Published in: Pattern Recognition journal.

This version of the article has been accepted for publication, after peer review and is subject to Elsevier terms of use.

The Version of Record is available online at: https://doi.org/10.1016/j.patcog.2022.108591

Head pose estimation: An extensive survey on recent techniques and applications

Andrea F. Abate[a], Carmen Bisogni[a,*], Aniello Castiglione[b], Michele Nappi[a]

[a] University of Salerno, Via Giovanni Paolo II, 132, Fisciano, Salerno 84084, Italy
[b] University of Naples Parthenope, Centro Direzionale di Napoli, Isola C4, Naples 80143, Italy

Keywords:
Biometrics
Head pose estimation
Face recognition
Frontalization

A B S T R A C T

Biometric based systems are involved in many areas, from surveillance to user authentication, from autonomous systems to human-robot interactions. Head pose estimation (HPE) is the task to support biometric systems in which any of the biometric traits of the head is involved, as face, ear or iris. This particular biometric branch finds its application in driver attention detection, surveillance for recognition, face frontalization, best frame selection and so on. The goal of HPE is to determine the head pose orientation (yaw, pitch, roll). The implemented methods use different techniques depending on the kind of input. In this survey we present an overview of involved datasets, recent techniques and applications. We evaluate and compare the different approaches with respect to their advantages and practical usage. In addition, we propose a technical comparison between training and training-free techniques for the most popular HPE methods.

1. Introduction

Head Pose Estimation (HPE) is the field that studies the rotation angles of the head. It can be seen in different purposes: as a preprocessing step to find the best frame to perform face recognition in a video; as a behavioral characteristic to estimate the intent of the subject; as a descriptor to help face frontalization and so on, as described in this survey. The study of HPE represents a subset of the wider biometrics field.

The choice of the biometric trait to use is mainly dictated by two aspects: the computational resources available and the applicability in terms of visible area. In this context, the advent of behavioral and soft biometrics has prepared the ground to the use of alternative biometrics as the head pose. In fact, HPE can be applied in both behavioral and soft biometrics.

Compared to the last comprehensive survey on HPE [1], in the proposed work we take under consideration the multitude of methods and advances born in the last years and also the impact of recent techniques as machine learning. In addition, compared to more recent surveys that are focused only to one kind of data [2], or for a particular purpose [3,4], the presented survey is proposed as an overview in datasets, methods and purpose of the recent techniques in HPE. The datasets we will explore are subdivided in mainly three categories, 2D data, depth data and video. From this starting point, we split the methods based on the type of data being processed. Section 2 introduces the fundamentals of HPE and basic concept to study this field. Section 3 presents the datasets in this field, and an exhaustive evaluation of their numerosity, labels, and popularity in literature. In Section 4 an overview of preprocessing techniques for HPE can be found. In Section 5 we describe recent techniques to perform HPE. In the same section, we also provide a technical comparison between training and training-free methods. Section 6 is focused on the application of those techniques in different. Finally, in Section 7 a summary of this survey is presented and related conclusions have been drawn. In Fig. 1 the application fields of HPE can be observed, in particular they are ordered by input, following the same architecture of this survey. Comparing the methods of the last five years, in which training techniques become so popular in this field, we can nevertheless claim that it is not possible to generalize that their performances are better than the performances of training-free techniques. A more detailed graph of method can be found in Fig. 3, in Section 5. In the same section, we will further discuss this observation on the methods performances.

2. Head pose estimation: the basics

HPE has a behavior assimilable to soft biometrics in terms of uniqueness. However, in terms of measurability it results stronger

* Corresponding author.
E-mail addresses: abate@unisa.it (A.F. Abate), cbisogni@unisa.it (C. Bisogni), castiglione@ieee.org (A. Castiglione), mnappi@unisa.it (M. Nappi).

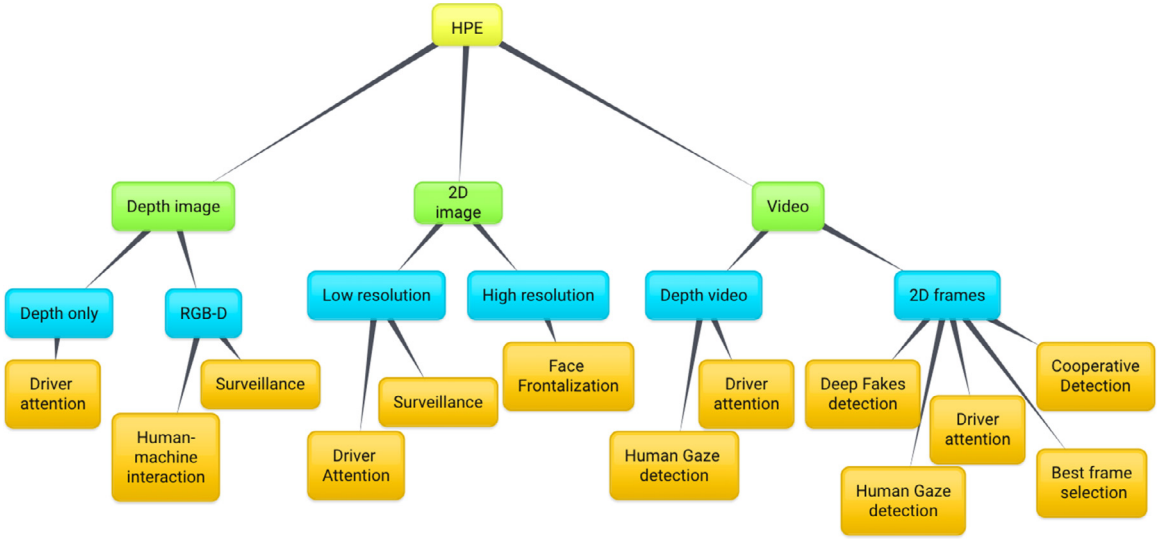

Fig. 1. Applications of HPE, ordered by input.

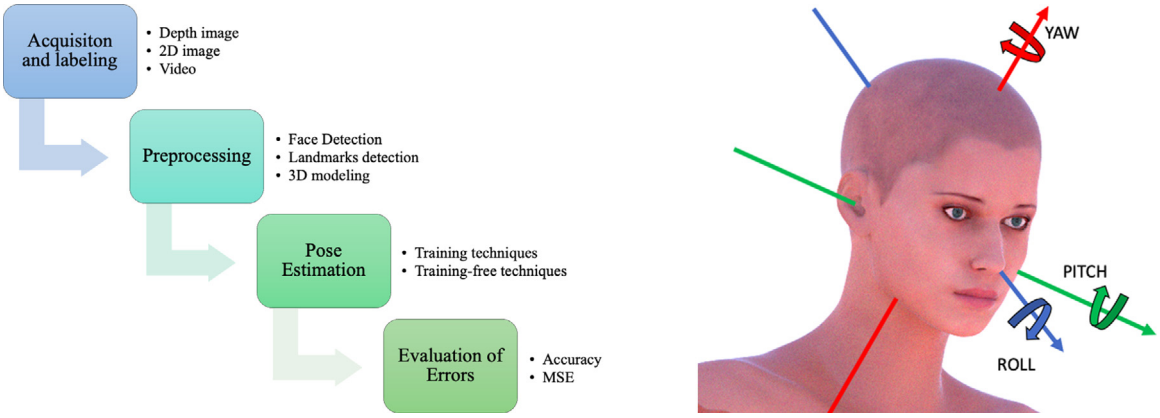

Fig. 2. On the left, the main steps of an HPE framework. On the right, pitch, yaw and roll.

than most soft biometric traits in which a subjective component can be observed. The head rotation is observable in each individual and it is not dependent on his/her age. The acceptability of HPE can be considered the same as the acceptability of face, or at the most of the ear because they are the only recognition traits involved. On the other hand, the shape of the face and of the head, that is different from an individual to another, can have an impact in HPE and its relevance depending on the particular characteristics of the used method. The main step of an HPE algorithm can be summarized as depicted in Fig. 2.

The head pose variation is measured in rotation angles. The center of the head, or the nose if 3D data are not available, is considered the centering point of the rotation $O(0, 0, 0)$. The Head is a 3-dimensional object, by nature, and therefore the possible angles of rotation are 3. The axes are, by convention, represented by the Motion Imagery Standards Board (MISB). MISB [5] as pitch, yaw and roll. The directions of pitch, yaw and roll can be seen in Fig. 2. If we consider the frontal view as a reference system, although there is individual variation, most people are able to turn their head $\pm 90^\circ$ in yaw, $\pm 45^\circ$ in roll and $\pm\ 30^\circ$ in pitch. Extreme poses of the head due to body movements (e.g. $\pm 180^\circ$ in yaw) are rarely considered, since the main scope of HPE is in most cases face recognition. There are various ways to represent a 3D rotation, however the most popular among the HPE datasets and algorithms are the Euler angles, the rotation matrix and the quaternions. We will introduce them in the followings.

2.1. The Euler angles

They were first introduced by Leonhard Euler to describe the orientation of a rigid body in space. The Euler angles can be split in two categories: Proper Euler angles and Tait–Bryan angles. Since, as previously claimed, the HP rotation follows the rules of MISB, the Tait-Bryan angles are properly used to describe the rotations. We define as x, y and z the original axes and X, Y and Z the axes after the rotation. The line that represent the intersection between plane xy and YZ is called the line of nodes N. With this conventions we can define the Euler angles as: ϕ the rotation angle between x and N, covering a range of 2π;θ the rotation angle between z and Z, covering a range of π; ψ the rotation angle between N and X, covering a range of 2π.

2.2. The rotation matrix

Using the rotation matrix, the rotation respect to the axis can be calculated using a single rotation angle θ and can be defined by three rotation matrices, one for each axes. If we define as α, β

and γ, the rotations in yaw, pitch and roll, respectively, the final rotation matrix will be

$$\begin{bmatrix} \cos\alpha\cos\beta & \cos\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma \\ \sin\alpha\cos\beta & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma \\ -\sin\beta & \cos\beta\sin\gamma & \cos\beta\cos\gamma \end{bmatrix} \quad (1)$$

Since this representation could be not convenient to use, the rotion matrix can be converted in rotation vector by the Rodrigues' formula and we obtain:

$$v_{rot}(\theta) = v\cos\theta + (k \times v)\sin\theta + k(k \cdot v)(1 - \cos\theta) \quad (2)$$

where v is a 3-D vector, k is a unit-vector describing an axis of rotation about which v rotates by an angle θ according to the right hand rule, $k \times v$ is a cross product and $k \cdot v$ is the scalar product.

2.3. The quaternions

. They are often known as versors. Firstly introduced by William Rowan Hamilton, are nowadays very popular among videogames developers. Basing on the concept of complex number and complex plane, we can introduce the general form to express quaternions as

$$q = s + xi + yj + zks, x, y, z \in \mathbb{R} \quad \text{and} \quad i, j, k \text{ in } \mathbb{C}, \text{ with} \quad (3)$$

$$i^2 = j^2 = k^2 = ijk = -1 \text{ and}$$

$$ij = k, jk = i, ki = j, ji = -k, kj = -i, ik = -j \quad (4)$$

The Euler angles results to be more human understandable, however they conduct to ambiguity problems and gimbal lock. It means that we will have infinite solutions of pose estimation for the same rotations that lead to an evident estimation problem. On the other hand, quaternions do not suffer from this problem and are simpler to compose. Compared to rotation matrix, their representation is more compact.

If on one hand, the Euler angles result to be more human understandable, on the other hand, they conduct to ambiguity problems and gimbal lock. In particular, the gimbal lock is the lost of one degree of freedom that occurs when two of the three axes of the system are parallel. In particular, in the case of head pose estimation, if we set the yaw to 90°, pitch and roll becames parallel (i.e. linearly dependent). It means that we will have infinite solutions of pose estimation for the same rotations that lead to an evident estimation problem.

Since different datasets may have different annotations for the angles, the testing methods usually choose a representation and, by the transformation formulas, they normalize the label of the dataset accordingly.

On the other hand, there is an homogeneity in papers concerning the evaluation of errors. In fact, the errors, separated in yaw, pitch and roll, are represented by the angular values of the differences between the estimated pitch, yaw and roll and the true pitch, yaw and roll for each head in the data. To achieve valuable results, the algorithms compute these three error for various heads in data, and then the mean absolute error is computed for each axis as:

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} |\theta_j - \hat{\theta}_j| \quad (5)$$

where θ_j is the ground truth, i.e the true angular value and $\hat{\theta}_j$ is the prediction, i.e the predicted angular value. The same formula is often used to compute also the total MAE along the three axes together.

3. Datasets with HP annotations

There are basically three kinds of input data to perform HPE: depth image, 2D images, video. For this reason we will split this section comparing datasets with similar kinds of data. As will be noticeable, some datasets have different kinds of data and, for the sake of clarity, we will add them in more than one section.

3.1. Depth images

Depth images datasets contain both information about the RGB and the depth from the same image. In the last decade many datasets have been proposed which are useful for benchmarking the applicability of HPE algorithms. All the characteristic of the depth datasets are summarized in Table 1.

One of the most popular depth dataset is undoubtedly the BIWI Kinect Head Pose Database (BIWI) [6]. BIWI contains 24 sequences of 20 people for a total of over 15 K images recorded with a Kinect 1. The variation of the head pose is between −75° and +75° in yaw and −60° and +60° in pitch. Faceshift has been used to annotate the head poses. ICT-3DHP [7] is a dataset of head pose collected with the Kinect, composed of 10 RGB-D video for a total of about 1400 frames. The labels were obtained using a Polhemus FASTRACK. SASE database [8
2. The total number of subjects is 50, and the average amount of frame per subject is 600. They provide pitch, yaw and roll using five blue stickers placed on the face. SASE covers a range between −75° and +75° in yaw and −45° and +45° in pitch. The ETH Face Pose Range Image Dataset [9] has over 10 K images and 20 subjects. The ground truth is provided by the 3D nose tip coordinates and the coordinates of a vector pointing in the face direction. ETH covers a range between −90° and +90° in yaw and −45° and +45° in pitch. The Pandora dataset [10] was captured with the Kinect 1. There are more than 250 K images in the dataset of 22 subjects, five recording per subject.

3.2. 2D images

Depth dataset, despite their reliable labels, are not the preferred input to develop and test HPE methods. This is because to obtain accurate depth images, the environment is controlled and HPE methods using only 2D RGB images are, on the contrary, conceived and developed to solve on-the-wild problems. In the following we will present the datasets without depth information and the way in which they were labeled. CMU-MultiPIE [11] is a dataset of RGB

Table 1
Depth HPE datasets that contain pose annotation. The popularity of each dataset is calculated using the amount of recent depth HPE method that use it, to the best of our knowledge ("Pop" is for popularity, i.e. the number of papers in this survey that use it). The methods are in Section 5.1.

Dataset	Year	RGB Res	Depth Res	#Subj	#Frms	Pop
BIWI	2013	640 × 480	640 × 480	20	+15 K	23
ICT-3DHP	2012	640 × 480	640 × 480	10	1400	6
SASE	2016	1080 × 1920	424 × 512	50	+30 K	3
ETH	2008	640 × 480	640 × 480	20	+10 K	2
Pandora	2017	1920 × 1080	512 × 424	22	+250 K	2

Table 2
2D RGB Datasets that contain Pose Annotation. The Popularity of each dataset is calculated using the amount of recent depth HPE method that use it, to the best of our knowledge ("Pop" is for popularity, i.e. the number of papers in this survey that use it). The methods are in Section 5.2. nd in the number of subjects column is for "not declared".

Dataset	Year	Res	#Subj	#Frms	Limit	Pop
BIWI	2013	640×480	20	+15 K	\	17
CMU MultiPIE	2013	3072×2048	337	+750 K	15 poses	4
Pointing'04	2004	384×288	15	2790	NO ROLL	14
AFLW	2011	various	20	25 K	\	12
AFLW2000	2018	450×450	nd	2000	\	14
AFW	2012	various	nd	205	NO PITCH	4
300W_lp	2016	various	nd	61,225	\	4
CASPEAL	2008	640×480	1040	99,594	27 poses	5
Youtube Faces	2011	100x100	1595	+600 K	\	2
McGill	2013	640×480	60	18 K	NO PITCH NO ROLL	1
GOTCHA-I	2020	512×512	62	137,826	\	1

images on which the head position is estimated by a stereo camera technique. This dataset is an extension of the previous CMU-PIE [12] and contains more than 750 K images of 337 subjects in 13 poses, in 4 recording sessions and 6 facial expressions. Pointing'04 Head Pose Image Database [13] were collected using 15 subjects by the PRIMA Lab. For each subject there are 2 series of 93 images, for a total of 2790 images. Roll rotation angle is not contemplated. The dataset has a limited number of poses, in particular 9 for pitch and 13 for yaw, and their combination between -90° and $+90^\circ$ degree. To obtain poses with known labels, the authors have put markers in a room and ask to stare at the 93 post-it notes without moving their eyes. The Annotated Facial Landmark in the Wild (AFLW) [14] is a dataset of about 25 K images collected from the web and, as a consequence, they have a large variation in pose, expressions, age, gender, ethnicity. This dataset provides face rectangle and the face ellipses, used in a POSIT algorithm [15] to estimate the head pose. As a more accurate version of AFLW respect to the HP, we can find AFLW2000 that include the first 2000 images of AFLW annotated using a 3DMM fitting and can be downloaded at [16]. The Annotated Face in the Wild (AFW) dataset [17] was collected from Flicks images. The dataset is quite small, containing only 205 images of about 468 faces. The 300W_lp Dataset [18] is expanded from 300W which includes 68 landmarks localization. 300W_lp collects different datasets, in particular AFW, LFPW, HELEN, IBUG and XM2VTS. There are 61,225 images, 1786 from IBUG, 5207 from AFW, 16,556 from LFPW and 37,676 from HELEN. The CAS-PEAL database [19] contains 99,594 images of 1040 subjects. There are 27 discrete poses in a controlled environment. The Youtube Faces database [20] is a video dataset collected from 3425 youtube videos of 1595 subjects. There are over 600 K extracted and annotated frames. Faces are detected with the Viola-Jones method we will introduce in Section 4. Another video dataset with annotated frames used for HPE is the McGill real world face video database [21]. This dataset contains 60 real word videos of 60 subjects, recorded at 30 fps. The amount of annotated frames are 18000. Finally, an emerging video dataset with precise HP annotations is GOTCHA-I [22]. GOTCHA-I was collected as video sequences of 62 subjects in 11 different environments, for a total of 682 videos of people walking. There are 137,826 labeled frames with 2223 HP per subject in the range of -40° and $+40^\circ$ in yaw and -30° and $+30^\circ$ in pitch and -20° and $+20^\circ$ in roll, with a step of 5°. In Table 2 there is a summarization of 2D datasets with HP annotations with some main characteristics.

3.3. Videos

The video datasets are in fact sequences of annotated frames. The majority of the datasets presented in this Section are used for HPE with tracking purpose, in which some characteristics are essential.

The UPNA Head Pose Database [23] is composed of 120 videos of 10 subjects. Since this dataset is born for head tracking and pose estimation, the authors collected 6 guided-movement sequences and 6 free-movement sequences. There are 300 frames per video. They used the initial frame of the frontal face as a keypoint to label the head pose. The Boston University Head Pose Database [24] is a set of videos recorded for tracking purposes, in particular under different illuminations. There are a total of 72 videos, split in two sessions of 45 videos of 5 subjects under uniform illumination and 27 videos of 3 subjects under time varying illumination. Each of them is 200 frames long. The Head Pose and Eye Gaze Dataset (HPEG) [25] is created in lab conditions and consists of 20 videos of 10 subjects. There are about 400 frames per video. The ground truth is collected using three LEDs and tracking their positions at each frame. In this dataset only yaw and pitch ground truth is available. The EYEDIAP Database [26] has been collected with Gaze Tracking purpose. They recorded 94 sessions with 16 subjects, the scene both with the Kinect than an HD camera placed as near as possible to the Kinect. Each video has about 4860 frames. Also the UBIPose dataset [27] was collected using a Kinect to obtain labels. There are 32 videos simulating a reception desk environment, but only 22 of the 32 videos are annotated. The head pose is available for about 10 K frames. Finally, a particular dataset composed by video of subjects driving, is the second Strategic Highway Research Program (SHRP2) [28]. The dataset is very wide, with over 3100 videos of the same number of subjects recorded during a period of 2 year. However, in successive studies only about 63 K frames of 41 videos have head pose annotations. Each video lasts about 15 min and is recorded at 15 fps, but the annotated frames, as claimed before, are about 1537 per video. All of the presented datasets were collected with the agreement of participants or by public available data online. In the first case, it is not uncommon to need a formal request to obtain the data.

4. Preprocessing techniques

To perform HPE, some preprocessing techniques are usually applied to find the head region or to detect some keypoint on the face. We can categorize those techniques in three main groups: face detection; landmark detection; 3D head modeling.

4.1. Face detection

It is a preliminary step that most HPE algorithm applies to exclude other parts of the body or the scene from the analysis. For longer, a very popular techniques has been the Viola-Jones method [29]. However, if the head pose is extreme (more than 60°), this

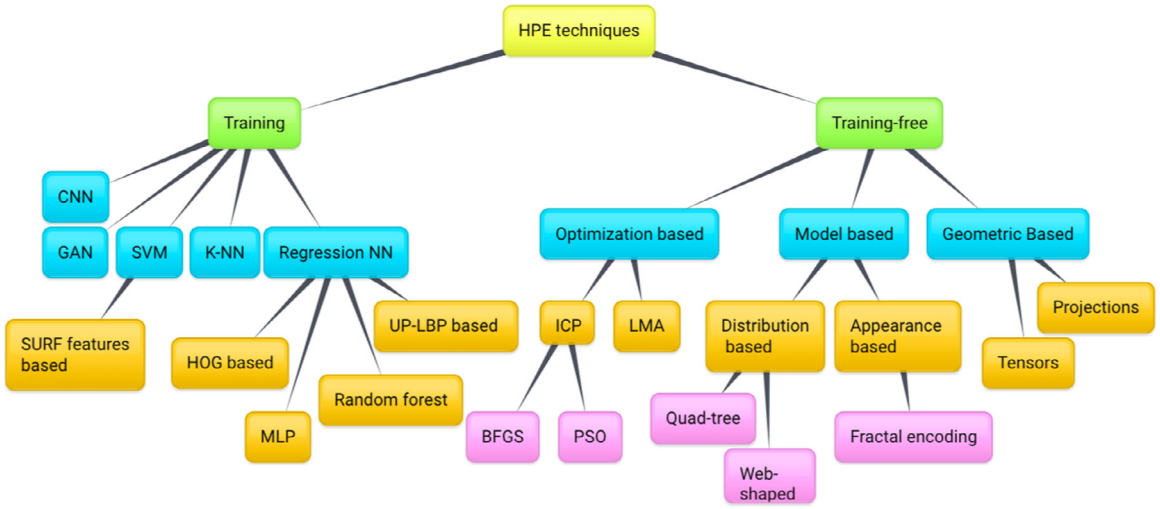

Fig. 3. Techniques used for HPE, ordered by training and training-free frameworks.

method would not succeed in finding the face location. For this reason it was firstly preferred the combination of Histogram of Oriented Gradients (HOG) and linear Support Vector Machine (SVM) proposed in Pang et al. [30], and then learning algorithms. As an example by training random forests [31] or deep architecture as ArcFace [32]. Of great relevance for recent times is also the impact of such existing detectors in the case of facial masks [33].

4.2. Landmark detection

It is a techniques that can be used after or in substitution to face detection. We refer to facial keypoints as the location of some particular features in the face, like the noise, the eyes, the mouth. Some authors developed adaptive boosting methods to search this keypoint, others used the above mentioned random forest classifier or SVM. As in the case of face detection, also here, in the last time the use of deep learning became popular. Recent CNN demonstrates to be robust on HP variations and able to operate in real time on webcams to detect facial landmarks [34]. For a comprehensive study on landmarks detection refer to Wu and Ji [35].

4.3. 3D modeling

It is the approach that aim to create a 3D model of the face to estimate the head pose. In this case, the availability of a depth image can be essential to create a realistic model. If a video is available, an adaptation process can be performed, also online, as proposed by Sheng et al. [36]. This modeling can also be made by using RGB images as proposed in Proença et al. [37] by using an Euclidean distance as a magnitude, or in combination with facial landmarks as proposed in Chang et al. [38].

5. HPE techniques

Following the observations made in the previous sections, it appears that the most relevant difference between HPE techniques is represented by the initial data. Regardless of the used methods, the available initial data radically change the approach to the problem. We summarized the approaches in the following sections, dividing them by the kind of initial data. In Fig. 3 we also made a hierarchical representation of HPE techniques, following the huge differentiation in this field, given by the presence or not of the training methods.

5.1. Depth images

The use of depth image is undoubtedly an advantage to solve HPE problems. However, differently from methods that use depth information in combination with the corresponding RGB image, a very recent method developed by Borghi et al. [10] uses only the depth data. The above mentioned method works in real time at 30 fps and it is based on a framework called POSEidon. The depth images are provided to a Convolutional Neural Network (CNN) that crop the image to the head region. This data is fed at the core of POSEidon. The pose is then estimated using CNNs that fuse the initial and built information and returns the head pose in pitch, yaw and roll degrees. The results are presented on BIWI, ICT-3DHP dataset and a homemade dataset called Pandora. A successor of POSEidon is undoubtedly the work in Borghi et al. [39], POSEidon^{+}. Here, in addition, is used a Motion images as support. Another work on the driving scenario is [40]. In this case the focus is to find a good HPE method with a non-invasive and calibration-free approach. The authors used depth images from a low cost sensor and tried to fit the head of the subject with a 3D model to perform a tracking system. Experiments were performed in a real-word scenario using homemade driving data composed of 2 drivers and over 4000 frames. On a different scenario but the use of the depth data alone, we can find the work in Papazov et al. [41]. This real time proposal for HPE introduces a triangular surface patch descriptor (TSP) that relate the shapes in a 3D synthetic modeled face with a triangular area. The depth map is matched with the synthetic models created offline in a training phase, using a fast nearest neighbor technique. This method has been evaluated on BIWI. In both [42–44], the landmark regression is used in an initial step. in the first, in cascade to estimate the pose and refine the head pose in each stage of the optimization process, testing on BIWI and a homemade dataset called 3DFEP. In the second, they explain the rotation effects by the use of tensor decomposition and trigonometric functions to obtain a linear-modeling that describes the structure of the rotation, testing on SASE dataset and BIWI. In the third they combined three different ways to search landmarks and then performing HPE using using geometric information, appearance and the dictionary-based HP. They tested on SASE. RGB-D images can be preferred as input, especially if there is some noise to deal with. In fact, in Li et al. [45], features are first detected using only the color image, and then the final HP is obtained combining this information with the depth data. A frontal image is used to build a subject template, the depth information are converted in a 3D point cloud building a template point cloud from

Table 3
Video datasets that contain pose annotation. The Popularity of each dataset is calculated using the amount of recent depth HPE method that use it, to the best of our knowledge ("Pop" is for popularity, i.e. the number of papers in this survey that use it). The methods are in Section 5.3. nd in the number of subjects column is for "not declared".

Dataset	Year	Resolution	#Subj	#Video	#FxV	Pop
UPNA	2016	1280 × 720	10	120	300	1
Boston University HP	2000	320 × 240	8	72	200	1
HPEG	2009	640 × 480	10	20	400	2
EYEDIAP	2014	1920 × 1080	16	94	4860	1
UBIPOSE	2016	640 × 480	nd	22	455	1
SHRP2	2012	360 × 420	41	41	1537	1

Table 4
The errors in degree for the methods using depth information of the BIWI dataset. Column "Code" is for code availability (accordingly to paperswithcode.com).

Depth on BIWI	$Err_P(°)$	Err_Y (°)	$Err_R(°)$	MAE(°)	Code
Methods that use BIWI in tests and training/calibration					
Borghi et al. [39] (2020)	1.6	1.7	1.7	1.67	No
Derkach et al. [43] (2019)	3.4	3.3	3.3	3.33	No
Luo et al. [31] (2019)	5.7	3.7	2.7	4.03	No
Wang et al. [42] (2017)	**1.16**	**1.3**	**1.51**	**1.32**	No
Sheng et al. [36] (2017)	2.0	2.3	1.9	2.07	No
Saeed et al. [50] (2015)	5.0	3.9	4.3	4.40	No
Methods that use BIWI only in tests					
Li et al. [45] (2017)	2.5	2.7	2.8	2.67	No
Yu et al. [48] (2017)	**1.5**	2.5	**2.2**	**2.07**	No
Li et al. [47] (2016)	1.7	2.2	3.2	2.37	No
Mukherjee et al. [49] (2015)	5.32	4.76	\	5.04	No
Meyer et al. [46] (2015)	2.1	**2.1**	2.4	2.20	No
Papazov et al. [41] (2015)	2.5	3.0	3.8	3.10	No

what they want to obtain the final pose called the target point cloud. Once the target point cloud is clear, the HP is estimated with Extended Levenberg–Marquardt (LMA). They performed their experiments on BIWI and on a homemade dataset of 12 subjects of over 10K RGB-D images. Some methods use the ICP modeling techniques. Luo et al. [31] obtaining the HPE by a discriminative random regression forest (DRRF) using a voting algorithm for the trees. Sheng et al. [36] estimates the rigid facial pose using as input the depth image and the probabilistic facial model built with the ICP. Meyer et al. [46] compose ICP with a particle swarm optimization by weighting the vertices of the morphable model to consider more important the ones on the visible part of the face. Li et al. [47] uses the ICP but also the RGB information are used, in this case to build a face template by fitting a 3D morphable model. All of them tests on BIWI and the second and the latter also on ICT-3DHP. It is clear that these techniques concern the quality of the reconstructed model only from the point of view of the HPE. However, if a multi-view observation of the subject is available, it is possible to build a more accurate 3D model also from a visible point of view. In [48] the authors propose an online reconstruction of a full 3D head model based on a variant of a technique called Kinect Fusion, obtaining smoothing synthetic samples. The results presented are experimented on the BIWI dataset. In [49] a CNN is used to estimate HP from multi-modal RGB-D data. The CNN is fine tuned replacing the last Softmax layer with an Euclidean loss layer that makes the NN a regression network. The BIWI dataset is used, as well as some other dataset like Caviar, HIIT, IDIAP head pose. Unfortunately, for this method, roll was not considered. In [50] the information from the RGB and the depth camera are used for different purposes. The RGB images are used to detect the human face and features. The depth information in this step is used to narrow down a location to find the face position. An SVM takes that data as input and predicts the HP. The results are presented for BIWI and ICT-3DHP datasets. In Table 4 we showed the results in terms of angular error of the above mentioned techniques on the most popular dataset: BIWI. As can be noticed, the best results were obtained by Wang et al. [42], both in mean than along the three axes. It is not a case that this framework benefits from two methods, landmark and regression, that are combined by developing more classifiers instead of a single one. The other considerable result in terms of MAE is by Borghi et al. [39]. Also here two powerful techniques are used, CNN and GAN, both of them are training based methods. However, in both cases, BIWI has been used in both training and testing. If we exclude this advantage, the resulting best method is in mean [48], with only a small exception for the yaw axis on which the method in Meyer et al. [46] is better. From the point of view of the angle representation used, in Luo et al. [31], Sheng et al. [36], Borghi et al. [39], Wang et al. [42], Saeed and Al-Hamadi [50] and Mukherjee and Robertson [49] it is preferred the Euler representation and in the other cases the matrix representation. We can thus observe that the matrix representation is preferred in methods that do not use the same dataset in training and test. From the point of view of the obtained performances, since the best ones use in one case the Euler and in the other the matrix, we can not claim that one representation is more convenient than the other.

5.2. 2D RGB images

The application of HPE at 2D RGB images is the most challenging and thriving field. For this reason, due to the numerous work published in the last five years, we will differentiate their presentation on the used techniques. The main difference that can be noticed is the involvement of training techniques. In Sections 5.2.2 and 5.2.1 we will present training and training-free methods respectively.

5.2.1. 2D RGB HPE using training-free techniques

As can be observed from the amount of recent work, the methods using training are more than the methods used training-free techniques. Also in the training-free methods some images are

used as reference, however, only as a reference for features, unlike training methods in which a part of the data from the same datasets to test undergoes further manipulation to obtain a trained NN. Methods without training present more heterogeneous characteristics. In [51], the core of the HPE is a quad-tree based method. The method proceeds work only on the landmarks after the detection. The same work is done for images of a reference synthetic model and the trees in binary vector form are then compared to detect pitch and yaw angles. Results are presented for the Pointing'04 dataset. It has been refined in Abate et al. [52] where the reference synthetic model has been changed with a more precise one. It also presented the best frame selection experiments on a youtube video in finding non-frontal faces. The experiments were performed on BIWI and AFLW. Also in Barra et al. [53] face and landmark are detected, however the core method produces cleaner data to analyse and better results. This method is based on the overlap of a web-shape on the detected landmark of the spider web. Also here a synthetic model is used as a reference and various web-shaped are tested to find the better configuration. This method has been tested on Pointing'04, BIWI and AFLW2000. In [54], very few keypoints are detected and the focus is about applications in mobile devices. The HP is calculated using the variation of three tracking points as a reference. The experiments, however, were made on a homemade dataset using a few smartphones and as a result cannot be comparable with the state-of-the-art. Using the property of 3D spheres during rotations, Peng et al. [55] solves the 3D HPE problem. The 3D sphere works as a model of the possible rotation of the head, supported by a method called Homeomorphic Manifold Analysis. The results of these methods are presented for the datasets CMU-MultiPIE, BU-4DFE, and AFW but only the total MAE is presented. In a similar way, using geometric properties on a synthetic model, the HP is estimated in Proença et al. [37]. The 2D points of the images are joined in the synthetic model using projective geometry. Convex energy minimization techniques are used to choose the set of landmarks in the model that result closer to the input. The dataset used for the test of HPE is AFLW, however there are no present the errors in pitch, yaw and roll. A method using feature-based technique is [50]. In this work a similarity kernel is learned using the feature correspondences of Geometric Blur features that results identity-invariant. The experiments are presented on the AFLW dataset, AFW, Youtube Faces, and McGill. However the pose error is discretized only in steps of 15° and presented only in the MAE form. Less focused on features but more on appearance based technique is [56]. In this method the preprocessing step is composed by face detection and 2D facial mask creation. The resulting image is then analysed to find the HP using the fractal encoding. Experiments were conducted on BIWI dataset, AFLW2000 and a subset of GOTCHA for the best frame selection. In [57] the difference with a classical HPE using training is enhanced by the challenge to recognize unseen head poses. Here a multivariate label distribution is used to represent the pose angle of a face image. The results are presented on Pointing'04, CASPEAL, and CMU-MultiPIE.

5.2.2. 2D RGB HPE using training techniques

Techniques based on training are, as already claimed, more explored in recent literature. In [58,59] the features extracted as in methods [53,56], respectively, are used in combination with regression. Various regression techniques are tested to improve the performance over BIWI, AFLW2000 and Pointing'04 Datasets. In [60] the proposed method that uses HOG-based descriptors to map the head pose starting from the face bounding boxes. The non linear regression is used to learn to map the features space in HP. The results are presented for BIWI and Pointing'04 dataset. Also in Liu et al. [61] we can find the use of regression, in particular regression forest, in an hybrid framework that introduces the concept of multi-structural features. The resulting method is quite robust and results are presented over Pointing'04, LFW, AFW and CCNU head pose datasets in the wide classroom. A completely different use of regression is made in Cao et al. [62] where the authors propose TriNet, a network architecture that regresses the head pose in 3 vectors by using an orthogonal loss function which punishes the model if the predicted ones are not orthogonal, as expected. Results were obtained over AFLW2000 and BIWI (trained on 300W-LP).

Regression seems to be particularly effective when the problem is related to low resolution images. As in the previous method, also in Chen et al. [63] the HOG features are combined with non-linear regression. In particular, here, the Support Vector Regression (SVR) is trained with extremely low resolution images. The authors also present improvements using depth information. For this reason the dataset tested is BIWI. We can find the use of the HOG also in Diaz-Chito et al. [64]. This approach is based on manifold learning-based methods and combines HOG, generalized discriminative common vectors, and continuous local regression. The experiments were conducted on the aforementioned Pointing'04, CMU-Pie, CASPEAL, and two other datasets called Taiwan and DrivFace. In [65], the initial data is in a higher resolution of the previous one, however authors focused on the advantage of the use of a small features vector. They propose to predict the bounding box of the face and its alignment together with the HP. The regression model is trained on partially observed output. The experiments were conducted on Pointing'04 and BIWI. HoG features can be also used in combination with other features techniques as the Uniform Pattern of Local Binary Pattern (UP-LBP). In [66] the image preprocessing involves the use of Second order HOG and UP-LBP that are fused through a normalized fusion to obtain the input of a classification method. The experiments are performed on CMU-Pie and CASPEAL, however no MSE of pitch, yaw and roll are presented. Another example of HOG and features combined can be observed in Alioua et al. [67]. The HOG is here fused with Haar features and Speed Up Robust Features (SURF) descriptor. The experiments were conducted only on Pointing'04, for this reason the roll component is not considered. Another tree-based algorithm using HOG is [68]. Here is presented a framework called Stacked Auto Encoder with Extreme Gradient Boosting (SAE-XGB). The results are presented only for Pointing'04, this means that the roll angle is not covered. Differently from the previous ones that extract features in a separate process, in Liu et al. [69] the regression is supported by a synthetic model. The head models used are 37 and the frames representing the poses are 74K. The experiments were performed on BIWI dataset. We introduced the linear and non-linear regression methods, however in Lathuiliére et al. [70] also a mixture of linear inverse regression is used. In particular a Convolutional Network (ConvNet) is used in combination with a Gaussian mixture of linear inverse regressions that can work also with relatively small datasets. The experiments were conducted on the BIWI dataset and two smaller homemade datasets of 20 K frames. The use of more than one regression method can be found also in Gou et al. [71]. They propose an unified method to detect landmarks and to estimate the head pose called Coupled Cascade Regression (CCR). The experimental results are presented on 300W_LP and the Boston University (BU) dataset. Multi-regression is also the strategy of Hsu et al. [72], that focused on the loss of the regression net. The loss regression function presented combines an L2 and an ordinal regression lost to train a CNN. The tests were performed on AFLW, AFLW2000, AFW and BIWI. Another very recent regression-based technique is presented in Yang et al. [73]. The method combines soft stagewise regression and features aggregation methods. Features maps are combined from different layers and it allows the method to learn meaningful intermediate features. The experiments were performed on AFLW2000, BIWI and 300W_LP. Of great

relevance is the recent use of regression applied to six degrees of freedom (6DoF) in Albiero et al. [74]. In this case there is no landmarks detection and the focus is to directly estimate the rigid transformation of all the face in the image by directly applying a proper trained R-CNN. Due to the nature of the problem, a lot of methods use regression, however, differently from the ones presented, some of them focus their attention on different steps of the process.

In [75] the image intensity is used in a multi loss-network using a different loss function for each angle with a classification and regression component. They also analyse the problem of low resolution. The results are discussed on AFLW2000 and BIWI. In opposition to the previous method, Gupta et al. [76] is based on the low precise location of facial keypoints that are subsequently used in the regression problem. The soft location is represented as five heatmap images, computed with a CNN to obtain the exact location. The experiments were conducted on BIWI and AFLW. Also in Cao et al. [77] the relationship between keypoints is explored to improve the accuracy. They used Convolutional Pose Machines (CPMs) to extract keypoints, confidence maps and feature maps, then leveraged several strategies to obtain the input for the HPE. The use of this simple CNN is justified by the target speed of the authors. The dataset used for the tests is AFLW. There are also other methods focused on a CNN approach. Xu and Kakadiaris [78] trained a GNet to obtain the face location,a preliminary pose and few landmarks. Then, an LNet refined this work learning local CNN features and predicting the final head pose. Results are shown for 300W_LP. The same approach to locate face and landmark as additional features, is adopted by the popular [79]. This algorithm performs at the same time face detection, landmark localization, pose estimation and gender recognition, all of them using CNNs. The HPE results are presented for the AFLW dataset. We can again find the landmarks as a support to CNNs in Xia et al. [80]. The authors focused on the generalization of CNNs, training and testing the network using different datasets. The HP is obtained by a Feed Forward CNN architecture. Experimental results are presented for AFLW2000, CASPEAL and BIWI. However, the results on BIWI are presented using depth information.

If on one hand the use of facial landmarks in the HPE process may help to improve the accuracy, on the other hand it can significantly and negatively impact on the computational time required. In [81] the use of HPE is focused for video, for this reason the aim is to obtain a very fast method. In fact, by the experiments conducted, the method takes only 21.8ms on a standard device to obtain the HPE. The head pose estimation is performed using the 3D mean of landmark location, in particular with 10 landmarks. The results are presented for AFLW2000. The correct face localization and its effects on HPE is analysed in Shao et al. [82]. The authors empirically analyse how the dimension of the face box impacts on the HPE and propose a new loss method in the CNN that they used. The results are presented for AFLW2000 and BIWI. The use of ResNet can be found also in Rieger et al. [83]. The original ResNet architecture is here adapted to work with images half the original size. The results are presented on AFLW and AFW. A completely different approach is proposed in Li et al. [84]. Here the HPE is decomposed in two problems. Anchor-Guided Pose Estimation (AGPE) and Task-Simplification oriented image Regularization (TSIR). These methods are combined in a unified end-to-end learning framework. The results are presented on the 300W_LP dataset, AFLW2000 and BIWI.

Other than the latter, there are various learning approaches that due to the specific methods they use in their framework, can not be assimilated to the previous works. In [85] the focus is again on the driver's attention. However, here an SSD object detection algorithm is here used to simultaneously classify and regress. Results are presented on AFLW2000 and 300W_*LP* dataset.

In [86] the core of the method is represented by the use of conditional random fields (CRF). The model trained in this way classifies each image in input using segmented face parts each of them giving a probability. The experiments were conducted on Pointing'04, AFLW2000 and two dataset better known for depth images, called BU-4DFE and ICT-3DH.

In [87] the focus is to make the method not sensitive to external conditions. Here the Peano–Hilbert space-filling curve is used to convert the images in one-dimensional vectors as a time series. The obtained sequences are used to train an encode-decode NN that generates labels for face orientation. The results in terms of HPE are presented on CASPEAL and Pointing'04 datasets. The method is, therefore, limited to pitch and yaw angles. In [88] is used a parametrized Multi-Variate Relevance Vector Machine (MVRVM) to learn the rotation angles. The author discussed the differences between their method and classical SVM and presented their results on Youtube Faces Dataset. In [89] the multivariate label approach is modified to alleviate problems derived from their application in unconstrained environments. They introduced regularization terms in the loss function using, to avoid over-fitting, the weighted Jeffreys divergence. The experiments were made on Pointing'04 and LFW datasets. In [90], to deal with occlusions and poor image quality, the proposed method is based on deep convolutional neural forests (D-CNF). A neurally connected split function (NCSF) is used as a new split node learning inside the D-CNF classification tree. The experiments were performed on Pointing'04, BU3D-HP and a dataset called CCNU-HP. All of them have only Pitch and Yaw annotations. In [91], the authors addressed the problem of the lack of sufficient training data for many poses, especially for large poses. In this case, they reformulate the facial pose estimation as a label distribution learning problem. They tested their methods on the popular AFLW2000, BIWI, AFLW and AFW. Another method that avoid landmark detection is in Zhang et al. [92]. Here the authors used a three-branch network architecture termed as FDN to perform HPE from features decoupling and cross category center loss. The experiments are provided for AFLW2000 and BIWI datasets. Finally, in Valle et al. [93], the authors proposed a multi-task approach to solve the face pose, alignment and visibility problems. The architecture they proposed is an encoder-decoder CNN with residual blocks and lateral skip connections. The HPE results are presented on AFLW, AFLW2000 and BIWI. In Tables 5–7 are shown the results obtained by the presented methods on the most popular datasets, BIWI, AFLW and Pointing'04 respectively.

In particular we consider inside the AFLW table both AFLW than AFLW2000 because the latter is a subset of AFLW and for this reason it presents the same characteristics in terms of heterogeneity and environment. We have highlighted in Tables the best results obtained in terms of angular error for pitch, yaw, roll and MAE. Since the use of the same dataset to perform both training and testing could be considered advantageous for the method, we split in two subtables each of the above mentioned tables to perform a fair comparison. As it can be noticed, on BIWI the minimum error is around 2.5°. Here the best results are quite similar between the two subtables, indeed, for both yaw and roll axes. The use of BIWI only in tests is advantageous. From the point of view of the representation used for the angles, the only method using quaternions is [72]. The matrix representation is preferred in Albiero et al. [74], Li et al. [84] and Cao et al. [62], the other methods on BIWI use Euler angles. In this case we do not notice a significance difference in terms of performances using one representation rather than another. In AFLW the use of the latter in both test and training has a significance difference in the best result. Here we can notice that the absolute best result is around 1.5° when AFLW is used in both training and test and around 4° in the other case. This because AFLW and AFLW2000 are quite small compared

Table 5
The errors in degree for the methods using 2D RGB images of the BIWI dataset (2011, available at http://data.vision.ee.ethz.ch/cvl/gfanelli/kinect_head_pose_db.tgz). Column "Code" is for code availability (accordingly to paperswithcode.com).

RGB on BIWI	Err_P(°)	Err_Y(°)	Err_R(°)	MAE (°)	Code
Methods that use BIWI in tests and training/calibration					
Bisogni et al. [56] (2021)*	6.23	4.05	3.30	4.53	No
Abate et al. [59] (2021)	5.29	6.58	3.80	5.28	No
Abate et al. [58] (2020)	**2.31**	**3.12**	**1.88**	**2.43**	No
Hsu et al. [72] (2019)	5.49	4.01	2.93	4.14	No
Gupta et al. [76] (2019)	3.49	3.46	2.74	3.23	Yes
Drouard er al. [65] (2017)	7.65	6.06	5.62	6.44	No
Lathuiliere et al. [70] (2017)	4.68	**3.12**	3.07	3.62	No
Chen et al. [63] (2016)	12.9	9.9	6.9	9.90	No
Drouard et al. [60] (2015)	5.9	4.9	4.7	5.17	No
Methods that use BIWI only in tests					
Valle et al. [93] (2021)	4.61	3.98	**2.39**	**3.66**	Yes
Albiero et al. [74] (2021)	5.03	3.42	3.27	3.91	Yes
Cao et al. [62] (2021)	4.75	**3.04**	4.11	3.97	No
Li et al. [84] (2020)	4.65	4.12	3.11	3.96	No
Barra et al. [53] (2020)*	**3.95**	6.21	4.16	4.77	No
Zhang et al. [92] (2020)	4.70	4.52	2.56	3.93	No
Abate et al. [52] (2019)*	7.51	4.07	5.50	5.69	No
Shao et al. [82] (2019)	7.25	4.59	6.15	6.00	Yes
Yang et al. [73] (2019)	4.96	4.27	2.76	4.00	Yes
Liu et al. [91] (2019)	5.61	4.12	3.14	4.29	No
Ruiz et al. [75] (2017)	6.60	4.81	3.27	4.89	Yes
Liu et al. [69] (2016)	4.3	4.5	2.4	3.73	No

Table 6
The errors in degree for the methods using 2D RGB images of the AFLW/AFLW2000 dataset (2016, avilable at https://www.cse.msu.edu/computervision/AFLW2000_FF-GAN.zip). Column "Code" is for code availability (accordingly to paperswithcode.com).

RGB on AFLW	Err_P(°)	Err_Y(°)	Err_R(°)	MAE (°)	Code
Methods that use AFLW in tests and training/calibration					
Bisogni et al. [56] (2021)*	7.46	6.28	5.53	6.42	No
Abate et al. [59] (2021)	6.90	6.70	4.48	6.02	No
Abate et al. [58] (2020)	5.43	4.31	2.62	4.09	No
Gupta et al. [76] (2019)	4.43	5.22	2.53	4.06	Yes
Ranjan et al. [79] (2019)	5.33	6.24	3.29	4.95	Yes
Xia et al. [80] (2019)	**2.05**	**0.63**	**1.70**	**1.46**	No
Khan et al. [86] (2019)	4.89	4.25	3.20	4.11	No
Rieger et al. [83] (2019)	6.5	8.5	3.9	6.30	No
Cao et al. [77] (2018)	7.14	7.04	3.86	6.01	No
Methods that use AFLW only in tests					
Valle et al. [93] (2021)	4.69	3.34	3.48	3.83	Yes
Albiero et al. [74] (2021)	3.54	4.56	3.24	3.78	Yes
Cao et al. [62] (2021)	5.76	4.19	4.04	4.66	No
Li et al. [84] (2020)	5.06	**2.78**	3.65	3.83	No
Sun et al. [85] (2020)	6.47	6.29	5.27	6.01	No
Wang et al. [81] (2020)	10.85	12.98	6.62	10.15	No
Barra et al. [53] (2020)*	4.82	3.11	**2.25**	**3.39**	No
Zhang et al. [92] (2020)	5.61	3.78	3.88	4.42	No
Abate et al. [52] (2019)*	7.60	7.60	7.17	7.46	No
Hsu et al. [72] (2019)	4.31	3.93	2.59	3.61	No
Yang et al. [73] (2019)	6.08	4.50	4.64	5.07	Yes
Xia et al. [80] (2019)	7.32	3.99	6.50	5.94	No
Shao et al. [82] (2019)	6.37	5.07	4.99	5.48	Yes
Liu et al. [91] (2019)	**3.02**	5.06	3.68	3.92	No
Ruiz et al. [75] (2017)	6.56	6.47	5.44	6.16	Yes

to other datasets used for the training (300W_LP in most cases). For this reason, training the method on another dataset results to be advantageous in this case. As in the previous table, also in this case the only method using quaterions is [72]. Methods in Albiero et al. [74], Xia et al. [80], Wang et al. [81] and Cao et al. [62] uses matrix and the other methods the Euler angle representation. We can notice that the best result is obtained for the method using matrix representation, however, since this representation is also used by the worst method, we can assume that the representation used is not indicative of the performance of the algorithm. In Pointing'04 the best result is around 1°. It is obtained in correspondence of a method using this dataset in both training that testing in a cross-fold-validation approach. Since this result is far better than the method only testing on Pointing'04, we can assume that in this case the use of the same dataset in training and testing represent a significance advantage. In this case, all the presented methods prefer the Euler representation to estimate the angles. All of those results should be considered under the specific information and framework used in the methods. In fact, methods denoted with $*$ are not using training techniques, as well as some methods in tables use manually annotated landmarks, facial annotations etc. For the majority of the presented methods, the computational time required is not reported in the papers, for this reason we can not add them in the presented tables. However, by the description of the methods in this section, we can conclude that methods that seem to have worst results in terms of angular error, are often associated with a low computational time required, since their focus is the speed instead of the accuracy. As a final consideration, in case of training free technique, the use of the dataset only in test means the support of a synthetic head to create a model. On the other hand, if the same dataset is not used and the method uses training, the training has been necessarily performed on another (and bigger) dataset that is in the majority of cases 300W_LP. In addition, methods that use a previously existing network in literature (as ResNet50 or VGGs) inherit also their starting weight that came from a long training process (usually performed on ImageNet [94]). This could be a significant advantage in terms of training time aiming to compensate the relatively small datasets available to train.

Table 7
The errors in degree for the methods using 2D RGB images of Ponting'04 dataset (2004, available at http://crowley-coutaz.fr/HeadPoseDataSet/HeadPoseImageDatabase.tar.gz). Column "Code" is for code availability (accordingly to paperswithcode.com).

RGB on Pointing'04	Err_P(°)	Err_Y(°)	MAE (°)	Code
Methods that use Pointing'04 in tests and training/calibration				
Abate et al. [58] (2020)	7.55	4.44	5.99	No
Bounoua et al. [87] (2020)	**0.82**	**1.78**	**1.30**	No
Vo et al. [68] (2020)	6.16	7.17	6.67	No
Khan et al. [86] (2019)	1.32	2.68	2.00	No
Xu et al. [89] (2019)	\	3.92	3.92	No
Diaz-Chito et al. [64] (2018)	9.6	8.1	8.85	No
Drouard et al. [65] (2017)	8.47	7.93	8.20	No
Liu et al. [61] (2017)	n	n	6.6	No
Alioua et al. [67] (2016)	4.6	6.1	5.35	No
Drouard et al. [60] (2015)	7.3	7.5	7.40	No
Methods that use Pointing'04 only in tests				
Barra et al. [53] (2020)*	**6.34**	**10.63**	**8.49**	No
Barra et al. [51] (2018)*	15	15	15.00	No

5.2.3. Training vs Training-free techniques

As introduced in the previous section, training techniques are more numerous than training-free techniques. The reason can be found in the popularity of training techniques as CNN, RNN etc., gained in last years, that lead to higher accuracies. In this section we want to evaluate the computational time required of a training vs. a training-free technique for, more or less, equal angular error. For this reason we will consider two methods, one representing the training technique, Hsu et al. [72] and one representing the training-free technique [53]. Both of those methods declare the computational time required to build the model and the devices on which the experiments were performed. The time required to test an image is more or less the same for both of them (30 fps) We reported this information, together with the angular errors on the AFLW2000 dataset, in Table 8. We can observe that the angular errors are quite similar, however there is a huge difference in the computational time required to build the model. It could be

Table 8
QuatNet (training) vs WSM (training-free).

Method and GPU	Time	$Err_P(°)$	$Err_Y(°)$	$Err_R(°)$	MAE(°)
WSM Barra et al. [53] (2020) Intel HD Graphics 515	0.16 h	4.82	3.11	2.25	3.39
QuatNet Hsu et al. [72] (2019) NVIDIA GTX 1080	4.5 h	4.32	3.93	2.59	3.61

said that the time required to build the model is not very relevant because it is performed only one time and not during the online tests. However, the model in QuatNet was built using a particular dataset as a reference, 300W_*LP*, it means that if the environment change or the characteristics of the subject are different in the application domain, 4.5 h of training will be again necessary to adjust the model to the new data. On the other hand, WSM uses a generic synthetic model as a reference, without a particular age, gender or ethnicity and without the use of the background during the computation. It means that the method has a good generalization property and, if it is still necessary, it can be remodeled in less than 10 min. Another thing to take under consideration is the devices on which those computational times were calculated. Following the results presented in a comparative website, specialized in GPU speed [95], in a ranking of the fastest GPU, from highest to lowest speed, the GPU used in QuatNet is 19th and the GPU used in WSM is 383th (data updated at 8th of July, 2020). This means that if we use a dumb proportion, on a NVIDIA GTX 1080 GPU, WSM will require about 28 s to rebuilt the model, in opposition to 4.5h required by QuatNet. It is clear that the more a method is generalizable, the less it will be necessary to rebuild it. Generalization is a very sensitive issue in training techniques and, in general, the choice of a method rather than another is conditioned by a higher request in one of generalization, speed or accuracy propriety.

5.3. Videos

Video are sequences of frames. From this point of view, they can be treated as depth images or 2D images. However, for some applications, it is necessary to take under consideration a temporal component together with the HPE solutions.

Some video applications are focused on the drive video study, in particular using the HPE for attention detection. In [96] different public available methods are compared over a challenging video dataset we already introduced, called SHRP2. The results is an assessment work of a new challenging dataset using known HPE methods. The use of video can be also justified to perform tracking during the HPE. In [97] a 2.5D Constrained Local Model (2.5DCLM) is developed focusing on devices with limited resources as tablets. Here the HP is estimated using the POSIT algorithm [15]. In addition, the initial position of the head in the camera acts like a reference system for the head rotations. The experiments for HPE are conducted on the UPNA dataset. Also in Kim et al. [98] the aim is to develop a HPE method for mobile devices. The face and facial features are detected, then the facial tracking is used to perform HPE with the previous features detected. The detection is Haar-based. The experiments were performed on the HPEG dataset, however does not declare the MAE for the estimated angles. We have then another tracking-based method, that also operates in real-time (40 fps) but not tested on mobile devices, Barros et al. [99]. Here also the 3D facial landmarks are refined using the temporal component, extracted from several frames through a Kalman filter. The authors present the results of their method on the Boston University HP dataset and also on an homemade dataset they made publicly available. The work of Cristina and Camilleri [100] uses a less constrained scenario compared to the previous ones, even if it performs only pitch and yaw. The method is training-free and the aim is to estimate the HP in real-time extracting the trajectories of few features points spread randomly over the face region. The experiments were performed on HPEG and EYEDIAP datasets. All of those methods use video with 2D frames, however some works focused also on video with RGB-D frames. In [47] the HPE is supported by an online face template reconstruction. This method uses up to 9 frames automatically selected from the video. The results, due to the need for depth video are presented on the BIWI and ICT-3DHP dataset. We can find online fitting also in Yu et al. [101]. Here the strength of a 3DMM model operating online is combined with the reconstruction of a full head 3D model without prior knowledge. The experiments were performed on the BIWI and the UbiPose dataset. In Table 9 there is a summary of the presented methods on Video. Since there is not a lot of recent work in this topic using video and the datasets used are quite heterogeneous, we added in the Table all the angular errors we detected in the related papers. For this reason we do not highlight the better results, because the environment in the dataset used are quite different.

As introduced in Section 1, if we analyse the techniques used for depth images, 2D-RGB images and Videos, we can find a prevalence of training techniques that are almost twice as the training-free techniques. This unbalance can be observed especially in applications at 2D-RGB images, the most popular data for HPE. The better results for 2D-RGB images are produced by Gupta et al. [76], Xia et al. [80] and Bounoua et al. [87], on BIWI, AFLW and Pointing'04, respectively. Each one of them uses training techniques, in particular, Gupta et al. [76] uses regression, Xia et al. [80] uses CNN, and [87] uses an encode-decode NN. However, if we compare the mean error of the training-free methods with the mean error of the training methods, we will obtain 4.99° for the training-free and 5° for the training methods. This result demonstrates that the use of training techniques can not be justified in needs of performances. In Section 5.2.3 we also compared the required time to find the best configuration of two methods, training and training-free, which have similar performances and both operating in real time, finding out that it is even smaller for training-free methods. From these considerations we can suppose that the training methods are preferred for their speed in the testing phase, resulting in most cases far below that the real-time, making it possible to elaborate a larger amount of frames. However, when we move to video applications, in which more than one frame must be computed, low computational requirements are preferred and we can observe a preference for features-based techniques as against training techniques. In addition, to compare performances in videos, is a more complex task, since purposes and datasets present different characteristics. The best result in mean is obtained by Ackland et al. [97] on the UPNA dataset, in which the movements of the head are guided. The mean error is about 3.6° lower than the better one obtained on the competitive SHRP2 by Paone et al. [96], using a Government-Out-Of-The-Shelf (GOTS) which includes face detection, tracking, landmark detection, pose estimation, and face recognition. On Depth images the best result is obtained by Wang et al. [42], that uses landmark regression in an optimization process involving a regression forest. Here again we find the involvement of a training step, however, we can observe that also in this case, training-free techniques reached very similar results, as in Yu et al. [48]. We have to mention that training-free techniques need more time to be developed, since it is not a feature-learning al-

Table 9
The errors in degree for the methods using videos of various datasets.Column "Code" is for code availability (accordingly to paperswithcode.com).

Method	Dataset	Err_p (°)	Err_y (°)	Err_r (°)	MAE (°)	Code
Ackland et al. [97] (2019)	UPNA	2.47	1.88	0.81	1.72	No
Diaz Barros et al. [99] (2018)	Boston University	3.23	4.12	2.16	3.17	No
Cristina [100] (2018)	EYEDIAP	2.39	2.56	\	2.48	No
Li et al. [47] (2018)	BIWI	3.2	3.0	3.0	3.07	No
Yu et al. [101] (2018)	BIWI	1.45	2.54	2.10	2.03	No
Yu et al. [101] (2018)	UbiPose	4.37	4.63	3.38	4.28	No
Cristina et al. [100] (2016)	HPEG	3.05	3.04	\	3.05	No
Li et al. [47] (2016)	ICT-3DHP	3.1	3.3	2.9	3.1	No
Paone et al. [96] (2015)	SHRP2	10.2	4.47	1.51	5.14	No

gorithm. In this case, the authors of the method have to spend a considerable effort to choose the features to be extracted that suits best the problem. In opposition to deep learning techniques, in which the network itself extracts and classifies what it consider the most relevant part/features of the image. We can, therefore, conclude that the use of training or training-free techniques should be determined by the environment, the available computational resources and the capability to perform a comprehensive training, more than the required performances.

6. Applications of HPE

From the previous sections we have observed that some works need to carefully detect the application field before to build a HPE method. This is because computational efficiency and accuracy play a different role depending on the application. In this section we will discuss some of the application field of HPE that emerged more or less distinctly from the recent works analysed in this survey.

6.1. Driver attention detection

The environment that emerged more clearly from the previous sections is the HPE for Driver Attention Detection. In past years a lot of technologies were developed to prevent car accidents like alcohol tests, speed measurement radar etc. However, those technologies are not always available on the vehicle and do not cover many other issues related to the driver's attention. The HPE, detected in real time, can be a very efficient support to understand if the driver is sleepy, if he/she looks at the phone instead of the road etc. The wider dataset specifically built for this purpose is the aforementioned SHRP2 [28]. The method operating in this field can use both video and depth images and not always presents their results on specifically built datasets. In method using depth image we want to remind [39] operating with both depth and RGB images at the same speed. For the RGB images, we find the work of Alioua et al. [67], where the speed is 20 fps. The only recent work using the mentioned SHRP2 dataset, is [96], having the same speed of the video recorded in the dataset, 15 fps.

6.2. Best frame selection

The selection of the best frame is the possibility in a video sequence to choose a specific head pose frame. This application of the HPE finds its usefulness in face recognition purposes. It is clear that the environment in which this kind of application is more required is surveillance. The work in this survey that specifically tests its HPE algorithm for this purpose is [53]. Here, starting from the sequences of the GOTCHA dataset, the authors selected in an automatic way the most frontal frame using their HPE algorithm with an integration for the selection. It is not necessary that the aim is to select the more frontal frame, sometimes if the purpose is to perform recognition from another part of the head, as an example the ear, it can be more convenient to select a specific head rotation. Those kinds of applications are not only time-saving but also space-saving. In fact, if we imagine studying many long surveillance videos covering a wide time span, we can save only the best frame for recognition in each video.

6.3. Face frontalization

The face frontalization is focused on the aim to obtain a frontal image starting from an image with a non-frontal pose. The motivations under the use of this approach is clearly identifiable in a more suitable face recognition. The HPE is used in this field of application as a preprocessing step to perform before the reconstruction of the missing parts [102]. Once the head pose is estimated it can successfully fit a 3D model that can rotate using the HP information to obtain the frontal view. If we combine this application (face frontalization) with the previous one (best frame selection) we can think to use HPE to build a 3D model of the face of a subject when he/she assumes different non-sequential head rotations during the video.

6.4. New challenges

Since in the head pose estimation field, the accuracies obtained in the last years are very high, the latter represent no more the only aim to reach performing HPE. In [103] the aim is drive attention, however the authors focused less on the accuracy, and more on some effect typically associated at driver video. They evaluate the distribution of head orientation during the drive and this can help methods in this field to spend their energy to improve performance in those ranges. A less explored field is the use of HPE in thermal infrared images. In [104] a combined modular system is proposed to solve the HPE problem on infrared images together with the face detection, face tracking and emotion recognition. The experiments were conducted on a homemade video dataset with a MAE of around 3°, demonstrating that it is possible to operate on this new kind of data reaching accuracies comparable with the state of the art. On the same line we find the work in Liu et al. [105], they proposed a CNN specifically built to be adapted to the IR HPE problem that outperform features-based HPE methods in this field. Instead of IR, Visible or Depth images, very recent works are focused on the use of 3D point clouds generated from depth information, reaching very interesting results on 36 classes of angles with step of 5° [106]. To avoid generalization problem, approach like [53] where a synthetic model is preferred to train the architecture, were also considered in Wang et al. [107] were, however, it is combined with deep learning. In [108], the innovation is represented by the application field. The HPE framework proposed by the authors is focused on the use of pose in human-machine interaction. To reach this aim, they combine different techniques we

explored in previous sections, in particular the combination of an online ad offline evaluation. The results are tested preliminarily on BIWI and ICT-3D dataset, and then on a homemade dataset built to suit the purpose of the research. The errors are in the mean of 2.18° on their dataset. A new kind of application of HPE can be found in Yang et al. [109]. Here the authors propose to use the anomalies in the head pose to detect AI-generated face images or video, named Deep Fakes. The motivation is that in a Deep Fake image the central region is from a synthesized image and a mismatch in these landmarks lead to a larger difference between the original and Deep Fake head pose.

7. Summary and conclusions

The head pose estimation is a wide problem explored from different inputs and application scenarios. In this survey we introduced the problem related to the HPE, starting from the definition of the rotations that a human head can perform in its mathematical form, in pitch, yaw and roll. The data used for HPE are mainly three, depth images, 2D RGB images and video. From this differentiation we presented the corresponding available dataset describing their characteristics and giving an estimate of their popularity among recent HPE techniques. Before introducing the method, to better understand the latter, we analysed the preprocessing techniques that seem to be in common to more frameworks and that can be resumed in face detection, landmark detection and 3D modeling. Following the same division adopted for the datasets, we presented the HPE methods developed in recent years, in particular from 2015, when the last comprehensive survey on HPE was presented, to the best of our knowledge. We finally analyzed the application on which the HPE is used and some new challenging on the topic. From this overview of recent HPE techniques we can claim that the accuracy reached from the last methods is impressive. We observed that recent methods try to focus on new challenging scenarios more than accuracy obtained on a constrained dataset built in a laboratory, however, obtain an exact head location, without the use of other methods to label the images, remain challenging. Taking a look at the new challenges we can conclude that HPE has, for its characteristics, the opportunity to be effective in more field and emerging problems than those in which has been submitted until now. HPE represents an interesting technique to be used in support of biometrics frameworks, which can boast on a thriving and excellent literature.

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] E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation in computer vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 607–626.
[2] Q. Dang, J. Yin, B. Wang, W. Zheng, Deep learning based 2D human pose estimation: asurvey, Tsinghua Sci. Technol. 24 (6) (2019) 663–676.
[3] B. Czupryński, A. Strupczewski, High accuracy head pose tracking survey, in: D. Ślzak, G. Schaefer, S.T. Vuong, Y.-S. Kim (Eds.), Active Media Technology, 2014, pp. 407–420.
[4] E. Amador, R. Valle, J.M. Buenaposada, L. Baumela, Benchmarking head pose estimation in-the-wild, in: M. Mendoza, S. Velastín (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 2018, pp. 45–52.
[5] M.I.S.B. (MISB), Misb standard 0601, UAS Datalink Local Metadata(2014).
[6] G. Fanelli, M. Dantone, J. Gall, A. Fossati, L. Van Gool, Random forests for real time 3D face analysis, Int. J. Comput. Vis. 101 (3) (2013) 437–458.
[7] T. Baltrušaitis, P. Robinson, L. Morency, 3D constrained local model for rigid and non-rigid facial tracking, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2610–2617.
[8] I. Lüsi, S. Escalera, G. Anbarjafari, SASE: RGB-depth database for human head pose estimation, in: Computer Vision - ECCV 2016 Workshops, 2016, pp. 325–336.
[9] M.D. Breitenstein, D. Kuettel, T. Weise, L. van Gool, H. Pfister, Real-time face pose estimation from single range images, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[10] G. Borghi, M. Venturelli, R. Vezzani, R. Cucchiara, Poseidon: face-from-depth for driver pose estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5494–5503.
[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, in: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition, 2008, pp. 1–8.
[12] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1618.
[13] N. Gourier, D. Hall, J.L. Crowley, Estimating face orientation from robust detection of salient facial structures, FG NET, Workshop on Visual Observation of Deictic Gestures, 2004.
[14] M. Koestinger, 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. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011.
[15] D.F. DeMenthon, L.S. Davis, Model-based object pose in 25 lines of code, in: G. Sandini (Ed.), Computer Vision — ECCV'92, 1992, pp. 335–343.
[16] Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou, Joint 3D face reconstruction and dense alignment with position map regression network, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, 2018, pp. 557–574.
[17] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2879–2886.
[18] X. Zhu, Z. Lei, X. Liu, H. Shi, S. Li, Face alignment across large poses: a 3D solution, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 146–155.
[19] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, D. Zhao, The CAS-PEAL large-scale Chinese face database and baseline evaluations, IEEE Trans. Syst., Man, Cybern. - Part A 38 (1) (2008) 149–161.
[20] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: CVPR 2011, 2011, pp. 529–534.
[21] M. Demirkus, J. Clark, T. Arbel, Robust semi-automatic head pose labeling for real-world face video sequences, Multimed. Tools Appl. 70 (2013) 495–523.
[22] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, M. Castrillón-Santana, Gotcha-I: a multiview human videos dataset, in: S.M. Thampi, G. Martinez Perez, R. Ko, D.B. Rawat (Eds.), Security in Computing and Communications, 2020, pp. 213–224.
[23] M. Ariz, J.J. Bengoechea, A. Villanueva, R. Cabeza, A novel 2D/3D database with automatic face annotation for head tracking and pose estimation, Comput. Vis. Image Underst. 148 (C) (2016) 201–210.
[24] M. La Cascia, S. Sclaroff, V. Athitsos, Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D models, IEEE Trans. Pattern Anal. Mach. Intell. 22 (4) (2000) 322–336.
[25] S. Asteriadis, D. Soufleros, K. Karpouzis, S. Kollias, A natural head pose and eye gaze dataset, in: Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots, AFFINE '09, 2009.
[26] K.A. Funes Mora, F. Monay, J.-M. Odobez, EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras, Association for Computing Machinery, 2014, pp. 255–258.
[27] S. Muralidhar, L.S. Nguyen, D. Frauendorfer, J.-M. Odobez, M. Schmid Mast, D. Gatica-Perez, Training on the job: behavioral analysis of job interviews in hospitality, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI '16, 2016, pp. 84–91.
[28] K. Campbell, The SHRP2 Naturalistic Driving Study, TR News, 2012, pp. 30–35.
[29] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[30] Y. Pang, Y. Yuan, X. Li, J. Pan, Efficient hog human detection, Signal Process. 91 (4) (2011) 773–781.
[31] C. Luo, J. Zhang, J. Yu, C.W. Chen, S. Wang, Real-time head pose estimation and face modeling from a depth image, IEEE Trans. Multimed. 21 (10) (2019) 2473–2481.
[32] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: additive angular margin loss for deep face recognition, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4685–4694.
[33] G. Jeevan, G.C. Zacharias, M.S. Nair, J. Rajan, An empirical study of the impact of masks on face recognition, Pattern Recognit. 122 (2022) 108308.
[34] S. Colaco, D.S. Han, Facial keypoint detection with convolutional neural networks, in: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2020, pp. 671–674.
[35] Y. Wu, Q. Ji, Facial landmark detection: a literature survey, Int. J. Comput. Vis. 127 (2018) 115–142.
[36] L. Sheng, J. Cai, T.-J. Cham, V. Pavlovic, K.N. Ngan, A generative model for depth-based robust 3D facial pose tracking, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4598–4607.
[37] H. Proença, J.C. Neves, S. Barra, T. Marques, J.C. Moreno, Joint head pose/soft label estimation for human recognition in-the-wild, IEEE Trans. Pattern Anal. Mach. Intell. 38 (12) (2016) 2444–2456.
[38] F.-J. Chang, A. Tran, T. Hassner, I. Masi, R. Nevatia, G. Medioni, Deep, landmark-free FAME: face alignment, modeling, and expression estimation, Int. J. Comput. Vis. 127 (2019) 930–956.

[39] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, R. Cucchiara, Face-from-depth for head pose estimation on depth images, IEEE Trans. Pattern Anal. Mach. Intell. 42 (3) (2020) 596–609.

[40] M. Breidt, H.H. Bülthoff, C. Curio, Accurate 3D head pose estimation under real-world driving conditions: a pilot study, in: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 2016, pp. 1261–1268.

[41] C. Papazov, T.K. Marks, M. Jones, Real-time 3D head pose and facial landmark estimation from depth images using triangular surface patch features, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4722–4730.

[42] J. Wang, J. Zhang, C. Luo, F. Chen, Joint head pose and facial landmark regression from depth images, Comput. Vis. Media 3 (2017) 229–241.

[43] D. Derkach, A. Ruiz, F. Sukno, Tensor decomposition and non-linear manifold modeling for 3D head pose estimation, Int. J. Comput. Vis. 127 (2019) 1565–1585.

[44] D. Derkach, A. Ruiz, F.M. Sukno, Head pose estimation based on 3-D facial landmarks localization and regression, in: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), 2017, pp. 820–827.

[45] C. Li, F. Zhong, Q. Zhang, X. Qin, Accurate and fast 3D head pose estimation with noisy RGBD images, Multimed. Tools Appl. 77 (2017) 14605–14624.

[46] G.P. Meyer, S. Gupta, I. Frosio, D. Reddy, J. Kautz, Robust model-based 3D head pose estimation, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3649–3657.

[47] S. Li, K.N. Ngan, R. Paramesran, L. Sheng, Real-time head pose tracking with online face template reconstruction, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2016) 1922–1928.

[48] Y. Yu, K.A.F. Mora, J. Odobez, Robust and accurate 3D head pose estimation through 3DMM and online head model reconstruction, in: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), 2017, pp. 711–718.

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

[50] A. Saeed, A. Al-Hamadi, Boosted human head pose estimation using kinect camera, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 1752–1756.

[51] P. Barra, C. Bisogni, M. Nappi, S. Ricciardi, Fast QuadTree-based pose estimation for security applications using face biometrics, in: M.H. Au, S.M. Yiu, J. Li, X. Luo, C. Wang, A. Castiglione, K. Kluczniak (Eds.), Network and System Security, 2018, pp. 160–173.

[52] A.F. Abate, P. Barra, C. Bisogni, M. Nappi, S. Ricciardi, Near real-time three axis head pose estimation without training, IEEE Access 7 (2019) 64256–64265.

[53] P. Barra, S. Barra, C. Bisogni, M. De Marsico, M. Nappi, Web-shaped model for head pose estimation: an approach for best exemplar selection, IEEE Trans. Image Process. 29 (2020) 5457–5468.

[54] E.N.A. Neto, R.M. Barreto, R.M. Duarte, J.P. Magalhaes, C.A.C.M. Bastos, T.I. Ren, G.D.C. Cavalcanti, Real-time head pose estimation for mobile devices, in: H. Yin, J.A.F. Costa, G. Barreto (Eds.), Intelligent Data Engineering and Automated Learning - IDEAL 2012, 2012, pp. 467–474.

[55] X. Peng, J. Huang, Q. Hu, S. Zhang, D.N. Metaxas, Three-dimensional head pose estimation in-the-wild, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, 2015, pp. 1–6.

[56] C. Bisogni, M. Nappi, C. Pero, S. Ricciardi, HP2IFS: head pose estimation exploiting partitioned iterated function systems, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 1725–1730.

[57] G.-l. Sang, H. Chen, G. Huang, Q.-j. Zhao, Unseen head pose prediction using dense multivariate label distribution, Front. Inf. Technol. Electron. Eng. 17 (2016) 516–526.

[58] A. Abate, P. Barra, C. Pero, M. Tucci, Head pose estimation by regression algorithm, Pattern Recognit. Lett. 140 (2020) 179–185.

[59] A.F. Abate, P. Barra, C. Pero, M. Tucci, Partitioned iterated function systems by regression models for head pose estimation, Mach. Vis. Appl. 32 (5) (2021) 1–8.

[60] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, R. Horaud, Head pose estimation via probabilistic high-dimensional regression, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 4624–4628.

[61] Y. Liu, Z. Xie, X. Yuan, J. Chen, W. Song, Multi-level structured hybrid forest for joint head detection and pose estimation, Neurocomputing 266 (2017) 206–215.

[62] Z. Cao, Z. Chu, D. Liu, V.Y. Chen, A vector-based representation to enhance head pose estimation, in: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 1187–1196.

[63] J. Chen, J. Wu, K. Richter, J. Konrad, P. Ishwar, Estimating head pose orientation using extremely low resolution images, in: 2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), 2016, pp. 65–68.

[64] K. Diaz-Chito, J. Martínez Del Rincón, A. Hernández-Sabaté, D. Gil, Continuous head pose estimation using manifold subspace embedding and multivariate regression, IEEE Access 6 (2018) 18325–18334.

[65] V. Drouard, R. Horaud, A. Deleforge, S. Ba, G. Evangelidis, Robust head-pose estimation based on partially-latent mixture of linear regressions, IEEE Trans. Image Process. 26 (3) (2017) 1428–1440.

[66] Z. Zhao, Q. Zheng, Y. Zhang, X. Shi, A head pose estimation method based on multi-feature fusion, in: 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB), 2019, pp. 150–155.

[67] N. Alioua, A. Amine, A. Rogozan, A. Bensrhair, M. Rziza, Driver head pose estimation using efficient descriptor fusion, EURASIP J. Image Video Process. 2016 (2016) 1–14.

[68] M.T. Vo, T. Nguyen, T. Le, Robust head pose estimation using extreme gradient boosting machine on stacked autoencoders neural network, IEEE Access 8 (2020) 3687–3694.

[69] X. Liu, W. Liang, Y. Wang, S. Li, M. Pei, 3D head pose estimation with convolutional neural network trained on synthetic images, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 1289–1293.

[70] S. Lathuiliére, R. Juge, P. Mesejo, R. Muñoz-Salinas, R. Horaud, Deep mixture of linear inverse regressions applied to head-pose estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7149–7157.

[71] C. Gou, Y. Wu, F. Wang, Q. Ji, Coupled cascade regression for simultaneous facial landmark detection and head pose estimation, in: 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 2906–2910.

[72] H. Hsu, T. Wu, S. Wan, W.H. Wong, C. Lee, Quatnet: quaternion-based head pose estimation with multiregression loss, IEEE Trans. Multimed. 21 (4) (2019) 1035–1046.

[73] T. Yang, Y. Chen, Y. Lin, Y. Chuang, FSA-Net: learning fine-grained structure aggregation for head pose estimation from a single image, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, 2019, pp. 1087–1096.

[74] V. Albiero, X. Chen, X. Yin, G. Pang, T. Hassner, img2pose: face alignment and detection via 6DoF, face pose estimation, CVPR, 2021.

[75] N. Ruiz, E. Chong, J. Rehg, Fine-grained head pose estimation without keypoints, in: 2018 IEEE CVF Conference on Computer Vision and Pattern Recognition Workshops, 2017.

[76] A. Gupta, K.C. Thakkar, V. Gandhi, P.J. Narayanan, Nose, eyes and ears: head pose estimation by locating facial keypoints, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 1977–1981.

[77] Y. Cao, O. Canévet, J. Odobez, Leveraging convolutional pose machines for fast and accurate head pose estimation, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 1089–1094.

[78] X. Xu, I.A. Kakadiaris, Joint head pose estimation and face alignment framework using global and local CNN features, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017, pp. 642–649.

[79] R. Ranjan, V.M. Patel, R. Chellappa, Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2019) 121–135.

[80] J. Xia, L. Cao, G. Zhang, J. Liao, Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks, IEEE Access 7 (2019) 48470–48483.

[81] W. Wang, X. Chen, S. Zheng, H. Li, Fast head pose estimation via rotation-adaptive facial landmark detection for video edge computation, IEEE Access 8 (2020) 45023–45032.

[82] M. Shao, Z. Sun, M. Ozay, T. Okatani, Improving head pose estimation with a combined loss and bounding box margin adjustment, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019, pp. 1–5.

[83] I. Rieger, T. Hauenstein, S. Hettenkofer, J.-U. Garbas, Towards real-time head pose estimation: exploring parameter-reduced residual networks on in-the-wild datasets, in: F. Wotawa, G. Friedrich, I. Pill, R. Koitz-Hristov, M. Ali (Eds.), Advances and Trends in Artificial Intelligence. From Theory to Practice, 2019, pp. 123–134.

[84] J. Li, J. Wang, F. Ullah, An end-to-end task-simplified and anchor-guided deep learning framework for image-based head pose estimation, IEEE Access 8 (2020) 42458–42468.

[85] J. Sun, S. Lu, An improved single shot multibox for videorate head pose prediction, IEEE Sens. J. 20 (2020) 1.

[86] K. Khan, N. Ahmad, F. Khan, I. Syed, A framework for head pose estimation and face segmentation through conditional random fields, Signal, Image Video Process. 14 (2019) 159–166.

[87] B.A. Mekami, H.S. Benabderrahmane, Leveraging deep learning with symbolic sequences for robust head poses estimation, Pattern Anal. Appl. 23 (2020) 1391–1406.

[88] M. Selim, A. Pagani, D. Stricker, Real-time head pose estimation using multivariate RVM on faces in the wild, CAIP, 2015.

[89] L. Xu, J. Chen, Y. Gan, Head pose estimation using improved label distribution learning with fewer annotations, Multimed. Tools Appl. 78 (2019) 1–22.

[90] Y. Liu, Z. Xie, X. Gong, F. Fang, Deep transfer feature based convolutional neural forests for-head pose estimation, in: Image and Video Technology, 2018, pp. 5–16.

[91] Z. Liu, Z. Chen, J. Bai, S. Li, S. Lian, Facial pose estimation by deep learning from label distributions, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 1232–1240.

[92] H. Zhang, M. Wang, Y. Liu, Y. Yuan, FDN: feature decoupling network for head pose estimation, AAAI, 2020.

[93] R. Valle, J.M. Buenaposada, L. Baumela, Multi-task head pose estimation in-the-wild, IEEE Trans. Pattern Anal. Mach. Intell. 43 (8) (2021) 2874–2881.

[94] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, CVPR09, 2009.

[95] Nvidia, GTX 1080 GPU vs. Intel HD graphics 515, (https://technical.city/en/video/GeForce-GTX-1080-mobile-vs-HD-Graphics-515).

[96] J. Paone, D. Bolme, R. Ferrell, D. Aykac, T. Karnowski, Baseline face detection, head pose estimation, and coarse direction detection for facial data in the

SHRP2 naturalistic driving study, in: 2015 IEEE Intelligent Vehicles Symposium (IV), 2015, pp. 174–179.
[97] S. Ackland, F. Chiclana, H. Istance, S. Coupland, Real-time 3D head pose tracking through 2.5Dconstrained local models with local neural fields, Int. J. Comput. Vis. 127 (6–7) (2019) 579–598.
[98] J. Kim, G. Lee, J. Jung, K. Choi, Real-time head pose estimation framework for mobile devices, Mob. Netw. Appl. 22 (2016) 634–641.
[99] J.M.D. Barros, B. Mirbach, F.R. Garcia, K. Varanasi, D. Stricker, Real-time head pose estimation by tracking and detection of keypoints and facial landmarks, VISIGRAPP, 2018.
[100] S. Cristina, K.P. Camilleri, Model-free non-rigid head pose tracking by joint shape and pose estimation, Mach. Vis. Appl. 27 (8) (2016) 1229–1242.
[101] Y. Yu, K.A.F. Mora, J. Odobez, Headfusion: 360° head pose tracking combining 3D morphable model and 3D reconstruction, IEEE Trans. Pattern Anal. Mach. Intell. 40 (11) (2018) 2653–2667.
[102] T. Hassner, S. Harel, E. Paz, R. Enbar, Effective face frontalization in unconstrained images, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4295–4304.
[103] S. Jha, C. Busso, Challenges in head pose estimation of drivers in naturalistic recordings using existing tools, in: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 2017, pp. 1–6.
[104] M. Kopaczka, J. Schock, J. Nestler, K. Kielholz, D. Merhof, A combined modular system for face detection, head pose estimation, face tracking and emotion recognition in thermal infrared images, in: 2018 IEEE International Conference on Imaging Systems and Techniques (IST), 2018, pp. 1–6.
[105] H. Liu, X. Wang, W. Zhang, Z. Zhang, Y.-F. Li, Infrared head pose estimation with multi-scales feature fusion on the IRHP database for human attention recognition, Neurocomputing 411 (2020) 510–520.
[106] Y. Xu, C. Jung, Y. Chang, Head pose estimation using deep neural networks and 3D point clouds, Pattern Recognit. 121 (2022) 108210.
[107] Y. Wang, W. Liang, J. Shen, Y. Jia, L.-F. Yu, A deep coarse-to-fine network for head pose estimation from synthetic data, Pattern Recognit. 94 (2019) 196–206.
[108] F. Madrigal, F. Lerasle, Robust head pose estimation based on key frames for human-machine interaction, EURASIP J.Image Video Process. 13 (2020) 1–19.
[109] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8261–8265.

Andrea F. Abate received the Laurea degree (summa cum laude) in computer science from the University of Salerno, Salerno, Italy, in 1991, and the Ph.D. degree in applied mathematics and computer science from the University of Pisa, Pisa, Italy, in 1998. Since 2006, he has been serving as an Associate Professor for the University of Salerno, where he is also the Co-Director of the Virtual Reality Laboratory. He has authored many scientific papers published in scientific journals and proceedings of refereed international conferences and co-edited one book. His current research interests include multibiometric systems, virtual/augmented/mixed reality, haptics, and human-computer interaction. Dr. Abate is a member of the IEEE Haptics Technical Committee and a member of the International Association for Pattern Recognition. He currently serves as an Associate Editor for Pattern Recognition Letters and IEEE Access.

Carmen Bisogni received the B.S. degree and M.S. degree (cum Laude) in Mathematics from University of Salerno in 2015 and 2017, respectively. She received the Ph.D. in Computer Science in 2021, from University of Salerno. She is currently a research fellow at the Biometric and Image Processing Laboratory (BIPLAB) at University of Salerno, Italy. Her research interests include applied mathematics for Machine Learning, Biometrics, Image Processing and Statistical Analysis. Dr. Bisogni is member of IEEE and GIRPR/IAPR, member of the Editorial Board of Electronics (MDPI) and Associate Editor of the IEEE Biometrics Council Newsletter.

Aniello Castiglione received the Ph.D. degree in computer science from the University of Salerno, Fisciano, Italy, in 2007. He is currently with the Department of Science and Technology, University of Naples Parthenope, Naples, Italy. He authored more than 200 papers in international journals and conferences. He has served in the organization of more than 200 international conferences. He served as a Reviewer for approximately 100 international journals and the Managing Editor for two ISI-ranked international journals. He was a Guest Editor for around 20 special issues and served as an Editor on more than ten Editorial Boards of international journals. One of his papers (published in the IEEE Transactions on Dependable and Secure Computing) was selected as the “Featured Article” in the “IEEE Cybersecurity Initiative” in 2014. His current research interests include information forensics, digital forensics, security and privacy on distributed systems, steganography, communication networks, applied cryptography, and sustainable computing.

Michele Nappi received the Laurea degree (cum laude) in computer science from the University of Salerno, Fisciano, Italy, in 1991, the M.Sc. degree in information and communication technology from the Istituto Internazionale per gli Alti Studi Scientifici “E.R. Caianiello,” Vietri sul Mare, Italy, in 1991, and the Ph.D. degree in applied mathematics and computer science from the University of Padova, Padua, Italy, in 1997. He is currently a Full Professor of Computer Science with the University of Salerno. He is also a Team Leader of the Biometric and Image Processing Lab. He has coauthored more than 200 papers in international conferences, peer-reviewed journals, and book chapters in his research interests. His research interests include multibiometric systems, pattern recognition, image processing, compression and indexing, multimedia databases, human-computer interaction, and virtual reality/augmented reality. Prof. Nappi was a member of the International Association for Pattern Recognition. He received several international Awards for Scientific and Research activities. He was the President of the Italian Chapter of the IEEE Biometrics Council from 2015 to 2017.