SN Computer Science (2023) 4:349
https://doi.org/10.1007/s42979-023-01796-z

SURVEY ARTICLE

Deep Learning for Head Pose Estimation: A Survey

Andrea Asperti[1] · Daniele Filippini[1]

Received: 18 April 2022 / Accepted: 16 March 2023 / Published online: 26 April 2023
© The Author(s) 2023

Abstract

Head pose estimation (HPE) is an active and popular area of research. Over the years, many approaches have constantly been developed, leading to a progressive improvement in accuracy; nevertheless, head pose estimation remains an open research topic, especially in unconstrained environments. In this paper, we will review the increasing amount of available datasets and the modern methodologies used to estimate orientation, with a special attention to deep learning techniques. We will discuss the evolution of the field by proposing a classification of head pose estimation methods, explaining their advantages and disadvantages, and highlighting the different ways deep learning techniques have been used in the context of HPE. An in-depth performance comparison and discussion is presented at the end of the work. We also highlight the most promising research directions for future investigations on the topic.

Keywords Head pose estimation · Head pose database · Face analysis · Deep learning · Convolutional neural networks

Introduction

The capacity to estimate the head pose of another person is a common human ability that presents a unique challenge for computer vision systems. People have the ability to quickly and effortlessly interpret the orientation and movement of a human head, thereby allowing one to infer the intentions of nearby people and to comprehend an important non-verbal form of communication.

In a computer vision context, *head pose estimation* (HPE) is the process of inferring the orientation of a human head from digital imagery. Like other facial vision tasks, an ideal head pose estimator must demonstrate invariance to a variety of image-changing factors, such as camera distortion, projective geometry, multi-source non-Lambertian lighting, as well as biological appearance, facial expression, and the presence of accessories like glasses and hats [1].

Head pose is an important cue in computer vision when using facial information and has a wide variety of uses in human-computer interaction, explaining the steadily increasing attention received by the scientific community over the last 3 decades.

Although many techniques have been developed over the years to address this issue, head pose estimation remains an open research topic, particularly in unconstrained environments [2].

Similarly to other applicative domains, HPE has greatly benefited in recent years by the exploitation of deep learning (DL) techniques, and the extensive use of Deep Neural Networks. In this article, we shall do a review of the topic from the distinctive perspective of deep learning, discussing and comparing the many different ways in which Deep Neural Networks contributed to the development of the field.

Motivation

HPE systems play an important role in the development of different intelligent environments, so that several computer vision applications rely on a robust HPE system as a prerequisite: for example, applications of gaze estimation [3], virtual/augmented reality [4], and human computer interaction [5], strongly benefit from knowing the exact position of the head in 3D space. Some application examples are:

- **Human Social Behaviour Analysis**: People use the orientation of their heads to convey rich, inter-personal

✉ Andrea Asperti
andrea.asperti@unibo.it

✉ Daniele Filippini
daniele.filippini2@studio.unibo.it

[1] Department of Informatics: Science and Engineering (DISI), University of Bologna, Mura Anteo Zamboni 7, 40126 Bologna, Italy

information. For example, there is important meaning in the movement of the head as a form of gesturing in a conversation [6] to indicate when to switch roles and begin speaking or to indicate who is the intended target subject [7, 8]. People nod to indicate that they understand what is being said, and they use additional gestures to indicate dissent, confusion, consideration, and agreement [9].

In addition to the information that is implied by deliberate head gestures, there is much that can be inferred by observing a person's head. For instance, quick head movements may be a sign of surprise or alarm, these could also trigger reflexive responses from other observers [10].

Therefore, HPE can be used in smart rooms to monitor participants in a meeting and to record their activities, in particular, their attention can be indirectly related to their head pose [11]. Systems exploiting head pose estimation to analyse people behaviour and human interaction in meeting and workplaces have been proposed in [12–14].

There are also studies on systems for automatic pain monitoring that show how including head pose can improve the performance for both person-specific and general classifiers [15].

- **Driving Safety & Assistance**: HPE systems are particularly useful for assisting drivers by providing contextual alert signals, for example in the case of pedestrians outside the driver's field of view [16].

  Moreover, the head pose can give clues about the intention of the pedestrian e.g. a pedestrian will wait for a stopped automobile driver to look at him before stepping into a crosswalk (this is an example of pattern recognition), very important also in the case of autonomous vehicles.

  Applications to infer the driver's pose are very important for safety, as they can provide insights about distraction, intention, sleepiness, awareness or detect blind spots of the driver [17], for this reason, in recent years many datasets that address this specific scenario have been published [18–20].
- **Surveillance and Safety**: Head pose estimation in surveillance video images is an important task in computer vision because it tracks the visual attention and provides insight on human behavioural intentions [21, 22]. Systems for direct an automated surveillance network have been proposed in [23, 24].
- **Targeted Advertisement**: Methods to track visual attention in wandering people have been proposed in the literature [25]. These systems count people looking at particular outdoor advertisements (targeted advertisement) and can determine what a person is looking at if movement is unconstrained. Systems like these can be used for behaviour analysis and cognitive science in real world applications also in indoor environments, such as TV viewers behaviour analysis [26].
- **Interface Design**: By perceiving the human attention when they look at an interface (e.g. the page of web or software), it is possible to evaluate the property and significance of the displayed visual elements and further guide the design or rearrangement of these elements [27] (see Fig. 1).

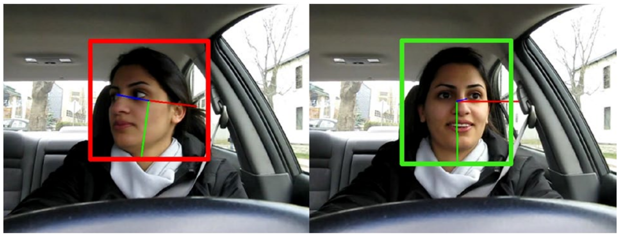


Fig. 1 An example of application to driver assistance. Right: Green box indicates yaw < ± 45° and potential awareness of vehicle. Left: Red box indicates possible inattention (image from [7])

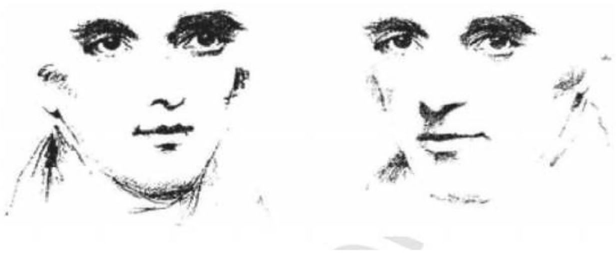


Fig. 2 Example of a task strongly linked to head pose estimation: Despite the eyes are in the same position in both face images, the perception is that the two gazes are differently oriented. Gaze prediction comes from a combination of both eyes and head pose direction [28]

Therefore, head pose estimation can be used to monitor human social activities, to observe the behaviour of specific targets, but also to enhance the function of some face-related tasks, including expression detection, gaze estimation (Fig. 2), full-body pose estimation and identity recognition.

The intrinsic interaction between head pose and other face parts is also confirmed in more recent research. Studies in [29–32] suggest that the mutual relationship between face parts can be exploited not only for HPE, but also for other visual tasks such as gender recognition, race classification, and age estimation making head pose estimation a useful and important task for many applications.

Contribution and Structure

The main contribution of the article are:

- a complete and updated review of all the available databases for the head pose estimation task , with a detailed comparison of the main characteristics (number of subjects, DoF, acquisition scenario) and the analysis of which are the most used and useful in the literature;
- a categorization and explanation of the different approaches used in the literature for head pose estimation, with a specific focus on modern deep learning approaches;
- report and discussion of modern head pose estimation methods and their comparative performance on common datasets, with a deep analysis of different evaluation pipelines and a clear tabular presentation of data;

The remainder of the article is organized as follows: Section "Head Pose Estimation" contains an introduction to the basic concepts of the head pose estimation field; Section "Datasets" presents a detailed list of available datasets and their characteristics; Section "Head Pose Rotations Representations" explains the main techniques for representing rotations used in the HPE field; Section "Methods" describes prominent deep learning based approaches for head pose estimation; Section "Evaluation Metrics" reports the most common evaluation metrics; Section "Evaluation" delineates most used evaluation pipelines; Section "Discussion" presents a discussion of datasets, evaluation metrics/pipelines and possible research directions; Section "Conclusion" concludes the paper summarizing the contribution of the proposed work.

Note: All numerical results reported in the following tables are borrowed from the original publications.

Head Pose Estimation

In the computer vision context, head pose estimation is most commonly interpreted as the ability to infer the orientation of a person's head relative to the view of a camera. More rigorously, head pose estimation is the ability to infer the orientation of a head relative to a global coordinate system, but this subtle difference requires knowledge of the intrinsic camera parameters to undo the perceptual bias from perspective distortion [1].

At the coarsest level, head pose estimation applies to algorithms that identify a head in one of a *few discrete orientations*, e.g. a frontal versus left/right profile view. At the fine (i.e., granular) level, a head pose estimate might be a *continuous angular measurement* across multiple Degrees of Freedom (DoF).

In particular, in the head pose estimation task, it is common to predict relative orientation with Euler angles—*pitch*, *yaw* and *roll*. They define the object's rotation in a 3D environment, if the right prediction about these three angles can be made, it can be found in which direction the human head will be facing (see Fig. 3).

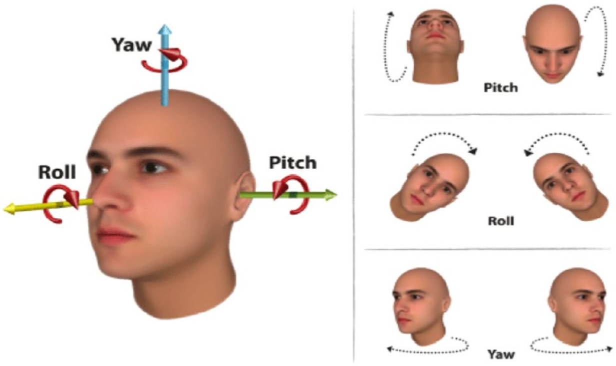


Fig. 3 Euler angles in Head Pose Estimation (image source [33])

Despite head pose estimation is an old and largely investigated problem, achieving acceptable quality on it has become possible only thanks to the recent advances in deep Learning. Challenging conditions like extreme pose, bad lighting, occlusions and other faces in the frame make it difficult for data scientists to detect and estimate head poses.

Nevertheless, SOTA methods for head pose estimation satisfy all the following criteria, firstly proposed by Erik Murphy-Chutorian in [1], on standard datasets:

- **Accurate**: the system should provide a reasonable estimate of pose with a mean absolute error of 5° or less.
- **Monocular**: the system should be able to estimate head pose from a single camera. Although accuracy might be improved by stereo or multi-view imagery, this should not be a requirement for the system to operate.
- **Autonomous**: there should be no expectation of manual initialization, detection, or localization, precluding the use of pure-tracking approaches that measure the relative head pose w.r.t. some initial configuration and shape/geometric approaches that assume facial feature locations are already known.
- **Multi-Person**: the system should be able to estimate the pose of multiple people in one image.
- **Identity & Lighting Invariant**: the system must work across all identities with the dynamic lighting found in many environments.
- **Resolution Independent**: the system should apply to near-field and far-field images with both high and low resolution.
- **Full Range of Head Motion**: the methods should be able to provide a smooth, continuous estimate of *pitch*, *yaw* and *roll*, even when the face is pointed away from the camera.

- **Real-Time**: the system should be able to estimate a continuous range of head orientation with fast (30fps or faster) operation.

Datasets

Most of the HPE models are trained and evaluated using publicly available datasets. These datasets significantly evolved during the last years, especially in terms of complexity of environmental conditions.

Most datasets provide rotation information by means of Euler angles, which define the orientation of a rigid body with respect to a fixed coordinate system; three rotations are always sufficient to express any target position. These rotation angles can be extrinsic or intrinsic, the former express the rotations with respect to the *xyz* axes of an original motionless coordinate system, the latter express rotations with respect to axes of a rotating *XYZ* coordinate system, rigidly attached to the moving body.

Since various formalisms exist to express a rotation in three dimensions beyond Euler angles, e.g. rotation matrices, unit quaternions, Rodrigues' formula, among others, the datasets contain different forms of representation (many of these formalisms use more than the minimum number of three parameters). More details about some of the representations exploited by the models to solve the HPE task can be found in Section "Head Pose Rotations Representations".

Head pose datasets can be categorized by different aspects, such as imaging characteristics, data diversity, acquisition scenario, annotation type, and annotation technique [18]. These aspects play an important role on whether and how the dataset identifies the challenges of the head pose estimation task.

- Imaging characteristics: relate to the image resolution, number of cameras, bit depth, frame rate, modality (RGB, grayscale, depth, infrared), geometric setup and field of view.
- Data diversity: incorporates aspects such as the number of subjects, the distribution of age, gender, ethnicity, facial expressions, occlusions (e.g. glasses, hands, facial hair) and head pose angles. Data diversity is essential for training and evaluating robust estimation models.
- Acquisition scenario: covers the circumstances under which the acquisition of the head pose takes place. The most important distinction is between *in-laboratory* vs. *in-the-wild* acquisition. While the former restricts the data by defining a rather well-defined, static environment, the latter offers more variety through being acquired in unconstrained environments, such as outside, thus covering many challenging conditions like differing illumination and variable background. Head movement can be staged by following a predefined trajectory or can be naturalistic by capturing head movement while the subject performs a different task, such as driving a car.
- Annotation type: describes what meta-information, such as head pose, comes alongside the image data and how it is represented. For example, head pose can be defined by a full 6 degrees of freedom (DoF) transformation from the camera coordinate system to the head coordinate system (covering 3 DoF for translation and 3 DoF in rotation) or only a subset of them can be provided. Annotation types can differ also in their granularity of sampling the DoF space: there are discrete annotation types that classify a finite set of head poses, and there are continuous annotation types that offer head pose annotations on a continuous scale for all the DoFs.
- Annotation technique: there are different methods for obtaining the head pose annotation (label) accompanying each image. The annotation technique has a large impact on data quality (see Table 1, 2 and 3).

Available Datasets

There are many available datasets in the literature:

- **300W-LP** [53]: The 300W-LP (Large Pose) is a synthetic extension of the 300W database [71], generated to augment the number of challenging samples with extreme poses. It includes 122 450 images with yaw angle in range ±89°.
- **AFLW** [45]: Annotated Facial Landmark in the Wild is a challenging dataset which was collected from the internet, in totally unconstrained conditions. It contains a collection of 25, 993 faces with head poses ranging between ± 120° for yaw and ± 90° for pitch and roll. The pitch, yaw and roll angles were obtained automatically from the labelled landmarks using the POSIT algorithm [72], assuming the structure of a mean 3D face, for this reason, several annotations errors were found [73].
- **AFLW2000-3D** [53]: This dataset contains the first 2000 identities of the in-the-wild AFLW [45] dataset which have been re-annotated with 68 3D landmarks using a 3D model which is fit to each face. Consequently, this dataset contains accurate fine-grained pose annotations and is a prime candidate to be used as a test in head pose estimation task. Yaw varies ±120°, while roll and pitch ±90°.
- **AFW** [47]: Annotated Faces in the Wild represents a small database (it's a subset of AFLW [45]), which is normally used for testing purposes only. AFW has 250 images and inside these images 468 faces in a very challenging environment are included. The yaw angles vary

Table 1 Available datasets for Head Pose Estimation

Database	Year	People	Images	Yaw	Pitch	Roll	DB type	GT method	Pose type
BU [34]	2000	5	200	✓	✓	✓	C	MS	C
PIE [35]	2000	68	40.000	✓			C	CA	D
IDIAP-HP [36]	2003	16	66.295	✓	✓	✓	C	MS	C
CAS-PEAL [37]	2004	1.040	99.594	✓	✓		C	CA	D
Pointing'04 [38]	2004	15	2.790	✓	✓		C	DS	D
FacePix [39]	2005	30	5.430	✓			C	CR	D
Bosphorus [40]	2008	105	4.652	✓	✓		C	DS	D
ETH [41]	2008	26	10.000	✓	✓		C	ICP	C
BJUT-3D [42]	2009	500	46.500	✓	✓		C		
Taiwan Rob.Lab [43]	2009	90	6.660	✓			C	CA	D
Multi-Pie [44]	2010	337	75.000	✓			C	CA	D
AFLW [45]	2011		25.993	✓	✓	✓	W	E	C
BIWI Kinect [46]	2011	20	15.000	✓	✓	✓	C	ICP	C
AFW [47]	2012	205	468	✓	✓	✓	W	M	D
ICT-3DHP [48]	2012	10	1.400	✓	✓	✓	C	IS	C
BioVid Heat Pain [15]	2013	90	9.000	✓	✓	✓	C	ICP	C
CAVE [49]	2013	56	5.880	✓			C	CA	D
McGill [50]	2013	60	18.000	✓			W	M	D
Dali3DHP [51]	2014	33	60.000	✓	✓	✓	C	IS	C
MTFL [52]	2014		12.995	✓			W	M	D
300W-LP [53]	2015		122.450	✓	✓	✓	$\mathrm{H}_{(W+S)}$	S	C
AFLW2000-3D [53]	2015		2.000	✓	✓	✓	W	E	C
AISL [54]	2015	20	6.480	✓	✓		C	CR†	D
CMU Panoptic$^\diamond$ [55]	2015		1.342.018	✓	✓	✓	C	P	C
CCNU [56]	2016	58	4.350	✓	✓		C	IS	C
GI4E-HP [57]	2016	10	36.000	✓	✓	✓	C	MS	C
Synthetic [58]	2016	37	74.000	✓	✓	✓	S	S	C
UMDFace [59]	2016	8.277	367.888	✓	✓	✓	W	E	C
DriveAHead [20]	2017	20	~ 1 M	✓	✓	✓	W*	O	C
Pandora [60]	2017	22	250.000	✓	✓	✓	C*	IS	C
SASE [61]	2017	50	30.000	✓	✓	✓	C	ICP	C
SyLaHP [62]	2017	30	~ 101 K	✓	✓	✓	S	S	C
SynHead [63]	2017	10	510.960	✓	✓	✓	S	S	C
UbiPose [64]	2018	22	10.400	✓	✓	✓	C	ICP	C
VGGFace2 [65]	2018	9.131	~ 3,31 M	✓	✓	✓	W	E	C
DD-Pose [18]	2019	27	~ 330 K	✓	✓	✓	W*	O	C
GOTCHA-I [66]	2019	62	137.826	✓	✓	✓	W	E	D
M2FPA [67]	2019	229	397.544	✓	✓		C	CA	D
AutoPOSE [19]	2020	20	1.018.885	✓	✓	✓	C*	O	C
MDM corpus [68]	2021	59	~ 10,5 M	✓	✓	✓	W*	ICP	C
UET-Headpose [69]	2021	9	12.848	✓	✓	✓	C	IS	C
DAD-3DHeads [70]	2022		44.898	✓	✓	✓	W	E	C

The most used in the literature are in bold

The legenda fot this table is in Table 2

between ± 90° with a step size of 15°. The ground-truth is manually annotated, so it may contain errors.

- **AISL** [54]: The Aisl head orientation database is a collection of small scale head images with various backgrounds of an indoor scene. This dataset contains 6480 images of 20 subjects under 36 yaw angles, 3 pitch angles and 3 different backgrounds. The orientation is determined by two categories: yaw angle in 360° with an interval of 10°, and pitch angle in the range ±45° with an interval of 45°.

Table 2 Legenda for Table 1

Database:

⋄ = Processing operations needed to extract head pose information from original data [7]

DB Type:

C = Constraint, faces of real people taken in a constraint environment (a lab, an office, etc.)

W = In-the-Wild, images of real people captured under any kind of conditions

S = Synthetic, synthetic generated images

H = Hybrid, a mixture of previous types

* = Dataset build for the driving context

Pose Type:

C = Continuous, pose estimate in continuous range

D = Discrete, few discrete orientations are acquired

GT Method: Ground Truth Acquisition Method

CA = Camera array

CR = Camera ring

CR† = It's not the camera that rotates around the person, but the seat that rotates on itself

DS = Directional suggestion

E = Estimation with neural networks or other algorithms

ICP = ICP algorithm

IS = Inertial sensor

L = Laser pointer directional suggestion

M = Manual annotation

MS = Magnetic sensor

O = Optical motion capture system

P = Panoptic studio

S = Synthetic images generation

- **AutoPOSE** [19]: It's a large-scale dataset that provides 1.1 million images taken from a car's dashboard view. AutoPOSE's ground-truth head orientation was acquired with a sub-millimetre accurate motion capturing system placed in a car simulator. The rotations are limited to the range [– 90°, + 90°], the average pitch angle is shifted in the negative values of the rotation angles, this is due to the placement of the camera in the dashboard.
- **BioVid Heat Pain** [15]: It contains videos and physiological data of 90 persons subjected to well-defined pain stimuli of 4 intensities, built for the development of automatic pain monitoring systems. It includes information about head pose of the recorded subjects for all 3 angles pitch, yaw, roll, all in the range ±50°.
- **BIWI Kinect** [46]: It's gathered in a laboratory setting by recording RGB-D video of different subjects across different head poses, using a Kinect v2 device. It contains roughly 15, 000 frames and the rotations are ±75° for yaw, ±60° for pitch and ±50° for roll. A 3D model was fit to each individual's point cloud and the head rotations were tracked to produce the pose annotations. This dataset is commonly used as a benchmark for pose estimation using depth methods that attests to the precision of its labels.
- **BJUT-3D** [42]: The database consists of 46 500 images collected from the 3D faces of 250 male and 250 female participants. The total number of poses in the database is 93. The pitch rotation is quantized into 9 angles [– 40°, +40°], where the difference between two consecutive poses is 10°. Similarly, the yaw rotation is divided into 13 angles [-60°, +60°], with the same angular step size as for the pitch.
- **Bosphorus** [40]: It contains 5 thousand high resolution face scans from 105 different subjects. The 3D scans are obtained by a commercial structured-light based 3D digitizer. It offers 13 discrete head pose annotations (seven yaw angles, four pitch angles, and two roll angles), with different facial expressions and occlusions.
- **BU** [34]: The Boston University Head Tracking dataset includes only 200 images and 5 subjects, which is the main drawback of this database. The acquisition process is repeated in two sessions: initially illumination conditions are uniform; then subject faces are exposed to rather complex scenarios with changing illumination. All three rotation angles were recorded thanks to a magnetic tracker attached to each participant's head. Pose variation is mainly less than 30°. Since the pres-

ence of facial occlusions (e.g., eyeglasses, facial hair, etc.) is very limited, most methods perform very well.

- **CAS-PEAL** [37]: The CAS-PEAL is a large dataset having 99 594 images, with a total number of 1040 participants, with 595 males and 445 female subjects. The CAS-PEAL dataset contains a total of 21 poses combining different yaw and pitch angles: the yaw orientation varies between – 45° and + 45° with an interval of 15° between two consecutive poses; the pitch orientation has only three poses – 30°, 0°, and + 30°. Although the dataset has sufficient data for evaluation and training, its complexity is low, as the number of poses is quite limited.
- **CAVE** [49]: The Columbia Gaze dataset contains a total of 5880 images of 56 different subjects (32 male, 24 female) of different ethnic groups and ages. The dataset is mainly created to solve the gaze estimation task, but contains also information about head pose of the participants, therefore it can be used to solve the discrete head pose estimation task. For each subject a combination of five horizontal head poses (0°, ± 15°, ± 30°), seven horizontal gaze directions (0°, ± 5°, ± 10°, ± 15°), and three vertical gaze directions (0°, ±10°) are available.
- **CCNU** [56]: All images in CCNU are low-resolution images collected in a classroom. The database consists of 58 participants, captured in 75 different poses, for a total number of 4 350 images. The face images are collected so that illumination conditions and facial expressions are changing, thus adding more complexity to the images. For obtaining the ground-truth data, SensoMotoric Instruments (SMI) eye tracking glasses are used. The head orientation changes from – 90° to + 90° in the horizontal direction, while the vertical direction spans in the range – 45° to + 90°.
- **CMU Multi-Pie** [44]: This is a database collected from subjects exhibiting multiple expressions under different illumination conditions in a constraint environment. All high-resolution images are captured using a system of 15 cameras for a total of 75 thousand images. The only angle of rotation available is the yaw with an incrementation step of 15°.
- **CMU Panoptic Dataset** [55]: It's a large scale dataset providing 3D pose annotations for multiple people engaging social activities. It contains 65 videos with multi-view annotations captured inside a dome from approximately 30 HD cameras. The panoptic dataset includes 3D facial landmarks and calibrated camera extrinsics and intrinsics, but does not include head pose information. Using landmarks and camera calibrations it is possible to locate and crop images of subjects' heads and compute the corresponding camera-relative Euler angles.

  After processing the dataset to address the head pose problem [7], it contains 1,342,018 images. The yaw angle distribution is almost uniform and ranges in ±179°, but at angles near 90° and – 90° there are fewer images due to the effect of Gimbal lock. For the two angles pitch and roll the magnitudes are in the range ± 89°.
- **CMU-PIE** [35]: The CMU Pose, Illumination, and Expression (PIE) dataset contains over 40,000 facial images of 68 people. Using the CMU 3D Room each person is imaged across 13 different poses, under 43 different illumination conditions and with 4 different expressions. The pose ground-truth was obtained with a 13 cameras array, each positioned to provide a specific relative pose angle. This consisted of 9 cameras at approximately 22.5° intervals across yaw, one camera above the centre, one camera below the centre, and one in each corner of the room.
- **DAD-3DHeads** [70]: This is an in-the-wild database that contains a variety of extreme poses, facial expressions, challenging illuminations, and severe occlusions cases. It consists of 44 thousand images annotated using a 3D head model, a non-linear optimization algorithm and a final manual adjustment. To validate head pose annotations the rotation matrices were compared to the ground-truth matrices from the BIWI dataset [46].
- **Dali3DHP** [51]: This database is an extreme head pose database collected from a camera mounted on a treadmill. The dataset was collected in two different sessions from 33 individuals. Ground-truth data is collected using Shimmer sensor 2 which was attached to each person's head. The database is large since it contains more than 60,000 depth and colour images. All the three rotation angles pitch, yaw and roll were defined at the time the acquisition took place, covering the following head angles: pitch [− 65.76°, + 52.60°], roll [−29.85°, + 27.09°], and yaw [− 89.29°, + 75.57°].
- **DD-Pose** [18]: It contains 330 thousand measurements from multiple cameras acquired by an in-car setup during naturalistic drives by 27 subjects. Large out-of-plane head rotations and occlusions are induced by complex driving scenarios, such as parking and driver-pedestrian interactions. Precise continuous 6 DoF head pose annotations are obtained by a motion capture sensor and a novel calibration device. The angles vary in the following ranges, ignoring outliers with less than 10 measurements in a 3° neighbourhood: pitch ∈ [– 69°, + 57°], yaw ∈ [– 138°, + 126°], roll ∈ [– 63°, + 60°].
- **DriveAHead** [20]: It's another driver head pose dataset, it contains frame-by-frame head pose labels obtained from a motion-capture system for 20 subjects (about 1 million of frames). It includes parking manoeuvres, driving on the highway and through a small town, different occlusions and illuminations, thus providing distributions

of head orientation angles and head positions which are typical for naturalistic drives. Images were collected with a resolution of 512×424 pixels, 6 DoF, the range of angles is [– 45°, + 45°] for pitch, [– 40°, + 40°] for roll and mainly [– 90°, + 90°] for yaw.

- **ETH** [41]: The ETH Face Pose Range Image Dataset contains more than 10 thousand images of 20 persons (3 of them being female) at a resolution of 640 × 480 pixels. Each person freely turned her head while the scanner captured range images at 28 fps. Yaw varies between -90° to + 90°, pitch between – 45° to +45°, whereas roll is not considered.
- **FacePix** [39]: The FacePix database is built depicting 30 individuals, for a total number of 5 430 images. It is an imbalanced dataset with 25 males and 5 females. Yaw rotation varies from – 90° (extreme left profile) to + 90° (extreme right profile), with a step size of 2°; no other rotation angles were considered.
- **GI4E-HP** [57]: It contains 36 thousand images from 10 subjects recorded with a web-cam in an in-laboratory environment. Head pose annotations are given in 6 DoF using a magnetic reference sensor. All transformations and camera intrinsics are provided. Head pose annotations are given relative to an initial subjective frontal pose of the subject.
- **GOTCHA-I** [66]: This dataset is a collection of 682 videos of 62 subjects in 11 different indoor and outdoor environments to address both security and surveillance problems. To obtain ground-truth a 3D head model is reconstructed and elaborated using Blender software. There are 137, 826 labelled frames with 2223 head pose per subject in the range of [– 40°, + 40°] in yaw, [-30°, +30°] in pitch and [– 20°, + 20°] in roll, with a step of 5°.
- **ICT-3DHP** [48]: It's a large dataset which was collected in-the-wild, i.e. captured in an unconstrained environment. All images were acquired through the Polhemus Fastrack[1] flock of birds tracker attached to a cap the participants that contains a magnetic sensor, so that the dataset contains both RGB and depth data. The database is evaluated for all three rotation angles including pitch, yaw and roll. No accurate information about the angle ranges is provided.
- **IDIAP Head Pose** [36]: It contains 66, 295 head images stemmed from a 8 video meeting recording, each approximately one minute in duration, of a few people in a meeting room. In each sequence, two subjects, which are always visible, were continuously annotated using a magnetic sensor. Therefore, each image has a complete annotation of a head pose orientation from pitch (range [– 60°, + 15°]), yaw (range ± 60°) and roll (range ± 30°) angles.
- **M2FPA** [67]: This dataset totally involves 397, 544 images of 229 subjects with 62 poses (including 13 yaw angles, 6 pitch angles and 44 yaw-pitch angles), 4 attributes and 7 illuminations. There are 6 classes for pitch in the range of [– 30°, +45°] with a step increment of 15° and 13 measurements for yaw in the range ±90° with a step increment of 15°.
- **McGill** [50]: The database consists of 60 videos of 60 different participants, in total it contains 18, 000 video frames. The videos were recorded in both indoor and outdoor environments. The participants were free to behave as they want during the video collection process, therefore arbitrary illumination conditions and background clutter are present, especially outdoor. Only yaw angles are estimated using a semi-automatic procedure, with variation in the range [– 90°, + 90°].
- **MDM corpus** [68]: The Multimodal Driver Monitoring database was collected with 59 subjects recorded while were diving a car and performing various tasks. To record the head pose the Fi-Cap device was used, this continuously tracks the head movement of the driver using fiducial markers, providing frame-based annotations to train head pose algorithms in naturalistic driving conditions. This set consists of 48.9 h of recordings (10, 541, 166 frames), it covers a large range of head poses along all three rotation axes due to the large number of subjects included, and the variety of primary and secondary driving activities considered during the data acquisition. Yaw angles range around the origin spanning between – 80° to 80°, pitch angles have an asymmetric range spanning from – 50° to 100°.
- **MTFL** [52]: The Multi-Task Facial Landmark dataset contains 12, 995 outdoor face images from the web. These images are from CUHK Face Alignment database and AFLW dataset. Each image is annotated with a bounding box and five facial landmarks. There are ground-truth annotations for gender, age, smiling, wearing glasses and head pose. For the latter, the images are manually categorized in 5 discrete classes: Left-profile, Left, Frontal, Right, Right-profile.
- **Pandora** [60]: It has been specifically created for head centre localization, head pose and shoulder pose estimation and is inspired by the automotive context. A frontal fixed device acquires the upper body part of the subjects, simulating the point of view of the camera placed inside the dashboard. Subjects also perform driving-like actions, such as grasping the steering wheel, looking to the rear-view or lateral mirrors, shifting gears and so on. Pandora contains more than 250 thousand full resolution RGB (1920× 1080 pixels) and depth images (512 × 424) acquired with a Microsoft Kinect 1 device. Subjects

[1] https://polhemus.com/motion-tracking/all-trackers/fastrak.

perform wide head movements: ± 70° roll, ± 100° pitch and ± 125° yaw. Garments as well as various objects are worn or used by the subjects to create head occlusions. The ground-truth annotations have been collected using a wearable Inertial Measurement Unit (IMU) sensor.

- **Pointing'04** [38]: It is one of the oldest databases, released in 2004, which was considered as the classical benchmark for HPE (in some studies is also called PRIMA database [74]). Despite its age, it's still used for research purposes, due to its challenging nature and a large variety in consecutive poses [29–32]. A total number of 15 participants (between 15 and 40 years) were involved for image acquisitions. Some of them wear eyeglasses or show facial hairs, thus increasing the task complexity. Images were collected in an indoor lab environment, with very low illumination conditions. Each participant is asked to look at some markers on the wall, and two rotation angles (yaw and pitch) are annotated through a subsequent manual labelling process (thus introducing some errors). The head orientation varies between ± 90° both in the horizontal and vertical directions, while the difference between two consecutive poses in horizontal and vertical orientation is kept at 15° and 30°, respectively.
- **SASE** [61]: This is a 3D database collected through Kinect 2 camera. It consists of both RGB and depth images of 32 male and 18 female subjects. The total number of frames is 30, 000. All subjects have different ethnicity and hairstyles, with an age range of 7–35 years. All three rotation angles pitch, yaw, and roll are considered. All participants have different facial expressions during image acquisition, so that, along with head pose estimation, the database may also be used for emotion recognition. For each person a large sample of head poses are included, within the bounds of yaw from – 45° to 45°, pitch – 75° to 75° and roll – 45° to 45° of rotation around each axis.
- **SyLaHP** [62]: The Synthetic dataset for Landmark based Head Pose estimation was proposed by Werner et al. [62] along with a benchmark protocol to learn head pose on top of any landmark detector (called HPFL). It contains about 101 thousand synthetic images from 30 subjects, with varying ethnicity, age and gender. The angles are in the ranges: ± 70° for pitch, ± 90° for yaw and ±55° for roll.
- **SynHead** [63]: This is a large-scale synthetic dataset for head pose estimation in videos containing 10 head models (5 female and 5 male), 70 motion tracks and 510 960 frames. Such synthetic dataset, which considers all Euler angles, generates 100% reliable ground-truth to compensate for errors existing in manually annotated datasets. The Euler angles are in the range of [– 100°, +100°].
- **Synthetic** [58]: The Synthetic image database is a large database of 74, 000 high quality images taken from head models. A total of 37 sequences have been considered, where each sequence includes 2000 frames. The head pose in face images covers ± 50° of roll, ± 75° for yaw, and ± 60° for pitch. The database is quite challenging as different ages, races, and facial expressions are included.
- **Taiwan RoboticsLab** [43]: It contains 6660 images of 90 subjects. For each subject there are 74 images, where 37 images were taken every 5 degrees from right profile (defined as + 90°) to left profile (defined as – 90°) in the yaw rotation using camera array and the remaining 37 images were generated (synthesized) by the existing 37 images using commercial image processing software in the way of flipping them horizontally.
- **UbiPose** [64]: This dataset relies on videos from the UBImpressed dataset, which has been captured to study the performance of students from the hospitality industry at their workplace. The data are recorded using a Kinect 2 sensor, however the ground-truth head pose is indirectly inferred from facial landmarks. The validated inferred head poses are 10.4 thousand, most frames fall within a [20°, 40°] interval.
- **UET-Headpose** [69]: The UET-Headpose dataset was created to capture the head pose of annotated people in many conditions, it includes 12, 848 images obtained from 9 people. The dataset has a uniform yaw angle distribution for all directions in the range [– 179°, 179°]. The dataset is obtained by having the annotated people rotated all yaw directions when collecting the dataset. Therefore, it is possible to learn all yaw angles within a 360° range.
- **UMD Faces** [59]: This dataset has 367, 888 annotated faces of 8277 subjects. It contains information about bounding boxes (verified by humans), twenty-one keypoint locations, Euler angles and the gender of the subject. These annotations have been generated using the All-in-one CNN model [75], therefore the dataset may contain erroneous annotations, especially for the pitch, yaw and roll angles.
- **VGGFace2** [65]: This is a very large HPE database which has been released in 2018. It contains 3.31 million images. The total number of participants to create this content are 9131, whereas the average number of images per subject is 362. The database is constructed with images downloaded from Google Image Search and shows large variations in pose, illumination, age, profession, and ethnicity. However, pose (pitch, yaw and roll) is estimated using pre-trained pose classifiers defining 5 classes for angles in ranges [– 100°, – 40°), [– 40°, – 10°), [– 10°, + 10°), [+ 10°, +40°) and [+ 40°, + 100°).

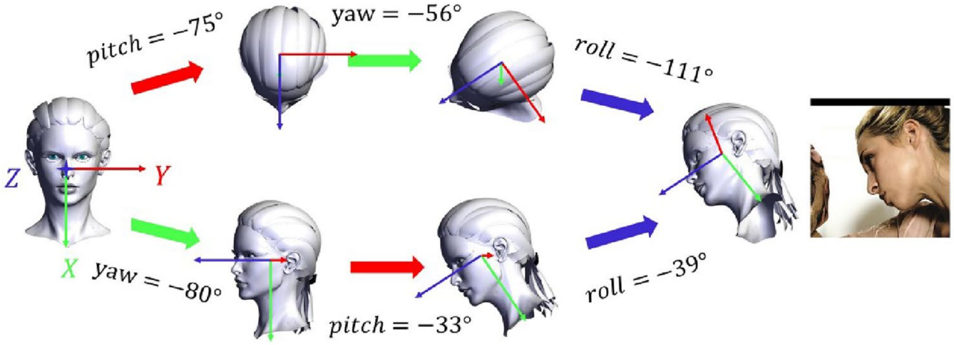


Fig. 4 Different processes from the same initial pose to the same final pose in different rotation order (image from [77])

Head Pose Rotations Representations

Many possible representations can be used to express rotations of rigid bodies. The widely used in the field of head pose estimation is that based on Euler angles, but other methods are exploited in the literature due to some problems of this specific representation.

Furthermore, it has been shown that any rotation representation in 3D with less than five dimensions is discontinuous, making the learning process harder [76]. We will further briefly review different rotation parametrizations, their pros and cons to see how they might affect the regression performance.

Euler Angles

The Euler angles were introduced by Leonhard Euler in rigid body dynamics to describe the orientation of a reference system attached to a rigid solid in motion. Three parameters are needed to describe an orientation in a 3 dimensional Euclidean Space $\mathbb{R}^3$.

Thus, the Euler angles are a set of three angular coordinates which specify the orientation of a reference system with orthogonal axes, usually mobile, with respect to another reference with known orthogonal axis called standard orientation. This standard initial orientation is normally represented by a motionless (fixed) coordinate system.

Euler angles can represent any rotation by means of three successive elemental rotations around three independent axes.

$$R_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

$$R_y(\beta) = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix}$$

$$R_z(\gamma) = \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

These three elemental rotations around distinct axes can be composed to obtain a single rotation matrix using matrix multiplication:

$$R = R_x R_y R_z.$$

Matrix multiplication is not commutative and the same thing applies to rotations, therefore the order of application of the three successive elemental rotation is important.

However, the definition of Euler angles is not unique, in the literature many different conventions are used, where varies the sequences of rotations and the axes about which the rotations are carried out (see Fig. 4).

Following the Trait–Bryan convention we can define as x, y and z the original axes and X, Y, and Z the axes after rotation. The line that represents the intersection between plane xy and YZ is called the line of nodes N, see Fig. 5. The Euler angles with this convention are: α the rotation angle between x and N, covering a range of 2π; β the rotation angle between z and Z, covering a range of π; γ the rotation angle between N and X, covering a range of 2π.

Many datasets have annotations of pitch, yaw and roll angles, but not all of them explicitly mention the order; the process of determining it become tedious and error-prone.

The main limitation of the Euler angles remains the **Gimbal lock**: when the second elemental rotation reaches 90 (or – 90) degrees, then first and third axes become parallel (i.e. linearly dependent), which gives an infinite number of solutions for the same rotation and the other axis can not be determined. This is a great limitation when wide ranges of rotations [– 180°, +180°] are considered (see FIg. 5).

Rotation Matrix

Each rotation can be uniquely described with a rotation matrix. The rotation matrix R is a special orthogonal 3×3 matrix, with a determinant equal to one, that represents a rotation in Euclidean space.

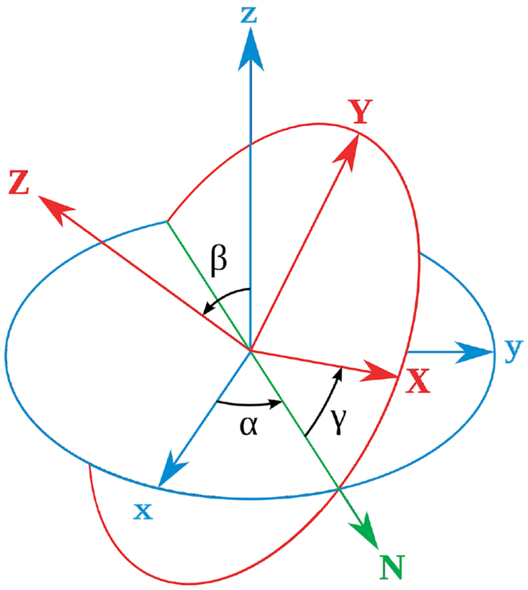


Fig. 5 Euler angles, image from Wikipedia [78]

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}, R^{\mathrm{T}}R = RR^{\mathrm{T}} = I, det(R) = 1.$$

Rotations can be composed using multiplication, and the resulting matrix will remain a rotation matrix. A rotation is represented using *nine* parameters.

To regress the parameters with back-propagation an orthogonality constraint must be enforced, otherwise something different from rotation matrix will be obtained during inference [79].

A complaint of rotation matrices is that they're less intuitive. In general, it's not easy to understand what the matrix is doing by simply looking at the matrix. This is why Euler angles sometimes are more favourable.

Let be the column vector v, the position of each point in the standard initial orientation and R the rotation matrix. Then, a rotated vector u is obtained by multiplying the rotation matrix with the vector.

$$u = R \cdot v.$$

The ease by which vectors can be rotated using a rotation matrix, as well as the ease of combining successive rotations, make the rotation matrix a useful and popular way to represent rotations, even though it is less concise than other representations [28].

Quaternions

Quaternions are a compact way to represent rotations, they have four parameters, which can be interpreted as a scalar component plus a three-dimensional vector component:

$$q = (s_0, \vec{v}) = (s_0, v_1, v_2, v_3).$$

Quaternions are quite popular because are more compact than matrix representation and it's simple to combine two individual rotations represented as quaternions using quaternion product.

Unlike Euler angles, quaternions are free from the Gimbal lock problem, but still they have an ambiguity caused by their anti-podal symmetry: q and $-q$ correspond to the same rotation.

Furthermore, it has been recently demonstrated that for 3D rotations, all representations are discontinuous in the real Euclidean spaces of four or fewer dimensions and empirical results suggest that continuous representation outperform discontinuous ones [76]. This means that Euler angles and quaternions representations might not be well suited for regression task.

Methods

The approaches used in the literature to solve the task of head pose estimation are quite different between them: they have different degrees of automation, different prerequisites and are based on different assumptions.

We try to arrange each system by the approach that underlies its implementation (taking as reference classifications proposed in previous works [1, 28]), by giving a description and evaluating advantages and disadvantages of each approach. Our taxonomy is briefly summarized in Fig. 6.

Since head pose estimation has been investigated for a long time, many methods have emerged during this period; however, starting from 2015, methods based on convolutional neural networks have been used more and more, highlighting a shift in methodology, from traditional machine learning (ML) methods towards deep learning (DL) approaches.

In the following sections, we first shortly review "classical methods" (Section "Classical Methods"), including all approaches that are little, or no longer, considered in

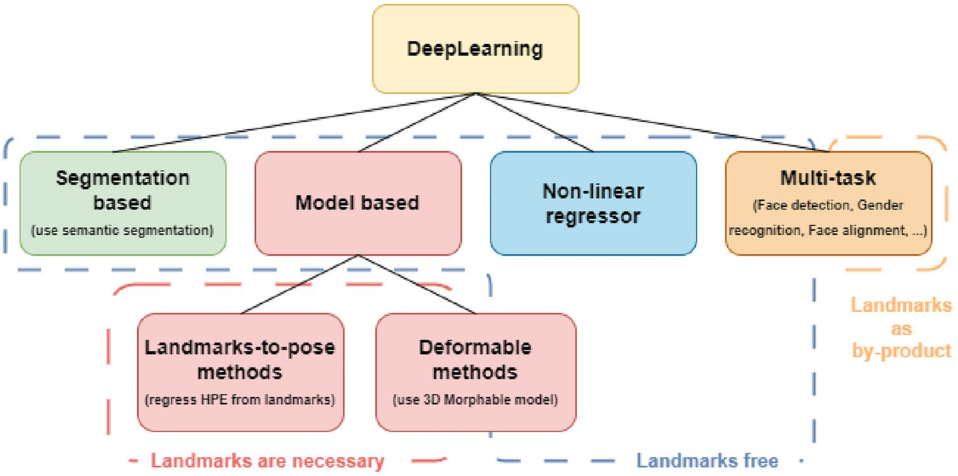


Fig. 6 Our taxonomy of deep learning approaches for head pose estimation problem

the most recent research, then shifting the focus on deep learning based models:

- Segmentation based models (Section "Segmentation Based Models"):
  compute head pose using probability maps produced by a face segmentation algorithm [29–32, 80];
- Model based methods (Section "Model Based Methods"):
  exploit facial keypoints, either for regressing head pose [62, 81–83] or for reconstructing 3DMM and learn its rotation parameters [84–87].
- Non-linear regression methods (Section "Non-linear Regression Methods"):
  use deep convolutional neural network to develop a mapping from the image to the head pose measurements [7, 8, 60, 63, 76, 88–91];
- Multi-task methods (Section "Multi-task methods"):
  jointly solve head pose with other correlated tasks (e.g. face detection or face alignment) to improve the overall performance [75, 92–103];

Additional details about classical methods can be found in [1, 104]. More recent surveys are [2, 28]; with respect to them, we will cover the parts relating to the state-of-the-art models in more detail, with a special focus on multi-task learning, 3DMM based and CNN based models.

Classical Methods

Here we briefly recall a short list of methods that played an important role for HPE but have been either outdated by most recent techniques, or are difficult to integrate with deep learning technology, that is the main focus of this survey:

- Appearance template methods: compare a face image to a set of exemplars template to find the most similar view [105, 106];
- Detector array: use a series of head detectors, each trained for a specific pose and assign the pose relative to the detector with the greatest support [107–109];
- Manifold embedding: embed an image into low-dimensional manifolds that model the continuous variation in head pose and use these for pose regression [110–119];
- Tracking methods: use temporal constraint to recover the pose from observed movements in video frames [51, 120–124];
- Hybrid classical approaches: combine one or more of the afronentioned methods in a single model [1, 104];

Segmentation Based Methods

These methods address the problem of head pose estimation by exploiting the strong relationship between the head pose and the position of various face parts. The idea is that the performance of the face pose predictor can be improved if a prior efficiently parsed image, having information about various facial features, is provided as input [29–32].

The first step is to perform *semantic segmentation* over the input image either by training a single segmentation model or multiple (discrete) pose specific models. Each model parses the face into different parts (e.g. nose, mouth, eyes, hair) and produces probability maps. Given a new image, the probabilities associated to face parts by the single model or the different pose-specific models are used as the only information for estimating the head pose by using specifically designed algorithms or by training a classifier (e.g. Random Forest, SVMs, etc...).

Huang et al. [125] were the first to exploit the relation between face segmentation and head pose estimation. In their method, initially, the face is segmented into three face parts (skin, hair, background) using traditional

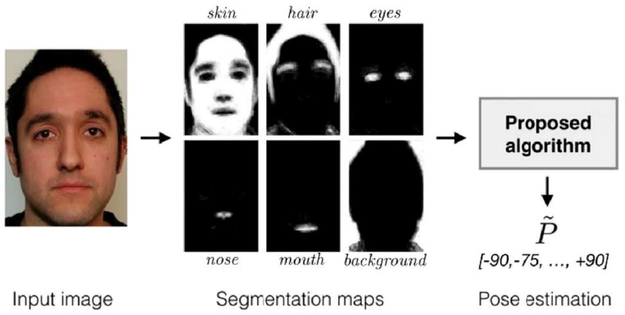


Fig. 7 Segmentation based method: perform face segmentation and from probability maps infer head pose (image from [29])

textural-based techniques, and then in a second stage, they estimate basic discrete head poses using a simple regressor: "frontal", "right-profile" and "left-profile".

More modern works address segmentation by means of Deep Neural Networks, that typically allow to consider a larger number of segmentation classes, and discrete poses (e.g. 13 poses [29, 31] or 93 poses [30, 32, 80]).

Khan et al. [29] proposed a simple algorithm to exploit probabilities associated to face parts to predict head pose: first, they run segmentation models for all different poses, obtaining probability maps; then, they consider the maximum of such probabilities to assign a pose to each pixel; finally, they count the total number of pixels associated to each discrete pose and assign to the face image that with the highest number. A similar approach was taken in [30], but relying on the concept of super-pixel, i.e. small meaningful patches belonging to the same object.

The estimation of the head pose after performing segmentation can be done by many traditional ML techniques, comprising multi-class linear SVM [31], Random Forest [32] and Soft-Max classifiers [80].

The main advantage of these methods is that are able to exploit the strong relationship between head pose and position of various face parts, which is useful for accurate pose estimation. Moreover, these methods do not require any landmark detection process or face alignment step. Finally, these systems are typically multi-task, they combine HPE, facial expression detection, gender recognition and age classification in a single framework (see Fig. 7).

A drawback of this technique is that manually segmented face images are needed for training, and creating supervised segmentation datasets is a notoriously onerous operation. On the other side, face segmentation has a lot of different applications, e.g. for editing [126, 127], so we may expect a steady improvement on this aspect of the task.

Surprisingly, only the coarse head pose classification task has been addressed so far. Testing these techniques on the more challenging continuous regression problem is an open issue, that could definitely help to assess the quality of the technique.

Model Based Methods

Model based methods require either a 3D head model or the localization of *facial keypoints* (landmarks), such as eyes, eyebrows, nose, lips, etc. (or both of them in some cases) and from these estimate the head pose. It is proven that these factors, such as the location of the face in relation to the contour of the head, strongly influence the human perception of the head [1]. For this reason, model based methods are particularly interesting, they can directly exploit properties which are known to influence human head pose estimation. Moreover, in recent years, with the development of deep learning and due to high availability of data, methods which directly extract facial landmarks have improved enormously their performance and have become the dominant approach in facial analysis tasks [8].

A by-product of face alignment is the ability to recover the 3D pose of the head in two different ways: (I) the *Landmark-to-Pose* approach and (II) by exploiting *deformable methods*.

In the *landmark-to-pose* approach the keypoints are given as input to a ML, or DL, algorithm that regress the head rotation angles.

Werner et al. [62] proposed a benchmark protocol to learn pose estimator on top of any landmark detector, called HPFL, that trains a Support Vector Regression (SVR) model using landmarks as features. To exploit the power of Deep Neural Networks not only to compute landmarks but also to obtain Euler angles Gupta et al. [81] proposed to use a deep learning architecture to regress head-pose giving as input *uncertainty maps* computed from 5 facial keypoints. Even Xia et al. [82] used a CNN, but they give as input a *heatmap* of 68 landmarks stacked with a transformed version of the input image, so that the neural network can focus on the area around facial landmarks while extracting features from the image, reducing interference from wild environment. Dapogny et al. [83] proposed an *attentional cascade model* that iteratively refines head pose and landmark estimates. The advantage is that using head pose information to refine landmark alignment provides more precise landmark estimates (as also stated in [128]), which in turn helps refine the head pose prediction, further advocating for an entwined landmark alignment and head pose prediction scheme. The disadvantage is that the network is bigger and requires a longer training time.

For this reason, recently, other researchers have tried to define methods that do not need training for estimating head pose once facial landmarks are detected. Abate et al. [129] used a quad-tree, i.e. a particular kind of unbalanced tree, that divides the image into smaller and smaller quadrants, to measure the distance between the representation of the input face with a reference model. Barra et al. [130] (2020) exploit a spider-web shaped model that uses the landmark

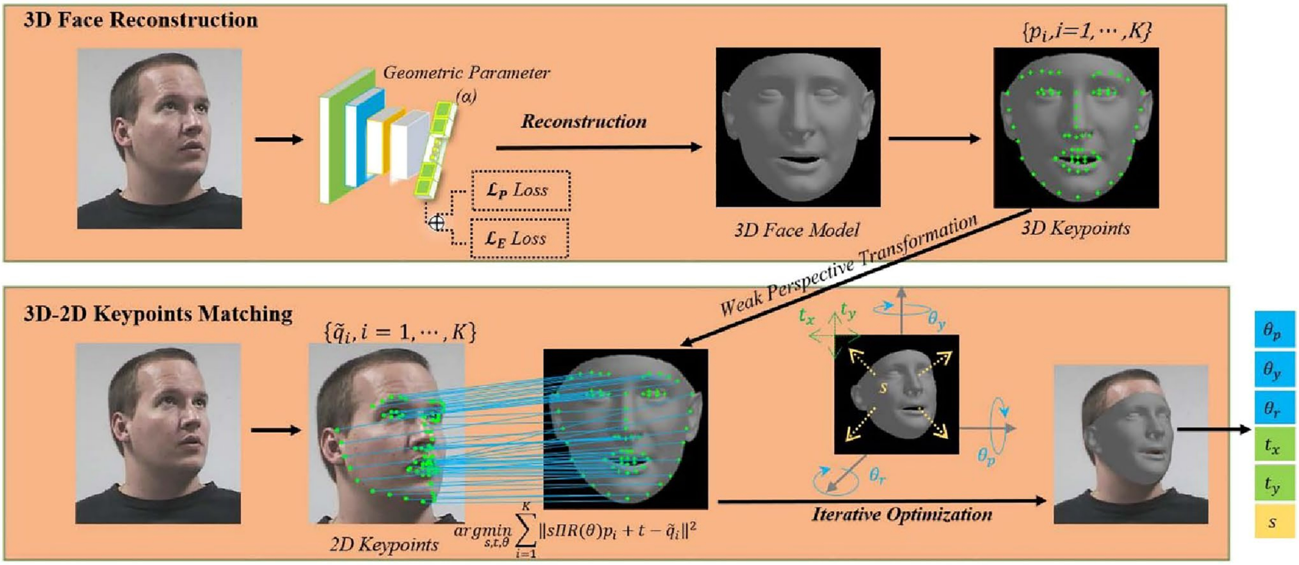


Fig. 8 An example of deformable model: A personalized 3D face is reconstructed from the input head image using a CNN, then keypoints matching is used to obtain the pose [85]

locations to build a feature vector, which in turn is compared to a set of prototypical vectors to determine the closest one and establish the pose. Unfortunately with these two methods only discrete pose can be obtained (with 5° of angular step), they are computationally efficient but less effective than other methods.

Deformable methods, instead, use a non-rigid face model and fit it to the image such that it conforms to the facial structure of each individual and estimate the head poses from the correspondence between feature points on a 2D face image and those on a 3D facial model.

The 3D pose information of the head can be inferred by solving the *Perspective-n-Point* (PnP) problem, i.e. the problem of estimating the pose of an object by finding the rotation matrix R and the translation vector t given intrinsic camera parameters, known locations of n 3D points and their corresponding 2D projection in the image. Indeed, by looking for the projection relation between a 3D facial model and a 2D face image, head pose angles can be calculated from the elements in the rotation matrix directly.

The most simple and commonly used pipeline involves a number of steps [8]: (1) face alignment; (2) definition of 3D human mean face model; (3) approximation of camera intrinsic parameters; (4) solving 2D-3D correspondence problem using one of the available PnP algorithms, such as POSIT [72] or DLS [131]. In their basic form, these methods do not need to include and train a pose estimation model; moreover, any method for face alignment can be used, such as Dlib [132] or FAN [133] (see [134] for a survey on face alignment methods). The drawback of PnP approach is that typically camera parameters are not known so they are approximated leading to errors in the final prediction.

Modern deformable approaches rely on a 3D face morphable model and learn to deform it to adapt to the person's head, then solve the 2D-3D correspondence more effectively.

Wu et al. [84] assumed to have a 3D deformable facial model and followed a cascade iterative procedure that iteratively updates the facial landmark locations, the head pose angles and non-rigid deformations. There is no learning involved for head pose that is estimated from the 3D deformable model by minimizing the projection error for all landmark points. Liu et al. [85] trained a CNN to reconstruct a personalized 3D face model from the input head image and through an iterative 3D-2D keypoints matching algorithm estimate head pose under constraint perspective transformation (see Fig. 8). Diaz Barros et al. [135] proposed a hybrid method that incorporates two strategies: (1) a temporal tracking scheme, which uses optical flow to compute the correspondences of a set of keypoints in every pair of frames; (2) a head pose estimation scheme which estimates pose independently in each frame by aligning 2D facial landmarks to every image; the head pose in each scheme is estimated by minimizing the reprojection error from the 3D-2D correspondences.

Unfortunately, these methods use deep learning only for face alignment and use some projection method to compute head pose, not exploiting its full potential. Instead, the state-of-the-art networks for head pose estimation follow a different approach, also based on 3DMM. In this case, the focus is on the *3DMM-based 3D dense alignment 3D dense reconstruction* task. The network can be directly

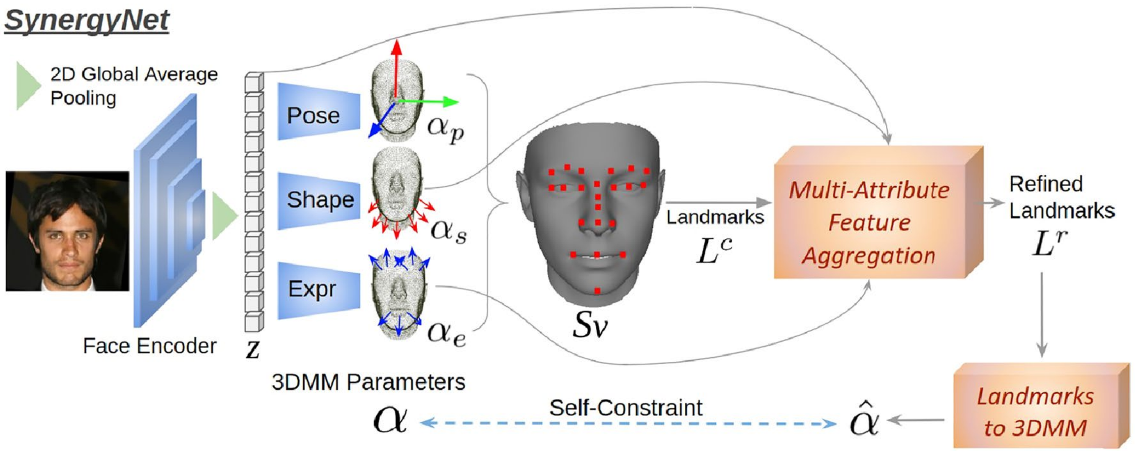


Fig. 9 In SynergyNet a backbone network learns to regress 3DMM parameters (pose, shape, expression) [86]

used for pose estimation, indeed, 3DMM regression contains pose, shape and expression parameters. There is no keypoints matching involved.

Zhu et al. [53] proposed an alignment framework termed 3D Dense Face Alignment (3DDFA), which directly fits a 3D face model to RGB images via convolutional neural networks. The primary task of 3DDFA is to align facial landmarks, even for the occluded ones, using a dense 3D model. As a result of their 3D fitting process, the 3D head pose is produced. SynergyNet [86] is a novel network designed to predict complete 3D facial geometry, including 3D alignment, face orientation and 3D face modelling. The network defines a synergy process that utilizes the relation between 3D landmarks and 3DMM parameters to improve the overall performance. Despite the large amount of work on 3DMM-based 3D dense alignment and the fact that many of the proposed approaches directly estimate rotation matrices, Wu et al. were the first to propose a discussion on the head pose estimation task, previous works only focus on the evaluation of landmarks and 3D faces. The authors, as well as evaluate SynergyNet, conducted extensive and detailed benchmarking on other 3DMM-based methods, such as 3DDFA-TAPAMI [136], 2DASL [137] and 3DDFA-V2 [138], highlighting the better performance of the proposed network due to the innovative synergy process introduced (see Fig. 9).

SADRNet is another network proposed very recently by Ruan et al. [87] that is one of the state-of-the-art models on AFLW2000 [53] dataset. This is an encoder-decoder-based architecture that regresses the deformation *D* and infers the pose parameters *f*, *R* and *t* to reconstruct the 3D face geometry from a single 2D face image. The most important novelty introduced in the network is the attention mechanism used to enhance the visible facial information and estimate the transformation matrix only with visible landmarks, giving robustness to occlusions and large pose variations.

Finally, with the development of consumer-level depth-image sensors, many studies have tried to exploit 3D-face model-based approaches using RGB-D data. These studies have developed in parallel with the others presented before and mainly use *optimization techniques*, such as the ICP algorithm [139], which aim to minimize the discrepancy between depth data and a parametrized 3D model. Martin et al. [140] proposed a real-time head pose estimation method that first creates a point-cloud based 3D head model from the input depth image and then registers the 3D head model with the iterative closest point (ICP) algorithm [139] for head pose estimation. Mayer et al. [141] proposed estimating head poses by registering a 3D morphable model (3DMM) to the input depth data through a combination of particle swarm optimization (PSO) and the ICP algorithm [139]. Higher pose estimation accuracy is achieved at the expense of a much higher computational cost. A 3D morphable model and online 3D reconstruction are used by Yu et al. [64] for full head pose estimation, thus also handling extreme poses. Although estimating the head poses on the depth image can avoid suffering from the cluttered background and illumination changes, that are common in RGB images, the main disadvantage is that depth image sensors are not available in most of the current real-world applications.

Summing up, we saw that there is a huge literature of approaches based on the facial keypoints, that are used as key elements of deformable methods, or given as input to neural networks (so used as features), or even are the only information needed in the PnP approach. It is evident that there is a close relationship between head pose and the distribution of the landmarks, so these are a valuable information to estimate head pose [82]. Moreover, there is a growing number of landmark detectors/trackers that can be used for research purposes for free and there is a rapid progress in improving the landmark quality, including unconstrained

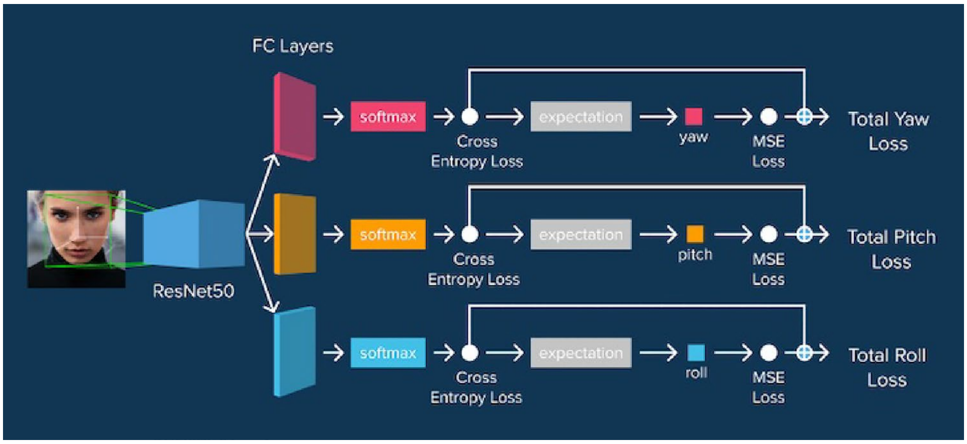


Fig. 10 Hopenet architecture [8]: ResNet50 with combined Mean Squared Error and Cross Entropy Losses (image from https://indatalabs.com/blog/head-pose-estimation-with-cv)

scenarios with difficult lighting, out-of-plane head poses, and occlusions [62].

PnP approach is one of the most used in the literature, but has a disadvantage: many parameters (such as camera pose) typically are approximated and this can lead to inaccuracies in the results. Moreover, when a mean face model is used, even with perfect registration, the images of two different people will not line up exactly, since the location of facial features varies between people, leading to errors in the final result [82]. For this reason, recently developed approaches rely on face reconstruction as previous step to 2D-3D keypoints matching [85]. These methods typically require high-resolution images and the position of landmarks must be initialized before the pose estimation.

Recent research has been focused on landmark-to-pose approaches that regress the head pose from landmark configuration using deep networks, and on 3DMM based approaches that reconstruct and align a 3D dense face model with the images. Less research has been devoted to the latter case, but this seems a very promising direction, able to achieve remarkable results, even if the head pose is only obtained as a by-product. The main drawbacks of 3DDFA approaches are that the networks are quite complex, and their training depend on costly face mesh annotations. Nevertheless, SADRNet [87] reconstructs the 3D model of the face (starting from a cropped image) in 13.5 ms. However, it is is not clear how these results could generalize in low resolution far-field imagery due to the difficulty in achieving good fitting and precise image feature location in those conditions.

Non-linear regression methods

The non-linear regression methods do not require keypoints detection, but directly predict the head pose angles through images. A model is trained in a supervised manner and learns a functional mapping from the image space to discrete/continuous pose directions. The main challenge is to train a model in a way to ensure that the regression tool will learn a proper mapping.

Early approaches used classical machine learning models such as Support Vector Regressor (SVR) [105], Localized Gradient Histograms (LCH) [142] or Random Forest (RF) [46, 56].

In the last decades, there was a drastical shift towards the deep learning paradigm, with an increasing use of convolutional neural networks to estimate the three-dimensional head pose with higher accuracy.

First attempts with deep models exploited simple architectures [143, 144] and common networks [73], such as AlexNet [145], VGG [146], ResNet [147]. Patacchiola et al. [148] improved the results by introducing dropout and adaptive gradient methods during the training of the network, and by training a different specialized network for each rotation angle (pitch, yaw, roll), that permits fine-tuning for a specific degree of freedom without loosing predictive power on another one. Work from Gu et al. [63] uses a recurrent neural network to regress the head pose Euler angles by exploiting the time dimension in video sequences. RNN has the ability to learn motion information implicitly, gaining robustness to large head pose variations and occlusions.

Ruiz et al. [8] proposed to use a three-branch convolutional neural network structure, that they called Hopenet, where each branch is responsible for one of the Euler angles. All branches share a backbone network that can be of arbitrary structure, e.g. ResNet50 [147], AlexNet [145], VGG [146]. This backbone network is augmented with a branch-specific fully-connected layer that predicts a specific angle. By having three cross-entropy losses, one for each Euler angle, three signals are backpropagated into the network, which improves learning (see Fig. 10).

The overall framework of Hopenet is adopted also by Zhou et al. [7] for their network WHENet. WHENet adopted a lighter backbone w.r.t. previous work, EfficientNet-B0 [149] was used (it incorporates Inverted Residual Blocks, from MobileNetV2, to reduce the number of parameters

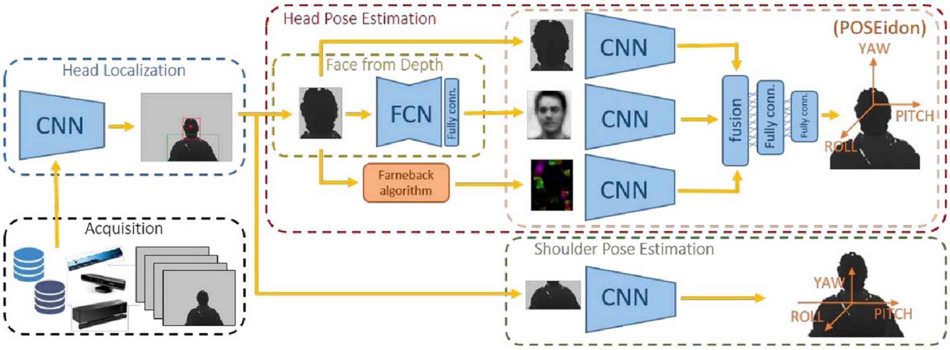


Fig. 11 POSEidon architecture [60]: depth images are provided to a head localization CNN, then the head region is given in input to the POSEidon network to obtain pitch, yaw and roll estimations (image from [60])

while adding skip connections). This network is optimized for the full range Euler angles (360°), not only for narrow range as the previous works (180°). This is achieved by careful choice of the wrapped loss function as well as by developing an automated labelling method for the CMU Panoptic dataset [55], that is used during the training of the network.

FSA-Net [88] introduced a feature aggregation method to improve pose estimation. QuatNet [89] proposed a Quaternion-based face pose regression framework which claims to be more effective than Euler angle-based methods. The quaternion representation is used also by Zeng et al. in their SRNet [150] where a specific Structural Relation-aware module is introduced, this module improved the prediction quality because discriminative pose features are learned from a global perspective (by capturing the valuable facial structure information) rather than low-level local details. TriNet [76] used a three vector-based representation that replaces Euler-based and Quaternion-based representations for increasing efficacy. RankPose [90] is another CNN that explored Siamese architecture and ranking loss to distinguish pose-related from a mixture of pose-related and irrelevant features, such as age, lighting and identity. Hempel et al. for 6DRepNet [151] efficiently regress a compressed 6D form of the rotation matrix. This representation has been reported to introduce smaller errors for direct regression then vector-based one and made 6DRepNet one of the SOTA models on popular datasets.

Given the fact that the bounding box significantly affects the quality of the trained NN for the HPE problem [152, 153], Sheka et al. [91] (2021) proposed to average the results of predictions of the same neural network, but with various bbox offsets, in what they call *offset ensemble*.

Not only bounding box affect the final result but also illumination and occlusion, for this reason Wang et al. in their FSEN [154] included low light enhancement, strong light suppression and face occlusion detection modules. This united with a four-branch CNN, in which three branches are used to extract three independent discriminative features of pose angles, and one branch is used to extract composite features corresponding to multiple pose angles, improved the results on benchmark datasets.

Recently, some attempts to propose lightweight networks that obtain good results at lower costs have been made, Berral-Soler et al. [155] and Dhingra [156] proposed respectively RealHePoNet and LwPosr networks. However, the results are less accurate than those obtained with more complex models.

Other researchers, to overcome the limitations of publicly available datasets, that are limited in size, resolution, annotation accuracy and diversity, used synthetic generated data from high-quality 3D facial models to train their networks [58, 63]. Wang et al. [157] proposed a coarse-to-fine network to predict head pose trained on synthetically rendered faces. However, they noticed that the difference (domain gap) between rendered (source domain) and real-world (target domain) images negatively affects the performance. For this reason in [158, 159] Domain Adaptation (DA) techniques are applied to reduce the influence of domain differences.

Recently, Liu et al. propose ARHPE model [160], a novel asymmetric relation-aware network albe to learn the discriminative representations of adjacent head pose images. Different weights are assigned to the yaw and pitch directions by introducing the half at half maximum of the Lorentz distribution. This has proven effective in extracting more discriminative features, even if it has been tested only with two DoF (see Fig. 11).

Finally, some researches leveraged depth data [46, 60, 161]. Among them the best performing is POSEidon [60], which is a network composed of three independent convolutional nets followed by a fusion layer, specially conceived for understanding the pose by depth. This is the state-of-the-art model on the BIWI database [46] (see Table 4).

The main advantage of head pose estimation derived from CNNs is the strong learning ability, especially for image processing, which make it possible to achieve the desired effects. These algorithms work properly with high and low resolution images, and they have demonstrated their representational ability in tolerating some errors in the training set data. They are not dependent on the head model chosen, the landmark detection method, the subset of points used for alignment of the head model or the optimization method used for aligning 2D to 3D points. Moreover, they can be computationally efficient, straightforward to implement and easily updated with the addition of new data (data-driven approach, the upper limit is high).

However, the performance of these methods drops drastically if the labelled face images are not properly annotated. There can be difficulties in obtaining sufficient data with head annotations for head pose estimation training, especially data with changes in appearance (such as sex, age group, and race attribute) or environmental interference (such as lighting conditions, shooting angle). Many datasets don't have a uniform distribution of data (many images contain frontal or near-frontal faces) causing difficulties in learning large pose variations. Moreover, powerful CNNs are complex, and can require a long training time. It is also worth to stress that all these methods rely on a face detection step, prior to pose estimation, that can heavily influence the result.

Multi-task Methods

The idea behind multi-task methods is to relate head pose estimation to other face image analysis problems, such as gender recognition, landmark detection, face expression recognition, race classification, etc. because it is proven that jointly solving multiple tasks can lead to better performance [52, 75, 92–96, 162–164].

The *multi-task learning* (MLT) paradigm encompasses a set of learning techniques that provide effective mechanisms for sharing information among multiple tasks. It enables the use of larger and more diverse datasets, improving the stability of training and the generalization of the final model.

Among multi-task methods adopting traditional machine learning frameworks there are [162, 163]. The former adopts the graph guided FEGA-MTL framework for head pose classification of mobile targets based on multi-view image source. The physical space is divided into a discrete number of planar regions and the model try to learn the pose appearance relationship in each region. The latter tried to do the same, but evaluating the SVM-MTL framework.

Multi-task methods have become particularly popular with the advent of deep learning because of the unique ability of neural networks to transfer and share knowledge among various tasks. MTL has been widely used to simultaneously learn related tasks, such as: face detection + head pose estimation [97, 102, 103, 165, 166], face alignment + head pose estimation [93, 94, 98–100], face detection + face alignment + head pose estimation [95, 96, 101], face detection + face alignment + head pose estimation + gender recognition [92, 167], or also in combination with other tasks such as face recognition and appearance attributes estimation (age, smile, etc.) [52, 75] and finally there is head pose estimation + gaze estimation [168].

Zhang et al. [52] were the first to investigate the possibility of optimizing multiple tasks using a Task-Constrained Deep Convolutional Neural Network (TCDCN) to jointly optimize facial landmark detection with a set of related tasks, such as head pose estimation. The proposed network learns a shared feature space that is optimized to solve all the tasks at the same time. The network does not perform face detection, therefore it requires an image of a face as input or an additional preprocessing step. A similar network was proposed also by Ahn et al. [165], but their focus was on real-time driving face detection and head pose estimation.

Ranjan et al. [92] proposed a new model called Hyperface that performs face detection, face alignment, pose estimation and gender recognition. The network is designed to exploit the fact that information contained in features is hierarchically distributed throughout the network, therefore lower layers respond to edges and corners, and hence contain better localization properties (are more suitable for face alignment and pose estimation tasks); on the other hand, higher layers are class-specific and suitable for learning complex tasks such as face detection and gender recognition. They make use of all intermediate layer features (called *hyperfeatures*) through a technique named *feature fusion*, which allows to transform features to a common subspace where these can be combined linearly or non-linearly. They show that fusing intermediate layers improves the performance for structure dependent tasks of pose estimation and landmarks localization, as the features become invariant to geometry in deeper layers of CNN.

Then, Ranjan et al. [75] proposed another model called All-in-One. It differs from Hyperface because (I) simultaneously performs a higher number of tasks and (II) domain-based regularization is adopted by training on multiple datasets, each one specific to a subset of the tasks.

Xu et al. [93] have brought into the field a new type of network, i.e. a cascaded architecture that is designed in a hierarchical way based on coarse-to-fine principles, which refines the shape and pose sequentially. Other cascaded architectures have been presented in the literature, the main difference among them is the number of stages, the type and the number of tasks addressed in each stage [96, 97] (see Fig. 12).

Kumar et al. [94] transformed the cascaded regression formulation into an iterative scheme, by proposing the KEPLER model. In each iteration, a regressor predicts visibility, pose and the corrections for the next stage, and a rendering module uses these corrections to prepare new rendered data employed in the next iteration. The network is trained on three tasks namely, pose, visibilities and the bounded error using ground-truth annotations. The joint training is helpful

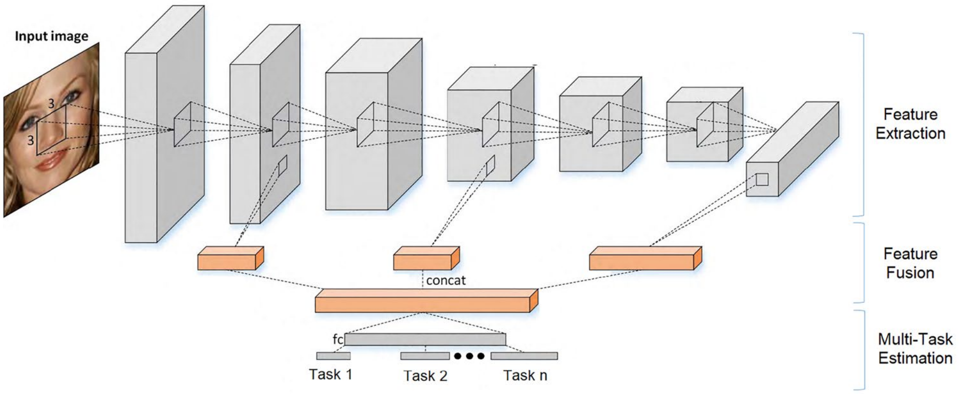


Fig. 12 A convolutional neural network with *feature fusion*, examples are Hyperface [92] and All-in-One [75] (image from [97])

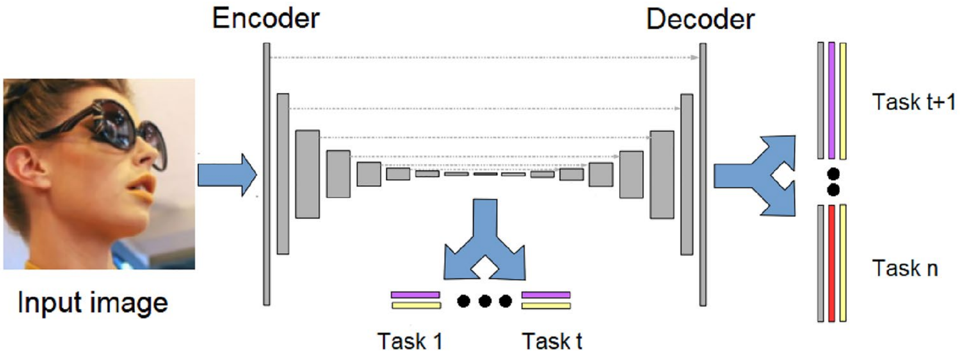


Fig. 13 Encoder-decoder network, called MNN, adopted in [98]

since it models the inherent relationship between the visible number of points, the pose and the amount of correction needed for a keypoint in a particular pose.

Many other researchers focused on improving the time needed for the network to resolve the tasks, indeed this is the main drawback of some of the presented models (e.g. Hyperface [92] or All-in-One [75]) that limits real-world applications. Cheng et al. [95] proposed a model that exploits single-shot object detection module (SSD) to perform multi-scale face detection, face alignment and head pose estimation at the same time at much higher speed. ASMNet [100] is a lightweight CNN assisted by an Active Shape Model (ASM) [169], used to guide the network towards learning, that achieves an acceptable performance for face alignment and pose estimation while having a significantly smaller number of parameters and floating point-operations. ATPN [99] and MOS [101] focused on defining a network structure with an even smaller number of parameters to augment efficiency. Other architectures, such as Multitask-net [102] and TRFH [103], leveraged the feature pyramid network to detect faces on different scales (see Fig. 13).

Valle et al. [98] proposed another type of architecture, an encoder-decoder CNN (see Fig. 13). They locate the head pose estimation task at the end of the encoder network, in this way the network bottleneck acts as embedding representing face pose. Instead, visibility and face alignment tasks are located at the end of the decoder, since they require information about the spatial location of landmarks in the image. This is the only paper to propose an encoder-decoder architecture. The presented model, called MNN, achieves results comparable to the state-of-the-art methods for the head pose estimation task; this is due to the network architecture and to a new training strategy that uses reannotated datasets.

Recently, Malakshan et al. [170] presented a completely different novel approach that jointly solves Face Super-Resolution (FSR) and HPE problems. To this end, a Multi-Stage Generative Adversarial Network (MSGAN) has been proposed: it benefits from the pose-aware adversarial loss and the head pose estimation feedback to generate super-resolved images that are properly aligned for HPE. Even if the network has not improved the results of SOTA methods on standard datasets, it significantly increased the pose estimation accuracy for the low resolution face images, obtaining at the same time very accurate results for original high-resolution images (on BIWI dataset MAE = 4.11).

The main advantage of the multi-task approach is that many tasks can be solved with a single model. Furthermore, all these tasks are strictly related, therefore the overall performance is improved due to the network's ability to learn correlations between data from different distributions in an effective way, so more discriminative features are learned. Also, some methods perform face detection with head pose estimation, reducing the time needed to perform preprocessing of the image. Another advantage is that multiple datasets can be used for training, increasing the amount of available data.

The main disadvantage of multi-task approach is the lack of public benchmark datasets with all the annotations for all the tasks. It's difficult to compare multi-task models among them and to other head pose estimation methods because they use a different combination of datasets for training and testing, therefore the better performance of a model could be due mainly to the training strategy rather than to the architecture of the proposed network. Moreover, some of the older models were not suited for real-world usage, e.g. Hyperface and All-in-One architectures took 3.5 s to process a single image [75]. Although newer models have managed to limit this problem, making it possible to obtain real-time systems.

Evaluation Metrics

A common informative metric used for evaluating HPE frameworks is the **Mean Absolute Error** (MAE) for all the three angles, i.e., pitch, yaw, and roll. MAE is quite popular (most of the papers discussed in this paper use it as main evaluation metric) since it provides a single statistics that gives a quick insight into the performance, for both fine or coarse pose estimations.

$$\text{MAE} = \frac{1}{n}\sum_{i=0}^{n}\left(\left|y_i - \hat{y}_i\right|\right).$$

However, in scenarios with large-range pose variations (360°), this evaluation method will not be reasonable. For example, when the actual angle is 170° and the predicted angle is – 170°, then the two angles are only 20° apart, but the MAE value calculated is 340°, making it bigger than its actual value [69].

For this reason, another measure has been proposed in the literature, called **Mean Absolute Wrapped Error** (MAWE) [7, 69]. The difference is clear by its definition:

$$\text{MAWE} = \frac{1}{n}\sum_{i=0}^{n}\min\left(\left|y_i - \hat{y}_i\right|, 360 - \left|y_i - \hat{y}_i\right|\right).$$

Another measure, mainly used for coarse head pose estimation, is the so-called **Pose Estimation Accuracy** (PEA). Being an accuracy measure, this metric depends on the number of poses, and therefore gives little information about the actual system performance. No recent work use it.

In recent studies on head pose estimation in the driving context, new evaluation metrics have been proposed [18–20]; however, no work on general head pose estimation use them.

The first metric is the **Balanced Mean Angular Error**, introduced to address the problem of the higher number of frontal pose images during evaluation, which leads to an unbalanced amount of different head orientations. The idea is to split the dataset in bins based on the angular difference from the frontal pose and average the MAE of each of the bins [18]

$$\text{BMAE} = \frac{d}{k}\sum_{i}\phi_{i,i+d} \qquad i \in d\mathbb{N} \cap [0, k],$$

where $\phi_{i,i+d}$ is the MAE of all hypotheses, the angular difference between the ground-truth and frontal pose is between i and $i + d$, d is the bin size and k is the maximum angle degree considered.

Other two metrics employed are the **Standard Deviation** (Std), that provides insights to the error distribution around the ground-truth, and finally the **Root Mean Squared Error**, to weight larger errors higher.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2}$$

RMSE takes the squared difference of the predicted value and the ground-truth value, weighing larger errors higher. Thus, high variation in predictions of an algorithm results in a higher overall error compared to the mean without squaring the values [19].

Evaluation

Comparing different methods is a complex and delicate problem, due to large number of different datasets that can be used for training and testing, and the different features that can be exploited by the models, such as depth information. The community is pushing for the adoption of well defined evaluation pipelines, discussed in the following section, that allows for a fair comparison between models; results relative to this group are given in Table 4 (no depth) and Table 5 (depth). In Table 6 we report figures relative to evaluation on the AFLW dataset [45], although the precise pipeline may be different or unknown. Finally, many systems uses ad-hoc datasets either for training, testing, or both tasks, as it is for instance the case for thematic scenarios like driving or video surveillance. Results relative this latter groups are provided in Tables 7 and 8, splitted in two parts for typographical reasons.

Evaluation Pipelines

Currently, in the state-of-the-art works [7, 8, 76, 82, 86, 87, 90, 91, 166, 187], there are two primary datasets for training: 300W-LP [53] and BIWI [46], corresponding two main datasets for testing AFLW2000-3D [53] and a part of BIWI [46].

The two most used evaluation protocols are [88]:

- *P1*: Training performed on a single dataset (300W-LP [53]), while BIWI [46] and AFLW2000-3D [53] are used as test sets. Only images with head rotation angles in range [– 99°, + 99°] are typically considered (in the case of AFLW2000 31 images are discarded);
- *P2*: Training and test sets are derived from the BIWI dataset [46], in some cases random split is applied (typically, 80% and 20% images), in others split by subject (18 and 2 subjects), recently the most common is the split by sequence (16-8 sequences for training and test respectively), but also *n*-fold cross-validation and leave-one-out cross-validation are used in the literature.

However, a major drawback of the considered evaluation pipelines is that the head pose angles (including pitch, yaw and roll) are all in the range [– 99°, + 99°], limiting the prediction of the models to a “narrow range” that makes them less effective on large-angle data, such as those acquired from security cameras [69].

For this reason, researchers frequently use additional head pose datasets. Zhou et al. for training the WHENet model [7] use the CMU Panoptic dataset [55] both to increase the amount of data and to get comprehensive yaw angles in range [– 179°, + 179°]. This is necessary to obtain a model optimized for the full range (360°) of face orientations, outperforming on such a task models exclusively trained with 300W-LP [53]. Albiero et al. [166] instead annotated the WIDER face database [189] using a deep learning regressor, and used it during training to increase the robustness of the model. Recently, Viet et al. [69] released the UET-Headpose dataset, also with uniform yaw angle in the range ±179°, that can be used as a new benchmark dataset for full range models.

Moreover, the semi-automatic pipeline used to label 300W-LP [53] and AFLW2000-3D [53] has been criticised for not producing accurate annotations for extreme poses and occluded faces [133]. Valle et al. [98] re-annotated AFLW2000-3D with poses estimated from the correct landmarks; this led to an improvement in model performance.

Other researchers employ synthetic datasets for training and tested on real ones [58, 63, 157–159]. Kuhnke et al. [158] propose novel benchmark datasets that are derived from BIWI [46] and SynHead [63], namely Biwi+, SynBiwi+, SynHead++. They propose these new datasets because SynHead was rendered using the Euler angles provided by BIWI, but with a different sequence of rotation axes. This rotation order, dissimilar to the BIWI one, causes that several SynHead images and BIWI images with the same label show different head rotations. For this reason, the reannotated SynHead+ contains SynHead images with correct angles. For every image in the BIWI dataset, SynBiwi+ has 10 corresponding images containing the 10 synthetic head models of SynHead. SynHead++ is the union of SynHead+ and SynBiwi+. To further improve the reproducibility manually collected bboxes for BIWI are provided in Biwi+ dataset.

Another dataset often used in the literature both for training and testing is the AFLW [45], however, there isn’t a common evaluation protocol used in the many studies published. The most common is:

- *P3*: Train and test set are defined by a random split, 23.386 images are used for training the model (of which typically 2.000 are employed as validation set) and 1.000 images for testing. More details about other evaluation pipelines for AFLW are in Table 6.

Discussion

Head pose estimation is an active research field of computer vision. It remains a challenging task due to several intrinsic and extrinsic problems, and the growing number of specialized contexts of application [2]. We organize this discussion in for parts: *datasets*, *methodologies*, *open problems*, and *research directions*.

Datasets

New databases are released every year because deep learning models require a huge quantity of data for training, but especially to overcome limitations of previous released datasets, such as limited head rotation angle ranges, non uniform distribution of angles, data captured in constraint environment, limited quality of ground-truth annotations, etc. (see FIg. 14)

Almost all most recent databases have annotations for all three rotation angles (*pitch*, *yaw* and *roll*), mainly acquired using depth cameras or optical motion capture systems. This is a major improvement with respect to earlier datasets that were acquired using direct suggestion or camera array methods, resulting in a discrete number of poses and annotations limited to one or two DoF.

The complexity of images has grown from simple faces on a flat background, to more complex scenarios with images acquired in-the-wild. However, a major drawback of the latter type is that pose is typically annotated manually or estimated with neural networks trained on other datasets,

Table 3 Head pose estimation publications most cited in recent literature

Year	Paper	Approach	DoF	Dataset
2011	Fanelli et al. [46]	Random Forest	3	BIWI
2012	Baltrusaitis et al. [48]	CLM-Z Model based	3	BIWI, BU, ICT-3DHP
2014	Ahn et al. [143]	DCNN	3	BIWI
2014	Martin et al. [140]	Model based	3	BIWI
2014	Peng et al. [117]	Manifold embedding	3	Multi-Pie
2014	Tulyakov et al. [51]	ML + Tracking	2	Dali3DHP
2014	Zhang et al. [52]	Multi-task DCNN	3	AFLW[a,c,d], AFW[a,c,d]
2015	Drouard et al. [171]	Gaussian locally-linear mapping	3	BIWI, Pointing'04
2015	Meyer et al. [141]	3DMM Model based	3	BIWI, ETH
2015	Papazov et al. [172]	3DMM Model based	3	BIWI, Synthetic data
2015	Saeed et al. [161]	ML: HoG + SVR	3	BIWI, ICT-3DHP
2015	Sundararajan et al. [115]	Manifold embedding	3	AFLW, AFW, McGill
2016	Gu et al. [63]	RNN	3	BIWI, ETH, SynHead
2016	Liu et al. [58]	DCNN	3	BIWI, Synthetic
2016	Xingyu et al. [144]	DCNN (VGG)	3	IDIAP-HP
2017	Amador et al. [73]	DCNN	3	300W, AFLW, AFW
2017	Barros et al. [173]	PnP Model based	3	BU
2017	Borghi et al. [60]	DCNN	3	BIWI, ICT-3DHP, Pandora
2017	Bulat et al. [133]	PnP Model based	3	300-VW[c], 300W-LP[c], AFLW2000[c], Menpo[c]
2017	Diaz-Chito et al. [118]	Manifold embedding	3	CAS-PEAL, CMU-Pie, DrivFace, Pointing'04, Taiwan RoboticsLab
2017	Gao et al. [174]	Deep label distribution learning	3	AFLW, BJUT-3D, Pointing'04
2017	Gou et al. [175]	Model based	3	300W[⋄], BU[a]
2017	Khan et al. [29]	Segmentation based	2	Pointing'04
2017	Kumar et al. [94]	Multi-task DCNN	3	AFLW[a,c], AFW[a,c]
2017	Lathuliere et al. [152]	DCNN	3	BIWI
2017	Patacchiola et al. [148]	DCNN	3	AFLW, AFW, Pointing'04
2017	Ranjan et al. [92]	Multi-task DCNN	3	AFLW[a,b], AFW[a,b,c], CelebA[d], FDDB[b], LFWA[d], Pascal[b]
2017	Ranjan et al. [75]	Multi-task DCNN	3	Adience[f], AFLW[a,b,c], AFW[a,b], CASIA[e], Chalern LAP2015[f], CelebA[d], FDDB[b], FG-NET[f], IJB-A[e], Morph[f], Pascal[b]
2017	Wu et al. [84]	Model based	3	BU4D-FE[g], BU[a], COFW[c], Multi-Pie[a,c]
2017	Xu et al. [93]	Multi-task DCNN	3	300W[a,c]
2017	Yu et al. [176]	Model based	3	BIWI, UbiPose
2018	Ahn et al. [165]	Multi-task DCNN	3	AFLW[a,b], BIWI[a,b], RCVFace[a,b], NDS[b]
2018	Barros et al. [135]	Model based + Tracking	3	BU
2018	Cai et al. [96]	Multi-task DCNN	3	300W[a,b,c]
2018	Chen et al. [95]	Multi-task DCNN	3	AFLW[a,b,c], AFW[a,c], FDDB[b], Pascal[b], WIDER[b]
2018	Gupta et al. [81]	Model based MLP	3	AFLW, BIWI
2018	Hong et al. [164]	Multi-task Multi-view + Manifold learning	3	BIWI, Pointing'04
2018	Ruiz et al. [8]	DCNN	3	300W-LP, AFLW, AFLW2000, BIWI
2018	Yu et al. [64]	Model based 3DMM	3	BIWI, UbiPose
2018	Zhang et al. [177]	Multi-task DCNN	3	AFLW[a,c]
2019	Abate et al. [129]	Model based Quad Tree	3	AFLW, BIWI
2019	Benini et al. [31]	Segmentation based SVM	2	Pointing'04
2019	Derkach et al. [119]	Manifold embedding	3	BIWI, SASE
2019	Hsu et al. [89]	DCNN	3	300W-LP, AFLW, AFLW2000, AFW, BIWI
2019	Khan et al. [30]	Segmentation based	3	AFLW, BU, ICT-3DHP, Pointing'04
2019	Khan et al. [32]	Segmentation based Random Forest	3	AFLW, BU, ICT-3DHP, Pointing'04
2019	Kuhnke et al. [158]	DCNN	3	Biwi+, SynBIWI+, SynHead++

Table 3 (continued)

Year	Paper	Approach	DoF	Dataset
2019	Liu et al. [178]	DCNN	3	300W-LP, AFLW, AFLW2000, AFW, BIWI
2019	Shao et al. [179]	DCNN	3	300W-LP, AFLW2000, BIWI
2019	Wang et al. [157]	DCNN	3	BIWI, BU, Pointing’04, Synthetic data
2019	Wang et al. [180]	DCNN	3	300W-LP, AFLW, AFLW2000, BIWI
2019	Xu et al. [181]	DCNN	3	CAS-PEAL, Multi-Pie, Pointing’04
2019	Xia et al. [82]	Model based DCNN	3	300W-LP, AFLW2000, BIWI, CAS-PEAL, DriveFace
2019	Yang et al. [88]	DCNN	3	300W-LP, AFLW2000, BIWI
2020	Barra et al. [130]	Model based	3	AFLW, BIWI, Pointing’04
2020	Cao et al. [76]	DCNN	3	300W-LP, AFLW2000, BIWI
2020	Dai et al. [90]	DCNN	3	300W-LP, AFLW2000, BIWI
2020	Dapongy et al. [83]	Model based	3	300W, 300W-LP, AFLW2000, CelebA, WFLW
2020	Ewaisha et al. [168]	Multi-task DCNN	3	CAVE
2020	Valle et al. [98]	Multi-task DCNN	3	300W-LP^{a,c}, AFLW^{a,c}, AFLW2000^{a}, BIWI^{a}, COFW^{c}, WFLW^{a,c}
2020	Wang et al. [182]	PnP Model based	3	300W, AFLW2000
2020	Zhang et al. [183]	DCNN	3	300W-LP, AFLW2000, BIWI
2020	Zhang et al. [167]	Multi-task DCNN	3	AFLW^{a,b,c}
2020	Zhou et al. [7]	DCNN	3	300W-LP, AFLW2000, BIWI, CMU Panoptic
2021	Albiero et al. [166]	Multi-task DCNN	3	300W-LP^{a}, AFLW2000^{a}, BIWI^{a}, WIDER^{a,b}
2021	Basak et al. [159]	DCNN	3	BIWI, SASE, Synthetic data
2021	Berg et al. [184]	DCNN	3	BIWI
2021	Berral-Soler et al. [155]	DCNN	3	AFLW, Pointing’04
2021	Fard et al. [100]	Multi-task DCNN + ASM	3	300W^{a,b}, WFLW^{a,b}
2021	Hu et al. [185]	DCNN	3	300W-LP, AFLW2000, BIWI
2021	Khan et al. [80]	Segmentation based Soft-max classifier	3	AFLW, BU, ICT-3DHP, Pointing’04
2021	Liu et al. [85]	Multi-task DCNN	3	AFLW^{c}, AFLW2000^{a}, WIDER*
2021	Naina Dhingra [186]	DCNN	3	300W-LP, AFLW2000, BIWI
2021	Ruan et al. [87]	Model based 3DMM + DCNN	3	300W-LP^{a,c,g}, AFLW2000$^{\diamond\circ*}$, Florence^{g}
2021	Sheka et al. [91]	DCNN	3	300W-LP, AFLW, AFLW2000, BIWI
2021	Viet et al. [102]	Multi-task DCNN	3	300W-LP^{a,b}, BIWI^{a,b}, CMU Panoptic^{a,b}
2021	Viet et al. [69]	DCNN	3	300W-LP, AFLW2000, CMU Panoptic, UET-Headpose
2021	Xia et al. [99]	Multi-task DCNN	3	300W-LP^{a}, 300VW^{c}, WFLW^{c}, WIDER^{b}
2021	Xin et al. [187]	Model based Graph CNN	3	300W-LP, AFLW2000, BIWI
2021	Wu et al. [86]	Model based 3DMM + DCNN	3	300W-LP^{a,c,g}, 300VW^{g}, AFLW^{c}, AFLW2000^{a,c}, Florence^{g}
2022	Cantarini et al. [188]	Model based DCNN	3	300W-LP, AFLW2000, BIWI
2022	Hempel et al. [151]	DCNN	3	300W-LP, AFLW2000, BIWI
2022	Liu et al. [160]	DCNN	2	AFLW2000, Pointing’04, HRIHP
2022	Martyniuk et al. [154]	Model based DCNN	3	300W-LP, AFLW2000, BIWI
2022	Naina Dhingra [156]	DCNN	3	300W-LP, AFLW2000, BIWI
2022	Wang et al. [154]	DCNN	3	300W-LP, AFLW2000, BIWI
2022	Zeng et al. [150]	DCNN	3	300W-LP, AFLW2000, BIWI
2023	Malakshan et al. [170]	Multi-task GAN	3	300W-LP, AFLW2000, BIWI, CelebA, WIDER

For multi-task models we annotated the specific tasks for which each dataset is used as follows: ^{a}head pose estimation, ^{b}face detection, ^{c}face alignment, ^{d}gender classification, ^{e}face recognition, ^{f}age estimation, ^{g}face reconstruction

leading to inaccuracies in the ground-truth annotations (see for example Fig. 15).

Another drawback of almost all the datasets is the data imbalance issue: the distribution between easy frontal faces and more challenging orientations is heavily unbalanced. Techniques to increase the number of hard faces [195] or to enhance the contribution of hard examples (such as HEM [150]) can be used to alter the data distribution space and

Table 4 Evaluation results of head pose estimation on AFLW2000 [53] and BIWI [46]

Name	Type	Eval pipeline *P1* test on BIWI				Eval pipeline *P1* test on AFLW2000				Eval pipeline *P2*					MB	Param 10^6	Extra training data	Data	Full range	Pre-process. step
		Pitch	Yaw	Roll	MAE	Pitch	Yaw	Roll	MAE	Pitch	Yaw	Roll	MAE	Split						
3DDFA [53]	MB	12.3	36.2	8.78	19.1	8.53	5.40	8.25	7.39									RGB	N	
KEPLER [94]	MT	17.2	8.8	16.2	13.9													RGB	N	
Dlib (68 landmarks) [132]	MB	13.8	16.8	6.19	12.2	13.6	23.1	10.5	15.8							6–24		RGB	N	
FAN (12 points) [133]	MB	7.48	8.53	7.63	7.89	12.3	6.36	8.71	9.12						183	~36.6		RGB	N	
Liu et al. [58]	D									6.10	6.00	5.70	5.90	Rnd				RGB	N	
Drouard et al. [171]	ME									5.90	4.70	4.10	5.20	Sbj				RGB	N	VJ^{fd}
MT-Net v2 (Euler) [102]	MT	7.23	4.64	6.23	6.03					5.33	6.02	5.11	5.48	Seq				RGB	Y	Direct
Shao et al. [179]	D	7.25	4.59	6.15	6.00	6.37	5.07	4.99	5.48						93	24.6		RGB	N	JCFDA^{fd}
HHP-Net [188]	MB	7.00	4.14	4.40	5.18	10.12	5.26	7.73	7.70	4.79	3.04	3.21	3.68	Seq	**0.4**	**0.1**		RGB	N	OpenPose^{kd}
VGG-IR-FT [152]	D									4.68	3.12	3.07	3.62	Seq	500			RGB	N	
Hopenet ($\alpha = 2$) [8]	D	6.98	5.17	3.39	5.18	6.56	6.47	5.44	6.16						95.9	23.9		RGB	N	FR^{fd}
Hopenet ($\alpha = 1$) [8]	D	6.61	4.81	3.27	4.90	6.64	6.92	5.67	6.41	3.39	3.29	3.00	3.23	Seq	95.9	23.9		RGB	N	FR^{fd}
RetinaFace R50 (5pnt) [166]	D	6.42	4.07	2.97	4.49	9.64	5.10	3.92	6.22									RGB	N	Direct
SSR-Net-MD [88]	D	6.31	4.49	3.61	4.65	7.09	5.14	5.89	6.01	4.35	4.24	4.19	4.26	Seq	1.1	0.2		RGB		MTCNN^{fd}
MT-Net v2 (Vecotr) [102]	MT	4.29	4.62	4.52	4.48					3.90	5.33	3.28	4.17	Seq				RGB	Y	Direct
LwPosr [156]	D	4.87	4.11	3.19	4.05	6.38	4.44	4.88	5.35	4.65	3.62	3.78	4.01	Seq		0.15		RGB	N	MTCNN^{fd}
FSA-Caps (1×1) [88]	D	5.15	4.56	2.94	4.31	6.19	4.82	4.76	5.25						1.1			RGB	N	MTCNN^{fd}
FSA-Caps-Fusion [88]	D	4.96	4.27	2.76	4.00	6.08	4.50	4.64	5.07	4.29	2.89	3.6	3.6	Seq	5.1	1.2		RGB	N	MTCNN^{fd}
FDN [183]	D	4.96	4.27	2.76	4.00	6.08	4.50	4.64	5.07	3.98	3.00	2.88	3.29	Seq	5.8			RGB	N	MTCNN^{fd}
QuatNet [89]	D	5.49	4.01	2.93	4.14	5.61	3.97	3.92	4.50									RGB		
HeadPosr EH38 [186]	D	5.10	4.08	3.02	4.06	4.86	4.60	2.87	4.11									RGB	N	MTCNN^{fd}
HeadPosr EH64 [186]	D	5.44	3.37	2.69	3.83	5.84	4.64	4.30	4.92	4.03	2.59	3.53	3.38	Seq				RGB	N	MTCNN^{fd}

Table 4 (continued)

Name	Type	Eval pipeline *P1* test on BIWI				Eval pipeline *P1* test on AFLW2000				Eval pipeline *P2*					MB	Param 10^6	Extra training data	Data	Full range	Pre-process. step
		Pitch	Yaw	Roll	MAE	Pitch	Yaw	Roll	MAE	Pitch	Yaw	Roll	MAE	Split						
HR-AT-nBG [185]	D									3.74	3.07	3.11	3.31	Seq				RGB	N	Direct
CNN+Heatmap [81]	MB									3.49	3.46	2.74	3.23	8FCV		3.2		RGB	N	
Gu et al. [63]	RNN									4.03	3.91	3.03	3.66	Seq	500	~136		RGB	N	
Gu et al. [63]	RNN									3.48	3.14	2.60	3.10	Seq	500	~136		RGB+Time	N	
Hybrid Coarse-fine [180]	D					6.23	4.82	5.14	5.40	**2.64**	3.43	2.98	3.02	8FCV	96.7	~24		RGB	N	FR^{fd}
img2pose [166]	MT	3.55	4.57	3.24	3.79	5.03	3.43	3.28	3.91						265		WIDER*	RGB	N	Direct
MNN [98]	MT	4.61	3.98	2.39	3.66	4.69	3.34	3.88	4.42									RGB	N	Direct
Ahn et al. [143]	D									3.40	2.80	2.60	2.93	Rnd				RGB	N	
TriNet [76]	D	4.76	**3.05**	4.11	3.97	5.77	4.20	4.04	4.67	3.04	2.44	2.93	2.80	3FCV	~26			RGB	N	MTCNN^{fd}
3DDFA-TPAMI [136]	MB					5.98	4.33	4.30	4.87									RGB	N	FTF^{fd}
MOS [101]	MT					5.42	3.91	3.98	4.43									RGB	N	Direct
FSA-Net-Wide [69]	D					5.69	4.59	2.85	4.37						2.91		UET, CMU	RGB	Y	
3DDFA-V2 [138]	MB					4.09	3.42	3.48	4.27									RGB	N	
2DASL [137]	MB					5.06	3.85	3.50	4.13								UMD	RGB	N	
SADRNet [87]	MB					5.00	2.93	3.54	3.82						60			RGB	N	
GLDL [178]	D	5.61	4.12	3.14	4.29	5.06	3.03	3.68	3.93									RGB	N	FR^{fd}
KD-ResNet152 [91]	D	4.73	3.50	2.87	3.70^{3}	4.52	2.97	3.48	3.48	2.88	2.61	2.37	2.62	Seq				RGB	N	Yolo-v5^{fd}
KD-ResNet18 [91]	D	5.07	3.96	3.06	4.03	4.69	3.00	3.22	3.64	2.82	2.59	2.15	2.58^{3}	Seq				RGB	N	Yolo-v5^{fd}
Direct Regression [184]	D									2.75	2.64	2.24	2.54^{2}	Seq				RGB	N	FR^{fd}
RankPose [90]	D	4.77	3.59	2.76	3.71	4.75	2.99	3.25	3.66									RGB	N	MTCNN^{fd}
SRNet [150]	D	3.81	4.36	2.77	3.65	3.75	5.10	3.46	4.10	3.01	2.78	2.86	2.88	Seq	8.9			RGB	N	
EVA-GCN [187]	MB	4.78	4.01	2.98	3.92	5.34	4.46	4.11	4.64	2.82	**2.01**	**1.89**	**2.24**^{1}	Seq	1.03	~3.3		RGB	N	FAN^{kd}
WHENet [7]	D	4.39	3.99	3.06	3.81	6.24	5.11	4.92	5.42						17.1	4.4	CMU	RGB	Y	Yolo-v3^{fd}
WHENet-V [7]	D	4.10	3.60	2.73	3.48^{3}	5.75	4.44	4.31	4.83						17.1	4.4		RGB	N	Yolo-v3^{fd}
6DRepNet [151]	D	4.48	3.24	**2.68**	3.47^{2}	4.91	3.63	3.37	3.97	2.92	2.69	2.36	2.66	Seq	17.1	4.4		RGB	Y	MTCNN^{fd}
FSEN [154]	D	**3.45**	3.34	3.29	**3.36**^{1}	4.37	3.24	3,41	3.67									RGB	N	MTCNN^{fd}

Table 4 (continued)

Name	Type	Eval pipeline *P1* test on BIWI				Eval pipeline *P1* test on AFLW2000				Eval pipeline *P2*					MB Param 10^6	Extra training data	Data	Full range	Pre-process. step
		Pitch	Yaw	Roll	MAE	Pitch	Yaw	Roll	MAE	Pitch	Yaw	Roll	MAE	Split					
SynergyNet [86]	MB					4.09	3.42	2.55	3.35^2						3.8		RGB	N	Direct
Xia et al. [82]	MB					**2.05**	**0.63**	**1.70**	**1.46**^{1}	2.52	2.83	2.86	3.74	5FCV			RGB	N	FAN^{kd}

For the evaluation protocol P2 many variants are reported in the literature: Random split, Split by subject (18 and 2 subjects), Split by sequence (16 and 8 sequences), *n*-fold cross-validation and Leave-one-out cross-validation, the splitting method is reported here when available. Model type: (D) deep learning regressor; (MB) Model based; (ME) Manifold embedding; (ML) Machine learning regressor; (MT) Multi-task; (RNN) Recurrent neural network. Narrow range models are optimized for ±99°, full range for ±180°, ▿ means ±120°. Extra training data used are CMU Panoptic [55], UET-Headpose [69], UMDFace [59] and WIDER [189] (* head pose are annotate with a deep learning regressor). In pre-processing *fd* means face detector, *kd* means keypoints (landmarks) detector. VJ is Viola-Jones face detector implemented in openCV [109]; FR is Faster-RCNN [190]; JCFDA [191]; openPose [192]; Yolo [193]; Dlib [132]; FTF is finding tiny faces detector [194]. Other training/testing strategies used for BIWI dataset are presented in Table 7

overcome this issue, making trained models more robust and with better a generalization capability (Fig. 14).

Among all the databases, Boston University [34] is still used to evaluate head pose estimation methods even if it is one of the oldest; some model-based and segmentation based methods obtain very accurate performance on it, as can be seen in Table 7. Also Pointing'04 [38] is still employed for research purposes, even if it was introduced back in 2004, due to its challenging nature and high image diversity.

BIWI Kinect [46] has become the de-facto benchmark dataset with a high number of publications that evaluate their models on it. However, this dataset has two main disadvantages: it's a *narrow range* dataset, head rotation angles go from – 75° to + 75°, making it not suitable to evaluate models optimized for *full range* (360°) head rotations; furthermore, it's a dataset with images acquired in a constraint environment, therefore less challenging than other captured with different lighting conditions, backgrounds or occlusions.

Nowadays *synthetic databases* [58, 62, 63] enable more precise evaluation and comparison of HPE methods because they contain nearly perfect ground-truth data. However, training solely on synthetic data can cause poor performance when testing on real-world data due to mismatch or shift of underlying data distribution (domain gap). For this reason, training on a combination of synthetic data and real ones can lead to an improvement of the final result, see for example FSA-Net [88] model tested on BIWI dataset [46] in Table 7.

Recently, the most active sub-field seems to be "driver head pose estimation", in the last five years five public datasets that address this specific scenario have been released, each with thousands or millions of images. This is mainly due to the increasing interest in driving assistance systems that aim to monitor the driver attention, behaviour and intention, and the fact that head pose is a key element to obtain accurate results [18, 19].

Methodologies

In parallel with the growing number and quality of available datasets, the number of head pose publications has constantly increased in the past few years. More and more people are interested in this area, leading to the development of many different and innovative approaches. Nowadays, deep learning and methods based on convolutional neural networks are the most pervasive: these are used to estimate head pose from monocular images, from a set of detected facial landmarks, from a combination of both in a multi-task approach, or even are used to perform 3D dense face alignment/reconstruction, from which the head pose information is obtained as by-product.

Segmentation based methods are the only recently developed methods that mainly rely on classical machine learning models. They proved the existence of a strong correlation

Table 5 Evaluation results of head pose estimation on AFLW2000 [53] and BIWI [46] for methods exploiting depth data

Name	Type	Eval pipeline *P2*					MB	Param 10^6	Extra training data	Data	Full range	Pre-process. step
		Pitch	Yaw	Roll	MAE	Split						
Fanelli et al. [46]	ML	8.50	7.90	8.90	8.43	Sbj				Depth	N	VJ^{fd}
Baltrusaitis et al. [48]	MB	5.10	11.30	6.30	7.60	Sbj				RGB+D	N	VJ^{fd}
Saeed et al. [161]	ML	5.00	4.30	3.90	4.40	Sbj				RGB+D	N	VJ^{fd}
LMK [119]	ME	3.80	3.60	5.20	4.20	L1O				Depth	N	
DESC [119]	ME	3.40	3.30	3.30	3.33	L1O				Depth	N	
Papazov et al. [172]	MB	2.50	3.80	3.00	3.20					Depth	N	VJ^{fd}
Martin et al. [140]	MB	2.50	2.60	3.60	2.90	Sbj				Depth	$\text{Y}^{\triangledown}$	Videmo^{fd}
Meyer et al. [141]	MB	2.40	2.10	2.10	2.20	Sbj				Depth	N	Custom^{fd}
Yu et al. [176]	MB	1.53	2.49	2.18	2.07^3					RGB+D	Y	$\text{Dlib}^{fd,kd}$
HeadFusion [64]	MB	**1.45**	2.54	2.10	2.03^2					RGB+D	Y	$\text{Dlib}^{fd,kd}$
POSEidon [60]	D	1.60	**1.70**	**1.80**	$\mathbf{1.70}^1$	Sbj	3.4			Depth	Y	CustomNN^{fd}

Evaluation protocols are typically based on variants of P2. Model type: (D) Deep learning regressor; (MB) Model based; (ME) Manifold embedding; (ML) Machine learning regressor; (MT) Multi-task. In pre-processing *fd* means face detector, *kd* means keypoints (landmarks) detector. VJ is Viola-Jones face detector implemented in openCV [109]; Dlib [132]. Other training/testing strategies used for BIWI dataset are presented in Table 7

between face segments and the corresponding pose, and that a precise face segmentation may lead to very accurate pose estimations [30]. However, a severe drop in performance is often registered when segmentation is applied in unconstrained environments [32], that hence remains a challenge for future research.

What emerges most from the literature is the strong correlation between face alignment and head pose estimation. This correlation is exploited in different ways in the literature. Among the best performing methods there are:

- Xia et al. [82] perform face alignment and then create a landmark *heatmap* that is given as input (along with the facial image) to a CNN. They obtain the best result on AFLW2000 dataset [53] because the heatmap generator improves the generalization ability by making the CNN focus on the area around facial landmarks and reducing the interference from background significantly. However, this method does not remarkably improve the performance on datasets taken under controllable conditions, such as BIWI [46].
- Valle et al. [98] combine face alignment and head pose estimation in a multi-task model improving the overall performance, obtaining the best result on AFLW dataset [45].
- Xin et al. [187] construct a landmark-connection graph to model the complex non-linear mapping between graph topologies and head pose angles. Their model has the lowest MAE when trained and tested on BIWI dataset [46] among the models that use only RGB data.
- Wu et al. [86] exploit facial landmarks to guide 3D facial geometry learning. Pose in this case is a by-product that a backbone network learns during 3DMM parameter regression. SynergyNet outperform all deep learning regressors on AFLW2000 dataset [53].

A different class of models that look particularly promising are those based on 3DMM. They focus on face reconstruction and incorporate occlusion aware mechanisms very useful in complex scenarios. Moreover, because these methods do not use any ground-truth head pose label during training, they do not suffer from the inaccuracy of head pose labels that exist in most publicly available training datasets. Room-of-improvement might exist by designing specialized loss function and addressing specifically the head pose estimation task.

From Table 3 we can see that almost all the models can estimate 3 DoF; actually, some of them (such as 3DMM based) can estimate 6 DoF, but databases are mainly equipped with 3 DoF or less. This highlights a great evolution, indeed until a few years ago, researchers focused more on yaw estimation, because of its importance in applications such as human attention, gaze estimation, etc. Deep learning changed the trend, all three rotation angles are currently being addressed in most works.

Table 6 Evaluation results of head pose estimation on AFLW [45] (ordered by training pipeline)

Name	Type	Train	Test	Evaluation pipeline	Pitch	Yaw	Roll	MAE	Data type	Pre-process. step
DLDL (KL) [174]	D	AFLW	AFLW	1	5.75	6.60			RGB	
AVM [115]	ME	AFLW	AFLW	2				17.48	RGB	VJ^{fd}
Dlib[a] [132]	MB	Not req.	AFLW	Unknown	13.6	23.1	10.5	15.7	RGB	
TRFH [103]	MT	AFLW	AFLW	Unknown	23.81	5.49	17.26	15.52	RGB	Direct
FAN[a] [133]	MB	Not req.	AFLW	Unknown	12.3	6.4	8.7	9.13	RGB	
3DDFA[a] [53]	MB	Not req.	AFLW	Unknown	8.2	5.4	8.7	7.43	RGB	
GLDL [178]	D	AFLW	AFLW	Unknown	5.31	6.00	3.75	5.02	RGB	FR^{fd}
LeNet-5 [148]	D	AFLW	AFLW	5-FCV	7.15	11.04	4.40	7.53	RGB	
MLP+Locations (5pnt.) [81]	MB	AFLW	AFLW	5-FCV	6.64	9.56	4.68	6.96	RGB	OpenPose^{kd}
CNN+Heatmaps (5pnt.) [81]	MB	AFLW	AFLW	5-FCV	5.58	6.19	3.76	5.18	RGB	OpenPose^{kd}
Segm+CNN [80]	SB	AFLW	AFLW	10-FCV	3.2	4.9			RGB	SSD^{fd}
HPE-MSF-CRFs [30]	SB	AFLW	AFLW	10-FCV	4.89	4.25	3.20	4.11	RGB	SSD^{fd}
HAG-MSF-CRFs [32]	SB	AFLW	AFLW	10-FCV	4.89	4.25	3.20	4.11	RGB	SSD^{fd}
QT_PYR [129]	MB	Not req.	AFLW	3	7.60	7.60	7.17	7.45	RGB	VJ^{fd}, Dlib^{kd}
Hybrid Coarse-fine [180]	D	300W-LP	AFLW	3	5.38	6.18	5.09	5.55	RGB	
4D_4S [130]	MB	Not req.	AFLW	3	4.82	3.11	**2.25**	3.39	RGB	Dlib^{kd}
KD-ResNet18 [91]	D	AFLW	AFLW	4	6.02	5.45	4.16	5.21	RGB	Yolo-v5^{fd}
KD-ResNet152 [91]	D	AFLW	AFLW	4	5.93	5.41	4.07	5.14	RGB	Yolo-v5^{fd}
QuatNet [89]	D	300W-LP	AFLW	5	4.32	**3.93**	2.59	3.61	RGB	Gt bbox
CCR [177]	MT	AFLW	AFLW	6	5.85	5.22	2.51	4.53	RGB	
KEPLER [14]	MB	AFLW	AFLW	P3	5.85	6.45	8.75	6.45	RGB	
Hyperface [92]	MT	AFLW	AFLW	P3	6.13	7.61	3.92	5.88	RGB	SSO^{fd}
Hopenet ($\alpha = 1$) [8]	D	AFLW	AFLW	P3	5.89	6.26	3.82	5.32	RGB	FR^{fd}
MLP+Locations (5pnt.) [81]	MB	AFLW	AFLW	P3	5.84	6.02	3.56	5.14	RGB	OpenPose^{kd}
VGG-16 [73]	D	AFLW	AFLW	P3	5.24	6.45	3.61	5.10	RGB	
AlexNet [73]	D	AFLW	AFLW	P3	5.21	6.40	3.47	5.02	RGB	
MOS [101]	MT	AFLW	AFLW	P3				4.89	RGB	Direct
ResNet-50 [73]	D	AFLW	AFLW	P3	5.02	6.03	3.22	4.75	RGB	
VGG-19 [73]	D	AFLW	AFLW	P3	4.93	5.99	3.15	4.69	RGB	
ResNet-101 [73]	D	AFLW	AFLW	P3	4.98	5.69	3.07	4.59	RGB	
ResNet-152 [73]	D	AFLW	AFLW	P3	4.88	5.92	2.98	4.58	RGB	
CNN+Heatmaps (5pnt.) [81]	MB	AFLW	AFLW	P3	4.43	5.22	2.53	4.06	RGB	OpenPose^{kd}
MNN [98]	MT	AFLW	AFLW	P3	**3.07**	4.16	2.43	**3.22**	RGB	

[a] Results taken from [28]. Evaluation pipeline: (1) Random split—15.561 images for training, 7.848 for testing; (2) Random split–—14.000 images for training, 7.041 for testing; (3) Test on all AFLW; (4) First 2.000 images for testing other for training; (5) Train on other dataset, test on 1.000 random sample from AFLW; (6) Random split - 20.000 images for training other for testing; (n-FCV) n-fold cross-validation. Model type: (D) Deep learning regressor; (MB) Model based; (ME) Manifold embedding; (ML) Machine learning regressor; (MT) Multi-task; (RNN) Recurrent neural network; (SB) Segmentation based model. In preprocessing *fd* means face detector, *kd* means keypoints (landmarks) detector. Not all papers specify the preprocessing applied, some are direct methods that incorporate a detection phase, other use face crop from gt bbox

From Table 5 we observe that methods that use *depth* data, alone or in conjunction with *RGB* information, can usually achieve better results. In particular, the use of depth data enhances the efficacy under challenging illumination conditions and occlusions, making the models suitable for particularly complex scenarios, such as automotive. From Table 7 we can see that, recently, also thermal infrared images (IR) are used as input for HPE algorithms, in some cases obtaining better results than with depth information. However, depth or infrared data are not always available in real-world contexts, and are also quite expensive; therefore,

Table 7 Evaluation results of head pose estimation on other databases

Name	Type	Train	Test	Evaluation pipeline	Pitch	Yaw	Roll	MAE	MAWE	BMAE	Data type
POSEidon [19]	D	AutoPose	AutoPose	18 sequences for train, 1 for test	**2.96**	**3.16**	**3.99**	**3.37**		**11.86**	IR
DLDL (KL) [174]	D	BJUT-3D	BJUT-3D	5-fold cross-validation	**0.02**	**0.07**					RGB
FSA-Net Caps-Fusion [159]	D	SynHead	BIWI	Test on all BIWI	8.58	6.04	9.82	8.29			RGB
KPM [85]	MB	Not req	BIWI	Test on all BIWI	7.94	5.81	6.74	6.83			RGB
FSA-Net Caps-Fusion [159]	D	SGD[a] (300k)	BIWI	Test on all BIWI	6.51	5.86	6.63	6.34			RGB
DANN [158]	D	SynHead++	BIWI+	Test on all BIWI+	8.08	6.17	3.91	6.05			RGB
QT_PYR [129]	MB	Not req	BIWI	Test on all BIWI	7.51	4.07	5.50	5.69			RGB
4C_4S [130]	MB	Not req	BIWI	Test on all BIWI	3.95	6.21	4.16	4.77			RGB
DC2F [157]	D	SGD[a] (208k) + BIWI	BIWI	Random split BIWI: 12k train, 3k test	5.48	4.76	4.26	4.54			RGB
FSA-Net Caps-Fusion [159]	D	SGD[a] (300k) + BIWI	BIWI	Random split BIWI: 14k train, 1k test	4.54	4.62	3.33	4.16			RGB
PADACO [158]	D	SynHead++	BIWI+	Test on all BIWI+	4.51	4.11	3.78	4.13			RGB
PADACO [158]	D	SynBiwi+	BIWI+	Test on all BIWI+	4.47	4.11	3.56	4.04			RGB
DAD-3DNet [70]	D	DAD-3DH	BIWI	Test on all BIWI	5.24	3.79	2.92	3.98			RGB
RT-MT-HPE [165]	MT	BIWI + RCVFace	BIWI	Random split 37 subjects for training, 10 for test	4.3	**3.4**	3.6	3.76			RGB
Liu et al. [58]	D	Synthetic	BIWI	Random 30 seq. from synthetic db for training, test on all BIWI	4.3	4.5	**2.4**	3.73			RGB
DANN [158]	D	SynBiwi+	BIWI+	Test on all BIWI+	**3.56**	3.43	3.03	**3.34**			RGB
CCR [175]	MB	Not req	BU	Test on 200 images of 5 subjects, uniform lighting conditions	4.8	5.1	3.3	4.4			RGB
S-FLD-HPE [84]	MB	Not req	BU	Test on 200 images of 5 subjects, uniform lighting conditions	5.3	4.9	3.1	4.4			RGB
CHM+PnP [173]	MB	Not req	BU	Test on 200 images of 5 subjects, uniform lighting conditions	4.58	4.87	2.80	4.08			RGB
EHM+PnP [173]	MB	Not req	BU	Test on 200 images of 5 subjects, uniform lighting conditions	3.39	3.99	2.56	3.31			RGB
HPE-FF [135]	MB	Not req	BU	Test on 200 images of 5 subjects, uniform lighting conditions	3.41	3.90	2.32	3.21			RGB
CLM-Z [48]	MB	BU	BU	Unknown	3.00	3.81	**2.08**	2.97			RGB+D
OpenFace+PnP [18]	MB	Not req	BU	Test on all BU				2.6			RGB
HPE-MSF-CRFs [30]	SB	BU	BU	10-fold cross-validation	2.9	**2.1**	2.2	**2.4**			RGB
HAG-MSF-CRFs [32]	SB	BU	BU	10-fold cross-validation	2.9	**2.1**	2.2	**2.4**			RGB
Segm+CNN [80]	SB	BU	BU	10-fold cross-validation	**2.0**	2.4					RGB
MSE-MR [118]	ME	CAS-PEAL-1	CAS-PEAL-1	5-fold cross-validation	**2.3**	**1.0**					RGB
MSE-MR [118]	ME	CAS-PEAL-2	CAS-PEAL-2	5-fold cross-validation	**30.6**	**2.9**					RGB
MSE-MR [118]	ME	CMU-Pie	CMU-Pie	5-fold cross-validation		**1.9**					RGB
Cascade Trees [51]	ML	Dali3DHP	Dali3DHP	Leave-one-out cross-val	**7.69**	**4.73**		**6.23**			RGB+D
OpenFace+PnP [18]	MB	Not req	DD-Pose	Test on all DD-Pose	**4**	**4**	**5**	**9**		**16**	RGB
HeHop [20]	ML	DriveAHead	DriveAHead	First 5 subjects for test, other 15 for train						26.3	Depth

Table 7 (continued)

Name	Type	Train	Test	Evaluation pipeline	Pitch	Yaw	Roll	MAE	MAWE	BMAE	Data type
OpenFace+PnP [20]	MB	Not req	DriveAHead	Test on first 5 subjects						20.6	IR
HPN [20]	D	DriveAHead	DriveAHead	First 5 subjects for test, other 15 for train						**13.4**	IR+D
Meyer et al. [141]	MB	Not req	ETH	Test on all ETH	**2.3**	**2.9**		**2.6**			Depth
Liu et al. [114]	ME	FacePix	FacePix	Leave-one-out cross-val		3.1					RGB
Balasubramanian et al. [112]	ME	FicePix	FacePix	8-fold cross-validation		**1.4**					RGB

[a] Syntehtic Generated Data. Model type: (D) Deep learning regressor; (MB) Model based; (ME) Manifold embedding; (ML) Machine learning regressor; (MT) Multi-task; (RNN) Recurrent neural network; (SB) Segmentation based model

methods based only on monocular images have more generalization abilities and simpler deployment.

Issues and Problems

The main problem that emerges from this analysis is that different experimental set-ups and different validation protocols are adopted for HPE algorithms, and this strongly influences the evaluation, making comparison difficult. Another source of noise comes from the preprocessing phase, that may easily result in the detection of different bounding boxes/facial keypoints eventually influencing further elaboration steps.

Coming to more technical problems, Shao et al. [179] discovered in their experiments that bounding box margin has a large impact on the final accuracy of the model; head pose estimators are vulnerable to changes in the background scene around the target face, as shown in image 16.

To solve this problem Xue et al. [153] propose a convolutional cropping module (CCM) that can learn to crop the input image to an attentional area for head pose regression, and a background augmentation technique that can make the network more robust to the background noise. In their experiment SSR-Net-MD [88] MAE error fell from 6.01 to 5.38 and FSA-Net [88] goes from 5.25 to 5.13 thanks to CCM and background augmentation. If on one hand, this shows how there are techniques that allow to improve the results obtained, on the other, hand differences in the ways of getting the bounding boxes do not allow for a valid comparison of the methods for HPE.

The same problem emerged for face landmark detectors, as shown by Xin et al. [187] in their experiments, as reported in Table 9.

Also, the impact of image quality is little studied in the literature. When few low-quality images are present in training data, networks can easily fail to cope with these under-represented cases. Using synthesized LR samples and data augmentation during training is a delicate tradeoff between the positive gain deriving from more diverse training instances, and the additional difficulty related to the higher problem complexity. It is proven that when the resolution variation increases, the performance on the original High-Resolution (HR) samples drops [8]. Little studies have been conducted on establish a resolution-agnostic HPE framework [170].

The last question that arises is about the evaluation metrics used. MAE is the standard evaluation metric employed, but is optimal only for narrow range models, as explained in section "Evaluation Metrics". It's worth noting that also Cao et al. [76] criticise the use of MAE of Euler angles as evaluation metric, as according to them it cannot correctly measure the performance on profile images. They propose to use the Mean Absolute Error of Vectors (MAEV) to assess

Table 8 Evaluation results of head pose estimation on other databases

Name	Type	Train	Test	Evaluation pipeline	Pitch	Yaw	Roll	MAE	MAWE	BMAE	Data type
POSEidon [60]	D	ICT-3DHP	ICT-3DHP	Unknown	5.0	7.1	3.5	5.2			Depth
OpenFace+PnP [18]	MB	Not req	ICT-3DHP	Test on all ICT-3DHP				3.2			RGB
CLM-Z [48]	MB	ICT-3DHP	ICT-3DHP	Unknown	3.14	2.90	3.17	3.07			RGB+D
HPE-MSF-CRFs [30]	SB	ICT-3DHP	ICT-3DHP	10-fold cross-validation	3.2	**2.6**	**2.7**	**3.0**			RGB
HPE-MSF-CRFs [32]	SB	ICT-3DHP	ICT-3DHP	10-fold cross-validation	3.2	**2.6**	**2.7**	**3.0**			RGB
Segm+CNN [80]	SB	ICT-3DHP	ICT-3DHP	10-fold cross-validation	**2.3**	2.9					RGB
AVM [115]	ME	AFLW	McGill	14k random AFLW images as train, 6833 McGill images as test				**16.29**			RGB
PointNet [68]	MLP	MDM Corpus	MDM Corpus	39 subjects for train, 10 as validation,10 for test	**6.33**	**5.84**	**5.77**	**5.98**			Depth
Reg-CNN [181]	D	Multi-Pie	Multi-Pie	3-fold cross-validation		**0.02**					RGB
POSEidon [60]	D	Pandora	Pandora	Subjects 10, 14, 16, 20 for test, the other for training	5.7	**4.9**	9.0	6.53			Depth
KPM [85]	MB	Not req	Pandora	Test on all Pandora	**4.99**	6.33	**3.87**	**5.06**			RGB
Pixel-based segmentation [29]	SB	Pointing'04	Pointing'04	People 1-7 for train, people 8-15 for test		3.75					RGB
Super-pixel segmentation [29]	SB	Pointing'04	Pointing'04	People 1–7 for train, people 8-15 for test		5.69					RGB
Khan et al. [31]	SB	Pointing'04	Pointing'04	10-fold cross-validation		2.79					RGB
Hopenet [85]	MB	Pointing'04 (reannot.)	Pointing'04 (reannot.)	Train-test split unknown	19.59	26.61		23.10			RGB
FSA-Net [85]	MB	Pointing'04 (reannot.)	Pointing'04 (reannot.)	Train-test split unknown	18.01	25.90		21.96			RGB
LeNet-5 [148]	D	Pointing'04	Pointing'04	Leave-one-out cross-val	10.71	7.74		9.23			RGB
MSE-MR [118]	ME	Pointing'04	Pointing'04	5-fold cross-validation	9.6	8.1		8.85			RGB
4C_4S [130]	MB	Not req	Pointing'04	Test on all Pointing'04	6.34	10.63		8.48			RGB
3DDFA [85]	MB	Not req	Pointing'04 (reannot.)	Test on all Pointing'04	7.38	6.18		6.77			RGB
KPM [85]	MB	Not req	Pointing'04 (reannot.)	Test on all Pointing'04	5.27	4.30		4.78			RGB
DLDL (KL) [174]	D	Pointing'04	Pointing'04	5-fold cross-validation	1.69	3.16		2.43			RGB
HPE-MSF-CRFs [30]	SB	Pointing'04	Pointing'04	10-fold cross-validation	1.32	2.68		1.94			RGB
HAG-MSF-CRFs [32]	SB	Pointing'04	Pointing'04	10-fold cross-validation	1.18	2.32		1.75			RGB
Segm+CNN [80]	SB	Pointing'04	Pointing'04	10-fold cross-validation	1.02	2.02		1.52			RGB
Reg-CNN [181]	D	Pointing'04	Pointing'04	5-fold cross-validation	0.76	1.74		1.25			RGB
ARHPE+ [160]	D	Pointing'04	Pointing'04	Train-test split: 80%-20%	**0.69**	**1.59**		**1.20**			RGB

Table 8 (continued)

Name	Type	Train	Test	Evaluation pipeline	Pitch	Yaw	Roll	MAE	MAWE	BMAE	Data type
LMK [119]	ME	SASE	SASE	28 subjects for training, 12 subjects for testing	7.07	6.50	6.06	6.54			Depth
FSA-Net Caps-Fusion [159]	D	SGD[a]	SASE	Test on all SASE	7.76	6.52	5.61	6.63			RGB
DESC [119]	ME	SASE	SASE	28 subjects for training, 12 subjects for testing	**6.64**	6.21	4.60	5.82			Depth
FSA-Net Caps-Fusion [159]	D	300W-LP	SASE	Test on all SASE	7.27	5.77	3.72	5.59			RGB
FSA-Net Caps-Fusion [159]	D	SGD[a] + SASE	SASE	Random split SASE: 1k for train, other for test	7.13	**5.10**	**3.64**	**5.29**			RGB
Gu et al. [63]	RNN	SynHead	SynHead	8 subjects for training, 2 for testing	**1.55**	**1.78**	**1.66**	**1.66**			RGB
Liu et al. [58]	D	Synthetic	Synthetic	Random split: 30 seq. for train, 7 for test	**3.4**	**2.7**	**2.2**	**2.76**			RGB
MSE-MR [118]	ME	Taiwan	Taiwan	5-fold cross-validation		**5.8**					RGB
OpenFace+PnP [64]	MB	Not req	UbiPose	Unknown	4.45	9.49	**3.83**	6.28			RGB
HeadFusion [64]	MB	UbiPose	UbiPose	Unknown	**4.37**	**4.63**	**3.83**	**4.28**			RGB+D
WHENet [69]	D	300W-LP + CMU Panoptic	UET-Headpose-val						53.65		RGB
FSA-Net-Wide [69]	D	300W-LP	UET-Headpose-val						52.76		RGB
FSA-Net-Wide [69]	D	300W-LP + CMU Panoptic	UET-Headpose-val						52.72		RGB
FSA-Net-Wide [69]	D	UET-Headpose-train	UET-Headpose-val						9.30		RGB
FSA-Net-Wide [69]	D	300W-LP + CMU Panoptic + UET-Headpose-train	UET-HEadpose-val						**7.29**		RGB

aSyntehtic Generated Data

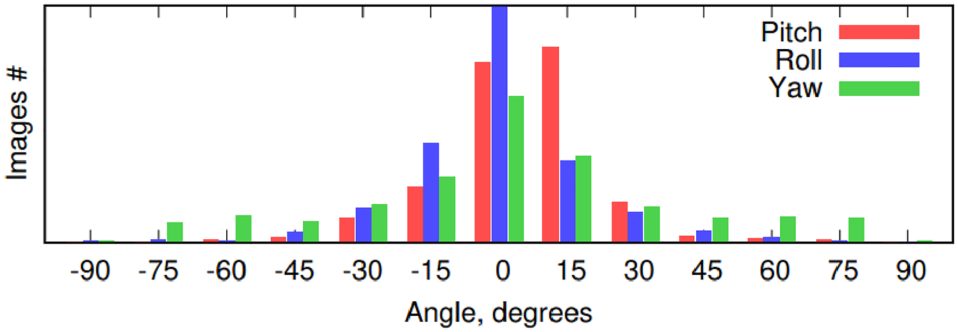


Fig. 14 Example of e distribution of the head rotation angles for the AFLW2000 dataset [53] (image from [195])

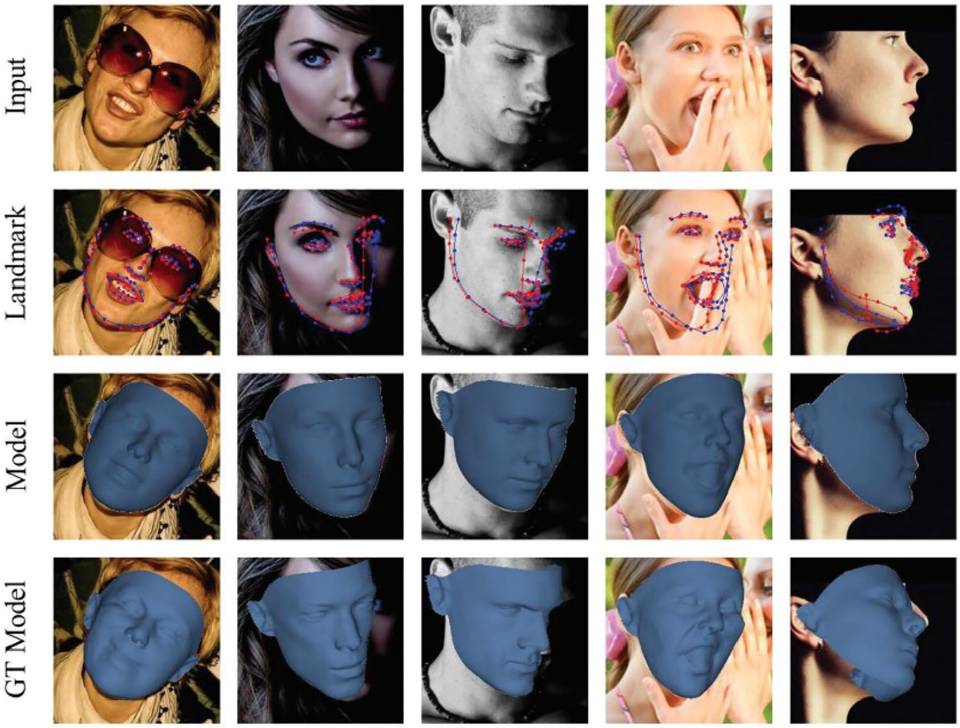


Fig. 15 Example of inaccuracies in ground-truth annotations on AFLW2000 dataset [53]. In some cases results from SADRNet [87] model are more accurate that the ground-truth. From the top row to the bottom row there are: the AFLW2000 [53] images, the sparse alignment results of SADRNet [87] and the corresponding ground-truth (blue for the former and red the latter), the reconstructed face models of SADRNet [87], and the ground-truth face models [87]. Vall et al. [98] reannotated AFLW2000 with poses estimated from correct landmarks and evaluated their MNN model, the MAE fell from 3.83 to 1.71 after the reannotation (image from [87])

the performance. They use three vectors, extracted from the rotation matrix, to describe head poses and compute the difference between the ground-truth vectors and the predicted ones. They showed how this representation is more consistent and how MAEV is a more reliable indicator for the evaluation of pose estimation results (see Fig. 17).

The MAWE metric (details in Section "Evaluation Metrics") could be a better choice: first, it can be used with Euler angles representation; second, if used to evaluate narrow range methods gives the same result as MAE; third, at this point narrow range methods have reached very high accuracy and it seems the time has come for a switch to full range methods with MAWE as main evaluation metric.

Research Directions

Due to the growing specialization of the field on ad-hoc contexts and tasks, it is natural to expect more and more investigation on topics like domain adaption, partial domain adaption, inaccurate semi-supervised learning, and knowledge transfer.

For similar reasons, we expect an increasing application of multi-task learning, which has seen a steady and strong development from 2017 to today. Head pose can be used as principal task to enhance other face-related subtasks, including gender classification, expression detection and identity recognition.

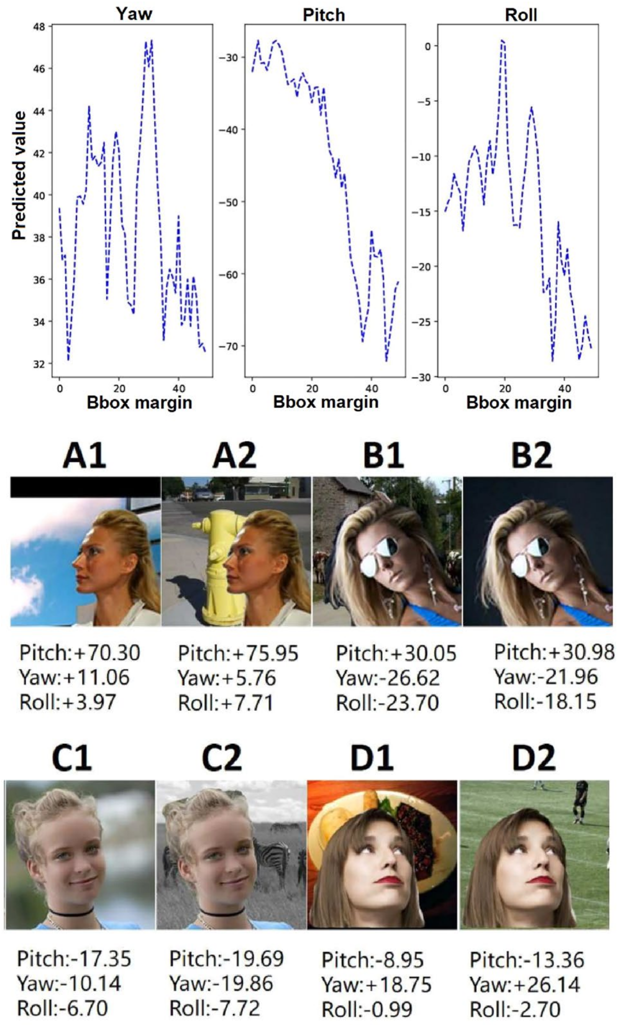


Fig. 16 Influence of bbox margin and background on head pose estimation: (a) Influence of bbox margin on head pose estimation. The values predicted by FSA-Net [88] change significantly with the change of bounding box size on all three axes. The network is not robust to the change of bbox margin; (b) Influence of background on head pose estimation. The values predicted by SSR-Net-MD [88] are not robust in different background, e.g. the offset of pitch and yaw between A1 and A2 is about 5° (images from [153])

Table 9 Influence of different landmark detectors for EVA-GCN performance

Landmark detector	Pitch	Yaw	Roll	MAE
EVA-GCN+OpenPose	5.52	7.25	4.78	5.85
EVA-GCN+Dlib	5.76	6.39	3.63	5.26
EVA-GCN+RetinaFace	5.33	5.02	4.26	4.87
EVA-GCN+FAN	5.34	4.96	4.11	4.64
EVA-GCN + GT*	4.15	3.23	3.05	3.48

GT* means ground-truth data (Table from [187])

For deformable models, an important improvement would be the ability to selectively ignore parts of the model that are self-occluded, overcoming a fundamental limitation in an otherwise very promising category, especially in unconstrained conditions.

Another interesting direction, not explored yet, is the use of deep learning in segmentation based methods. A possibility is to use convolutional neural networks to regress pose angles from segmented faces, or alternatively, segmentation based methods can be extended through geometric/deformable methods, where the feature extraction and classification could exploit specific deep learning architectures.

Finally, only Malakshan et al. [170] explored the use of generative models, showing that HPE can be effectively solved in conjunction with other face-related tasks typically associated with the generative field. This seems a very interesting possibility that showed promising result in another partially unexplored area of HPE task the extreme low-resolution images. We expect the development of a specific sub-filed that studies these techniques.

Although general head pose estimation will continue to be an exciting field with a lot of room for improvement, we expect an even stronger development of specific sub-fields that address thematic areas of application, such as the “security and surveillance” problem, recently addressed with the release of GOTCHA-I [66] database, or the “driver head pose estimation” which is already a very active field [16–20, 68]. Indeed, the role of head pose estimation in driving systems is becoming more and more important. By monitoring the head pose of the driver in real-time and analysing the behaviour of the driver, it will be possible to determine whether the driving status of the driver is good, having a profound impact on the future of automotive safety.

We expect new datasets will continue to be released with an increasing focus on 6 degrees of freedom and full range head angles, thanks to the development of new cheap and powerful RGB-D cameras (such as Microsoft Kinect), and other acquisition techniques.

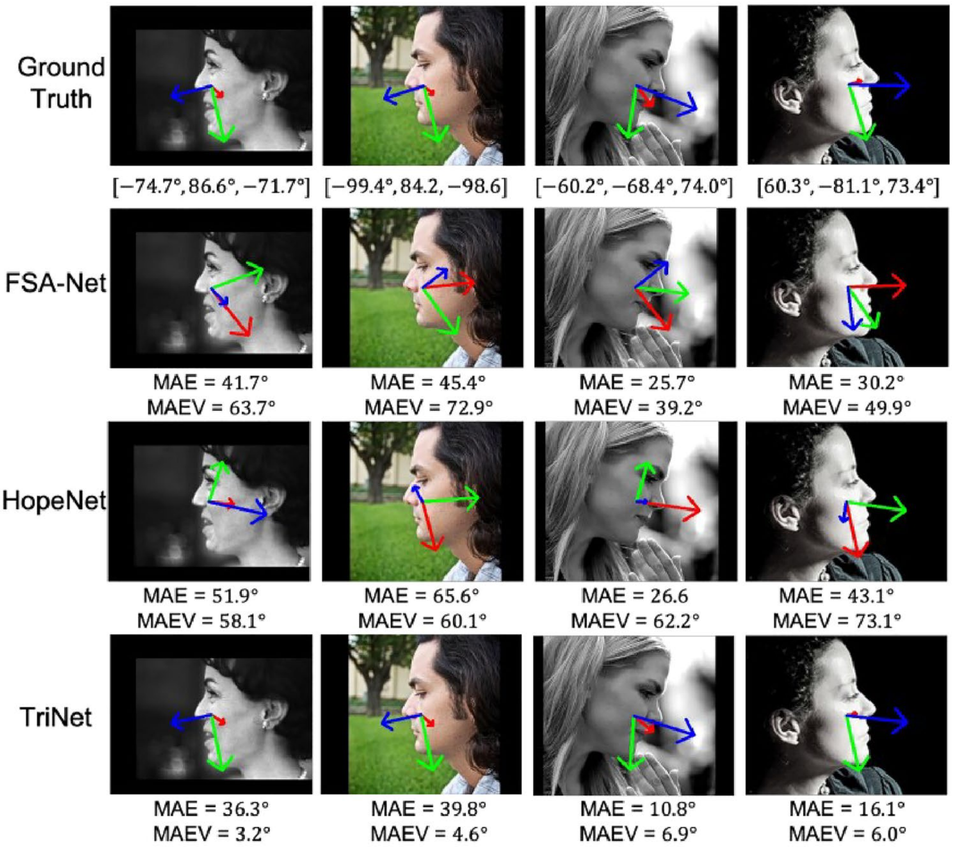
Conclusion

Head pose estimation is a very important task for human-computer interaction, since it provides rich information about the intent, motivation and visual attention of people.

Despite the extensive research in this field, especially during the last years, HPE still remains challenging when images are collected under unconstrained conditions.

In this article, we presented a detailed list of publicly available databases, and gave an in-depth survey of head pose estimation methods, briefly mentioning oldest and no more used classical approaches, and then providing an

Fig. 17 Comparison of pose estimation results with MAE and MAEV metrics on AFLW2000 profile images. All models are trained on 300W-LP (image from [76])



extensive analysis of modern techniques, mainly based on deep learning. Indeed, most current heads pose estimation methods exploit convolutional neural networks, from direct regressors to deformable based approaches passing through multi-task learning. We have also presented a comparative analysis of the state-of-the-art performance obtained so far in the field by providing organized and informative tables.

The article also discusses and suggests possible directions for future work. In particular, we expect the introduction of new light DL architectures that can perform well on challenging datasets, i.e., those collected in unconstrained environments.

We also expect the development of new sub-fields with dedicated databases and evaluation pipelines, such as the "driver head pose estimation" that is already very active.

An important trend observed is that the number of head pose publications has constantly increased in the past few years. This is a sign that more and more people are interested in this area, which means that the development cycle of new methods will be faster. A constant and periodic updating of the literature is therefore important.

We hope that this survey may help to clarify the evolution of the field, its evaluation methodologies and techniques thanks to the provided comprehensive list of datasets, methods and algorithms.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. Research partially supported by the Future AI Research (FAIR) project of the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 funded from the European Union-NextGenerationEU.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Murphy-Chutorian E, Trivedi MM. Head pose estimation in computer vision: a survey. IEEE Trans Pattern Anal Mach Intell. 2008;31(4):607–26.

2. Shao X, Qiang Z, Lin H, Dong Y, Wang X. A survey of head pose estimation methods. In: 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), IEEE, 2020; pp. 787–96.
3. Valenti R, Sebe N, Gevers T. Combining head pose and eye location information for gaze estimation. IEEE Trans Image Process. 2011;21(2):802–15.
4. Grinshpoon A, Sadri S, Loeb GJ, Elvezio C, Feiner SK. Hands-free interaction for augmented reality in vascular interventions. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), IEEE, 2018; p. 751–52.
5. Wang K, Zhao R, Ji Q. Human computer interaction with head pose, eye gaze and body gestures. In: 2018 13th IEEE International Conference on automatic face & Gesture recognition (FG 2018), IEEE, 2018; p. 789.
6. Munhall KG, Jones JA, Callan DE, Kuratate T, Vatikiotis-Bateson E. Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychol Sci. 2004;15(2):133–7.
7. Zhou Y, Gregson J. Whenet: Real-time fine-grained estimation for wide range head pose. 2020. arXiv preprint arXiv:2005.10353
8. Ruiz N, Chong E, Rehg JM. Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE Conference on computer vision and pattern recognition Workshops, 2018; p. 2074–83.
9. Morency L-P, Sidner C, Lee C, Darrell T. Head gestures for perceptual interfaces: the role of context in improving recognition. Artif Intell. 2007;171(8–9):568–85.
10. Langton SR, Bruce V. You must see the point: automatic processing of cues to the direction of social attention. J Exp Psychol Hum Percept Perform. 2000;26(2):747.
11. Ba SO, Odobez J-M. A study on visual focus of attention recognition from head pose in a meeting room. In: International Workshop on Machine Learning for Multimodal Interaction, Springer; 2006, p. 75–87.
12. Thomas C, Jayagopi DB. Predicting student engagement in classrooms using facial behavioral cues. In: Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education, 2017; p. 33–40.
13. Afroze S, Hoque MM. Classification of attentional focus based on head pose in multi-object scenario. In: International Conference on intelligent computing & optimization, Springer; 2019, p. 349–60.
14. Li D, Liu H, Chang W, Xu P, Luo Z. Visualization analysis of learning attention based on single-image pnp head pose estimation. In: 2017 2nd International Conference on Education, Sports, Arts and Management Engineering (ICESAME 2017), Atlantis Press; 2017, p. 1508–12.
15. Walter S, Gruss S, Ehleiter H, Tan J, Traue HC, Werner P, Al-Hamadi A, Crawcour S, Andrade AO, da Silva GM. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In: 2013 IEEE International Conference on Cybernetics (CYBCO), IEEE, 2013; p. 128–131.
16. Perdana MI, Anggraeni W, Sidharta HA, Yuniarno EM, Purnomo MH. Early warning pedestrian crossing intention from its head gesture using head pose estimation. In: 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA), IEEE, 2021; p. 402–7.
17. Ye M, Zhang W, Cao P, Liu K. Driver fatigue detection based on residual channel attention network and head pose estimation. Appl Sci. 2021;11(19):9195.
18. Roth M, Gavrila DM. Dd-pose-a large-scale driver head pose benchmark. In: 2019 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2019; p. 927–34.
19. Selim M, Firintepe A, Pagani A, Stricker D. Autopose: large-scale automotive driver head pose and gaze dataset with deep head orientation baseline. In: VISIGRAPP (4: VISAPP), 2020; p. 599–606.
20. Schwarz A, Haurilet M, Martinez M, Stiefelhagen R. Driveahead-a large-scale driver head pose dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017; p. 1–10.
21. Yamaura Y, Tsuboshita Y, Onishi T. Head pose estimation for an omnidirectional camera using a convolutional neural network. In: 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2018; p. 1–5.
22. Wang X. Intelligent multi-camera video surveillance: a review. Pattern Recogn Lett. 2013;34(1):3–19.
23. Benfold B, Reid I. Guiding visual surveillance by tracking human attention. In: BMVC, 2009; vol. 2, p. 7.
24. Sankaranarayanan K, Chang M-C, Krahnstoever N. Tracking gaze direction from far-field surveillance cameras. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), IEEE, 2011; p. 519–26.
25. Smith K, Ba SO, Odobez J-M, Gatica-Perez D. Tracking the visual focus of attention for a varying number of wandering people. IEEE Trans Pattern Anal Mach Intell. 2008;30(7):1212–29.
26. Wu S, Liang J, Ho J. Head pose estimation and its application in tv viewers' behavior analysis. In: 2016 IEEE Canadian Conference on electrical and computer engineering (CCECE), IEEE, 2016; p. 1–6.
27. Itoh TD, Kubo T, Ikeda K, Maruno Y, Ikutani Y, Hata H, Matsumoto K, Ikeda K. Towards generation of visual attention map for source code. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2019; p. 951–4.
28. Khan K, Khan RU, Leonardi R, Migliorati P, Benini S. Head pose estimation: a survey of the last ten years. Signal Process Image Commun. 2021;99: 116479.
29. Khan K, Mauro M, Migliorati P, Leonardi R. Head pose estimation through multi-class face segmentation. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2017; p. 175–180.
30. Khan K, Ahmad N, Khan F, Syed I. A framework for head pose estimation and face segmentation through conditional random fields. SIViP. 2020;14(1):159–66.
31. Benini S, Khan K, Leonardi R, Mauro M, Migliorati P. Face analysis through semantic face segmentation. Signal Process Image Commun. 2019;74:21–31.
32. Khan K, Attique M, Syed I, Sarwar G, Irfan MA, Khan RU. A unified framework for head pose, age and gender classification through end-to-end face segmentation. Entropy. 2019;21(7):647.
33. Neto ENA, Barreto RM, Duarte RM, Magalhaes JP, Bastos CA, Ren TI, Cavalcanti GD. Real-time head pose estimation for mobile devices. In: International Conference on intelligent data engineering and automated learning, Springer; 2012, p. 467–74.
34. La Cascia M, Sclaroff S, Athitsos V. Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3d models. IEEE Trans Pattern Anal Mach Intell. 2000;22(4):322–36.
35. Sim T, Baker S, Bsat M. The cmu pose, illumination, and expression (pie) database. In: Proceedings of Fifth IEEE International Conference on automatic face gesture recognition, IEEE, 2002; p. 53–8.
36. Ba SO, Odobez J-M. A video database for head pose tracking evaluation. Technical report, IDIAP; 2005.

37. Gao W, Cao B, Shan S, Chen X, Zhou D, Zhang X, Zhao D. The cas-peal large-scale Chinese face database and baseline evaluations. IEEE Trans Syst Man Cybern-Part A Syst Humans. 2007;38(1):149–61.
38. Gourier N, Hall D, Crowley JL. Estimating face orientation from robust detection of salient facial features. In: ICPR International Workshop on Visual Observation of Deictic Gestures, Citeseer; 2004.
39. Little G, Krishna S, Black J, Panchanathan S. A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. In: Proceedings.(ICASSP'05). IEEE International Conference on acoustics, speech, and signal processing, 2005, IEEE, 2005; vol. 2, p. 89.
40. Savran A, Alyüz N, Dibeklioğlu H, Çeliktutan O, Gökberk B, Sankur B, Akarun L. Bosphorus database for 3d face analysis. In: European Workshop on Biometrics and Identity Management, Springer; 2008, p. 47–56.
41. Breitenstein MD, Kuettel D, Weise T, Van Gool L, Pfister H. Real-time face pose estimation from single range images. In: 2008 IEEE Conference on computer vision and pattern recognition, IEEE, 2008; p. 1–8.
42. Yin B, Sun Y, Wang C, Ge Y. Bjut-3d large scale 3d face database and information processing. J Comput Res Dev. 2009;46(6):1009.
43. Chen J-C, Lien J-JJ. A view-based statistical system for multiview face detection and pose estimation. Elsevier; 2009. p. 1252–71.
44. Gross R, Matthews I, Cohn J, Kanade T, Baker S. Multi-pie. Image Vis Comput. 2010;28(5):807–13.
45. Koestinger M, Wohlhart P, Roth PM, Bischof H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011; p. 2144–2151.
46. Fanelli G, Weise T, Gall J, Gool LV. Real time head pose estimation from consumer depth cameras. In: Joint Pattern Recognition Symposium, Springer; 2011, p. 101–10.
47. Zhu X, Ramanan D. Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012; p. 2879–86.
48. Baltrušaitis T, Robinson P, Morency L-P. 3d constrained local model for rigid and non-rigid facial tracking. In: 2012 IEEE Conference on computer vision and pattern recognition, IEEE, 2012; p. 2610–17.
49. Smith BA, Yin Q, Feiner SK, Nayar SK. Gaze locking: passive eye contact detection for human-object interaction. In: Proceedings of the 26th Annual ACM Symposium on user interface software and technology, 2013, p. 271–80.
50. Demirkus M, Clark JJ, Arbel T. Robust semi-automatic head pose labeling for real-world face video sequences. Multimed Tools Appl. 2014;70(1):495–523.
51. Tulyakov S, Vieriu R-L, Semeniuta S, Sebe N. Robust real-time extreme head pose estimation. In: 2014 22nd International Conference on pattern recognition, IEEE, 2014; p. 2263–68.
52. Zhang Z, Luo P, Loy CC, Tang X. Facial landmark detection by deep multi-task learning. In: European Conference on computer vision, Springer; 2014, p. 94–108.
53. Zhu X, Lei Z, Liu X, Shi H, Li SZ. Face alignment across large poses: a 3d solution. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2016; p. 146–155.
54. Dewantara BSB, Miura J. The aisl head orientation database and preliminary evaluations. In: 2015 International Electronics Symposium (IES), IEEE, 2015; p. 140–4.
55. Joo H, Liu H, Tan L, Gui L, Nabbe B, Matthews I, Kanade T, Nobuhara S, Sheikh Y. Panoptic studio: a massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on computer vision, 2015; p. 3334–42.
56. Liu Y, Chen J, Su Z, Luo Z, Luo N, Liu L, Zhang K. Robust head pose estimation using Dirichlet-tree distribution enhanced random forests. Neurocomputing. 2016;173:42–53.
57. Ariz M, Bengoechea JJ, Villanueva A, Cabeza R. A novel 2d/3d database with automatic face annotation for head tracking and pose estimation. Comput Vis Image Underst. 2016;148:201–10.
58. Liu X, Liang W, Wang Y, Li S, Pei M. 3d head pose estimation with convolutional neural network trained on synthetic images. In: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016; p. 1289–93.
59. Bansal A, Nanduri A, Castillo CD, Ranjan R, Chellappa R. Umdfaces: an annotated face dataset for training deep networks. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2017; p. 464–73.
60. Borghi G, Venturelli M, Vezzani R, Cucchiara R. Poseidon: Face-from-depth for driver pose estimation. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017; p. 4661–70.
61. Lüsi I, Junior JCJ, Gorbova J, Baró X, Escalera S, Demirel H, Allik J, Ozcinar C, Anbarjafari G. Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases. In: 2017 12th IEEE International Conference on automatic face & gesture recognition (FG 2017), IEEE, 2017; p. 809–13.
62. Werner P, Saxen F, Al-Hamadi A. Landmark based head pose estimation benchmark and method. In: 2017 IEEE International Conference on image processing (ICIP), IEEE, 2017; p. 3909–13.
63. Gu J, Yang X, De Mello S, Kautz J. Dynamic facial analysis: from Bayesian filtering to recurrent neural network. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017; p. 1548–57.
64. Yu Y, Mora KAF, Odobez J-M. Headfusion: 360° head pose tracking combining 3d morphable model and 3d reconstruction. IEEE Trans Pattern Anal Mach Intell. 2018;40(11):2653–67.
65. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A. Vggface2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on automatic face & gesture recognition (FG 2018), IEEE, 2018; p. 67–74.
66. Barra P, Bisogni C, Nappi M, Freire-Obregón D, Castrillón-Santana M. Gotcha-i: a multiview human videos dataset. In: International Symposium on Security in computing and communication, Springer; 2019, p. 213–24.
67. Li P, Wu X, Hu Y, He R, Sun Z. M2fpa: a multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis. In: Proceedings of the IEEE/CVF International Conference on computer vision, 2019; p. 10043–10051.
68. Jha S, Marzban MF, Hu T, Mahmoud MH, Al-Dhahir N, Busso C. The multimodal driver monitoring database: A naturalistic corpus to study driver attention. IEEE Trans Intell Transport Syst. 2021;23:10736–52.
69. Viet LN, Dinh TN, Minh DT, Viet HN, Tran QL. Uet-headpose: a sensor-based top-view head pose dataset. In: 2021 13th International Conference on knowledge and systems engineering (KSE), IEEE, 2021; p. 1–7.
70. Martyniuk T, Kupyn O, Kurlyak Y, Krashenyi I, Matas J, Sharmanska V. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2022; p. 20942–20952.
71. Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 faces in-the-wild challenge: The first facial landmark localization

challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013; p. 397–403.

72. DeMenthon DF, Davis LS. Model-based object pose in 25 lines of code. Int J Comput Vis. 1995;15(1):123–41.
73. Amador E, Valle R, Buenaposada JM, Baumela L. Benchmarking head pose estimation in-the-wild. In: Iberoamerican Congress on Pattern Recognition, Springer; 2017, p. 45–52.
74. Drouard V, Horaud R, Deleforge A, Ba S, Evangelidis G. Robust head-pose estimation based on partially-latent mixture of linear regressions. IEEE Trans Image Process. 2017;26(3):1428–40.
75. Ranjan R, Sankaranarayanan S, Castillo CD, Chellappa R. An all-in-one convolutional neural network for face analysis. In: 2017 12th IEEE International Conference on automatic face & gesture recognition (FG 2017), IEEE, 2017; p. 17–24.
76. Cao Z, Chu Z, Liu D, Chen Y. A vector-based representation to enhance head pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision, 2021; p. 1188–1197.
77. Bai J, Peng C, Li Z, Du S, Li Y. A study of general data improvement for large-angle head pose estimation. In: International Conference on computer analysis of images and patterns, Springer; 2017, p. 199–209.
78. Euler angles. Wikimedia Foundation. Accessed: January 2022.
79. Kostyaev D. Better rotation representations for accurate pose estimation. Towards Data Science; Accessed: December 2021.
80. Khan K, Ali J, Ahmad K, Gul A, Sarwar G, Khan S, Thanh Hoai Ta Q, Chung T, Attique M. 3d head pose estimation through facial features and deep convolutional neural networks. Comput Mater Contin. 2021;66(2):1757–70.
81. Gupta A, Thakkar K, Gandhi V, Narayanan P. 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), IEEE, 2019; p. 1977–1981.
82. Xia J, Cao L, Zhang G, Liao J. Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks. Ieee Access. 2019;7:48470–83.
83. Dapogny A, Bailly K, Cord M. Deep entwined learning head pose and face alignment inside an attentional cascade with doubly-conditional fusion. In: 2020 15th IEEE International Conference on automatic face and gesture recognition (FG 2020), IEEE, 2020; p. 192–8.
84. Wu Y, Gou C, Ji Q. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017; p. 3471–80.
85. Liu L, Ke Z, Huo J, Chen J. Head pose estimation through keypoints matching between reconstructed 3d face model and 2d image. Sensors. 2021;21(5):1841.
86. Wu C-Y, Xu Q, Neumann U. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In: 2021 International Conference on 3D Vision (3DV), IEEE, 2021; p. 453–463.
87. Ruan Z, Zou C, Wu L, Wu G, Wang L. Sadrnet: self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. IEEE Trans Image Process. 2021;30:5793–806.
88. Yang T-Y, Chen Y-T, Lin Y-Y, Chuang Y-Y. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2019; p. 1087–1096.
89. Hsu H-W, Wu T-Y, Wan S, Wong WH, Lee C-Y. Quatnet: quaternion-based head pose estimation with multiregression loss. IEEE Trans Multimed. 2018;21(4):1035–46.
90. Dai D, Wong W, Chen Z. Rankpose: learning generalised feature with rank supervision for head pose estimation. 2020. arXiv preprint arXiv:2005.10984
91. Sheka A, Samun V. Knowledge distillation from ensemble of offsets for head pose estimation. 2021. arXiv preprint arXiv:2108.09183
92. Ranjan R, Patel VM, Chellappa R. Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans Pattern Anal Mach Intell. 2017;41(1):121–35.
93. Xu X, Kakadiaris IA. 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), IEEE, 2017; p. 642–649.
94. Kumar A, Alavi A, Chellappa R. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In: 2017 12th Ieee International Conference on automatic face & gesture recognition (fg 2017), IEEE, 2017; p. 258–265.
95. Chen J-C, Lin W-A, Zheng J, Chellappa R. A real-time multi-task single shot face detector. In: 2018 25th IEEE International Conference on image processing (ICIP), IEEE, 2018; p. 176–180.
96. Cai Z, Liu Q, Wang S, Yang B. Joint head pose estimation with multi-task cascaded convolutional networks for face alignment. In: 2018 24th International Conference on pattern recognition (ICPR), IEEE, 2018; p. 495–500.
97. Wu H, Zhang K, Tian G. Simultaneous face detection and pose estimation using convolutional neural network cascade. IEEE Access. 2018;6:49563–75.
98. Valle R, Buenaposada JM, Baumela L. Multi-task head pose estimation in-the-wild. IEEE Trans Pattern Anal Mach Intell. 2020;43(8):2874–81.
99. Xia J, Zhang H, Wen S, Yang S, Xu M. An efficient multitask neural network for face alignment, head pose estimation and face tracking. 2021. arXiv preprint arXiv:2103.07615
100. Fard AP, Abdollahi H, Mahoor M. Asmnet: a lightweight deep neural network for face alignment and pose estimation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2021. p. 1521–30.
101. Liu Y, Gu Z, Gao S, Wang D, Zeng Y, Cheng J. Mos: a low latency and lightweight framework for face detection, landmark localization, and head pose estimation. 2021. arXiv preprint arXiv:2110.10953
102. Viet HN, Viet LN, Dinh TN, Minh DT, Quac LT. Simultaneous face detection and 360 degree head pose estimation. In: 2021 13th International Conference on knowledge and systems engineering (KSE), IEEE, 2021; p. 1–7.
103. Chen S, Zhang Y, Yin B, Wang B. Trfh: towards real-time face detection and head pose estimation. Pattern Anal Appl. 2021;24(4):1745–55.
104. Czupryński B, Strupczewski A. High accuracy head pose tracking survey. In: International Conference on active media technology, Springer, 2014; p. 407–20.
105. Ng J, Gong S. Multi-view face detection and pose estimation using a composite support vector machine across the view sphere. In: Proceedings International Workshop on recognition, analysis, and tracking of faces and gestures in real-time systems. In: Conjunction with ICCV'99 (Cat. No. PR00378), IEEE, 1999; p. 14–21.
106. Ng J, Gong S. Composite support vector machines for detection of faces across views and pose estimation. Image Vis Comput. 2002;20(5–6):359–68.
107. Huang J, Shao X, Wechsler H. Face pose discrimination using support vector machines (svm). In: Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170), IEEE, 1998; vol. 1, p. 154–56.
108. Zhang Z, Hu Y, Liu M, Huang T. Head pose estimation in seminar room using multi view face detectors. In: International Evaluation Workshop on Classification of Events, Activities and Relationships, Springer; 2006, p. 299–304.

109. Viola P, Jones MJ. Robust real-time face detection. Int J Comput Vis. 2004;57(2):137–54.
110. McKenna SJ, Gong S. Real-time face pose estimation. Real-Time Imaging. 1998;4(5):333–47.
111. Raytchev B, Yoda I, Sakaue K. Head pose estimation by nonlinear manifold learning. In: Proceedings of the 17th International Conference on pattern recognition, 2004. ICPR 2004., IEEE; 2004. vol. 4, pp. 462–66.
112. Balasubramanian VN, Ye J, Panchanathan S. Biased manifold embedding: A framework for person-independent head pose estimation. In: 2007 IEEE Conference on computer vision and pattern recognition, IEEE, 2007; p. 1–7.
113. Huang D, Storer M, De la Torre F, Bischof H. Supervised local subspace learning for continuous head pose estimation. In: CVPR 2011, IEEE; 2011. p. 2921–2928.
114. Liu X, Lu H, Li W. Multi-manifold modeling for head pose estimation. In: 2010 IEEE International Conference on image processing, IEEE, 2010; p. 3277–80.
115. Sundararajan K, Woodard DL. Head pose estimation in the wild using approximate view manifolds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015; p. 50–58.
116. Wang C, Guo Y, Song X. Head pose estimation via manifold learning. Manifolds-Current Research Areas; 2017.
117. Peng X, Huang J, Hu Q, Zhang S, Metaxas DN. Head pose estimation by instance parameterization. In: 2014 22nd International Conference on pattern recognition, IEEE, 2014; p. 1800–1805.
118. Diaz-Chito K, Del Rincon JM, Hernández-Sabaté A, Gil D. Continuous head pose estimation using manifold subspace embedding and multivariate regression. IEEE Access. 2018;6:18325–34.
119. Derkach D, Ruiz A, Sukno FM. Tensor decomposition and nonlinear manifold modeling for 3d head pose estimation. Int J Comput Vis. 2019;127(10):1565–85.
120. Morency L-P, Rahimi A, Darrell T. Adaptive view-based appearance models. In: 2003 IEEE Computer Society Conference on computer vision and pattern recognition, 2003. Proceedings., IEEE, 2003; vol. 1.
121. Yao P, Evans G, Calway A. Using affine correspondence to estimate 3-d facial pose. In: Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), IEEE, 2001; vol. 3, p. 919–922.
122. Ohayon S, Rivlin E. Robust 3d head tracking using camera pose estimation. In: 18th International Conference on Pattern Recognition (ICPR'06), IEEE, 2006; vol. 1, p. 1063–1066.
123. Lu L, Zhang Z, Shum H-Y, Liu Z, Chen H. Model and exemplar-based robust head pose tracking under occlusion and varying expression. In: Proc. of CVPR. 2001.
124. Malciu M, Prêteux F. A robust model-based approach for 3d head tracking in video sequences. In: Proceedings Fourth IEEE International Conference on automatic face and gesture recognition (Cat. No. PR00580), IEEE, 2000; p. 169–174.
125. Huang GB, Narayana M, Learned-Miller E. Towards unconstrained face recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2008; p. 1–8.
126. Chen P, Xiao Q, Xu J, Dong X, Sun L. Facial attribute editing using semantic segmentation. 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD &IS), 2019. p. 97–103.
127. Lee C-H, Liu Z, Wu L, Luo P. Maskgan: towards diverse and interactive facial image manipulation. In: IEEE Conference on computer vision and pattern recognition (CVPR). 2020.
128. Yang H, Mou W, Zhang Y, Patras I, Gunes H, Robinson P. Face alignment assisted by head pose estimation. 2015. arXiv preprint arXiv:1507.03148.
129. Abate AF, Barra P, Bisogni C, Nappi M, Ricciardi S. Near real-time three axis head pose estimation without training. IEEE Access. 2019;7:64256–65.
130. Barra P, Barra S, Bisogni C, De Marsico M, Nappi M. Web-shaped model for head pose estimation: an approach for best exemplar selection. IEEE Trans Image Process. 2020;29:5457–68.
131. Hesch JA, Roumeliotis SI. A direct least-squares (dls) method for pnp. In: 2011 International Conference on Computer Vision, IEEE, 2011, p. 383–90.
132. Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2014; p. 1867–74.
133. Bulat A, Tzimiropoulos G. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on computer vision, 2017; p. 1021–30.
134. Hui X. A survey for 2d and 3d face alignment. In: 2019 International Conference on machine learning, big data and business intelligence (MLBDBI), IEEE, 2019; p. 57–63.
135. Barros JMD, Mirbach B, Garcia F, Varanasi K, Stricker D. Fusion of keypoint tracking and facial landmark detection for real-time head pose estimation. In: 2018 IEEE Winter Conference on applications of computer vision (WACV), IEEE, 2018; p. 2028–2037.
136. Zhu X, Liu X, Lei Z, Li SZ. Face alignment in full pose range: a 3d total solution. IEEE Trans Pattern Anal Mach Intell. 2017;41(1):78–92.
137. Tu X, Zhao J, Xie M, Jiang Z, Balamurugan A, Luo Y, Zhao Y, He L, Ma Z, Feng J. 3d face reconstruction from a single image assisted by 2d face images in the wild. IEEE Trans Multimed. 2020;23:1160–72.
138. Guo J, Zhu X, Yang Y, Yang F, Lei Z, Li SZ. Towards fast, accurate and stable 3d dense face alignment. In: European Conference on computer vision, Springer; 2020, p. 152–168.
139. Besl PJ, McKay ND. A Method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Intell. 1992;14:239–56.
140. Martin M, Van De Camp F, Stiefelhagen R. Real time head model creation and head pose estimation on consumer depth cameras. In: 2014 2nd International Conference on 3D vision, IEEE, 2014; vol. 1, p. 641–48.
141. Meyer GP, Gupta S, Frosio I, Reddy D, Kautz J. Robust model-based 3d head pose estimation. In: Proceedings of the IEEE International Conference on computer vision, 2015; p. 3649–3657.
142. Murphy-Chutorian E, Doshi A, Trivedi MM. Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In: 2007 IEEE Intelligent Transportation Systems Conference, IEEE, 2007; p. 709–714.
143. Ahn B, Park J, Kweon IS. Real-time head orientation from a monocular camera using deep neural network. In: Asian Conference on computer vision, Springer; 2014, p. 82–96.
144. Liu X. Head pose estimation using convolutional neural networks. 2016.
145. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012;25.
146. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint arXiv:1409.1556.
147. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2016; p. 770–78.

148. Patacchiola M, Cangelosi A. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. Pattern Recogn. 2017;71:132–43.
149. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, PMLR; 2019, p. 6105–114.
150. Zeng Z, Zhu D, Zhang G, Shi W, Wang L, Zhang X, Li J. Srnet: structural relation-aware network for head pose estimation. In: 2022 26th International Conference on pattern recognition (ICPR), IEEE, 2022; p. 826–32.
151. Hempel T, Abdelrahman AA, Al-Hamadi A. 6d rotation representation for unconstrained head pose estimation. In: 2022 IEEE International Conference on image processing (ICIP), IEEE, 2022; p. 2496–2500.
152. Lathuilière S, Juge R, Mesejo P, Munoz-Salinas R, Horaud R. Deep mixture of linear inverse regressions applied to head-pose estimation. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017; p. 4817–825.
153. Xue A, Sheng K, Dai S, Li X. Robust landmark-free head pose estimation by learning to crop and background augmentation. IET Image Proc. 2020;14(11):2553–60.
154. Wang B-Y, Xie K, He S-T, Wen C, He J-B. Head pose estimation in complex environment based on four-branch feature selective extraction and regional information exchange fusion network. IEEE Access. 2022;10:41287–302.
155. Berral-Soler R, Madrid-Cuevas FJ, Munoz-Salinas R, Marín-Jiménez MJ. Realheponet: a robust single-stage convnet for head pose estimation in the wild. Neural Comput Appl. 2021;33(13):7673–89.
156. Dhingra N. Lwposr: lightweight efficient fine grained head pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision, 2022; p. 1495–1505.
157. Wang Y, Liang W, Shen J, Jia Y, Yu L-F. A deep coarse-to-fine network for head pose estimation from synthetic data. Pattern Recogn. 2019;94:196–206.
158. Kuhnke F, Ostermann J. Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In: Proceedings of the IEEE/CVF International Conference on computer vision, 2019; p. 10164–173.
159. Basak S, Corcoran P, Khan F, Mcdonnell R, Schukat M. Learning 3d head pose from synthetic data: a semi-supervised approach. IEEE Access. 2021;9:37557–73.
160. Liu H, Liu T, Zhang Z, Sangaiah AK, Yang B, Li Y. Arhpe: asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction. IEEE Trans Ind Inf. 2022;18(10):7107–17.
161. Saeed A, Al-Hamadi A. Boosted human head pose estimation using kinect camera. In: 2015 IEEE International Conference on image processing (ICIP), IEEE, 2015; p. 1752–1756.
162. Yan Y, Ricci E, Subramanian R, Lanz O, Sebe N. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In: Proceedings of the IEEE International Conference on computer vision, 2013; p. 1177–1184.
163. Yan Y, Subramanian R, Ricci E, Lanz O, Sebe N. Evaluating multi-task learning for multi-view head-pose classification in interactive environments. In: 2014 22nd International Conference on pattern recognition, IEEE, 2014; p. 4182–87.
164. Hong C, Yu J, Zhang J, Jin X, Lee K-H. Multimodal face-pose estimation with multitask manifold deep learning. IEEE Trans Ind Inf. 2018;15(7):3952–61.
165. Ahn B, Choi D-G, Park J, Kweon IS. Real-time head pose estimation using multi-task deep neural network. Robot Auton Syst. 2018;103:1–12.
166. Albiero V, Chen X, Yin X, Pang G, Hassner T. img2pose: face alignment and detection via 6dof, face pose estimation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2021; p. 7617–7627.
167. Zhang C, Hu X, Xie Y, Gong M, Yu B. A privacy-preserving multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. Front. Neurorobot. 2020;112.
168. Ewaisha M, Shawarby ME, Abbas H, Sobh I. End-to-end multitask learning for driver gaze and head pose estimation. Electron Imaging. 2020;2020(16):110–1.
169. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models-their training and application. Comput Vis Image Underst. 1995;61(1):38–59.
170. Malakshan SR, Mostofa M, Soleymani S, Nasrabadi NM, et al. Joint super-resolution and head pose estimation for extreme low-resolution faces. IEEE Access. 2023;11:11238–53.
171. Drouard V, Ba S, Evangelidis G, Deleforge A, Horaud R. Head pose estimation via probabilistic high-dimensional regression. In: 2015 IEEE International Conference on image processing (ICIP), IEEE, 2015; p. 4624–28.
172. Papazov C, Marks TK, Jones M. Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2015; p. 4722–30.
173. Barros JMD, Garcia F, Mirbach B, Stricker D. Real-time monocular 6-dof head pose estimation from salient 2d points. In: 2017 IEEE International Conference on image processing (ICIP), IEEE, 2017; p. 121–5.
174. Gao B-B, Xing C, Xie C-W, Wu J, Geng X. Deep label distribution learning with label ambiguity. IEEE Trans Image Process. 2017;26(6):2825–38.
175. Gou C, Wu Y, Wang F-Y, Ji Q. Coupled cascade regression for simultaneous facial landmark detection and head pose estimation. In: 2017 IEEE International Conference on image processing (ICIP), IEEE, 2017; p. 2906–10.
176. Yu Y, Mora KAF, Odobez J-M. 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), Ieee; 2017; p. 711–18.
177. Zhang W, Zhang H, Li Q, Liu F, Sun Z, Li X, Wan X. Cross-cascading regression for simultaneous head pose estimation and facial landmark detection. In: Chinese Conference on biometric recognition, Springer; 2018, p. 148–156.
178. Liu Z, Chen Z, Bai J, Li S, Lian S. Facial pose estimation by deep learning from label distributions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
179. Shao M, Sun Z, Ozay M, Okatani T. 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), IEEE, 2019; p. 1–5.
180. Wang H, Chen Z, Zhou Y. Hybrid coarse-fine classification for head pose estimation. 2019. arXiv preprint arXiv:1901.06778.
181. Xu L, Chen J, Gan Y. Head pose estimation with soft labels using regularized convolutional neural network. Neurocomputing. 2019;337:339–53.
182. Wang W, Chen X, Zheng S, Li H. Fast head pose estimation via rotation-adaptive facial landmark detection for video edge computation. IEEE Access. 2020;8:45023–32.
183. Zhang H, Wang M, Liu Y, Yuan Y. Fdn: Feature decoupling network for head pose estimation. In: Proceedings of the AAAI Conference on artificial intelligence, 2020; vol. 34, p. 12789–796.
184. Berg A, Oskarsson M, O'Connor M. Deep ordinal regression with label diversity. In: 2020 25th International Conference on pattern recognition (ICPR), IEEE, 2021; p. 2740–47.

185. Hu Z, Xing Y, Lv C, Hang P, Liu J. Deep convolutional neural network-based Bernoulli heatmap for head pose estimation. Neurocomputing. 2021;436:198–209.
186. Dhingra N. Headposr: end-to-end trainable head pose estimation using transformer encoders. In: 2021 16th IEEE International Conference on automatic face and gesture recognition (FG 2021), IEEE, 2021; p. 1–8.
187. Xin M, Mo S, Lin Y. Eva-gcn: head pose estimation based on graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2021; p. 1462–71.
188. Cantarini G, Tomenotti FF, Noceti N, Odone F. Hhp-net: a light heteroscedastic neural network for head pose estimation with uncertainty. In: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision, 2022; p. 3521–30.
189. Yang S, Luo P, Loy C-C, Tang X. Wider face: a face detection benchmark. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2016; p. 5525–33.
190. Jiang H, Learned-Miller E. Face detection with the faster r-cnn. In: 2017 12th IEEE International Conference on automatic face & gesture recognition (FG 2017), IEEE, 2017; p. 650–57.
191. Chen D, Ren S, Wei Y, Cao X, Sun J. Joint cascade face detection and alignment. In: European Conference on Computer Vision, Springer; 2014, p. 109–22.
192. Cao Z, Simon T, Wei S-E, Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017; p. 7291–99.
193. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2016, p. 779–88.
194. Hu P, Ramanan D. Finding tiny faces. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017; p. 951–59.
195. Sheka A, Samun V. Rotation augmentation for head pose estimation problem. In: 2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), IEEE, 2021; p. 0308–0311.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.