chist-era

XAIface

Measuring and Improving Explainability for AI-based Face Recognition

Module-level explainability report (V1)

Deliverable number: D5.2

Version: 2.0

EURECOM Sophia Antipolis

JOANNEUM RESEARCH DIGITAL

universität wien

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/

Deliverable

Editor:	[responsible partner name]
Deliverable nature:	Report (R);
Dissemination level: (confidentiality)	Public (PU);
Contractual delivery date:	June 30, 2023
Actual delivery date:	20 July 2023
Keywords:	Module-level explainability, performance assessment, integration
Author(s): (names and affiliations)	Chiara GALDI, EURECOM; Nelida MIRABET-HERRANZ, EURECOM; Martin Winter, JRS, Yuhang LU, EPFL

Short abstract

Deliverable D5.2 “Module-level explainability report” collects the descriptions of the explainability modules to be integrated into the XAIface pipeline. The document serves in particular to guide integration into the pipeline by identifying and resolving possible incompatibilities between modules.

Table of content

Abbreviations

AI	Artificial Intelligence
FR	Face Recognition
XAI	Explainable Artificial Intelligence
FMR	False Match Rate
DEX	Deep EXpectation
LIBF	Locally Interpretable Boosted Features

Executive summary

Deliverable D5.2 – “Module-level explainability report” – reports on the development and the performance evaluation of the explainability modules to be integrated into the XAIface face recognition (FR) pipeline. The deliverable will be published in two versions, one in month 26 and another one in month 34.

The first version (v1) summarises the work of the consortium towards the integration of the developed explainability techniques presented in deliverable D4.1 – “Explainability protocol and methods” – into the FR pipeline developed under WP5. This includes the selection and the description of the methods to be actually integrated into the pipeline and the mapping (as well as the potentially required adaptations) of the explainability techniques into the modules that will build the final FR pipeline.

Version 2 (v2) of this document will include the module-level performance evaluation of the explainability solutions proposed.

1. Introduction

The objective of deliverable D5.2 “Module-level explainability report” is to plan, report, and track the consortium’s work towards the integration of the explainability techniques into modules of the FR pipeline. In deliverable D4.1 “Explainability protocol and methods”, the consortium reported several explainability techniques, developed by both consortium researchers and other researchers from outside the project. Among the presented techniques in D4.1, a subset of methods will be selected for the final integration into the XAIface FR pipeline.

This deliverable is published in two versions, one in month 26 and another in month 34 of the XAIface project. The second version, in addition to collecting the description of the progress of the explainability module development, will also include the module-level performance assessment. There the modules will be evaluated separately (as opposed to the performance evaluation that will be reported in deliverable D5.3 (v2) where the overall performances of the end-to-end system will be reported instead of).

This document contains the following sections:

- Section 1: Introduces the general motivation for the document;
- Section 2: Provides an overview of the mapping of the selected explainability techniques into the FR pipeline;
- Section 3: Provides a description of the explainability techniques in each system module;
- Section 4: (In version 2 of the document) reports on the performance assessment of the explainability modules;
- Section 5: Highlights any possible conflict in the integration of the different modules, for example, if their integration into the FR pipeline requires a major modification or the exclusion of other techniques.

2. Explainability Tools to System Modules Mapping

This section provides an overview of the mapping of the selected explainability techniques into the FR pipeline.

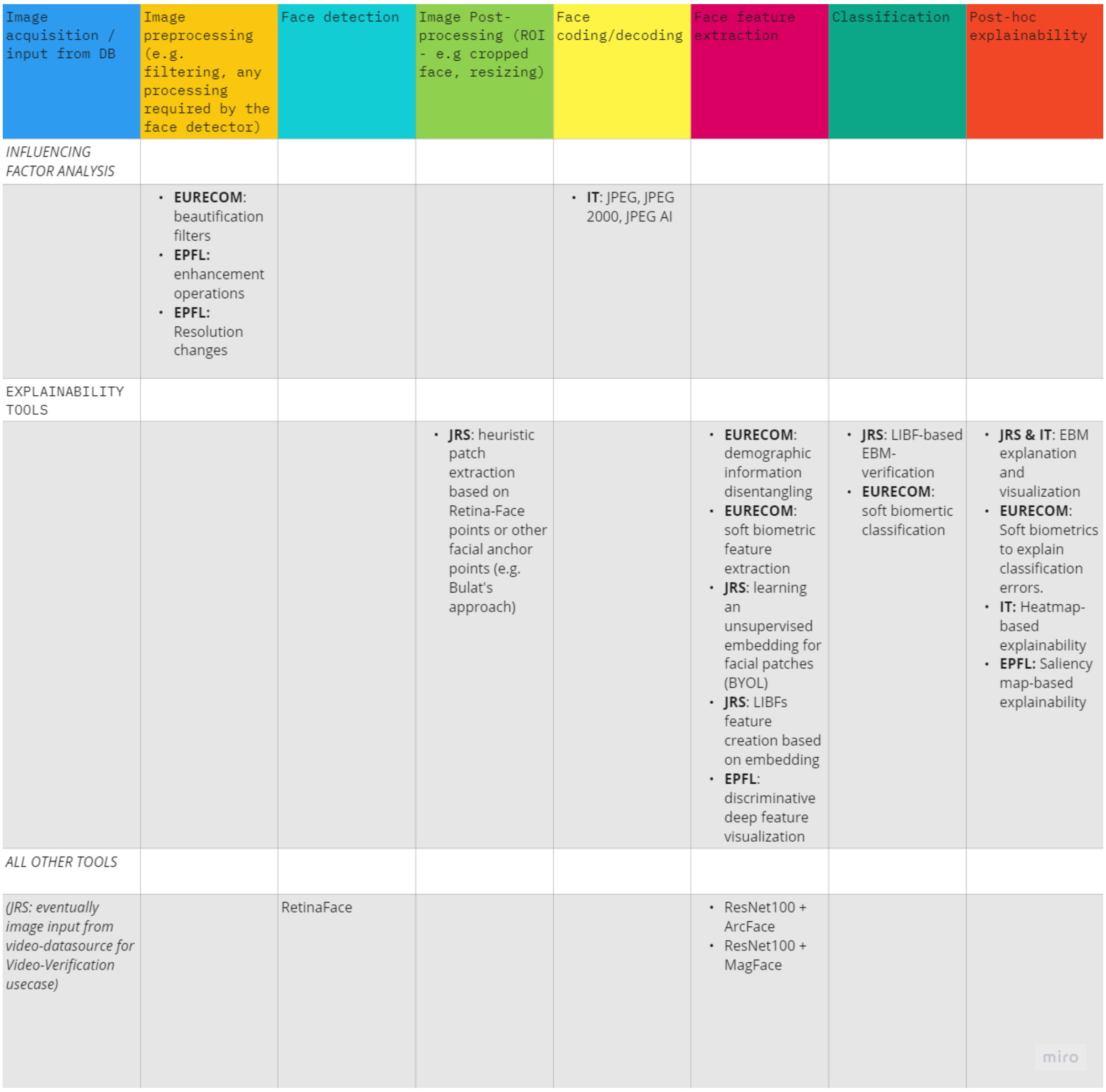

Image acquisition / input from DB	Image preprocessing (e.g. filtering, any processing required by the face detector)	Face detection	Image Post-processing (ROI - e.g cropped face, resizing)	Face coding/decoding	Face feature extraction	Classification	Post-hoc explainability
INFLUENCING FACTOR ANALYSIS							
	• **EURECOM**: beautification filters • **EPFL:** enhancement operations • **EPFL:** Resolution changes			• **IT**: JPEG, JPEG 2000, JPEG AI			
EXPLAINABILITY TOOLS							
			• **JRS**: heuristic patch extraction based on Retina-Face points or other facial anchor points (e.g. Bulat's approach)		• **EURECOM**: demographic information disentangling • **EURECOM**: soft biometric feature extraction • **JRS**: learning an unsupervised embedding for facial patches (BYOL) • **JRS**: LIBFs feature creation based on embedding • **EPFL**: discriminative deep feature visualization	• **JRS**: LIBF-based EBM-verification • **EURECOM**: soft biomertic classification	• **JRS & IT**: EBM explanation and visualization • **EURECOM**: Soft biometrics to explain classification errors. • **IT:** Heatmap-based explainability • **EPFL:** Saliency map-based explainability
ALL OTHER TOOLS							
(JRS: eventually image input from video-datasource for Video-Verification usecase)		RetinaFace			• ResNet100 + ArcFace • ResNet100 + MagFace		

Figure 2.a: Explainability Tools to System Modules Mapping. This table shows the overall FR pipeline structure and their explainability related mapping.

Two FR implementations have been initially selected for XAIface, namely using ArcFace and MagFace, which had their own implementation structure and consequently their structure served as a template for the implementation of the FR pipeline (without explainability mechanism). More details about this can be found in Deliverable D5.1 "End-to-end System Design and Assembling (v1)". Please note, that this initial version of the FR pipeline was thus used for the first part of the project, for example to study the impact of the influencing factors and to develop some of the explainability techniques.

As for now, the FR pipeline needs to be updated to include also the explainability mechanisms developed in XAIface. Figure 2.a shows a complete overview of the modules of the FR pipeline to be developed and integrated in the XAIface demonstration pipeline, including not only the explainability tools but also all other required tools and the analysis framework to perform performance evaluations XAIface. This table helps to understand the difference between the work carried out in WP2 and WP4 and the integration in the final FR pipeline in WP5.

In the following we shortly describe the individual modules in more detail.

Influencing Factor Analysis

The explainability process starts by identifying the AI-based face recognition influencing factors, understanding their impact on the overall performance of AI-based facial recognition systems. The work related to these studies has been undertaken in WP2. Figure 2.a illustrates where the Influencing Factor Analysis takes place: in Image Processing and Face Coding/Decoding modules.

Although this analysis will be featured in the final demo developed to show the XAIface results, the influencing factors are not ***"tools"*** to be integrated into the XAIface FR online-demonstration pipeline (although their study indirectly contributes to explainability).

Explainability Tools

For a detailed description of the XAIface contributions to AI-based Face Recognition Explainability, the reader is referred to Deliverable D4.1. The ultimate goal of XAIface is to develop new and collect existing tools for face recognition explainability. As a proof of concept, a face recognition pipeline will be implemented and will integrate some of these tools. In this document, the selection of tools and their mapping in the FR pipeline are reported.

As it can be seen, the modules where explainability tools will be integrated are: Image Post-Processing, Face Feature Extraction, Classification, and Post-Hoc Explainability.

As can be seen, the implementation of some explainability tools can have an impact on several pipeline modules. For example, the study of soft biometrics requires a modification of the pipeline at the feature extraction level, in order to extract the information related to soft biometrics classification, which is different from identity classification, and the result will be presented in the post-hoc explainability module.

All Other Modules

The entire FR pipeline will also integrate other tools that serve for recognition but do not necessarily contribute to FR explainability. We do not explicitly mention them here, and thus in the following, the process of integrating the explainability tools into the modules that make up the FR pipeline will be described in detail.

3. Module-level Explainability Methods

In this section, we describe the explainability tools that will be integrated into the final FR pipeline. In particular, we explain which modules will be integrated into the pipeline. If applicable, we explain whether they impact more than one module and how they will contribute to the explainability of face recognition

It is worth noting that this is different from what has been described in the previous deliverables. In fact, previously each tool was developed separately by each project partner without any constraints. In this case instead, each project partner had to think about how to integrate its solution into the FR pipeline in harmony with all the other tools.

3.1. Image Postprocessing

3.1.1. Heuristic patch extraction

This module serves as the basic step for explainability using Locally Interpretable Boosted Features (LIBF) where the importance of distinctive regions in the face are highlighted. In particular the algorithm uses five natural anchor points (facial landmarks) in the face (namely eyes, nose and corners of the mouth) as base patches and aligns six additional heuristically selected patches - namely (left and right) forehead, cheek and chin - to capture other potentially important parts of the face. The position and size of the additional patches are determined based on the position of the base patches (such as distance between left eye - right eye or mouth - nose) to be independent of resolution, scale and rotation of the face. Hence in total 11 patches are extracted from a face image. In the case of extreme rotation around the vertical axis (e.g.) causing invisible/hidden patches, this information is stored as an additional feature (binary information) for subsequent training and verification modules.

3.2. Face Feature Extraction

3.2.1. Demographic information disentangling

Demographic information disentangling can indirectly contribute to face recognition explainability. In fact, by removing, for example, the gender information from a feature vector extracted from a face picture, it is possible to assess the relevance of this information for the FR systems under evaluation.
This technique will be integrated into the XAIface FR pipeline at the feature extraction level. Just after extracting the features, according to any of the selected algorithms, the feature vector can be further processed in order to disentangle the demographic information.

3.2.2. Learning an unsupervised embedding for face patches

In order to specifically tune the heuristic face-part-patches extracted (see Heuristic patch extraction), to the nature of the problem and coevally minimising the number of dimension of the ideally compact feature vectors, we propose to use a self-supervised technique for generating specific representations (embeddings) without the need for annotated and labelled data. Thus this module implements an unsupervised technique for finding such embeddings based on a two-stream competitive neural network architecture (Siamese network) learning from various augmented views of single images only.

3.2.3. LIBF-feature creation on learned embedding

In order to create an entire, explainable representation of the face (the LIBF-features) we project each face patch of a novel face to the embeddings learned in Learning an unsupervised embedding for face patches (JRS) and obtain a 16-dimensional feature vector for each patch plus the binary information on visibility of face patches. The individual representations are simply stacked and per-patch normalised to form a 11 x 17 = 187 dimensional description for faces.

3.2.4. Discriminative deep feature visualisation method for face recognition

The main goal of this method is to investigate the connection between the deep face representation captured by the feature extractor and the original input face image. It introduces an additional face reconstruction module to alternatively study the affinity of deep features and a feature-reconstructed face. Moreover, a saliency explanation algorithm has been developed to visualise the connection, which forward-propagates the channel-wise deep features to the face reconstructor and produces heat-maps that are capable of highlighting the most discriminative regions of input faces. The heatmaps can be further disentangled into similarity map and dissimilarity map to provide more detailed explanations. This method requires an end-to-end training along the original ArcFace/MagFace face recognition pipeline and an additional face reconstruction module.

3.2.5. Soft biometrics feature extraction

In order to implement the explainability method proposed in Section 3.4.1, different soft biometrics need to be computed from the probe and gallery face images. The soft biometrics traits proposed are gender, age and weight. The estimation of those traits is based on deep learning models that automatically learn the extraction of the relevant features for a successful human trait estimation. All the models are built through a Convolutional Neural Networks though the specific structure implemented in each case varies. In the case of gender estimation, the open-source cvlib Python library is used, which contains an AlexNet model specifically trained for gender classification on the Adience dataset. Deep EXpectation (DEX) is selected as a model for apparent age estimation based on an ensemble of 20 VGG-16 architectures pre-trained on ImageNet. It extracts predictions from the ensemble of 20 age estimator networks from the subject's cropped face without explicitly using facial landmarks. Finally, the features extracted for the weight estimation task, are learned using a ResNet architecture with 50 layers.

3.3. Classification

3.3.1. LIBF-feature based EBM verification

The usage of LIBF-features for solving the verification procedure needs a specific adaptation of the matching-vector to provide both the query and the template LIBF representation at the same time. Therefore this module encodes the comparison of the 187-dimensional LIBFs for query (L_q) and template features (L_t) directly in the feature representation resulting in a 374 dimensional verification-feature. For the training the label (match or no match) is encoded as [0, 1] and thus the EBM is able to provide the user with a probability for correct classification.

3.3.2. Soft biometric classification

The features extracted in Section 3.2.5, different for each soft biometrics estimation task, are passed through a final classification layer that delivers the final trait prediction for the probe and gallery images. The output layer in the gender prediction network is of softmax type with 2 nodes indicating the two classes “Male” and “Female” while the proposed weight network includes a final regression layer for this trait´s estimation. DEX, the proposed age estimator network, works with an ensemble of 20 networks each one extracting its own features. When the network is trained for regression, the output layer consists of a single neuron while when training for classification, the output layer is adapted to 101 output neurons corresponding to natural numbers from 0 to 100, the year discretization used for age class labels. The final prediction is based on a weighted average of each network´s forecast.

3.4. Post-Hoc Explainability

3.4.1. Soft biometrics as an explainability method for classification errors

Previous research [1] has analysed the False Match Rate (FMR) of face recognition for comparisons in which one or both of the images are gender-misclassified. In our pipeline, we want to give more insights into how the difficulty to estimate certain traits from images can be related to a False Acceptance and a False Rejection output by an FR system. More specifically, given a pair of images and a FR model, in the case of False Acceptance, we want to further investigate the possible cause of the misclassification. Thus, we estimate different soft biometric traits such as gender, age, or weight and determine if the probe and the gallery images have such estimated traits in common. Analogously, when the FR model delivers a False Rejection for a pair of images, different soft biometrics will be computed and the study will focus on the differences between the estimated traits for each image.

3.4.2. EBM explanation and visualisation using the LIBF-features

In a first version, the “importance” of each entry of the 374 dimensional verification-feature as well as binary verification-feature correlation (pairs) are visualised by this module using standard visualisation techniques from Microsoft’s InterpretML framework. Thus it is possible

to see the importance of single features and pairs between the source and query image for forming an optimal classifier (global explanation), as well as visualising their contributions for a specific decision thus allowing for a basic insight and explanation of the verification process.
In future versions of this module it is planned to provide improved visualisations e.g. overlaying semi-transparent coloured regions to the original image to increase the explainability experience.

3.4.3. Correlation-based RISE algorithm for explainable face recognition

This technology aims to provide explanation saliency maps as a way to interpret the decision of the deep face recognition system. In general, this explanation method depicts facial regions that the deep face recognition system believes are similar and dissimilar between two given faces via the produced saliency maps. Then, we can leverage the saliency maps to analyse why the FR system believes two given facial images are a good matching or not, why the FR system believes the faces match even when they are occluded or heavily compressed, and why the FR system fails to give correct predictions in specific scenarios. This technology is model-agnostic, post-hoc and it runs independently from the face recognition pipeline.

4. Module-level Explainability Assessment

This section will report the module-level explainability assessment in version two of the deliverable.

5. Explainability Tools Interactions Analysis

In the first version of this document, this section makes a preliminary analysis of the face feature extraction explainability modules and post-hoc explainability tools proposed by the consortium. More specifically, their condition of usage, the potential interference with the FR pipeline, and possible conflicts with other explainability methods are summarised and analysed.

Explainability Modules for Face Feature Extraction

The explainability modules in face feature extraction level mainly focus on explaining the face embeddings.

The demographic information disentangling method aims at concealing the soft-biometrics traits by directly studying the deep face representation. This technique can be integrated into the FR pipeline right after the feature extraction module and before the classification module. The analysis of deep features relies on a separate autoencoder network and will not alter the original representation. Thus, it can be applied independently after the feature extraction and will not affect the original FR pipeline.

The discriminative deep feature visualisation method aims to investigate the connection between the deep face representation and the facial image. It leverages an additional face reconstruction network to build the mapping between the deep feature and specific facial regions. To use this method, the original FR pipeline needs to be trained together with the reconstruction network in an end-to-end manner and there will be a minor modification which replaces the last fully-connected (max-pooling) layer. Therefore, this approach interferes with the pre-defined FR pipeline and may also conflict with other explainability approaches due to a possible change of deep feature distribution. Solving this conflict-situation is currently under research by the consortium and will be described in the next version of this deliverable.

The BYOL method learns face-patch representation in a self-supervised manner and the coefficient of the embedding forms the explainable LIBF features of the face. Currently, this module does not directly use ArcFace/MagFace pipeline, but it is possible to adapt and explain an ArcFace/MagFace pipeline when the verification process is good enough. On the other hand, this approach can also be regarded as a parallel approach in addition to the current FR pipeline. This will be discussed in the next version of the document.

Post-Hoc Explainability Modules

As for post-hoc explainability modules, multiple explanation approaches have been proposed to interpret the deep model's decision via saliency maps and visualise the soft biometrics.

The correlation-based RISE algorithm is a perturbation-based "black-box" explanation method. It recursively forwards perturbed images to the FR pipeline and measures the impact on the predicted similarity score in order to estimate the saliency importance map. This method is model-agnostic and does not require access to the intrinsic network architecture nor the gradient information. It is often applied after the FR pipeline and will not affect the FR pipeline or other explainability tools.

Explainable soft biometrics estimation method provides explanations by predicting and comparing different traits such as gender, age, and weight of the input facial images. It can be switched on if the recognition fails and the disparities between the soft biometrics information will be analysed to provide insights for the mismatching. This technique is employed after the FR pipeline, so it will not affect the recognition process nor other explainability methods.

The EBM explanation and visualisation method leverages the internal visualisation of Microsoft's InterpretML framework. Currently, the visualisation approach is used for EBM-based explanation methods. However, it can be also adapted to other recognition pipelines. The generalised use cases of the visualisation method will be discussed in the next version of the document.

In conclusion, various explainability modules have been proposed by the consortium, which can be categorised into two groups. One type of method works independently by directly studying the extracted features or the classification results of the current ArcFace/MagFace pipeline. The other type adopts a new face recognition pipeline and learns new feature embeddings. Although the latter conflicts with the currently selected FR pipeline, they can be regarded as in-parallel approaches. This issue about parallel XFR frameworks will be researched in the future and there will be a detailed description in the next version of this document.

6. Conclusions

In this document we summarised the XAIface consortium work towards the implementation of the XAIface FR pipeline, which will include the explainability mechanisms selected and developed by the project partners. In particular this document highlighted the organisation and design of the FR *modules*, with the aim of better planning the consortium's collaborative work for the actual implementation of the FR pipeline.
In summary, at this stage we studied in depth possible overlapping, synergies, incompatibilities, and in general any sort of issue or consideration beneficial for the integration of the different tools in a single FR pipeline.

This effort has allowed the consortium to better understand each partners' point of view and converge towards a common solution that will enable the XAIface project to move forward its successful completion.

References

[1] GBEKEVI, Afi Edem Edi, et al. Analyzing the Impact of Gender Misclassification on Face Recognition Accuracy. En *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023. p. 332-339.