UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Study on 3D Head Understanding Benchmarks: spectrum of tasks, existing methods and challenges

Author:
Yana KURLIAK

Supervisor:
Tetiana MARTYNIUK

*A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science*

in the

Department of Computer Sciences
Faculty of Applied Sciences

УКРАЇНСЬКИЙ КАТОЛИЦЬКИЙ УНІВЕРСИТЕТ

APPLIED
SCIENCES
FACULTY

Lviv 2022

Declaration of Authorship

I, Yana KURLIAK, declare that this thesis titled, “Study on 3D Head Understanding Benchmarks: spectrum of tasks, existing methods and challenges” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Study on 3D Head Understanding Benchmarks: spectrum of tasks, existing methods and challenges

by Yana KURLIAK

Abstract

The face plays a key role in communication of both verbal and non-verbal information in human interaction. A correct 3D head representation helps computer visual system identify key facial features to serve a large range of use cases both in and beyond Computer Vision field. This work includes an overview and analysis of existing 3D Head Understanding area datasets and benchmarking methods. Based on the analysis of their benefits and limitations a novel set of data and metrics are concluded.

Acknowledgements

I want to thank the Armed Forces of Ukraine for providing security to perform this work.

I am grateful to my supervisor Tetiana Martyniuk and also Orest Kupyn and Igor Krashenyi for fun and supportive work on the DAD-3DHeads project together.

Contents

List of Figures

List of Tables

List of Abbreviations

PCA	**P**rincipal **C**omponent **A**nalysis
3DMM	**3D** **M**orphable **M**odel
CNN	**C**onvolutional **N**eural **N**etwork

Dedicated to people who live in a three-dimensional space

Chapter 1

Introduction

The face plays a key role in communication of both verbal and non-verbal information in human interaction. By looking at a face we can read such non-verbal details like emotional state of a human, their identity and intentions.

The initial research question behind the area of 3D Head Understanding is how a visual system, of a human or a computer can extract a detailed information about the object based on its appearance and to cope with the high variety of of resemblant image data this object can generate.

The area consists of such head and facial analysis subtasks as Facial Landmark Localisation, Head Pose Estimation, Facial Alignment in 2D [SOURCES], and finally 3D Face or Head Modeling and Reconstruction.

Why 3D?

Due to the 3D nature of the face, a 2D image is insufficient to accurately capture its geometry, as it collapses one dimension. Furthermore, 3D imaging provides a representation of the facial geometry that is invariant to pose and illumination, which are two of the major inconveniences of 2D imaging.[Morales, Piella, and Sukno, 2021]

Building a precise 3D head model is a center topic of recent research as it solves all of the listed subtasks as facial landmarks, head pose and 2D image projection is easily retrievable from the set of accurate 3D coordinates.

Increasingly powerful methods for precise 3D head model building from casually captured imagery have an abundant scope of use cases in real world. The scope includes: building of highly believable human avatars for further application in game development and virtual reality; new face imagery synthesis and modification; video content creation and animation; advanced face recognition and face detection algorithms development; semantic information extraction for medical observation of emotional and general health state of a subject.

Challenges

Such rich extent of 3D Head models application requires not only precise 3D geometry but also raises a question of how scalable existing solutions are in terms of handling age, gender, and ethnicity diversity and of how generalisable they are beyond the data they were trained on.

To cope with these challenges the research community has made a tremendous effort in collecting and annotating large sets of image data and proposed various benchmarks with metrics for efficient approaches development.

Nevertheless, 3D head analysis tasks from a single image in the wild remain an open challenge. The difficulty comes from (1) lack of 2D-3D ground-truth data and, as a result, (2) ambiguity of the task and reliance on 3D shape priors. Many methods have been developed to fill the gap of missing 2D-3D annotations (1), primarily using 2D landmarks datasets for fitting, or exploring extra knowledge such as identity invariance [Sanyal et al., 2019], or co-training with related face detection [Deng et al., 2020], [Chang et al., 2017] tasks to drive the recovery of 3D face geometry. Up until now, evaluation of the efficiency of these approaches has been problematic due to the lack of ground-truth data.[Martyniuk et al., 2022]

Another challenge for the research community is to agree on common standards of annotated data collections and evaluation methods, which would simplify an efficient development of new approaches, and enable us to better test and compare them.

Goal

The goal of this work is to perform a detailed analytical study of existing data and benchmarks for the task of 3D Head Alignment and other related human head analysis tasks. Discuss their advantages and disadvantages in relevance to the challenges stated above, and to present a novel benchmark for 3D head understanding based on this analysis.

Chapter 2

Background

2.1 History

Representation and recognition of human face features has been in the focus of computer vision research for a long time. The first solid contribution to the field of human face understanding was the Eigenfaces approach [Sirovich and Kirby, 1987, Turk and Pentland, 1991].

2.1.1 Eigenfaces

Eigenfaces approach was developed to learn an explicit face representation from the set of samples. Having the distribution of face images treated as a vector space, the aim was to find the principal component of the distribution with the eigenvectors performing as the main modes of variation in that space. Such approach operated on gray-scale images and was not only restricted to certain pose and illumination, but also did not cope with shape differences: represented as a linear combination of eigenfaces, continuous face shape change would look like fading structures rather than a shape shift on an image plane. An example is shown on Figure 2.1.

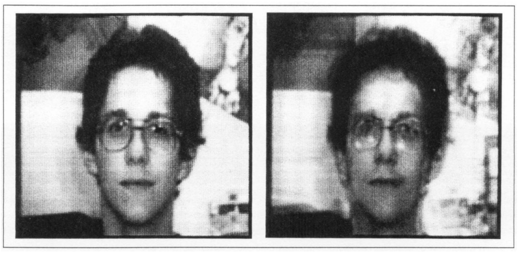

FIGURE 2.1: An original face image and its projection onto the face space defined by eigenfaces [Turk and Pentland, 1991]

The Eigenfaces technique was extended to operate in 3D space by Atick, Griffin, and Redlich, 1996, to represent shape variations, but ended up with roughly the same limitation.

The reason for that is: The only preprocessing used by the similar methods to those described above is matching eye position. Faces generated using just Principal Component Analysis (PCA) will likely have the ghost artifacts of a face shape, because all face features, except eyes, differ in locations from sample to sample.

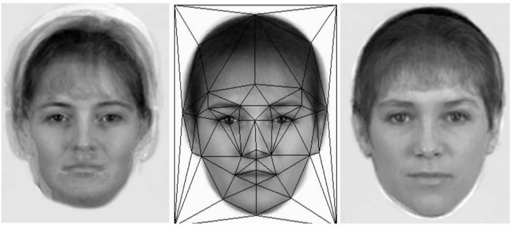

FIGURE 2.2: A face generated by random combination of components from originals matched only for location of eyes. b) The average shape-free face, showing the points used to locate features. c) A face generated by random combination of shape-free components. [Craw and Cameron, 1991]

This issue was overcomed in the following work of Craw and Cameron, 1991: prior to PCA, they mark the location of key features around each face and morph them to an average shape as shown on Figure 2.2

This landmark-based approach was the first step to statistical shape models, advancement of which are widely used in modern 3D head modeling algorithms.

2.1.2 Statistical Shape Models

The main idea behind Statistical Shape Models is that each shape in the training set of t shapes can be represented by a set of N labelled landmark points, $\{(x_i, y_i)\}$. Each shape example can then be represented as the $2N$ element vector (2D image), $\mathbf{x}$, where

$$\mathbf{x} = (x_1, \ldots, x_N, y_1, \ldots, y_N)^T$$

Landmarks must be consistent from one shape to the next.

If we apply PCA to the data, we can then approximate any shape of the training set, $\mathbf{x}$ using

$$\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{Pb}$$

Where $\overline{\mathbf{x}}$ is the mean of the aligned training examples, $\mathbf{Pb} \in \mathrm{R}^{2N \times t}$ is a matrix whose columns are unit vectors along the principal axes of the data cloud, and $\mathbf{b}$ is a t-element vector of shape parameters.

By varying the shape parameters within limits learnt from the training set, we can generate new examples.

Such approach was used for face shape representation in Active Appearance Models framework presented by Cootes, Edwards, and Taylor, 2001.

The shape model, on contrast to Eigenfaces and other correspondence-based approaches, provides a powerful and compact representation of shape differences by shifting pixels in the image plane. However, compared to the simple linear projection in Eigenfaces, the image analysis task is transformed into a more challenging nonlinear model fitting problem [Egger et al., 2020].

2.1.3 3D Morphable Models

A 3D Morphable Model (3DMM) is a parametric face shape modeling technique that was presented by Blanz and Vetter, 1999.

It was introduced to assist in two main problems that existed in 3D face modeling: (1) inconsistencies in 3D face landmark annotation; (2) the separation between natural and generated faces;

That both are dependant on painstaking human-involved work.

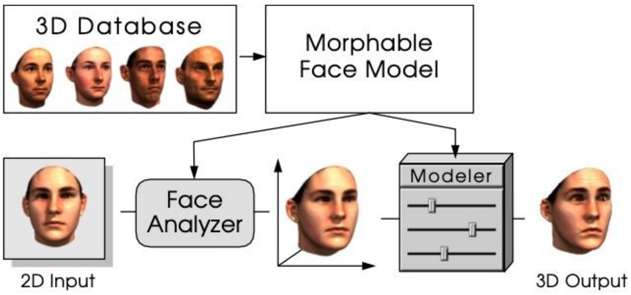

FIGURE 2.3: Derived from a dataset of prototypical 3D scans of faces, the morphable face model contributes to two main steps in face manipulation: (1) deriving a 3D face model from a novel image, and (2) modifying shape and texture in a natural way. [Blanz and Vetter, 1999]

3DMM would solve it by:
(1) all face scans are preregistered in common topology, and maintained throughout any further processing steps; **(2)** due to this correspondence, morphologically realistic faces may be defined as linear combinations of other in a set.

Building 3D facial model:

(1) Dense correspondence: A set of meshes is reparametrized to a form where each mesh has the same number of vertices that are connected into a triangulation; the form is shared by all meshes.

The geometry of a 3D facial mesh is defined by the vector ("shape vector"):

$$\mathbf{S_i} = (x_1, y_1, z_1, \ldots, x_N, y_N, z_N)^T$$

where N is the number of vertices in each mesh and $[x_i, y_i, z_i]^T$ are the coordinates of an i-th vertex in 3D space.

In classic approach by Blanz and Vetter, 1999, texture vector is also constructed:

$$\mathbf{C_i} = (R_1, G_1, B_1, \ldots, R_N, G_N, B_N)^T$$

where N is the number of vertices in each mesh $[R_i, G_i, B_i]^T$ are the R (red), G (green), and B (blue) values of the RGB colour model of an i-th vertex.

Separating facial shape and color helps disentangle them from external factors such as illumination and camera parameters. As the focus of this work is mainly the 3D geometry, I will not further cover the texture counterpart.

(2) Statistical modeling: The idea behind 3DMM is that, if the set of M 3D faces is sufficiently large, any new shape can be expressed as a linear combination of the shapes of the training 3D faces:

$$\mathbf{S}_{\text{new}} = \sum_{i=1}^{M} a_i \mathbf{S}_i$$

with $a_i \in \mathbb{R}, \forall m = 1, \cdots, M$. As a result, any new face shape can be parametrised by its shape parameter vector $\mathbf{S}_{\text{new}} = (a_1, \cdots, a_M)^{\text{T}}$.

To optimise the computation for large set of raw scan data, 3DMM was made to act as a linear basis of the 3D face shapes. For this Principal Component Analysis is applied.

Represented by its shape vector, each scan in the set of M faces is a data point in $\mathbb{R}^{3N}$, where N is the number of vertices in each scan.

By using the average of all the shape vectors and adding together weighted offsets from the feature points, by tuning the parameter vector, any new face that is a linear combination of the faces in the database can be generated.

$$\mathbf{S}_{\text{model}} = \bar{\mathbf{S}} + \sum_{i=0}^{m-1} a_i \Delta\mathbf{S_i}, \quad \Delta\mathbf{S_i} = \mathbf{S_i} - \bar{\mathbf{S}}$$

By using Principal Component Analysis one can reduce the dimensionality of the source data and ensure that all shape vectors are orthogonal.

In PCA we take all the sample points which will lead to a certain distribution of the whole data sets and compute the mean of all the data points. PCA will result in principal component vectors which correspond to the features of most variance in the original dataset. And by limiting the number of principal components used to represent the variance in the source data, we can build a new basis with lower dimensionality than the source data itself. This makes it possible to take the mean shape and apply the offsets along the principal component vectors to morph the shapes where variance is highest.

2.2 Available 3D Head Models and Modeling Approaches

The quality of 3D Morphable Model is highly dependant on the set of shapes it is constructed from and on the shape acquisition method. In this section I will give a brief analysis of most common models available in the literature. The summary of the analysis is shown on the Table 2.1.

Blanz and Vetter, 1999: a morphable model was constructed with 200 laser scans of heads of "young adults" (100 females and 100 males). Dense point-to-point correspondence was made using an optical flow algorithm after process of UV-unwraping of 3D shapes.

BFM (Basel Face Model) [Paysan et al., 2009]: the training set contained scans of 200 subjects (aged between 8-62 years) mostly of European ethnicity. Scans were made using much more advanced laser scanner, that improved detalization of meshes. Paysan et al. computed point-to-point correspondence directly between 3D surfaces by applying nonrigid iterative closest point (NICP) algorithm.

FaceWarehouse [**cao2013facewarehouse**]: shapes of 150 subjects (aged between 7-80 years) were presented as depth maps. In order to perform mesh correspondence, Blanz and Vetter, 1999 model was fitted to the depth-maps. Captured geometry contained 20 expressions, which improved naturality and generalization of a model.

BFM (Basel Face Model) [Gerig et al., 2018]: improvement of 2009's version incorporated adding 160 expression scans to the training set and advancing the scanning procedure to full head with multi-resolution.

LSFM (Large Scale Facial Model) [Booth et al., 2018]: built from scans of 9663 individuals of highly diverse age, gender (48% male) and ethnicity (82% White, 9% Asian, 5% mixed heritage, 3% Black and 1% other). Because of the scale of this dataset, simpler models based on forms of a specified age range and ethnicity might be created.

FLAME (Faces Learned with Articulated Model and Expressions) [Li et al., 2017a]: the identity shape space from the heads of roughly 4000 CAESAR body scans [Robinette et al., 2002] spanning a wide range of ages, ethnicities, and both genders. To model pose and expression variation we use over 400 4D face capture sequences from the D3DFACS dataset [Cosker, Krumhuber, and Hilton, 2011] and additional 4D sequences that we captured, spanning more expression variationuses linear transformations to describe identity and expression dependent shape variations, and standard vertex based linear blend skinning (LBS) with corrective blendshapes, with N = 5023 vertices, K = 4 joints. [Li et al., 2017b]

LYHM (Liverpool-York Head Model) [Dai et al., 2017]: full head model, the identity learned from 1212 3D scans. Expression variations are learned separately from identity.

CFHM (Combined Face & Head Model) [Ploumpis et al., 2019]: merged from LYHM [Dai et al., 2017] and LSFM [Booth et al., 2018]. They proposed two methods for the combination of models. The first utilizes the latent PCA parameters and solves a linear least squares problem to approximate the full head shape, whereas the second constructs a combined covariance matrix that is later utilized as a kernel in a Gaussian Process Morphable Model (GPMM).

Figure 2.4 shows how these models incorporate identity and/or expression change by shifting parameters along the principal axes computed from the data sets.

model	**geometry**	**# subjects**	**%gender (male/female)**	**age**	**ethnicity**	**expressions**
Blanz and Vetter, 1999	shape	200	50/50	"young adults"	-	neutral
BFM (Basel Face Model) [Paysan et al., 2009]	shape	200	50/50	8-62	"most Europeans"	neutral
FaceWarehouse [Cao et al., 2014]	shape & expression	150	-	7-80	“various”	20 expressions
LSFM (Large Scale Facial Model) [Booth et al., 2018]	shape	9663	48/52	0-80	82% white	neutral
BFM (Basel Face Model) [Gerig et al., 2018]	shape & expression	200	50/50	-	-	160 expression scans
FLAME (Faces Learned with Articulated Model and Expressions) [Li et al., 2017a]	shape & expression	3800	48/52	"wide range"	"wide range"	21000 frames for expression
LYHM (Liverpool-York Head Model) [Dai et al., 2017]	shape	1212	50/50	1-85	-	neutral

TABLE 2.1: Overview of publicly available 3D head shape models.

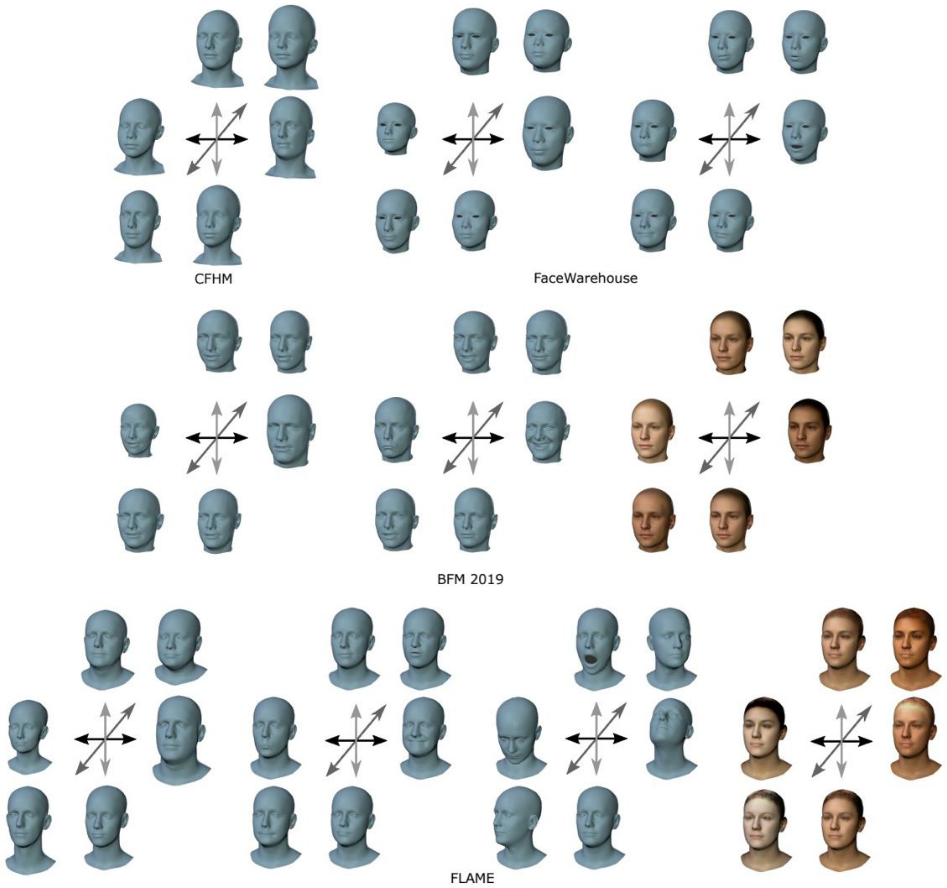

FIGURE 2.4: Model variations of existing face models. Top left: CFHM [Ploumpis et al., 2019] shape variations. Top right: FaceWarehouse [**cao2013facewarehouse**] shape and expression variations. Middle: BFM 2019 [Gerig et al., 2018] shape, expression, and appearance variations. Bottom: FLAME [Li et al., 2017a] shape, expression, pose, and appearance variation. For shape, expression, and appearance variations, three principal components are visualized at ±3 standard deviations. The FLAME pose variations are visualized at $\pm\pi\ /6$ (components three and four) and at 0, $\pi/8$ (component six). Visualisation is made by Egger et al., 2020

2.3 Spectrum of Facial Analysis Tasks

As stated in the introductory Chapter 1, 3D Head Understanding area includes a range of head related analysis tasks apart from the 3D Head Modeling itself. This section includes an overview of the related tasks such as Face Landmarks Detection (Subsec. 2.3.1), Head Pose Estimation (Subsec. 2.3.2), 3D Head Reconstruction (Subsec. 2.3.3). Figure 2.5 shows schematic essense of the tasks.

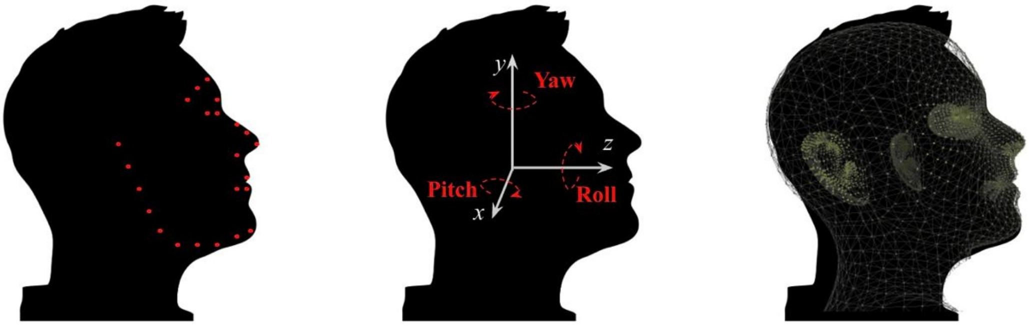

FIGURE 2.5: Facial analysis tasks schematic explanation. From left to right: facial landmarks detection; roll, pitch, yaw parameters representation for head pose estimation; overlined 3D mesh as output of 3D head reconstruction methods

2.3.1 Face Landmarks Detection

The task of Facial Landmark Detection involves identification of the locations of key facial landmark points on images or videos. The number of landmarks varies from only 7 up to thousand based on the exact problem statement and availability of annotated data. Key points describe the unique location of facial features (such as nose edge or eye corner) or interpolated locations that connect those feature points around facial components (such as facial contour). Formally, face landmark detection algorithm is trained to predict the location of N keypoints $(x_1, y_1, \ldots, x_N, y_N)$, where (x_i, y_i) represent the coordinates of a point on a 2D image I.

Precise facial landmarks extraction enables application of them for: tracking faces in images and video; analysing facial expressions; detecting dysmorphic facial signs for medical diagnosis; biometrics / face recognition tasks.

Facial landmark detection is challenging for several reasons: Firstly, facial appearance changes significantly across subjects under different facial expressions and head poses; secondly, facial occlusions by other objects, or self-occlusion due to extreme head poses, leads to incomplete facial appearance information; and thirdly, environmental conditions such as illumination can affect the appearance of the face on facial images.[Bodini, 2019]

2.3.2 Head Pose Estimation

Head pose estimation aims to approximate viewing direction of human head given an input image or video. A head pose estimator's output comprises of the yaw and pitch angles in 2D space, as well as the roll angle in 3D space if desired i.e. the Euler angles (see Fig. 2.5).

The main challenge behind accurate head pose estimation is availability of large-scale and accurate data for training. Labeling head pose parameters is a hard piece of manual work, while automated methods still produce noisy data (see Sec. 3.3). See a detailed overview of available data in the Chapter 3.

Head pose estimation that has a large amount of applications such as aiding in gaze estimation, modeling attention, fitting 3D models to video or performing face alignment.

2.3.3 3D Head Reconstruction

3D Head Reconstruction is a task of estimating the geometry of the face/head from an uncalibrated 2D image and is an attractive alternative to acquiring a 3D scan. This is because 3D-from-2D reconstruction tries to combine the ease of gathering 2D image datasets with the benefit of a 3D representation of the facial geometry.

Solving the 3D Head Alignment problem solves both th problem of (1) Face Landmark Detection and (2) Head Pose estimation. This is because: (1) methods for 3D Head Reconstruction commonly use 3DMM fitting techniques, which results in consistent topology of meshes and simplicity of extraction specific face landmark points; (2) fitting a 3D facial model to a photograph is done by estimating the 3D posture and lighting of the resultant 3D face, in addition to the model parameters, so that the projection onto the image plane of the resulting 3D face generates an image that is as close to the provided picture as possible.

Following the 3D Head Understanding area overview, next chapters include description of common 2D/3D datasets (Ch. 3), and benchmarks (Ch. 4) for the area problems, addressing their benefits and limitations.

Chapter 3

Datasets

Crucial ingredient for 3D head analysis tasks is a representative set of 3D annotations presented along with corresponding appearance image data.

Generally construction of such sets can be done in two ways:

(1) using scanners or depth-cameras to acquire a precise 3D model of a subject,

(2) manually or automatically annotating existing 2D image data,

In this chapter, I will list and analize publicly available 2D and 3D face and head data sets that are most frequently used for 3D Head Understanding methods training and benchmarking. Section 3.1 describe publicly available 2D face data sets, and Section 3.2 is focused on 3D annotated data.

3.1 2D Face Datasets

Generally, there are two types of datasets: image datasets and video datasets. Photographs can be shot under confined settings, such as regulated lighting conditions and specified positions; otherwise, images can be taken under uncontrolled conditions, which are frequently referred to as *in-the-wild* datasets.

Multi-PIE, introduced by [Gross et al., 2008], contains 337 subjects spread across 15 viewpoints, 19 lighting settings, and six distinct emotions. The 39-point and 68-point face landmarks are identified.

300-W (300 Faces in-the-Wild Challenge), introduced by Sagonas et al., 2016 is a large-scale dataset that combines several, such as Helen [Le et al., 2012], LFPW [Belhumeur et al., 2013], AFW [Zhu and Ramanan, 2012], and a newly introduced challenging dataset, iBug. It consist of 3837 photos as well as a separate test set of 300 indoor and outdoor images. Annotations come as 68 facial keypoints.

Menpo. In the training set, there are 6679 semi-front view face photos with 68 points and 5335 profile view face images with 39 points. There are 12006 front view photographs and 4253 profile view images in the test set. It was added to the Menpo challenge in 2017 to provide a more challenging task for testing the robustness of face landmark extraction algorithms, since it encompasses a wide range of poses, light conditions, and occlusions.

Sequential facial landmark extraction is done using video-based annotated datasets. The 300-VW dataset, introduced by Shen et al., 2015, contains the largest number of annotated images and video frames. The dataset consists of 50 training and 64 testing videos, which are then separated into three scenarios based on light conditions, emotions, head postures, and occlusions. Semi-automatically, all of the frames are annotated with 68 facial landmarks. The Menpo 3D [Zafeiriou et al., 2017] dataset is the only one in which uses the 3DMM fitting method [Booth et al., 2017] to annotate 3D face landmarks in video. In addition to it, all of the photos in 300W, and Menpo were re-annotated in the same way in 3D.

3.2 3D Face Datasets

Existing 3D face datasets differ based on a registration of a 3D model.

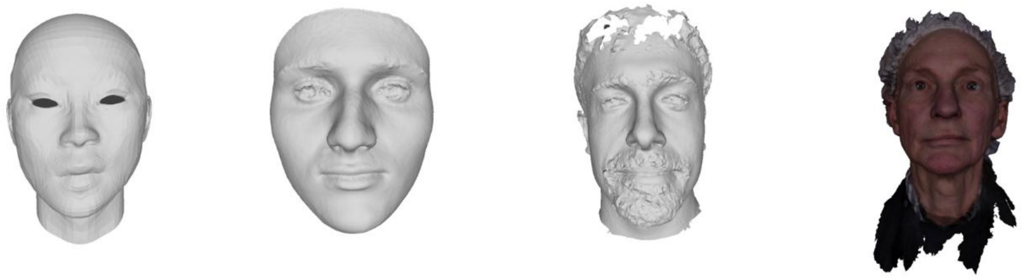

FIGURE 3.1: Visualisation of 3D face databases shape representation difference. From left to right: BIWI Face Database [Fanelli et al., 2011], NoW Face Dataset [Sanyal et al., 2019], Florence 2D/3D Face Dataset [Bagdanov, Del Bimbo, and Masi, 2011], Stirling/ESRC 3D Face Database

3.2.1 3D Model fitting registration

AFLW2000-3D, introduced by Zhu et al., 2016, contains 2000 images annotated with image-level 68-point 3D face keypoints. The goal of the Zhu et al. work was to solve the problem of 3D face alignment accross large poses. The term large poses implies such head positions, in which significant part of main facial keypoints are hidden from the camera viewpoint e.g. the big head rotation angle.

A 6-layers CNN was trained iteratively to output the update of the pose, shape, and expression parameters and then the sample of 3D points was projected to estimated locations on a 2D image.

AFLW2000-3D is commonly used for evaluation of 3D facial landmark detection and head pose estimation methods (see Sec. **??**).

However, the dataset has couple of limitations:

(+) diverse; (+) contains large poses;

(-) shape detalization is lower than of scanned 3D shapes; (-) some labels in the dataset are inconsistent (see Fig.) 3.2.

3.2.2 Depth camera registration

BIWI, introduced by Fanelli et al., 2011, contains over 15000 images of 20 people (6 females and 14 males - 4 people were recorded twice). For each frame, a depth image, the corresponding rgb image (both 640x480 pixels), and head pose annotation is provided. The head pose range covers about +-75 degrees yaw and +-60 degrees pitch. Ground truth is provided in the form of the 3D location of the head and its rotation.

The dataset is commonly used for evaluation of head pose estimation methods (see Sec. 4.2.1).

(+) accurate and detalized shapes; (+) diverse in terms of poses;

(-) lacks subjects diversity; (-) not in-the-wild (i.e. subjects captured under constrained conditions).

3.2.3 Multi-View Camera System registration

Stirling/ESRC Face Database is a large 2D/3D face database that was funded by ESRC project ES/J010081/1 to Peter Hancock at Stirling and Bernie Tiddeman at Aberystwyth.

The 3D collection of the database contains 3D scans of subjects (45 male and 54 female), captured with a DI3D camera system, with neutral, smile mouth closed, smile mouth open, anger, disgust, fear, sad and surprised expressions. The export format is Wavefront obj.

Part of the database is used for the Liu, Tran, and Liu, 2019 3D Face Reconstruction benchmark (see Sec. 4.1.2).

(+) large-scale; (+) accurate and detalized shapes; (+) diverse;

(-) not in-the-wild (i.e. subjects captured under constrained conditions).

NoW (Not quite in-the-Wild) **Face Dataset**, introduced by Sanyal et al., 2019, is dataset that contains 2054 2D images of 100 subjects, captured with an iPhone X, and a separate 3D head scan for each subject. The subjects are selected to contain variations in age, BMI, and sex (55 female, 45 male). The export format is Wavefront obj.

For each subject they capture a raw head scan in neutral expression with an active stereo system (3dMD LLC, Atlanta). The multi-camera system consists of six gray-scale stereo camera pairs, six color cameras, five speckle pattern projectors, and six white LED panels. The reconstructed 3D geometry contains about 120K vertices for each subject. Each subject wears a hair cap during scanning to avoid occlusions and scanner noise in the face or neck region due to hair. [Now Face Challenge Webpage, Sanyal et al., n.d.]

Main purpose of the dataset is to serve evaluation of the 3D Head Reconstruction methods (see Sec. 4.1.1).

(+) accurate and detalized shapes; (+) diverse subjects;

(-) not in-the-wild (i.e. subjects captured under constrained conditions); (-) contains 3D scans in neutral only expressions.

FaceWareHouse is a database of 3D facial expressions for visual computing applications. It consists of 150 individuals aged 7-80 from various ethnic backgrounds scans captured using a Kinect RGBD camera.

For each person, they capture the RGBD data in different expressions, including the neutral expression and 19 other expressions such as mouth-opening, smile, kiss, etc. For every RGBD raw data record, a set of facial feature points on the color image such as eye corners, mouth contour and the nose tip are automatically localized, and manually adjusted if better accuracy is required. They then deform a template facial mesh to fit the depth data as closely as possible while matching the feature points on the color image to their corresponding points on the mesh. From these fitted face meshes, they construct a set of individual-specific expression blendshapes for each person. [Cao et al., 2014]

(+) accurate and detalized shapes; (+) diverse subjects and expressions;

(-) not in-the-wild (i.e. subjects captured under constrained conditions);

Florence Face Database introduced by Bagdanov, Del Bimbo, and Masi, 2011, consists of high-resolution 3D scans of 53 human faces (14 female and 39 male) aged between 20 and 70 years (50% 20-30 years) of caucasian ethnicity along with several video sequences of varying resolution and zoom level.

Each subject is recorded under three settings: constrained indoor HD video recording, unconstrained indoor using a standard PTZ surveillance camera, and unconstrained outdoor setting environment under challenging recording conditions.

This dataset is being constructed specifically to support research on techniques that bridge the gap between 2D, appearance-based recognition techniques, and fully 3D approaches. It is designed to simulate, in a controlled fashion, realistic surveillance conditions and to probe the efficacy of exploiting 3D models in real scenarios.[Bagdanov, Del Bimbo, and Masi, 2011]

(+) accurate and detalized shapes; (+) various scanning conditions;

(-) lacks subjects diversity.

3.3 Main Challenges

None of these most popular **3D head shape datasets** can be considered a diverse, accurate and large scale at the same time. Collecting such data set with 3D head annotations is still an open challenge as:

(1) Manual annotation of existing 2D datasets it is expensive, time-consuming, and will not produce such high accuracy of annotation as scanning subjects;

(2) Scanning subjects is still expensive and time-consuming procedure but result in high-quality 3D meshes. However, scanning methods can not help with constructing large-scale and diverse dataset, as scanning process requires presence of a subject in a laboratory or under more or less constrained conditions.

(3) Automatic methods of annotation (e.g. 3D model fitting) can solve both problem of manual annotation and constrained scanning conditions, but require of advanced methods of fitting or reconstruction, which are still challenging to develop to achieve the accuracy of the scans.

Head pose annotated datasets are also limited because of two reasons:

(1) Although there exists large facial databases, it is almost impossible to manually annotate these collected images with an exact head orientation.

(2) Available solution adopted by the public benchmarking datasets is to ask the participants to look at a set of markers that are located in predefined direction in the measurement room (e.g., 93 direction marks of Pointing'04 dataset [Gourier, Hall, and Crowley, 2004]). However, the noisiness of the annotated labels obtained with the pose direction in fact define the direction of gaze instead of the head pose direction. [Huttunen et al., 2015]

(3) Another challenge is head pose representation:

The most widespread rotation representation in head pose datasets is *Euler angle system* (i.e. three rotation axes (e.g. Euler Angles, degrees of freedom) around which object is rotated (see Fig. 2.5)). This representation has one vulnerability called "gimbal lock": rotation is applied orderly (e.g. $xyz/yzx/...$), and whenever the second rotation reaches over $\pi/2$ in any direction, i.e., large head poses, other rotation axes become linearly dependent, resulting in loosing one deegre of freedom and yielding an infinite number of representations for the same transformation [Cao et al., 2021]. This problem is seen in AFLW2000-3D dataset (see Fig. 3.2).

The solution to this may be representing head pose rotation transformation in relevance to "zero pose" in terms of $3x3$ rotation matrix.

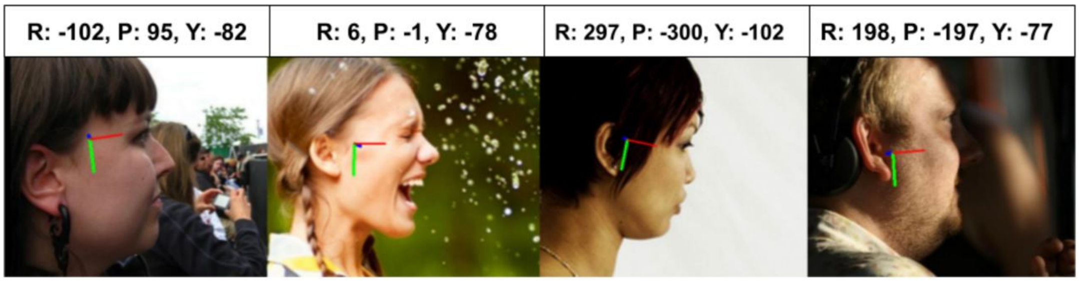

FIGURE 3.2: AFLW2000-3D large poses label inconsistencies due to the gimbal lock problem in Euler angles pose representation system.

Chapter 4

Benchmarks and Metrics

High-quality assessment of proposed methods is decisive for any field to develop. The quality of the benchmark depends on two aspects:

(1) data quality (i.e. to contain a representative and diverse set of objects under various scenarios to test methods' generalizability);

(2) efficient metrics.

This chapter contains overview of four common benchmarks for 3D Head Reconstruction (Sec. 4.1) and Head Pose estimation (Sec. 4.2)

4.1 3D Head Reconstruction

4.1.1 NoW Face Challenge

Benchmark introduced by Sanyal et al., 2019, is designed for the task of 3D face reconstruction from single monocular face image. The NoW dataset (see Sec. 3.2.3) was collected to serve this purpose.

Data: NoW dataset is divided into a validation set and a test set. The validation set consists of 20 subjects and the test set consists of 80 subjects. They provide the ground truth scans for the validation set and their corresponding seven facial landmarks (see Fig.4.1) for rigid alignment of the meshes. Sanyal et al. also provide bounding boxes of faces on the images, to avoid bias produced by different head detectors.

The captured data is categorized in four challenges: neutral (620 images), expression (675 images), occlusion (528 images) and selfie (231 images). Expression contains different acted facial expressions such as happiness, sadness, surprise, disgust, and fear. Occlusion contain images with varying occlusions from e.g. glasses, sunglasses, facial hair, hats or hoods. For the selfie category, participants are asked to take selfies with the iPhone, without imposing constraints on the performed facial expression. The images are captured indoor and outdoor to provide variations of natural and artificial light.[Sanyal et al., 2019]

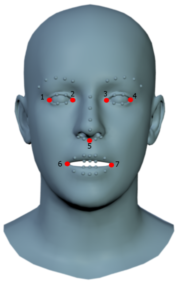

FIGURE 4.1: Seven facial landmark points

Evaluation Metrics:

(1) mesh prediction should be performed on a cropped images given ground-truth bounding boxes;

(2) predicted mesh should be neutralised before passing through the evaluation process, locations of seven facial keypoints should be passed along with the predicted mesh;

(3) before computing the metrics predicted mesh to a ground-truth scan is rigidly aligned (rotation, translation, and scaling) using Procrustes analysis[1], based on seven landmark points, as the predicted mesh and ground-truth scan appear in different local system of coordinates;

(4) ground-truth scan is cropped with a mask, radius of which is computed as $r = 1.4 \times (outereyedistance + nosedistance)/2)$, to avoid differing outer face area artifacts produced during scanning;

(5) the distance is calculated from each vertex of the scan mask to the closest point on a mesh surface;

(6) the metrics are the ***mean, median***, and ***standard deviation*** of the distances; computed metrics claimed in millimeters as ground-truth scans remain in real-world scale.

Limitations: This benchmark is not suitable to test expression parameters of the reconstructed model, as, although a big set of images containing facial expressions is provided, only neutral 3D face scans are provided as ground-truth and all the predicted meshes should be neutralized before computing the metrics. Neutral face reconstruction has another limitation: images and scans of the subjects were captured under different conditions. As it is not possible for a person to maintain the same facial expression, ground-truth scans differ from the actual 3D appearance of the person on an image.

4.1.2 FG2018 3D face reconstruction challenge

Benchmark introduced by Liu, Tran, and Liu, 2019, uses a subset of Stirling/ESRC 3D face database (see Sec. 3.2.3) as the test dataset for the challenge.

Data: The test set has 2000 2D images, including 656 high-quality and 1344 low-quality images, of more than one hundred subjects. The 2D and 3D faces in the Stirling ESRC dataset were captured under 7 different expression variations. The high quality images were captured in constrained scenarios with good lighting conditions. Whereas, low resolution images or video frames were captured with a variety of image degradation types such as image blur, low resolution, poor lighting, large scale pose rotation etc.[Liu, Tran, and Liu, 2019]

Evaluation Metrics:

(1) - (5) steps are the same as in NoW Face Challenge (4.1.1)

(6) 3D Normalized Mean Error ***3D-NME***:

$$\frac{\sum_{i=1}^{n} \sqrt{(x_{scan_i} - x_{pred_i})^2}}{n}$$

of distances is calculated on High-/Low-quality subsets a full dataset.

[1] Procrustes analysis determines a linear transformation (translation, reflection, orthogonal rotation and scaling) of the points in Y to best conform them to the points in matrix X, using the sum of squared errors as the goodness of fit criterion.[Liu, Tran, and Liu, 2019]

Limitations: FG2018 3D face reconstruction challenge has the same limitation of neutralizing expression. However, the dataset contains images captured at the exact same moment as 3D scan of a face, that is why it is better to test neutral face reconstruction methods. Another limitation is that dataset contains mostly images captured under constrained conditions and can not test methods' generalizability to the in-the-wild settings.

4.2 3D Head Pose Estimation

4.2.1 BIWI & AFLW2000-3D

Data: a detailed overview of the datasets is presented in the sec. 3.2.2. Both BIWI and AFLW2000-3D report the ground truth of yaw, pitch and roll angles. BIWI dataset also consists the head center and the calibration matrix.

Evaluation Metrics: the Mean Absolute Error (MAE) of each Euler angle is computed.

Limitations: The original BIWI [Fanelli et al., 2011] paper does not report the adopted split between training and testing sets; fair comparisons are thus not guaranteed[Amador et al., 2017]. AFLW contains ambiguous labels.

Chapter 5

Proposed Benchmark

This chapter describes a benchmark that addresses limitations of all the above listed 3D Head Understanding benchmarks. The main contribution of it is a novel combination of metrics that assess both quality of Face Landmark detection, Head Pose estimation, and 3D Head reconstruction of the methods.

Proposed dataset consist of 44,898 images collected from various sources (37,840 in the training set, 4,312 in the validation set, and 2,746 in the test set). For each image, we provide 5,023 vertices of the FLAME mesh, 3,669 of which are accurately labeled (neck and eyeballs excluded). It is well balanced over a wide range of poses, face expressions, and occlusions. For a detailed overview of the annotation process and accuracy as well as dataset statistics refer to Martyniuk et al., 2022.

Data and metrics combination was proposed in the DAD(Diverse Accurate and Dense)-3DHeads[1] project [Martyniuk et al., 2022] that I was involved in. My contribution to the project consisted of experimental evaluation of the proposed 3D head alignment method on existing and novel DAD-3DHeads benchmark. we use head bounding box size for normalization. The metric is computed on 68 landmarks

5.1 Reprojection NME

To assess the reprojection quality the normalized mean error of the reprojected 3D vertices onto the image plane, taking X and Y coordinates into account.

$$\text{Error}_{\text{reprojection}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|x_i - x_i^*\|_2^2}{L}$$

where $N = 68$ is a number of landmark points, L is the head bounding box size for a given image that is used for normalization.

5.2 Z_n Accuracy

As the annotation scheme is conditioned only upon model prior and the reprojection onto the image, we do not guarantee the absolute depth values to be as accurate as sensor data. For each of K vertices v_i of the ground-truth mesh, n closest vertices $\{v_i^1, \ldots, v_i^n\}$, are chosen, and calculated which of them are closer to (or further from) the camera. Then, we compare if for every predicted vertex w_i this configuration is the same:

$$Z_n = \frac{1}{K}\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n}\left(\left(v_i \succeq_z v_i^j\right) == \left(w_i \succeq_z w_i^j\right)\right)$$

[1] https://github.com/PinataFarms/DAD-3DHeads

The Z_n metric is valid only for predictions that follow FLAME mesh topology (sec.2.2)

5.3 Chamfer Distance

To measure the goodness of fit for the methods that follow different mesh topology, the Chamfer distance is added. To ensure generalization to any number of predicted vertices, we measure a one-sided Chamfer distance from our ground-truth mesh to the predicted one. The chamfer distance is computed by summing the squared distances between nearest neighbor correspondences of two point clouds:

$$d_{CD}(X, X^*) = \sum_{x \in X} \min_{x^* \in X^*} \|x - x^*\|_2^2$$

where X is a set of ground-truth vertices, and X^* is the set of predicted ones. Following a standard approach, explained in Sec. 4.1.1, the mesh is rigidly aligned to the ground-truth scan before the computation based on the seven facial keypoints (Fig.4.1)

5.4 Pose Error

To avoid the "gimbal lock" problem (sec. 3.3), pose error is computed based on rotation matrices. The distance between rotations represented by rotation matrices R_1 and R_2 is the angle of the difference rotation represented by the difference rotation matrix $R_D = R_1 R_2^T$. Larochelle, Murray, and Angeles, 2007 proposed a metric:

$$\text{Error}_{\text{pose}} = d(R_1, R_2) = \left\| I - R_1 R_2^T \right\|_F$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the matrix, as a distance measure between two rigid body displacements. Metric outputs values in the range $[0, 2\sqrt{2}]$[Huynh, 2009].

Chapter 6

Conclusion

In this work I performed an analytical study on 3D Head Understanding benchmarks. 3D Head Understanding area includes Face Landmark detection, Head Pose Estimation, and mainly 3D Head Reconstruction as it solves both previous ones.

Analysis included the most frequently avaliable in the literature datasets as a crucial part of training and validation of newly appearing methods; and addressed key requirements:

(1) accuracy of the annotations;

(2) diversity of the subjects and conditions;

The focus of benchmarks analysis was to describe methods of assessment and their limitations.

Based on this analysis a novel benchmark with representative data and novel combination of metrics was described to serve the advancements in 3D Head Understanding area.

Bibliography

Amador, Elvira et al. (2017). “Benchmarking head pose estimation in-the-wild”. In: *Iberoamerican Congress on Pattern Recognition*. Springer, pp. 45–52.

Atick, Joseph J., Paul A. Griffin, and A. Norman Redlich (1996). “Statistical Approach to Shape from Shading: Reconstruction of Three-Dimensional Face Surfaces from Single Two-Dimensional Images”. In: *Neural Comput.* 8.6, 1321–1340. ISSN: 0899-7667. DOI: 10.1162/neco.1996.8.6.1321. URL: https://doi.org/10.1162/neco.1996.8.6.1321.

Bagdanov, Andrew D., Alberto Del Bimbo, and Iacopo Masi (2011). “The Florence 2D/3D Hybrid Face Dataset”. In: *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*. J-HGBU ’11. Scottsdale, Arizona, USA: Association for Computing Machinery, 79–80. ISBN: 9781450309981. DOI: 10.1145/2072572.2072597. URL: https://doi.org/10.1145/2072572.2072597.

Belhumeur, Peter N et al. (2013). “Localizing parts of faces using a consensus of exemplars”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.12, pp. 2930–2940.

Blanz, Volker and Thomas Vetter (1999). “A Morphable Model for the Synthesis of 3D Faces”. In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH ’99. USA: ACM Press/Addison-Wesley Publishing Co., 187–194. ISBN: 0201485605. DOI: 10.1145/311535.311556. URL: https://doi.org/10.1145/311535.311556.

Bodini, Matteo (2019). “A Review of Facial Landmark Extraction in 2D Images and Videos Using Deep Learning”. In: *Big Data and Cognitive Computing* 3.1, p. 14. ISSN: 2504-2289. DOI: 10.3390/bdcc3010014. URL: http://dx.doi.org/10.3390/bdcc3010014.

Booth, James et al. (2017). “3d face morphable models" in-the-wild"”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 48–57.

Booth, James et al. (2018). “Large scale 3d morphable models”. In: *International Journal of Computer Vision* 126.2, pp. 233–254.

Cao, Chen et al. (2014). “FaceWarehouse: A 3D Facial Expression Database for Visual Computing”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.3, pp. 413–425. DOI: 10.1109/TVCG.2013.249.

Cao, Zhiwen et al. (2021). “A vector-based representation to enhance head pose estimation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1188–1197.

Chang, Feng-Ju et al. (2017). “FacePoseNet: Making a Case for Landmark-Free Face Alignment”. In: *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp. 1599–1608. DOI: 10.1109/ICCVW.2017.188. URL: https://doi.org/10.1109/ICCVW.2017.188.

Cootes, Timothy F., Gareth J. Edwards, and Christopher J. Taylor (2001). “Active appearance models”. In: *IEEE Transactions on pattern analysis and machine intelligence* 23.6, pp. 681–685.

Cosker, Darren, Eva Krumhuber, and Adrian Hilton (2011). "A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling". In: *2011 international conference on computer vision*. IEEE, pp. 2296–2303.

Craw, Ian and Peter Cameron (1991). "Parameterising images for recognition and reconstruction". In: *BMVC91*. Springer, pp. 367–370.

Dai, Hang et al. (2017). "A 3d morphable model of craniofacial shape and texture variation". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3085–3093.

Deng, Jiankang et al. (2020). "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5202–5211. DOI: 10.1109/CVPR42600.2020.00525.

Egger, Bernhard et al. (2020). "3d morphable face models—past, present, and future". In: *ACM Transactions on Graphics (TOG)* 39.5, pp. 1–38.

Fanelli, Gabriele et al. (2011). "Real time head pose estimation from consumer depth cameras". In: *Joint pattern recognition symposium*. Springer, pp. 101–110.

Gerig, Thomas et al. (2018). "Morphable face models-an open framework". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pp. 75–82.

Gourier, Nicolas, Daniela Hall, and James L. Crowley (2004). "Estimating Face orientation from Robust Detection of Salient Facial Structures". In.

Gross, Ralph et al. (2008). "Multi-PIE". In: *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–8. DOI: 10.1109/AFGR.2008.4813399.

Huttunen, Heikki et al. (2015). "Computer vision for head pose estimation: review of a competition". In: *Scandinavian conference on image analysis*. Springer, pp. 65–75.

Huynh, Du Q (2009). "Metrics for 3D rotations: Comparison and analysis". In: *Journal of Mathematical Imaging and Vision* 35.2, pp. 155–164.

Larochelle, Pierre M, Andrew P Murray, and Jorge Angeles (2007). "A distance metric for finite sets of rigid-body displacements via the polar decomposition". In.

Le, Vuong et al. (2012). "Interactive facial feature localization". In: *European conference on computer vision*. Springer, pp. 679–692.

Li, Tianye et al. (2017a). "Learning a model of facial shape and expression from 4D scans." In: *ACM Trans. Graph.* 36.6, pp. 194–1.

— (2017b). "Learning a model of facial shape and expression from 4D scans." In: *ACM Trans. Graph.* 36.6, pp. 194–1.

Liu, Feng, Luan Tran, and Xiaoming Liu (2019). "3d face modeling from diverse raw scan data". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9408–9418.

Martyniuk, Tetiana et al. (2022). "DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image". In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Morales, Araceli, Gemma Piella, and Federico M. Sukno (2021). "Survey on 3D face reconstruction from uncalibrated images". In: *Comput. Sci. Rev.* 40, p. 100400.

Paysan, Pascal et al. (2009). "A 3D face model for pose and illumination invariant face recognition". In: *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, pp. 296–301.

Ploumpis, Stylianos et al. (2019). "Combining 3d morphable models: A large scale face-and-head model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10934–10943.

Robinette, Kathleen M et al. (2002). *Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary*. Tech. rep. Sytronics Inc Dayton Oh.

Sagonas, Christos et al. (2016). "300 faces in-the-wild challenge: Database and results". In: *Image and vision computing* 47, pp. 3–18.

Sanyal, Soubhik et al. (n.d.). *Now Face Challenge Webpage*. URL: https://now.is.tue.mpg.de/.

Sanyal, Soubhik et al. (June 2019). "Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision". In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 7763–7772. URL: https://ringnet.is.tue.mpg.de/.

Shen, Jie et al. (2015). "The first facial landmark tracking in-the-wild challenge: Benchmark and results". In: *Proceedings of the IEEE international conference on computer vision workshops*, pp. 50–58.

Sirovich, Lawrence and Michael Kirby (1987). "Low-dimensional procedure for the characterization of human faces." In: *Journal of the Optical Society of America. A, Optics and image science* 4 3, pp. 519–24.

Turk, M.A. and A.P. Pentland (1991). "Face recognition using eigenfaces". In: *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591. DOI: 10.1109/CVPR.1991.139758.

Zafeiriou, Stefanos et al. (2017). "The 3d menpo facial landmark tracking challenge". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2503–2511.

Zhu, Xiangxin and Deva Ramanan (2012). "Face detection, pose estimation, and landmark localization in the wild". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE, pp. 2879–2886.

Zhu, Xiangyu et al. (2016). "Face alignment across large poses: A 3d solution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 146–155.