D9.7 Benchmarking and evaluation methodology 1

Document Due Date:	28/02/2021 (M18)
Document Submission Date:	08/04/2021

Work Package 9: Field/Operational Environment Testing

Document Dissemination Level:
Public

D4FLY - **D**etecting **D**ocument frau**D** and i**D**entity on the fly
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833704

Abstract

This deliverable provides the initial specification of the D4FLY benchmarking framework whose score will be continuously updated through the entire lifetime of the project.

More specifically, an overview of the evaluation/validation events and activities is included. The overview will report on current benchmarking scores that will be achieved throughout the benchmarking process alongside with all improvements and updates that will be achieved during the project lifecycle.

The initial benchmarking score (before the first evaluation event), is formed based on two factors:

1) an acceptance protocol and criteria provided by end users within the task related to user requirements of WP2 (Task 2.2), and
2) recorded benchmarking scores achieved within relevant well-respected worldwide grand challenges (i.e. MBGC [75], MBE [76], FRGC [77], FRVT [78], FERET [79]).

The corresponding publicly available databases of the grand challenges (along with the disjoint development and evaluation sets) have guided the evaluation methodology of the D4FLY benchmarking framework.

In order to evaluate the biometric systems that have been developed under D4FLY, two benchmarking events are planned. These events are designed based on different methods types of feedback gathering and targeted results. The interpretation of these results should drive the extraction of guidelines for iterative system design.

Project Information

Project Name	Detecting Document frauD and iDentity on the fly
Project Acronym	D4FLY
Project Coordinator	Veridos GmbH
Project Funded by	European Commission
Under the Programme	Horizon 2020 Secure Societies
Call	H2020-SU-SEC-2018
Topic	SU-BES02-2018-2019-2020 Technologies to enhance border and external security
Funding Instrument	Research and Innovation Action
Grant Agreement No.	833704

Document Information

Document reference	**D9.7**
Document Title	**D9.7 Benchmarking and evaluation methodology 1**
Work Package reference	WP9
Delivery due date	28/02/2021
Actual submission date	31/03/2021
Dissemination Level	Public
Lead Partner	**NCSRD**
Author(s)	**Nikos Argyreas (NCSRD)**
Contributor(s)	**Stelios C. A. Thomopoulos (NCSRD)** **Lemonia Argyriou (NCSRD)**
Reviewer(s)	**Stelios C. A. Thomopoulos (NCSRD – Internal review),** **Antonios Danelakis (NTNU – Consortium member review),** **Dimitris Kyriazanos (NCSRD - SAB review)**

Document Version History

Version	Date created	Beneficiary	Comments
0.1	26/03/21	NCSRD	First consolidated version after internal review
0.2	29/03/21	NCRSD	Updates after second review
0.3	31/03/21	NCSRD	Updates to section 2
1.0	31/03/21	NCSRD	Final updates and approval at all levels

List of Acronyms and Abbreviations

ACRONYM	EXPLANATION
EC	European Commission
EU	European Union
D4FLY	Detecting Document frauD and iDentity on the fly
TPR	True Positive Rate
TNR	True Negative Rate
FRR	False Rejection Rate
FAR	False Acceptance Rate
FNMR	False Non-Match Rate
FMR	False Match Rate
ROC	Receiver Operating Characteristic
DET	Detection Error Trade-off
EER	Equal Error Rate
DoA	Description of Action

Table of Contents

1 INTRODUCTION

1.1 Background

D4FLY benchmarking framework targets the exploitation of biometric datasets for the development and evaluation of new biometric techniques. The D4FLY benchmarking framework seeks maximum portability so that algorithm implementers can use the operating system and development tools of their choice. The key elements of the framework are: (a) open-source software which consist mostly of portable scripts; publicly available biometric databases; well defined evaluation protocols; and (b) additional information (such as How-to documents) that allow the reproducibility of the proposed benchmarking experiments.

1.2 Aim of this document

D4FLY benchmarking framework is available for four biometric modalities: 3D face, iris, thermal-to-visible, and somatotype for performing unimodal evaluation and benchmarking, however, the availability of a multimodal dataset enables multimodal benchmarking as well.

1.3 Input / Output to this document

Input

Input to this deliverable constitutes the deliverables D2.2 (User Requirements) and D5.1 – D5.4 (Biometric on-the-move recognition methodologies), and the citations therein.

Output

The output of this deliverable is the initial D4FLY benchmarking framework. This deliverable serves as input for D9.13.

2 D4FLY Benchmarking Evaluation framework

2.1 D4FLY Benchmarking Framework for Biometric Verification Systems

This section is dedicated on the biometric verification procedure which is the most important to meet the main objectives of the D4FLY project (contactless, on-the-move biometric data capturing). Biometric verification (or authentication) compares data for the person's characteristics to that person's biometric "template" to determine resemblance, assuming that the reference model is captured and stored during the enrolment. The typical output of this procedure is Boolean (returning TRUE if the person's characteristics match the biometric template and FALSE otherwise), however the majority of biometric verification systems are producing as output a numeric value, the **verification score** which is an indicator of how well the captured biometric sample matches the biometric template (at most cases, the higher score means better matching), then the Boolean decision is up to a predefined **verification threshold** which is the basic parameter for such systems, so if the score exceeds this threshold (greater or equal) then we get a successful matching (TRUE), otherwise we get a non – successful matching (FALSE).

2.1.1 The reference architecture

The field of biometric research combines three different scientific areas to deal with a. Biometric sensors and their technology, b. Implementation of biometric verification algorithms and c. the integration of biometric sensors and algorithms into fully operational systems. For the construction of an integrated information system that aims at the recognition based on biometric data, the development should be approximately equal in all three areas. However, the D4FLY benchmarking framework is focused on the area of algorithmic implementation which means the evaluation and the benchmarking of biometric verification algorithms.

A generic algorithmic experiment involves the consumption of data from the algorithm where this data corresponds to distinct snapshots belonging to the problem domain, then the algorithm processes this data in order to produce the result which is usually a solution that belongs to the solution domain.

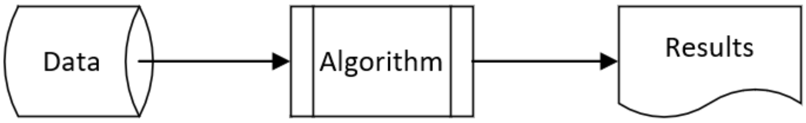

FIGURE 1: FLOW DIAGRAM OF A GENERIC ALGORITHMIC EXPERIMENT

In the most common case of approximation algorithms where the exact solutions are probably unavailable, the evaluation suite must provide the best-known solutions for each input so that the result of the algorithm can be compared with them based on predefined metrics. In this way, statistical evaluation of the algorithm will be possible by executing the algorithmic experiment for each distinct input and then extracting the appropriate statistical measurements (Mean value, standard deviation etc.).

However, biometrics could be seen as an example of the pattern recognition field where biometric algorithms are designed to work on biometric data and that point introduces a series of problems related to biometric datasets and usually, more data is always better from the point of view of pattern recognition. Another point that characterizes pattern recognition field is that at least two distinct datasets are needed in the different phases of the design of classifiers, the development dataset which is intended to be used during the implementation of the classifier, and the evaluation dataset which is intended to be used for the evaluation of its performance. In the most common case where the classifier is based on machine learning technologies the development dataset coincides with the training set and the evaluation dataset coincides with the test set, but because classifiers do not have to follow this methodology, the terms development and evaluation datasets will be used instead.

Moreover, the biometric recognition process is not a trivial pattern recognition process but is a rather complex one that requires the execution of separate procedures such as enrolment (registration) verification or/and identification where it is possible to involve multiple biometric samples for each one.

Since multiple biometric recognition technologies are used where each one of them is based at different sensing data and taking into consideration that D4FLY is targeted at both unimodal and multimodal biometric verification schemes, it is of course impossible to use just one evaluation metric. Each biometric modality introduces several performance and evaluation metrics with some of them being focused on the verification and some of them being focused on the identification procedure as well. In the most complex case of multimodal biometrics where some mathematical techniques for fusing biometric scores from multiple modalities and hence considerably boosting biometric recognition accuracy are being involved, novel evaluation metrics must be introduced. For all the above reasons the need to define a comprehensive evaluation protocol is imperative.

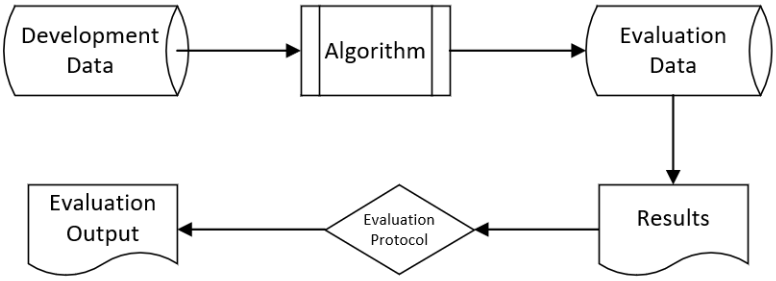

FIGURE 2: FLOW DIAGRAM OF A GENERIC BIOMETRIC EXPERIMENT

Thus, based on these characteristics, we come to a more complicated evaluation scheme that includes two separate datasets and a rather composite evaluation protocol whose characteristics we will analysed later in the text.

2.1.2 Biometric Datasets

The collection of biometric samples is rather a difficult task mostly related to the restrictions about personal data and although that the European Union has introduced a modern and comprehensive legal framework the General Data Protection Regulation (GDPR), the personal data protection laws are different between countries outside EU which may create incompatibilities between the biometric datasets.

Moreover, there are numerous cases of published work where researchers report their biometric experiments which could not be compared to other published work in the domain. Sometimes, the biometric databases that underlie the experiments are private databases but even in the case were publicly available databases are being used, the evaluation protocols are poorly defined and each publisher is free to choose a customized evaluation protocol because there are no standardized ones. There are also cases where the use of a highly customized evaluation is a way for the researcher for showing some particularities of the proposed system as well.

For all the reasons mentioned above, D4FLY benchmarking framework is providing a clear and solid evaluation protocol as well as specific (common) development and evaluation datasets to fully ensure the comparability of the results between the different algorithms to be benchmarked. However there is another major issue to be solved: Publicly available biometric databases does not mean "free" or "open source", most of them requires (even a) small license fee and their unconditional distribution is of course not permitted and although that a research team that intends to participate to the benchmarking event owns a licensed copy of the specific dataset the issue is still unresolved because this dataset should be divided into two parts where one will be the development and the other will be the evaluation dataset. This means that the dataset has to be altered by the D4FLY benchmarking framework, and the distribution of an altered version of this dataset to another entity is still illegal even if both entities (D4FLY consortium and the participant to the benchmarking event) are licensed for the original dataset as this is provided by its owner.

2.1.3 Development Datasets

For bypassing this incompatibility, within D4FLY four unimodal biometric datasets were created including 3D face, thermal face, iris, and somatotype by combining existing public datasets. These datasets are used internally for training and validation of the individual biometric verification technologies.

- The D4FLY 3D Face Unimodal Dataset combines several publicly available datasets with the aim of maximizing the number of identities, which has great influence on training and testing. In addition, all unique characteristics of each dataset are kept and combined, thus creating a more challenging dataset. The created D4FLY 3D face unimodal dataset contains six 3D facial mesh instances for each identity of the initial datasets, selected randomly. In total, there are 671 identities and 4,026 3D faces. Five out of the six instances per identity are randomly selected as the training set. The remaining instance belongs to the testing set. Furthermore, some identities are set aside and considered as unknown; in which case all six instances of the identity belong to the testing set. Due to the terms of use of the three initial datasets, the unified dataset cannot be directly re-distributed. Instead, the filenames or indices of the selected 3D faces are provided so that the unified dataset can easily be reproduced from the initial datasets by a third party.
- The D4FLY Thermal face Unimodal dataset.
- The Iris Unimodal Dataset.
- The Somatotype Unimodal dataset combines several publicly available datasets including 3,000 subjects in total. This dataset also contains six image instances for each identity of the initial datasets, selected randomly. In total, there are 2,939 identities (identities with less than six instances available were skipped) and 17,634 images. The resolution of the images is resized to 293x293 pixels which is the standard dimension of the Inception V3

Network input, which is expected to be employed for the training process. Once again, five out of the six instances per identity are randomly selected as the training set, while the remaining instance belongs to the testing set. Furthermore, the case where an identity is unknown is also taken into account (all six instances of the identity belong to the testing set in this case). Due to the terms of use of the four initial datasets, the unified dataset cannot be directly re-distributed. Subsequently, the filenames or indices of the selected images are provided, so that one can easily reproduce the dataset.

The publicly datasets, as being detailed referenced in the D5.6 are:

TABLE 1: PUBLICLY AVAILABLE DATASETS

Dataset	Modality
BU-3DFE	3D Face
Bosphorus	3D Face
FRGC v2.0	3D Face
CARL	Thermal-to-visible face
Laval	Thermal-to-visible face
TUFTS	Thermal-to-visible face
University of Notre Dame - UND X1	Thermal-to-visible face
ND-CrossSensor-Iris-2013	Iris
Face and Ocular Challenge Series (FOCS)	Iris
CASIA Iris-v4-distance	Iris
CASIA Iris-v4-thousand	Iris
The Notre Dame LivDet-Iris 2017	Iris
CUHK03	Somatotype
Market-1501	Somatotype
RGBD-ID	Somatotype
RAiD	Somatotype

All candidate competitors are to be advised to choose and acquire any of the proposed datasets mentioned in deliverable D5.6 (or combination of them based on their preferred biometric modalities) and use it as the Development Dataset. The Development Data set may be used for both training and testing the classifier following the methodology already described above (five out of the six instances per identity are randomly selected as the training set, while the remaining instance belongs to the testing set). Alternatively, a dataset partition of about 80% for training and 20% for testing can be used. Of course, this dataset is intended to be used for the internal performance testing and fine tuning for each competitor. It is also advisable for each competitor to deal with the fine-tuning process especially for the unimodal dataset due to its construction conditions (contactless, on-the-move biometric data capturing).

2.1.4 Evaluation Dataset

D4FLY will provide a new multimodal biometric dataset (see Figure 3) which consists of biometric data of the same individuals for the four modalities: 3D face, iris, thermal face, and somatotype while the biometric samples being captured using D4FLY paradigm (contactless, on-the-move). This dataset will be the evaluation dataset for the D4FLY benchmarking framework. It follows the same rules as the Development Data set, of course it is clearly smaller than that and contains different identities. However, the diversity rules are exactly the same on both datasets. Both datasets (Development and Evaluation) are being assembled having as a key priority the maintenance of the maximum possible diversity. The D4FLY multimodal dataset will be exploited for both unimodal and multimodal benchmarking, where at the unimodal case, the specific modality (subset/database view) will be initialized.

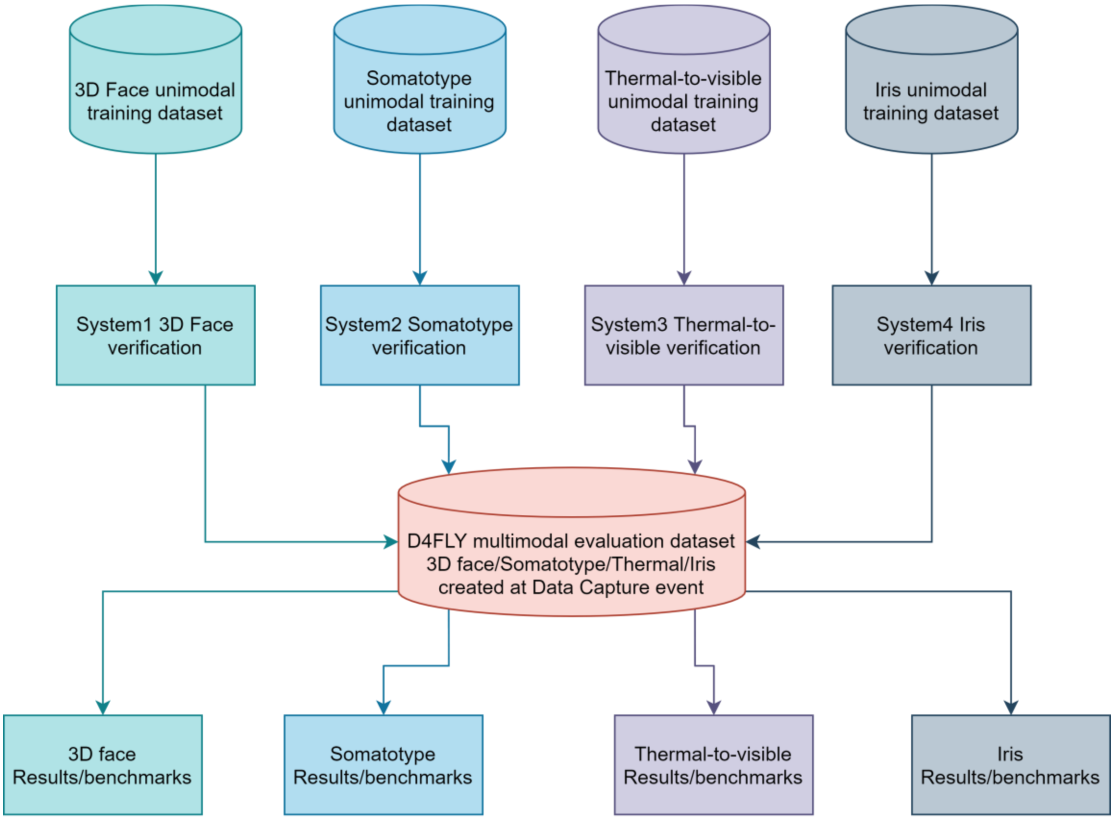
3D Face unimodal training dataset
Somatotype unimodal training dataset
Thermal-to-visible unimodal training dataset
Iris unimodal training dataset
System1 3D Face verification
System2 Somatotype verification
System3 Thermal-to-visible verification
System4 Iris verification
D4FLY multimodal evaluation dataset 3D face/Somatotype/Thermal/Iris created at Data Capture event
3D face Results/benchmarks
Somatotype Results/benchmarks
Thermal-to-visible Results/benchmarks
Iris Results/benchmarks

FIGURE 3: D4FLY EVALUATION DATASET

2.2 Evaluation Framework

The evaluation framework includes two distinct evaluation procedures:

- **Biometric Verification Accuracy Evaluation** which is the standard evaluation practice for the biometrics industry. It is done by performing many genuine and impostor comparisons and analysing produced similarity scores or match decisions. Furthermore, the Biometric Verification Accuracy Evaluation Metrics are more or less standardized. In the context of D4FLY benchmarking framework the utilized metrics are also proposed at ISO/IEC 19795-1:2006 specification.
- **Biometric Verification Performance Evaluation** which is also a key factor for the D4FLY paradigm (contactless, on-the-move). Execution time, memory consumption, storage requirements and network traffic requirements will be benchmarked for specific setups corresponding to the use cases of the D4FLY system for fulfilling the user requirements.

2.2.1 Biometric Verification Accuracy Metrics

A biometric verification system derives a match score by comparing biometric data from a person attempting to authenticate with enrolment data for the identity they claim. The closer the match, the higher the match score will be. If the match score exceeds a threshold, the person authenticating is accepted. If the threshold is set too high, genuine users will be rejected. If it is set too low, impostors will be authenticated. Ideally, the lowest score from a genuine user would be higher than the highest impostor score. Then the threshold would be set somewhere between the two.

For measure the performance of any binary classifiers including of course the biometric verification ones, two fundamental statistical metrics are used:

The **Sensitivity** (**True Positive Rate**) which measures the proportion of positives that are correctly identified.

$$TPR = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{number of false negatives}}$$

and the **Specificity (True Negative Rate)** measures the proportion of negatives that are correctly identified.

$$TNR = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{number of false positives}}$$

However, in reality, the genuine and impostor scores overlap. Then the system will generate a greater or lesser number of two types of errors. In order to quantify these errors, the **False Rejection Rate** (**FRR**) and the **False Acceptance Rate** (**FAR**) for a biometric device are defined as:

$$FRR = \frac{\text{Number of failed attempts at verification by authorized users}}{\text{Number of attempts at verification by authorized users}}$$

$$FAR = \frac{\text{Number of succesful verifications by impostors}}{\text{Number of attempts at verification by impostors}}$$

Both FAR and FRR depend on threshold. A higher threshold will generally reduce FAR, but at the expense of increased FRR, and vice versa.

At a sufficiently low threshold, few or no users will be rejected, so the FRR will be low. Most or all impostors will be accepted, so the FAR will be high. Then as the threshold is increased more genuine users will be rejected and less impostors will be accepted.

Receiver Operating Characteristic (ROC) curves is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

However, **Detection Error Trade-off (DET) curves**, which plot FRR on the y axis vs. FAR on the x axis, are preferred because they treat both types of error in the same way. If desired, the error rates can be plotted on logarithmic scales to cover a wider range of errors. These curves allow a relatively complete view of the characteristics of a biometric system.

At some point FRR and FAR will be equal. The value of the FAR and FRR at this point is the E**qual Error Rate (EER).** The equal error rate may be useful as a single value to allow comparison between different biometric verification systems. However, it can be misleading because systems will seldom be operated at the EER. In some cases, it will be more important to keep impostors out, even at the expense of rejecting authorized users, and in other cases it will be more important to avoid rejecting authorized users. Using a detection error trade-of curve, one can find the EER, the FAR corresponding to any desired FRR, or the FRR corresponding to any desired FAR.

The common accuracy evaluation metric in D4FLY (as defined in D5.1-D5.4 deliverables) is:

Accuracy rate: (TP+TN) / (TP+TN+FP+FN)

Where TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

Regarding Iris: Since the work was mainly focused on optimising the hardware setup, evaluation metrics are expected to be provided for the follow up deliverable (D9.13). Nonetheless, accuracy and processing time are to be expected.

2.2.2 Multimodal Evaluation Metrics

Two methods of combining multiple biometric authentication systems that in practice exhibit good performance will be considered, **sum rule** and **weighted majority vote**. Usually, the sum rule may be expected to outperform majority voting. However, the sum rule requires the extraction of a numeric biometric score for each modality (and the scores have to be normalized in order to be combined) and not just a Boolean (Hard) result.

Majority voting: This is the most common and intuitive method in which the input sample is assigned to that identity on which majority of the matchers agree.

- For odd number of matchers, the final result will be the result of the maximum matchers, i.e., results of at least half of (the total no. of matchers + 1) matchers.
- For even matchers, (this is the D4FLY case with four biometric modalities) final output will be the result of at least half of the total no. of matchers plus one. When the match and mismatch output come out from same no. of matchers, we can then choose any one from match or mismatch.

However, the recognition accuracy of different matchers varies significantly, so weighted majority voting approach is preferred for the D4FLY multimodal evaluation framework. In this approach different weights are assigned to the decision of different matchers. Higher weights are assigned to the decisions made by the more accurate matchers. The recognition procedure is almost similar to the majority voting approach procedure, except that the weights of the decision of individual matchers are also considered in this case.

The multimodal evaluation metrics can be the same as the ones used for benchmarking the verification accuracy, that is accuracy rate and performance metrics, such as execution time. The reason being, that the goal in the fusion case (multimodal case) is the same as in the unimodal case, i.e., the correct verification.

2.2.3 Biometric Verification Performance Basics

The performance measures that will be benchmarked are strictly bonded with the verification procedure. This implies that considering that we are focused at the algorithmic performance during this process, the evaluated part of the process is considered to begin after the capture of the biometric sample and the acquisition of the biometric prototype. Subsequently, some reference system architectures will be proposed regarding:

- Hardware specifications: CPU type, Memory (RAM) Size, GPU type (in case of massive parallel processing powered by high end GPUs), storage type together with their performance indicators.
- Networking Specifications: Connection type (wired/wireless), performance rating.
- Implementation Specifications: Compiler type and version, optimization levels etc.
- Generic Software Specifications: Host Operating System, Encryption Networking Protocol (if applicable)
- Generic Hardware Specification: Host Type (Server/Workstation or Mobile Device for on – the – go operation)

2.3 D4FLY Benchmarking Framework for Biometric Identification Systems

Biometric Identification procedure, in contrast with biometric verification which is a one – to – one matching procedure with a binary result, is a one – to many – procedure where the captured biometric sample for the specific person is compared with more than one biometric sample, usually stored into a biometric database. In short, the biometric verification answers the question “Are you who you claim to be?” while the biometric identification answers the question “Who are you?”

The obvious difference between these two methods is that the biometric identification requires much longer execution time than biometric verification. A relatively simplistic approach but not far from reality is that the identification based on a biometric database with N records (biometric samples) could be broken into N – verification procedures, thus the required execution time is multiplied by the size of the database, so it goes without saying that the feasibility of such a process is not a given and although it can be relatively easily implemented in a parallel processing environment it takes powerful (and of course very expensive) computer systems to deal with the vast processing effort. Such systems are usually available in police authorities where there are databases with biometric data of convicts and these data is being exploited for forensics purposes.

2.3.1 The major differences in relation to the verification process

The logic behind the identification process is not far from that applied to verification. The main difference is the definition of an additional, relatively small biometric database containing a small number of reference biometric sampled with which the matching will be made. The solution chosen is the creation of the “identification” database with existing data from the D4FLY unimodal and multimodal datasets, in this way, four unimodal and one multimodal “identification” databases are introduced.

- Each unimodal identification database contains a specific percentage of records relative to the source dataset (i.e., if this percentage is 5%, and the D4FLY 3D Face Unimodal Dataset contains 671 identities, then the corresponding identification database contains 33 identities)

- The same stands also for the unimodal “identification” database as well.
- The diversity of the identification data should be similar to that of the source data base. For this reason, the selection of the identities to be part of the identification database will be random but also representative of all possible sample categories shown in the original.

In accordance with the two basic biometric verification metrics (FAR and FRR) the two basic biometric identification metrics are:

- The **False Non-Match Rate (FNMR)**: The rate at which a biometric process miss-categorizes two captures from the same individual as being from different individuals. False Non-Match Rate Normally, measures the classification subsystem performance and False Reject Rate is an overall system.

- The **False Match Rate (FMR):** The rate at which a biometric process mismatches biometric signals from two distinct individuals as coming from the same individual.

- The **identification time**, i.e., the time it takes to converge to an identification output (either correct identification or missed) normalized by the size of the database against which the identification is performed.

3 BENCHMARKING EVENTS PLAN

Since the datasets have not been published yet, the evaluation events have not been finalized. Nonetheless, a preliminary planning gathering all potential evaluation events and corresponding feedback types has been sketched:

1. The first benchmarking event is the experiments performed by each partner on their corresponding unimodal datasets. Thus, the initial unimodal baseline, for each biometric modality will be provided.
2. Within the context of the Scenario 2 field-tests and demonstrator results for both the unified fused D4FLY biometric system as well as the individual biometric modalities will be extracted. Thus, the latter events are three additional (two field-tests and one demonstrator) potential benchmarking events that will provide updated results for both the unimodal and the multimodal case. The results will be recorded with respect to the evaluation metrics presented in Section 2 as well as to the feedback of the end-users that will participate.

4 Conclusion and Next steps

The D4FLY Benchmarking and Evaluation Framework for benchmarking unimodal and multimodal biometric verification and identification algorithms, devices and systems has been established using the D4FLY multimodal database, which can also be used for unimodal biometric testing. Widely used and accepted accuracy and performance metrics have been included in the benchmarking framework, alongside with instructions on how to conduct the benchmarking tests.

Baseline results will be reported in the next deliverable D9.13 Benchmarking and evaluation methodology 2.

D9.13 will also record the corresponding evaluation events chosen from the possible candidate events as mentioned in previous section.

Moreover, D9.13 will refer to any results of potential open challenges as well as the procedural format followed by those challenges.

REFERENCES

D4FLY-D2.2	D4FLY Deliverable D2.2 - Requirement Analysis Report
D4FLY-D5.1	D4FLY Deliverable D5.1 – 3D face recognition on-the-move 1
D4FLY-D5.2	D4FLY Deliverable D5.2 – Thermal-to-visible face recognition on-the-move 1
D4FLY-D5.3	D4FLY Deliverable D5.3 - Iris recognition on-the-move using LF tech 1
D4FLY-D5.4	D4FLY Deliverable D5.4 - Somatotype recognition 1
D4FLY-D5.6	D4FLY Deliverable D5.6 - D5.6: Biometric datasets
D4FLY-D9.13	D4FLY Deliverable D9.13 – D9.13: Benchmarking and evaluation methodology 2
Petrovska-Delacrétaz, D., et al. (2009)	Petrovska-Delacrétaz, D., Chollet, G. and Dorizzi, B., 2009. Guide to biometric reference systems and performance evaluation (p. 405). Berlin: Springer.