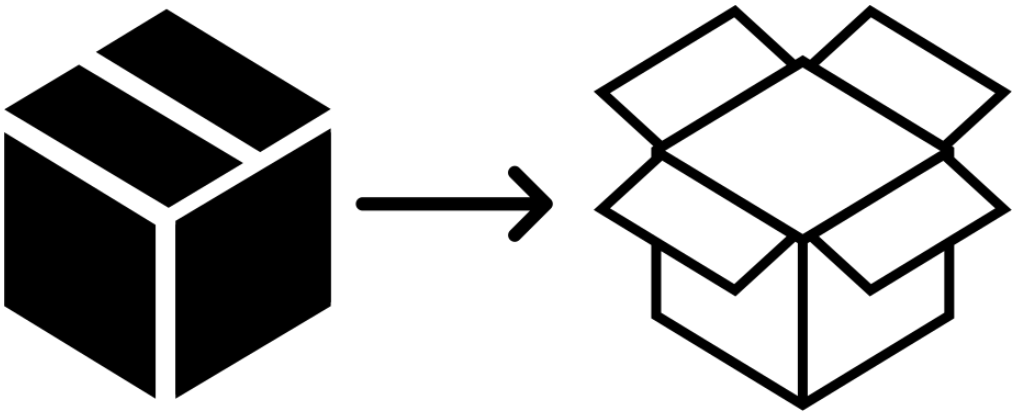
chist-era



XAIface

Measuring and Improving Explainability for AI-based Face Recognition

Face image dataset

Deliverable number: D3.2

Version: 1.0

EURECOM Sophia Antipolis

JOANNEUM RESEARCH DIGITAL

universität wien

it

EPFL

Acronym of the project: XAIface
Title of the project: Measuring and Improving Explainability for AI-based Face Recognition.
Grant: CHIST-ERA-19-XAI-011
Web site of the project: https://xaiface.eurecom.fr/

Short abstract
This document reports the activity carried out under T3.1 “Database collection and generation” (WP3). In particular, in this document we will report the advancement in the process of collecting the data needed for the development and testing of the explainability techniques. As such, this document will be continuously updated during the project as new data may be needed as the project advances and evolves.

Table of content

Definitions

ACC	Accuracy
AgeDB	Age Database
ArcFace	Additive Angular Margin Loss
AUC	Area Under the Curve
AWGN	Additive White Gaussian Noise
CALFW	Cross-Age Labeled Faces in the Wild
CASIA	Chinese Academy of Sciences' Institute of Automation
CMC	Cumulative Match Characteristic
DCNNs	Deep Convolutional Neural Networks
DiveFace	Dataset for Diversity-Aware Face Recognition
FAR	False Accept Rate
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FPR	False Positive rate
FPIR	False Positive Identification Rate
FR	Face Recognition
GB	Gaussian Blur
GN	Gaussian Noise
IJB	IARPA Janus Benchmark
JPEG	Joint Photographic Experts Group
LFW	Labeled Faces in the Wild
LR	Low Resolution
MagFace	Magnitude Face
ROC	Receiver Operating Characteristic
TAR	True Accept Rate
TN	True Negative
TP	True Positive
TPIR	True Positive Identification Rate
TPR	True Positive Rate
YTF	YouTube Faces

1. Introduction

An important step in the development and benchmarking of face recognition systems is data selection. Particularly for systems based on artificial intelligence, data selection is crucial as it can introduce bias in the system's behavior, as the system's rules are inferred from the data and the responses provided by the human programmer. The illustration in Figure 1a shows the difference between human learning, classic programming, and machine learning.

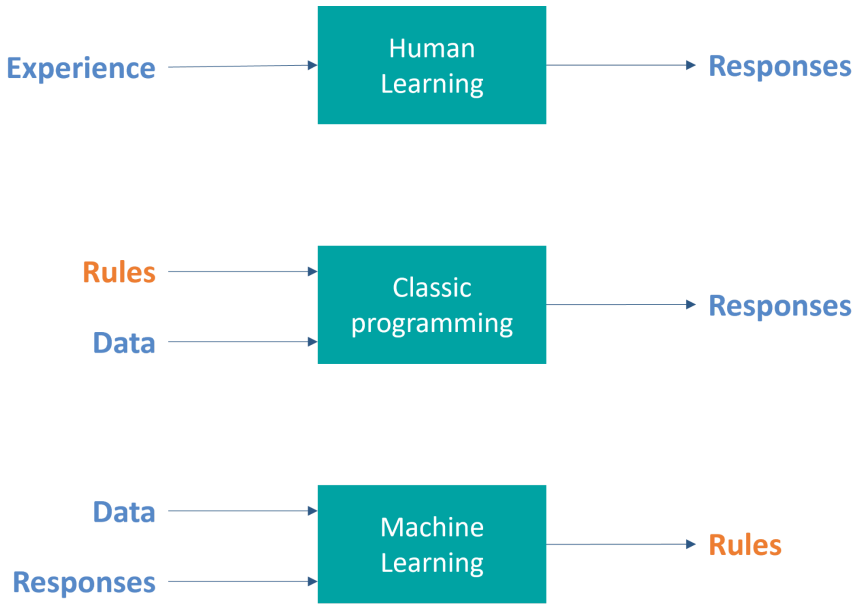


Figure 1a: Comparison of different earning processes.

The database selection phase for the XAIface project is therefore a critical phase to which much attention has been devoted and which will be monitored throughout the project to ensure that the data available to the consortium is sufficient and valid for the project's purposes. In addition, much attention and work has also been put into studying the legislation of the countries involved in the consortium so that the biometrics used comply with current legislation, including the GDPR. This is discussed in more detail in section 4, where a detailed study of ethical and legal issues is presented.

2. Database collection and generation

In this section we report on the process that led to the choice of the databases to be used in the project. We detail the criteria used by the consortium for database selection and we provide a description of the databases collected.

2.1. Collected databases

A number of databases were selected according to a list of criteria defined by the consortium (see Table 2.1a). The objective of the selected criteria is on the one hand to ensure that the database has the necessary characteristics for the development of the techniques envisaged in XAIface, and on the other hand to ensure the reproducibility of the experiments.

Table 2.1a: list of criteria for database selection.

Criterion - type	**Criterion - definition**	**Criterion - values**
Abstract	important features/information about the database	text
Availability	In case the database is no longer available, are there available models trained on that database?	text
Database composition	# classes	#
	# samples per class	#
	Total # of samples	#
	Male / female %	male: #%, female #%
	Ethnicity %	%, not provided
	Age / Age groups %	%, not provided
	# of acquisition sessions	#
	Time span between acquisition sessions	# days
	PIE* variations	yes / no (notes)
	DB size	# GB
	test/train/val split?	yes / no (notes)
Documentation and baseline evaluation	Documentation available and quality assessment	yes (rating 1-5) / no
	Used with ArcFace?	yes / no (notes, ref. to article(s))
	Used with MagFace?	yes / no (notes, ref. to article(s))
Sample features	Data type	e.g. images, videos
	Data format	e.g. TIFF, JPG, PNG, AVI
	Faces are aligned?	yes / no (alignment method)

	Faces are cropped?	yes / no (cropping method)
	Other processing?	yes / no (notes)
	Sample size	#x# pixels
Acquisition	Acquisition sensor	
	Acquisition modality	e.g. visible, thermal, NIR**, etc.
	Multimodal acquisition?	yes / no (list of acquisition modalities and sensors)
	Acquisition conditions	e.g. controlled, uncontrolled
Annotation	Annotation	list of annotated features
	Annotation method	e.g. manual, automatic
GDPR	GDPR compliance	notes
	Collection	yes / no
	Use	yes / no
License	Publicly available	yes / no (notes)
	Commercial usage and changes allowed?	yes / no (notes)
About	Link to DB	url
	Provider	name of provider
	Associated article	reference
Use in XAIface	Used for what task in XAIface?	Task number (notes)

*Pose Illumination Expression, **Near InfraRed

2.2. Description of Datasets

2.2.1. AgeDB

The AgeDB[1] dataset is used in age-invariant face verification in the wild experiments since it is a manually collected database with a large range of ages for each subject. This property makes AgeDB highly beneficial when training models for age progression experiments. Every image is annotated with identity, age, and gender attributes.
AgeDB-30, which is a subset of AgeDB, has been used for validation with MagFace and ArcFace, which are the selected FR pipelines in XAIface.

[1] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia and S. Zafeiriou, "AgeDB: The First Manually Collected, In-the-Wild Age Database," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1997-2005, doi: 10.1109/CVPRW.2017.250.

2.2.2. Labeled Faces in the Wild

Labeled Faces in the Wild (LFW)[2] is a public benchmark for face verification. It is a database of face photographs designed for studying the problem of unconstrained face recognition. The dataset contains more than 13,000 images of faces collected from the web. Each face has been labeled with the name of the portrayed person. The faces were detected by the Viola-Jones face detector[3]. LFW is often selected as a standard reference, however, for some identities only a small number of samples is provided and contains a relatively small proportion of women (according to authors).
LFW is used for validation of ArcFace and MagFace.

2.2.3. Cross-Pose LFW

The Cross-Pose LFW (CPLFW)[4] is an improved version of the LFW face dataset, where more pose variations of the same persons were added while keeping the same identities as in the LFW dataset. The CPLFW dataset is used to achieve face verification. The evaluation of multiple DL face recognition models on CPLFW showed that the accuracy drops by about 15%-20% compared to LFW (see Table 3.4a).
CPLFW is used for validation of ArcFace and MagFace.

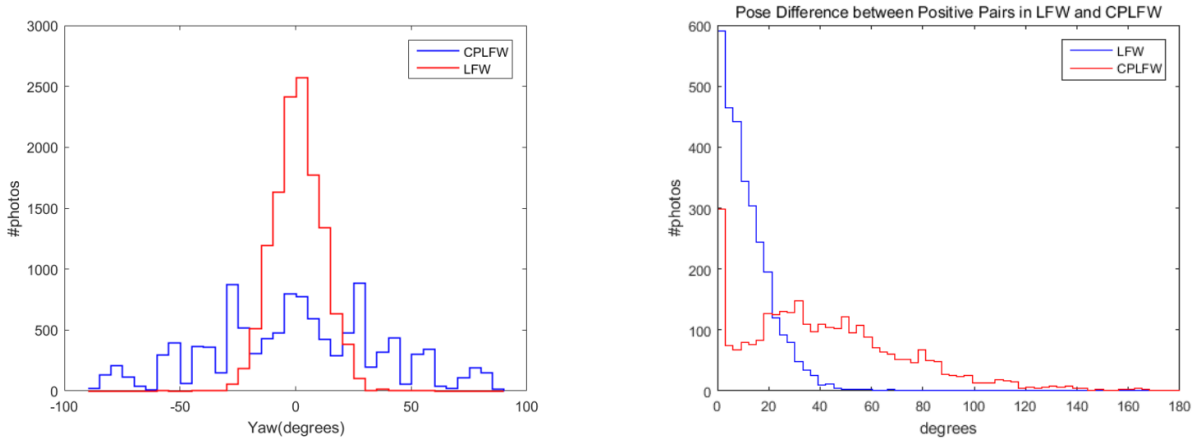


Figure 3.3a: Pose variation comparison between LFW and CPLFW.

2.2.4. Cross-Age LFW

The Cross-Age LFW (CALFW)[5] is an improved version of the LFW face dataset, where more face pairs with age gaps were added to add age variation and intra-class variance while keeping the same identities as in the LFW dataset. The CALFW dataset is used to achieve

[2] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.
[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, pp. I-I, doi: 10.1109/CVPR.2001.990517.
[4] T. Zheng and W. Deng, Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments, Beijing University of Posts and Telecommunications, Technical Report 18-01, February, 2018.
[5] T. Zheng, W. Deng, and J. Hu, Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments,CoRR, vol. abs/1708.08197, 2017.

face verification. The evaluation of multiple DL face recognition models on CPLFW showed that the accuracy drops by about 10%-17% compared to LFW (see Table 3.4a).
CALFW is used for validation of ArcFace and MagFace.

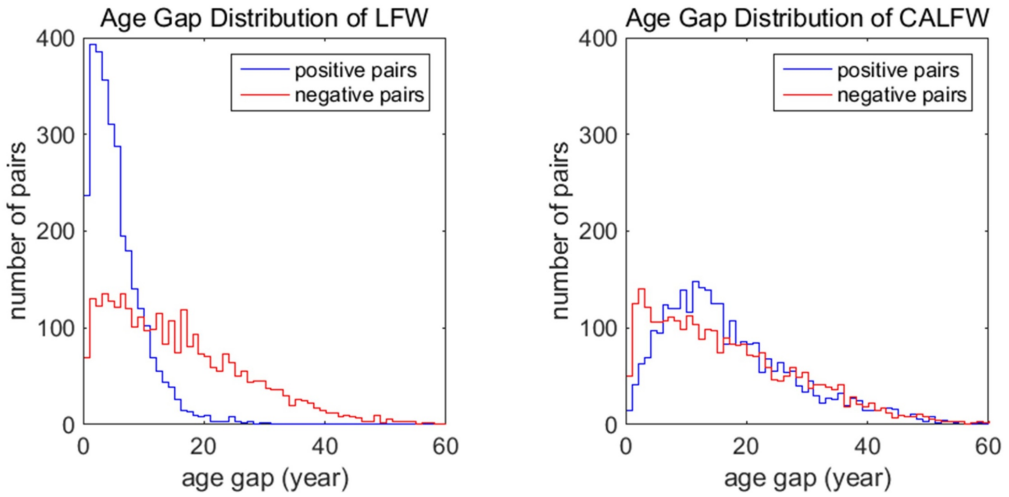


Figure 3.4a: Age gap comparison between LFW and CALFW.

Table 3.4a: comparison of verification accuracy (%) on LFW and CPLFW using ArcFace.

Method	**LFW**	**CPLFW**	**CALFW**
ArcFace	99.82%	92.08%	95.87%

2.2.5. DiveFace

DiveFace[6] is a dataset designed for bias analysis. It is obtained by extracting balanced sets of face images, according to gender and ethnicity, from the MegaFace database. MegaFace contains images from Flickr. Apparently the MegaFace has been recently decommissioned.
DiveFace contains annotations equally distributed among six classes related to gender and ethnicity (male, female and three ethnic groups). Gender and ethnicity have been annotated following a semi-automatic process. There are 24K identities (4K for class). The average number of images per identity is 5.5 with a minimum number of 3 for a total number of images greater than 150K. Although DiveFace is no longer available, it has been selected as an example of a balanced dataset that might be useful to recreate for experiments in XAIface.

2.2.6. The IARPA Janus Benchmark-C (IJB-C)

Despite the importance of rigorous testing data for evaluating face recognition algorithms, all major publicly available faces-in-the-wild datasets are constrained by the use of a commodity face detector, which limits, among other conditions, pose, occlusion, expression,

[6] A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana. SensitiveNets: Learning Agnostic Representations with Application to Face Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

and illumination variations. In 2015, the NIST IJB-A dataset, which consists of 500 subjects, was released to mitigate these constraints[7].

In 2017, IARPA Janus Benchmark-B (NIST IJB-B) database was released, a superset of IJB-A. IJB-B consists of 1,845 subjects with human-labeled ground truth face bounding boxes, eye/nose locations, and covariate metadata such as occlusion, facial hair, and skin tone for 21,798 still images and 55,026 frames from 7,011 videos. IJB-B was also designed to have a more uniform geographic distribution of subjects across the globe than that of IJB-A.

IJB-C[8], released in 2018, adds 1,661 new subjects to the 1,870 subjects released in IJB-B, with increased emphasis on occlusion and diversity of subject occupation and geographic origin with the goal of improving the representation of the global population. Annotations on IJB-C imagery have been expanded to allow for further covariate analysis, including a spatial occlusion grid to standardize the analysis of occlusion. Due to these enhancements, the IJB-C dataset is significantly more challenging than other datasets in the public domain and will advance the state of the art in unconstrained face recognition.

IJB-C has been used for evaluation of ArcFace and MagFace[9].

Note: It is well known, that IJB-C shows gender- and skin tone-wise bias[10].

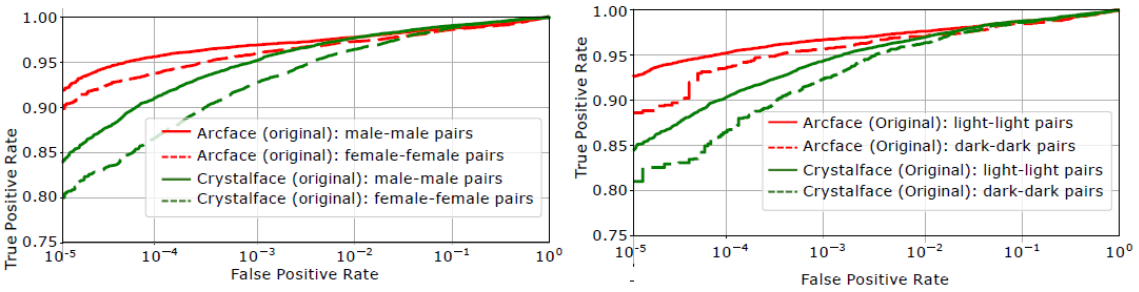


Figure 3.6a: IJB-C shows gender- and skin tone-wise bias.

Anyway, the authors took care to select a large variation of “geographic regions” and did not use “celebrity-only” media. Amazon Mechanical Turk has been used to get good metadata

[7] C. Whitelam et al., "IARPA Janus Benchmark-B Face Dataset," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 592-600, doi: 10.1109/CVPRW.2017.87.
[8] B. Maze et al., "IARPA Janus Benchmark - C: Face Dataset and Protocol," 2018 International Conference on Biometrics (ICB), 2018, pp. 158-165, doi: 10.1109/ICB2018.2018.00033.
[9] Q. Meng, S. Zhao, Z. Huang and F. Zhou, "MagFace: A Universal Representation for Face Recognition and Quality Assessment," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14220-14229, doi: 10.1109/CVPR46437.2021.01400.
[10] Prithviraj Dhar u. a., „PASS: Protected Attribute Suppression System for Mitigating Bias in Face Recognition“, 2021, 15087–96, https://openaccess.thecvf.com/content/ICCV2021/html/Dhar_PASS_Protected_Attribute_Suppression_System_for_Mitigating_Bias_in_Face_ICCV_2021_paper.html.

(occlusion, facial hair, gender, capture environment, skin tone, age, and face yaw) so that it should be possible to mitigate and analyze bias-issues.

2.2.7. CASIA-WebFace

The CASIA-WebFace[11] is the second largest public dataset available for face verification and recognition problems. This database is used for face verification and face identification tasks and any individual or group is allowed to use this database for educational or non-commercial use free of charge. The face images in the database are crawled from the Internet, more specifically from IMDB by the Institute of Automation, Chinese Academy of Sciences (CASIA). Image collection and identity annotation have been performed following a semi-automatic process. The dataset contains 494,414 face images of 10,575 real identities. This database has been used for training MagFace and the trained model is available online.

2.2.8. MS1MV2 (cleaned version of MS1M, provided by InsightFace)

The MS1MV2 is a refined version of the MS-Celeb-1M[12]. This large-scale database is used for training face recognition systems and even though the official dataset is no longer available, trained models are public to all internet users. The original images present in the database were collected from the Internet and the subjects collected were selected according to their popularity on the web. MS1MV2 consists of 5.8M images of 85K different identities. It has also been used to train MagFace and the trained model is publicly available.

2.2.9. FairFace

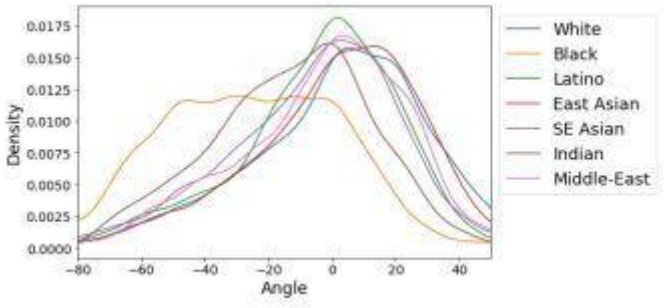


Figure 3.9a:. Individual Typology Angle (ITA), i.e., skin color, distribution of different races measured in our dataset.

FairFace[13] is a dataset focused on race balance for bias estimation. In order to mitigate the race bias, the authors emphasize a balanced race composition in the dataset by defining 7 race groups: White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino,

[11] YI, Dong, et al. Learning face representation from scratch. arXiv preprint arXiv:1411.7923, 2014.
[12] Guo, Yandong, et al. "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition." European conference on computer vision. Springer, Cham, 2016.
[13] KÄRKKÄINEN, Kimmo; JOO, Jungseock. Fairface: Face attribute dataset for balanced race, gender, and age. arXiv preprint arXiv:1908.04913, 2019.

and ensuring, as shown in Figure x, equal representation. The images were collected from the YFCC-100M Flickr dataset[14] and labeled with race, gender, and age groups thus making possible a bias estimation for all 3 categories. FairFace contains 108,501 images not currently available and just pretrained models on this dataset are still publicly available.

2.2.10. FairFaceRec

The FairFaceRec dataset is a superset of the IJB-C[15] dataset created for ChaLearn challenge. The participants of this challenge were asked to develop fair face verification methods aiming for a reduced bias in terms of gender and skin color. The new superset consists of 13k images from 3k new subjects along with a reannotated version of IJB-C (140k images from 3.5k subjects), totaling ~153k facial images from ~6.1k unique identities. The new database was annotated for gender and skin color as well as for age group, eyeglasses, head pose, image source and face size. Although DiveFace is no longer available, it has been selected as an example of a balanced dataset useful in future experiments in XAIface.

2.2.11. WIDERFace

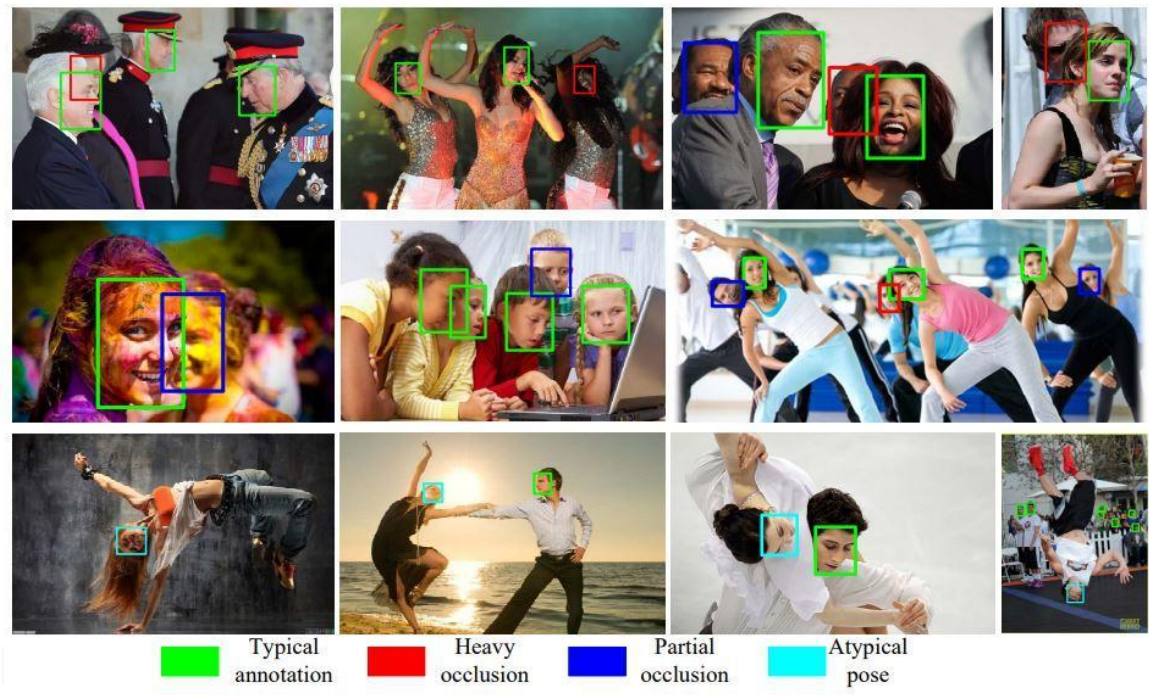


Figure 3.11a:. Examples of annotation in WIDER FACE dataset (Best view in color).

[14] THOMEE, Bart, et al. YFCC100M: The new data in multimedia research. Communications of the ACM, 2016, vol. 59, no 2, p. 64-73.
[15] MAZE, Brianna, et al. Iarpa janus benchmark-c: Face dataset and protocol. En 2018 international conference on biometrics (ICB). IEEE, 2018. p. 158-165.

WIDERFace[16] is a database designed for face detection purposes. It contains rich annotations, including occlusions, poses, event categories, and face bounding boxes and it is composed of 32,203 images, labeling 393,703 faces with a high degree of variability in scale, pose and occlusion. The authors suggest a dataset division into training (40%), validation (10%) and testing (50%) sets. Although the database is not publicly available anymore, XAIface members have had access to it. Furthermore, several pre-trained models can be found on the Internet.

2.2.12. VIP_attribute_extended (extended by EURECOM)

The VIP_attribute is a dataset composed of facial images, annotated for gender, body height, weight and BMI which has been used to prove that facial images contain discriminatory information pertaining to those traits. The database is publicly available under request to the authors and consists of mainly frontal face images of celebrities (mainly actors, singers and athletes) collected from the WWW. It contains one image of each of the 1026 subjects enrolled in it, specifically 513 female and 513 male celebrities. The VIP_attribute_extended is an extension of the annotation of the VIP_attribute database performed by EURECOM. The original database was extended by adding for every subject annotations of their hairstyle, presence and type of facial hair and presence of glasses thus making possible further studies of those categories.[17]

[16] YANG, Shuo, et al. Wider face: A face detection benchmark. En *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 5525-5533.
[17] DANTCHEVA, Antitza; BREMOND, Francois; BILINSKI, Piotr. Show me your face and I will tell you your height, weight and body mass index. En 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018. p. 3555-3560.

3. Addressing ethical and legal issues

The legal aspects regarding the choosing of an image data set for the training and evaluation of artificial intelligence facial image recognition technology are manifold and are largely still being discussed in the literature. All of these legal aspects have been addressed and can be referred to in "**Annex I Legal Aspects (Image Dataset)**" which is part of D 3.2.

4. Conclusions

Nowadays, numerous large-scale face image datasets have been proposed for different tasks such as automated face detection, alignment, recognition, generation, modification, and attribute classification. When choosing the most appropriate database for face recognition tasks, different criteria should be considered. Face databases, when used for face recognition, need to be annotated by identity although other traits such as gender, age or ethnicity might be useful since they can help recognize the subject by acting as auxiliary tasks. The database selected should also consist of at least an amount of annotated face images to successfully train and test deep learning models. Moreover, there exist two types of tasks for face recognition. On the one hand, we have the so-called face *verification*, which is to determine whether two given face images belong to the same person. On the other hand, there is face *identification* where given a face image in the query set the model seeks to find the most similar faces in the gallery image set. According to the selected task, the database needs a different type of annotation. Despite the vast amount of available data, existing public face datasets might also be biased toward a specific category, causing other groups to be significantly underrepresented. This means the model may not apply to some subpopulations and its results may not be compared across different groups without calibration.

All the above-mentioned factors raised the need of creating selection criteria for database choosing. In Section 2, the database selection process as well as a description of the databases chosen by the consortium is depicted. Likewise, in Section 3 as well as in the Annex document to this deliverable, guidelines on how to choose data sets in a GDPR-compliant manner are presented.