again marginally better, which is in agreement with all previous experiments.

In Table 4 we also show aligned results of previous methods. In all of them we provide a more accurate estimation that significantly reduces the Geodesic Error (GE) and, what is more interesting, the ranking of algorithms also changes.

In this section we confirmed experimentally some of the methodological results presented in Section 3. In a SRHP configuration Euler angles and 6D representations have similar performance. However, for cross-data set evaluation, it is required to apply an alignment procedure to remove systematic errors.

#### 5.5. WRHP results

In this section we perform experiments on a WRHP configuration to confirm the results obtained using synthetic images in Section 5.3. Here we also experiment with our novel Opal loss that allows us to emphasize certain error ranges while training.

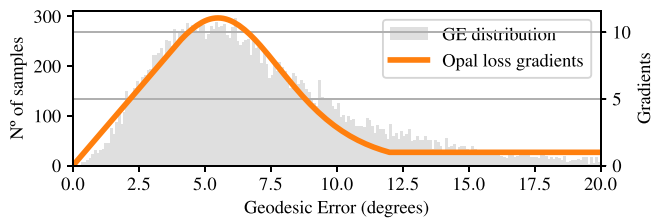


Fig. 10. Gradients of Opal loss (in orange) plotted over the GE distribution of the validation set of CMU Panoptic using a 6D model with geodesic loss (in gray). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

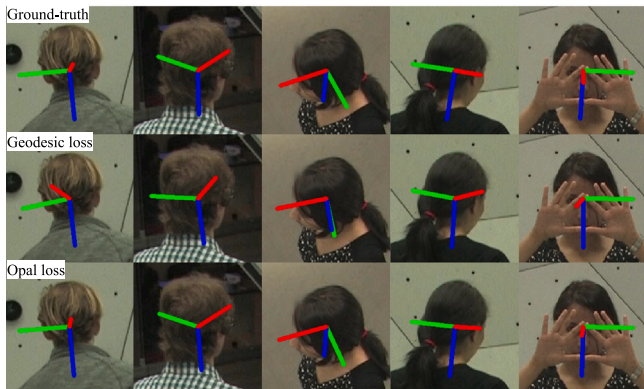


Fig. 11. Qualitative comparison between annotations (first row) and predictions using Geodesic loss (second row) and Opal loss (third row) in CMU Panoptic test set.

The first challenge with WRHP experiments is the availability of annotated real images. Zhou et al. [13] proposes a WRHP protocol using a combined data set of 300W-LP and Panoptic. However, annotations are not public. We propose a new benchmark based exclusively on Panoptic images (see Section 5.1).

In our experiments we train with Euler, quaternion and 6D representations. Additionally, we use Opal to improve the GE of the 6D network. Since what we estimate in Eq. (2) is the mean error of all training samples, we set the parameters of Opal so that its influence function fits the distribution of errors of the validation set (see Fig. 10). In this way our model concentrates more in the GE range around 5.5, and does not waste its learning capacity trying to improve very easy or atypical cases, below 2 or above 12 degrees respectively.

Results are shown in Table 5, aggregated (Mean column) and broken down into frontal ( $|\alpha_x| \in [0^\circ, 60^\circ]$ ), profile ( $|\alpha_y| \in [60^\circ, 120^\circ]$ ) and back views ( $|\alpha_x| \in [120^\circ, 180^\circ]$ ). Note that, since WHENet does not provide training code, we used the public pre-trained weights to evaluate their network. In contrast, we trained 6DRepNet with their training code, choosing the epoch with minimum validation loss. The first observation in this experiment is that here the mean GE is greater than in the SRHP case shown in Tables 3 and 4, which means that this problem is more difficult. As expected from the synthetic experiments, and due to the discontinuity effect, the Mean error of Euler and quaternion representations is respectively 35% and 21% larger than that of the 6D representation. Here again, like in the synthetic tests, Euler provides the worst results. 6DRepNet results show a similar effect, since they are close to our 6D network, and also outperform our Euler and Quaternion networks. We can conclude that a key factor in WRHP estimation is choosing a continuous representation.

Furthermore, Opal loss is capable of improving the mean GE of the 6D network by focusing more on the most frequent errors (mostly related to frontal and profile views) and decreasing the importance of small and atypical errors. In Fig. 11 we show some examples for which Opal significantly improved the estimation.

Table 5

WRHP results in CMU Panoptic benchmark. Mean GE and GE in frontal ( $|\alpha_x| \in [0^\circ, 60^\circ]$ ), profile ( $|\alpha_y| \in [60^\circ, 120^\circ]$ ) and back ( $|\alpha_x| \in [120^\circ, 180^\circ]$ ) views. **first, second and third** are shown respectively in blue, green and red colors.

Method	Representation	Mean	Frontal	Profile	Back
WHENet [13]	Euler	24.83	29.34	24.73	20.33
6DRepNet [20]	6D	8.08	5.80	8.07	10.40
<b>Ours</b>	Euler	10.47	7.69	10.08	13.68
<b>Ours</b>	Quaternion	9.32	7.04	8.98	11.98
<b>Ours</b>	6D	7.70	5.81	7.15	10.18
<b>Ours + Opal loss</b>	6D	7.45	5.40	6.75	10.25

## 6. Conclusions

This paper addressed the problem of estimating the orientation of a head in an image both in the profile to frontal SRHP and in the full 360° rotation WRHP setting. In our analysis we distinguish the representation used for the pose from the distance between two orientations used as error metric. This enabled us to shed some light on the adequacy of each representation and error metric. Contrary to common belief, we show that for HPE the gimbal lock is not a drawback for Euler angles representation. In fact, given its minimal dimensionality, intuitive interpretation, and marginal edge over other alternatives, it is the preferred representation in a SRHP setting involving frontal and profile heads. However, to solve the HPE in the full WRHP setting we need a continuous representation, like the 6D obtained with two columns of the rotation matrix [14].

Unlike with the representation problem, the gimbal lock prevents the use of Euler angles for measuring the distance between two orientations. So, the popular MAE metric should be avoided both in SRHP and WRHP settings. We suggest the use of proper angular metric such as the geodesic error, GE [15,27], because of its interpretability.

For a proper cross-data set evaluation we must consider the fact that it is very likely that the reference systems used in the annotations of the training and test data sets are different. As a consequence, we cannot directly compare the orientation estimated by the trained model with the ground truth annotation in the test data set. We have also introduced a procedure to align the estimations with the ground truth annotations and empirically show that it does not bias the estimation in cases where the train and test annotations share the same reference system. With this procedure we set a new, more accurate, SOTA in the 300W-LP/Biwi cross-data set benchmark.

We also generalize the geodesic distance to introduce Opal, a new distance conceived for controlling the contribution of images in the train data set to the learning process. In the future we plan to use this distance in other regression problems, such as facial landmark estimation and to optimize other particular ranges of error, instead of the usual average.

Finally, we propose a new benchmark based on the Panoptic data set and the annotations proposed in WHENet [13].

## Data availability

Most data sets are public. Our own annotations to CMU Panoptic will public upon publication of the paper.

## Acknowledgments

This work was partially funded by project PID2022-137581OB-I00 from MCIN/AEI/10.13039/501100011033/FEDER, UE. Alejandro Cobo was also funded by the Comunidad de Madrid grant PEJ-2020-AI/TIC-17682, and with a doctoral contract from Universidad Politécnica de Madrid, UPM. José M. Buenaposada was partially funded by “AYUDA PUENTE 2022”, Universidad Rey Juan Carlos (ref. M3037). Luis Baumela and José M. Buenaposada are members of ELLIS Unit Madrid, funded by the Autonomous Community of Madrid. The authors gratefully acknowledge the UPM for providing computing resources on the Magerit supercomputer.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- [1] A.F. Abate, C. Bisogni, A. Castiglione, M. Nappi, Head pose estimation: An extensive survey on recent techniques and applications, *Pattern Recognit.* 127 (2022) 108591.
- [2] R. Valle, J.M. Buenaposada, L. Baumela, Multi-task head pose estimation in-the-wild, *PAMI* 43 (8) (2021) 2874–2881.
- [3] V. Albiero, X. Chen, X. Yin, G. Pang, T. Hassner, img2pose: Face alignment and detection via 6DoF, face pose estimation, in: *Proc. CVPR*, 2021, pp. 7617–7627.
- [4] T. Martyniuk, O. Kupyn, Y. Kurylyak, I. Krashenyi, J. Matas, V. Sharmanska, DAD-3DHeads: A large-scale dense, accurate and diverse dataset for 3D head alignment from a single image, in: *Proc. CVPR*, 2022, pp. 20910–20920.
- [5] P. Barra, C. Bisogni, M. Nappi, S. Ricciardi, Fast QuadTree-based pose estimation for security applications using face biometrics, in: *Proc. Network and System Security*, 2018, pp. 160–173.
- [6] P. Barra, S. Barra, C. Bisogni, M.D. Marsico, M. Nappi, Web-shaped model for head pose estimation: An approach for best exemplar selection, *IEEE Trans. Image Process.* 29 (2020) 5457–5468.
- [7] R. Ranjan, V.M. Patel, R. Chellappa, HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, *PAMI* 41 (1) (2019) 121–135.
- [8] M.F. Valstar, E. Sánchez-Lozano, J.F. Cohn, L.A. Jeni, J.M. Girard, Z. Zhang, L. Yin, M. Pantic, FERA 2017 - addressing head pose in the third facial expression recognition and analysis challenge, in: *Proc. International Conference on Automatic Face and Gesture Recognition*, 2017, pp. 839–847.
- [9] S. Jha, C. Busso, Challenges in head pose estimation of drivers in naturalistic recordings using existing tools, in: *Proc. IEEE International Conference on Intelligent Transportation Systems*, 2017, pp. 1–6.
- [10] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B.C. Nabe, I.A. Matthews, T. Kanade, S. Nobuhara, Y. Sheikh, Panoptic studio: A massively multiview system for social interaction capture, *PAMI* 41 (1) (2017) 190–204.
- [11] W. Rahmaniari, Q.M. ul Haq, T.-L. Lin, Wide range head pose estimation using a single RGB camera for intelligent surveillance, *IEEE Sensors J.* 22 (11) (2022) 11112–11121.
- [12] L. Beyer, A. Hermans, B. Leibe, Biternion nets: Continuous head pose regression from discrete training labels, *Pattern Recognit.* 9358 (2015) 157–168.
- [13] Y. Zhou, J. Gregson, WHENet: Real-time fine-grained estimation for wide range head pose, in: *Proc. BMVC*, 2020.
- [14] Y. Zhou, C. Barnes, J. Lu, J. Yang, H. Li, On the continuity of rotation representations in neural networks, in: *Proc. CVPR*, 2019, pp. 5745–5753.
- [15] D.Q. Huynh, Metrics for 3D rotations: Comparison and analysis, *J. Math. Imaging Vision* 35 (2) (2009) 155–164.
- [16] X. Zhu, X. Liu, Z. Lei, S.Z. Li, Face alignment in full pose range: A 3D total solution, *PAMI* 41 (1) (2019) 78–92.
- [17] S. Yang, P. Luo, C.C. Loy, X. Tang, WIDER FACE: A face detection benchmark, in: *Proc. CVPR*, 2016, pp. 5525–5533.
- [18] G. Fanelli, M. Dantone, J. Gall, A. Fossati, L. Van Gool, Random forests for real time 3D face analysis, *IJCV* 101 (3) (2013) 437–458.
- [19] H. Hsu, T. Wu, S. Wan, W.H. Wong, C. Lee, QuatNet: Quaternion-based head pose estimation with multi-regression loss, *IEEE Trans. Multimed.* 21 (4) (2019) 1035–1046.
- [20] T. Hempel, A.A. Abdelrahman, A. Al-Hamadi, 6D rotation representation for unconstrained head pose estimation, in: *Proc. International Conference on Image Processing*, 2022, pp. 2496–2500.
- [21] Z. Cao, Z. Chu, D. Liu, Y.V. Chen, A vector-based representation to enhance head pose estimation, in: *Proc. WACV*, 2021, pp. 1187–1196.
- [22] H. Li, B. Wang, Y. Cheng, M. Kankanhalli, R.T. Tan, DSFNet: Dual space fusion network for occlusion-robust 3D dense face alignment, in: *Proc. CVPR*, 2023, pp. 4531–4540.
- [23] X. Jiang, D. Li, H. Chen, Y. Zheng, R. Zhao, L. Wu, Uni6D: A unified CNN framework without projection breakdown for 6D pose estimation, in: *Proc. CVPR*, 2022, pp. 11164–11174.
- [24] G. Poniatkin, Y. Labbé, B.C. Russell, M. Aubry, J. Sivic, Focal length and object pose estimation via render and compare, in: *Proc. CVPR*, 2022, pp. 3815–3824.
- [25] N. Ruiz, E. Chong, J.M. Rehg, Fine-grained head pose estimation without keypoints, in: *Proc. CVPRW*, 2018, pp. 2074–2083.
- [26] T. Yang, Y. Chen, Y. Lin, Y. Chuang, FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image, in: *Proc. CVPR*, 2019, pp. 1087–1096.
- [27] R. Hartley, J. Trampf, Y. Dai, H. Li, Rotation averaging, *IJCV* 103 (3) (2013) 267–305.
- [28] A. Torralba, A.A. Efros, Unbiased look at dataset bias, in: *Proc. CVPR*, 2011, pp. 1521–1528.
- [29] J. Manton, A globally convergent numerical algorithm for computing the centre of mass on compact Lie groups, in: *Proc. International Conference on Control, Automation, Robotics and Vision*, 2004, pp. 2211–2216.
- [30] Z. Feng, J. Kittler, M. Awais, P. Huber, X. Wu, Wing loss for robust facial landmark localisation with convolutional neural networks, in: *Proc. CVPR*, 2018, pp. 2235–2245.
- [31] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 Faces in-the-wild challenge: database and results, *Image Vis. Comput.* 47 (2016) 3–18.
- [32] M. Köstinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, in: *Proc. ICCVW*, 2011, pp. 2144–2151.
- [33] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, *IEEE Signal Process. Lett.* 23 (10) (2016) 1499–1503.
- [34] A. Sheka, V. Samun, Rotation augmentation for head pose estimation problem, in: *2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology, USBERIT*, 2021, pp. 308–311.
- [35] M. Sandler, A.G. Howard, M. Zhu, A. Zhmoginov, L. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: *Proc. CVPR*, 2018, pp. 4510–4520.
- [36] T. Li, T. Bolkart, M.J. Black, H. Li, J. Romero, Learning a model of facial shape and expression from 4D scans, *ACM Trans. Graph.* 36 (6) (2017) 194:1–194:17.
- [37] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, A 3D face model for pose and illumination invariant face recognition, in: *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.
- [38] J. Li, J. Zhang, S.J. Maybank, D. Tao, Bridging composite and real: Towards end-to-end deep image matting, *IJCV* 130 (2) (2022) 246–266.
- [39] C. Zhang, H. Liu, Y. Deng, B. Xie, Y. Li, TokenHPE: Learning orientation tokens for efficient head pose estimation via transformers, in: *Proc. CVPR*, 2023, pp. 8897–8906.
- [40] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, J. Wang, MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation, *IEEE Trans. Multimed.* 24 (2021) 2449–2460.

**Alejandro Cobo** received the Computer Science BS in 2019 and the Data Science MS degree in 2020, both from the Universidad Politécnica de Madrid (UPM). He is currently attending a Ph.D. course in Artificial Intelligence at UPM. His research interests include face image analysis, generative vision models and deep fake detection.

**Roberto Valle** received the B.Sc. and M.Sc. degrees in 2012 and 2013 respectively, both at Universidad Rey Juan Carlos (URJC) and the Artificial Intelligence Ph.D. degree in 2020 from the Universidad Politécnica de Madrid (UPM). Since 2022 he is an Assistant Professor at the UPM and is a member of the Computer Perception Group. His research interests concern computer vision and deep learning algorithms for facial analysis problems.

**José M. Buenaposada** received the BS and MS degrees in 1999 and the Ph.D. degree in 2005 both in computer science from the Universidad Politécnica de Madrid (UPM). From 2008 to 2018 he has been Associate Professor and since July 2023 he is a Professor of computer science both at Universidad Rey Juan Carlos (URJC), Spain. He is member of the GAVAB research group at URJC and the Computer Perception Group at UPM. His research interests include image alignment, face image analysis, object detection and efficient computer vision.

**Luis Baumela** received the BS and MS degrees in 1989 and the Ph.D. degree in 1995, all in computer science from the Universidad Politécnica de Madrid. From 1989 to 1992, he was an engineer at Telefonía's R&D labs. From 1997 to 2016, he has been an Associate Professor, and since 2016 he is a Professor of computer science at the ETSI Informáticos of the Universidad Politécnica de Madrid, where he leads the Computer Perception Group. His research interests include image alignment, face image analysis, and medical imaging.