European Commission

JRC SCIENCE FOR POLICY REPORT

Study on Face Identification Technology for its Implementation in the Schengen Information System

Administrative Arrangement

JRC-34751

Galbally, J
Ferrara, P
Haraksim, R
Psyllos, A
Beslay, L

2019

EUR 29808 EN

Joint Research Centre

This publication is a Science for Policy report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

Contact information
Name: Laurent Beslay
Address: Joint Research Centre, Via Enrico Fermi 2749, 21027 Ispra, Italy
E-mail: laurent.beslay@ec.europa.eu
Tel.: +39 0332 78 5998

EU Science Hub
https://ec.europa.eu/jrc

JRC116530

EUR 29808 EN

PDF ISBN 978-92-76-08843-1 ISSN 1831-9424 doi:10.2760/661464

Luxembourg: Publications Office of the European Union, 2019

How to cite this report:

J. Galbally, P. Ferrara, R. Haraksim, A. Psyllos, L. Beslay, *Study on Face Identification Technology for its Implementation in the Schengen Information System*, EUR 29808 EN, Publication Office of the European Union, Luxemburg, 2019, ISBN 978-92-76-08843-1, doi:10.2760/661464, JRC116530.

Table of contents

PART I. Overview of ABIS-Face technology

PART II. ABIS-Face within CS-SIS

Abstract

The present report assesses the readiness and availability of automatic face recognition technology for its integration in the Schengen Information System (SIS). This functionality has been introduced in the latest SIS Regulation adopted on the 28th of November 2018. The legislation determines the use of this technology first in the context of regular border crossing, however it also foresees its possible use in the near future in the context of police and judicial cooperation. The first part of the report introduces automatic face recognition technology, presenting a thorough review of the state of the art, which concludes with the lessons learnt and the challenges faced by automatic face recognition systems. The second part makes an analysis of how face recognition technology can be integrated within CS-SIS and presents the different use-cases in which the functionality will be exploited. A number of recommendations for the successful implementation of face processing techniques in CS-SIS are then proposed.

Acknowledgements

This report was carried out by members of the DG JRC team in charge of the "*Biometric Research Applied to Security in the Schengen Area* (BRASSA)" work package, working at the "Cyber and Digital Citizens' Security" Unit E.3 of the "Space, Security and Migration" Directorate E, at DG Joint Research Centre.

The study would not have been possible without the help, dedication and active involvement of a number of people working in different institutions all over Europe and beyond. With our apologies to all those people who actively contributed to the study and are not explicitly mentioned hereafter, the authors would like to give a special recognition to:

DG JRC and DG HOME

We would like to thank the following colleagues from the European Commission (EC) for their determinant support, comments and contribution:

Jean-Pierre Nordvik, DG JRC Directorate E, Cyber and Digital Citizens' Security Unit

Valerio Scandiuzzi, Philippe Van Triel, Richard Rinkens, Rob Rozenburg, Michael Flynn DG HOME Information Systems for Borders and Security unit.

European Agency for the operational management of large-scale IT systems in the area of freedom, security and justice (eu-LISA)

We would like to thank all the colleagues from eu-LISA who always provided us information and explanations on the operational management of some of the large-scale IT systems currently working in Europe, namely EURODAC, Visa Information System, and more specifically the Schengen Information System.

Representatives from EU Member States, Norwegian, Israeli and United States of America authorities

We would like to express our gratitude to the representatives of the six EU Member States visited and contacted during the fulfilment of the study: Estonia, France, Germany, Netherlands, Poland and Sweden.

We would like to thank the representatives of the associated EU Member State Norway that also welcomed us during the development of the study.

We would like express our gratitude to the colleagues from the Israeli Division for Identification and Forensic Sciences in Jerusalem.

We would like as well to thank the USA representatives from the National Institute for Standards and Technology (NIST), Department of Justice and the Federal Bureau of Investigation (FBI) for their great hospitality, openness and full cooperation.

European Border and Coast Guard Agency (FRONTEX) and INTERPOL

We are grateful to the colleagues from FRONTEX who provided us clear and detailed information regarding the operational use cases at border crossings in the Schengen area.

We would like to express our gratitude to the representatives of the INTERPOL, who welcomed us during our study and provided us further insight into the use of biometric technology in operational scenarios.

Biometric developers and vendors

For the study we also received input from different vendors that produce automatic biometric identification systems, we would like to thank them for their time and availability to discuss with us topics regarding the actual deployment of this type of systems: AnyVision, Cognitec, Gemalto, Idemia, Microsoft and NEC

External Experts Board

We would also like to thank the renowned international experts who joined the external scientific board of the study for reviewing its results and conclusions:

Prof. Christoph Busch Hochschule Darmstadt, (DE)

Dr. Patrick Grother US National Institute for Standards and Technology, USA)

Prof. Josef Kittler University of Surrey, (UK)

Dr. Sébastien Marcel IDIAP Research Institute, (CH)

Prof. Raymond Veldhuis University of Twente, NL)

Executive summary

This report details the results of the DG JRC study on the readiness and availability of Automatic Biometric Identification System Face (ABIS-Face) technology for its introduction in the Central Schengen Information System (CS-SIS). The study was carried out for DG HOME via an Administrative Arrangement.

Policy context

Created as a compensatory measure for the abolition of internal border checks within the Schengen area, the SIS was established with two intentions: to contribute to police and law enforcement cooperation between the Member States and to support external border control. In its first generation, the SIS was the first large-scale IT system launched by the EU Member States in 1995. It was followed by EURODAC (asylum seekers' database) in 2003 and the Visa Information System (VIS) in 2011. The second-generation of the SIS entered into operation on 9 April 2013. The Central Schengen Information System (CS-SIS) offers the possibility to store biometric data in alerts related to persons. In addition to alphanumeric data, alerts related to persons should contain as well the fingerprints and facial image of the subject of the alert, whenever they are available.

However, while the storage of fingerprints and facial images of persons was allowed in the CS-SIS Database, in the original version of SIS, these could not be used to search the Data Base in order to identify a person. All searches were performed based on alphanumeric data. Then, fingerprints and facial images would be used to *verify* a given identity, in case the search based on alphanumeric data resulted in a positive identification of a subject. As such, the use of only alphanumeric data to perform searches in the database resulted in the introduction in the system of duplicated identities belonging to the same subject who would provide false alphanumeric data at the time of the creation of the alerts.

This situation would change with Articles 22.c of CS-SIS Decision[1] and Regulation[2] from 2007, which stated that the CS-SIS could also be used to *identify* a person on the basis of his/her *fingerprints*. This option required the implementation of an Automatic Fingerprint Identification System (AFIS) "*once it becomes technically possible*" and when the Commission had presented "*a report on the availability and readiness of the required technology on which the European Parliament is consulted*". In October 2015 DG JRC provided such a report supporting the final decision of integrating 10-prints fingerprint identification technology within the functionalities of CS-SIS[3]. The CS-SIS AFIS went into production in March 2018.

In December 2016, following the decision to introduce 10-print fingerprint identification technology in CS-SIS, a revision of the Regulatory framework was proposed which was finally approved on the 28th of November 2018, for **police use**[4] for **border use**[5] and for

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007D0533&from=EN

[2] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:381:0004:0004:EN:PDF

[3] https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/fingerprint-identification-technology-its-implementation-schengen-information-system-ii-sis

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1862&from=EN

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1861&qid=1544694006055&from=EN

the **return** of illegally staying third country nationals[6]. The ways face related data stored in alerts can be processed are described in articles 33.4 of the new SIS-Border regulation and article 43.4 of the new SIS-Police regulation.

In support of this newly adopted 2018 Regulation, the DG JRC study presents an assessement on whether face recognition technology is mature enough for its integration into the context of the SIS with the aim of achieving the following objectives:

OBJECTIVE 1: Determine the readiness of facial recognition technology, to be integrated in CS-SIS for the identification of a person.

OBJECTIVE 2: Provide recommendations on the best way to integrate facial recognition technology in CS-SIS based on: 1) the current state of the art of this technology; 2) the particularities and constraints of CS-SIS and its dual use for law-enforcement and border management.

The JRC conducted an in-depth analysis of the face recognition technology including: a review of the scientific literature, visits to forensic laboratories in EU Member States and third countries; consultations with eu-LISA, FRONTEX and INTERPOL; and concluded with interviews of technology providers. An external scientific board of renowned international experts reviewed the results and conclusions of the study. The report presents the main findings of the study together with a series of recommendations for the successful implementation of ABIS-Face technology in CS-SIS. The complete technical specifications of the ABIS-Face system to be integrated in the CS-SIS should be subjected to further study, ideally in the form of a benchmark test linked to the call-for-tenders issued to the vendors of the aforementioned technology.

The report is structured in two parts:

- Part I introduces automatic face recognition technology, presenting a thorough review of the state of the art, including a review of quality metrics and biometric standards that have been developed in this field. This first part is concluded with a summary section of the lessons learnt and the challenges faced by automatic face recognition systems, which should be addressed during its integration in SIS (Part II of the report).

- Part II makes an analysis of how face recognition technology can be integrated within SIS. For this purpose, the document presents the different use-cases in which the functionality will be exploited in SIS and, building upon the lessons and challenges identified in Part I. With this objective, it makes a number of recommendations for the successful implementation of face processing techniques in SIS.

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R1860

Key conclusions

Given all the information presented in the study, the conclusion reached in the study with respect to objectives 1 and 2 is that:

CONCLUSION

Given the great boost in accuracy that face recognition technology has experimented since 2014 with the advent of deep learning-based systems, it is the conclusion of the present study that: ABIS-Face systems have reached a sufficient level of readiness and availability for its integration into CS-SIS, provided that the recommendations listed in the present report are implemented and respected, to the largest extent possible, during the rollout and utilisation of this new functionality.

Here below the reader can find the summary of all recommendations given throughout the report for the integration of an ABIS-Face system in CS-SIS.

List of recommendations

RECOMMENDATION 1: Searchable database logical separation.

We recommend that the searchable database containing the templates extracted from face images of the eventual SIS ABIS-Face is logically separated into two quality types: 1) portraits; 2) other type of images.

This will enable to perform the searchers on a given subset of images depending on the context. This will allow the system to provide as output a "match/no-match" response (portrait images) or a "list of candidates" (all images) depending on the quality of the images that the search has been performed on.

RECOMMENDATION 2: Single ABIS-Face search engine.

Even though we recommend to have two logically separated databases according to image quality (i.e., "portraits" and "other", see Recommendation 1), we recommend to have only one unique ABIS-Face search engine to perform the automatic consultations on either of the two face image quality types.

Having two dedicated search engines for "portraits" and "other", presents the high risk of overfitting the systems to a specific type of images. In this case, if the images used in the comparison differ slightly from the images expected by the system, the accuracy will drastically drop. Therefore, such systems will not be able perform well in cases where it is needed a comparison between images belonging to different quality classes (i.e., portraits VS other)

On the other hand, current deep-learning based technology has shown that, if trained on a sufficiently large quantity of data, it is able to generalise well to different types of images. Therefore, one single system trained on a significant large quantity of variable data will perform similarly to systems tuned to a specific quality type, without presenting the accuracy drop in case the compared images do not comply with the expected quality.

RECOMMENDATION 3: Definition of parameters for portrait images.

We recommend to clearly set the parameter(s) defining what consitutes an image that can be flagged as "portrait" quality.

These parameters should be set together with the supplier of the ABIS-Face system, as they will depend on the system accuracy. The requirements should be set to guarantee a certain accuracy on the subset of "portrait" images. For example, False Positive Identification Rate, FPIR=0.001% over a DB of 1 million entries.

RECOMMENDATION 4: Development of an overall face quality metric.

Linked to recommendation 3, we recommend to promote the development of a vendor-independent, robust and reliable, face quality metric to be integrated in the ABIS-Face as soon as it becomes available.

This quality metric could be the result of: 1) the combination of a number of individual values estimating human-defined features such as illumination, sharpness, pose, background, etc. 2) deep-learning derived features; or 3) a combination of both hand-crafted and deep-based features.

The development of a face quality metric should contribute and get feedback from the currently under review standard: "ISO/IEC TR 29794-5 Information Technology – Biometric sample quality – Part 5: Face image data".

Whenever such a quality metric, ideally subjected to its incorporation into an international standard becomes available, we recommend the following actions:

- Integration in CS-SIS. We recommend to include in the CS-SIS ABIS-Face the quality metric algorithm. The quality metric at central level can be of great utility to: 1) as monitoring tool of the face images stored in CS-SIS; 2) to automatically classify images between the two quality types 'portrait' and 'other'; 3) to give feedback to the MSs regarding the quality of the face images submitted to CS-SIS.

Integration at MS level. We recommend to implement the quality metric also at the level of the MS. In this case the quality metric can be useful to incorporate in an acquisition loop/recapture procedure to be carried out until satisfactory quality face images have been obtained both at the time of enrolment and of querying the system. This procedure should contemplate alternative acquisition processes, according to the sample quality, and should include human intervention, where appropriate.

RECOMMENDATION 5: Evaluation of the ABIS-Face on operational data.

We recommend to perform an evaluation of the ABIS-Face on the operational data already present in CS-SIS in order to determine:

- The accuracy that should be expected for portrait images (or, if available, the minimum quality required to categorise an image as portrait).
- Decision thresholds to produce match/no-match responses (see recommendation 12).

We recommend that, in addition to the initial evaluation on operational data to determine certain parameters of the system (e.g., minimum quality for portrait images, threshold to determine match/no-match), a similar evaluation is performed on a regular basis in order to adapt the parameters to possible changes in the accuracy of the system due to an increase/decrease of the enrolled data or to an update of the system.

Any evaluation of FR systems should follow as close as possible the directives given in: ISO/IEC 19795-1 2006 "Information Technology – Biometric Performance and reporting – Part 1: principles and framework".

RECOMMENDATION 6: High-quality enrolment process.

We recommend that, whenever a cooperative data subject is available at the enrolment process, that is, in most of the cases, the enrolment phase should favor the use of high quality cameras, in fully controlled conditions, to adhere as much as possible to the ICAO Standard specifications or to the ISO/IEC 19794-5 specifications, under the supervision of experienced operators, as is usually the case in a law enforcement context.

This should result in the production of high-quality portrait-like face images which are to be stored in the CS-SIS database.

In order to promote this high-quality enrolment process, we recommend that best practices for face acquisition are compiled and distributed to the Member States in order to obtain a central database as homogeneous as possible.

RECOMMENDATION 7: Need for complementary statistics.

We recommend that eu-LISA identifies the best possible ways to include in its annual report the statistics of CS-SIS:

- The number of consultations per year related to persons at border checks. In order to complement this assessment at central level, we also recommend that Member States report annually on the number of consultations related to persons that have been carried out on their national copies.
- Once the ABIS-Face is running, the number of consultations performed based on the ABIS-Face.
- The number of person related alerts that contain face images.
- The number of hits obtained based on ABIS-Face.
- The number of duplicated alerts detected based on ABIS-Face.
- The quality of the enrolled face images in CS-SIS.

The quality of the live images submitted to perform queries in CS-SIS.

RECOMMENDATION 8: Use of live captured images.

For consultation of CS-SIS at border crossings we recommend to use in all cases a live picture of the traveller, carefully designing the set-up of the capture points (see recommendation 10). The additional use of the face image stored in the passport chip can be optional although it is not recommended as this image:

- Is in general of lower resolution than the images captured live.
- Face Recognition systems have shown to obtain worse accuracy using passport images than live-captured images.

Increases the vulnerability of the system, since it cannot be guaranteed that the image in the passport belongs to the traveller (e.g., morphing attacks).

RECOMMENDATION 9: Quality of capture points.

Supervision by an operator. Adequate operator training is recommended, in order to:

- Train the operator to capture good quality face images (e.g., indicate him the best position for the capture subject, pose, face expression, presence of glasses).
- As supervision of biometric acquisition is a repetitive task and requires additional attention in the case of centralised enrolment stations. The aim is to avoid tiredness and boredom adversely affecting the process.
- Train the operator to detect Presentation Attacks.

In case of automatic ABC gates, they should be thoroughly tested in each location where they will be deployed to ensure their ability to capture good quality face images. ABC gates should in all cases be equipped with Presentation Attack Detection measures.

Adequate sensor. We recommend to use performant cameras (e.g. in speed, imaging sensor and resolution), offering also enhanced capabilities to acquire good quality images in sub-optimal environments.

Enhanced graphic user interface (GUI). We recommend that capture points have large displays and provide real-time feedback regarding the quality of the acquired data.

Proper user interaction. The enrolment process should be user-friendly with clear step-by-step procedures properly explained. The use of good ergonomics should be considered to support better acquisition practices. The user should receive some feedback from the system as where to locate himself.

Adequate environment. The acquisition environment should be appropriate in terms of illumination, temperature and backgrounds both for the subject and the operator. These elements are recommended mainly for fixed stations but similar considerations are instrumental as well for mobile stations. It is especially relevant to pay attention to the illumination factor, as it is key to the acquisition of good quality face images.

Sensor maintenance. There should be regular and systematic maintenance of the enrolment stations to avoid a decrease in performance, especially in the case of consultation processes taking place in heavily used check points (e.g., high-traffic airports).

RECOMMENDATION 10: Common exchange standard.

At the moment, the exchange of face data in the SIS system is done on slightly modified version of the ANSI/NIST ITL 1-2011 containers type 10, as required by the SIRENE manual. These containers seem to provide an appropriate basis regarding the exchange of face data. We recommend that an automatic check for their mandatory and complete implementation should be developed in order to appropriately support the deployment of the SIS ABIS-Face functionality.

A transition between the NIST container to the ISO/IEC 39794-5 standard (which will soon be available) could be explored. Two main reasons for this possibility:

- The ISO/IEC 39794 standard is an extensible data format that guarantees both backward and forward compatibility (in case that future versions of the standard require further data fields to be included in the containers).
- The ISO/IEC 39794 standard allows for human annotated points to be encoded in facial images. These points can help to enhance the accuracy of FR systems under certain contexts.

The ANSI/NIST ITL 1-2011 is mainly a forensic-based standard. This could be seen as user-unfriendly in order to process the data of travellers.

RECOMMENDATION 11: Computation of the match/no-match threshold.

We recommend to set the threshold that defines the match/no-match output of the system based on the acceptable number of false matches to be produced by the system.

This rate is defined by the False Positive Identification Rate (FPIR) of the system and determines the number of subjects that will be sent to the second line of inspection due to a mistake of the system. Therefore, the FPIR is a determinant factor to set the amount of workload and manpower that will be needed for the second line of inspection (based only on face consultations).

We recommend to perform an evaluation of the ABIS-Face on the real operational data where it will be used in order to set the threshold for the match/no-match reply according to the FPIR predefined (see recommendation 6).

While the FPIR may be the determinant factor to determine the accuracy of the system, a lower FPIR necessarily implies a higher FNIR, that is, the number of non-detected subjects in SIS will increase.

RECOMMENDATION 12: Accuracy evaluation across ethnicities and gender.

We recommend to perform an evaluation of the ABIS-Face on data coming from different ethnicities (e.g., caucasian, black, asian) and also gender (i.e., male, female).

Current face recognition technology based on deep learning has shown that, if not properly trained on data that represents the variability and nature of the data that the system will operate on, it can be biased towards a specific ethnicity or gender. It is important to test that the system has been trained on data that models to the largest extent possible, the variability present in human faces. The ABIS-Face in SIS will be used (initially) in border control, therefore it should be able to perform consistently on all ethnicities and genders.

RECOMMENDATION 13: Need to study the age effect.

We recommend to analyse the difference in accuracy of face recognition technology between different age groups (e.g., children, adults, elderly). Some initial studies have shown that accuracy drops drastically for children below 13 years of age, although these results need further confirmation. There may be an age limit to be set for the accurate use of face recognition technology. Alternatively specific algorithms may have to be developed to cope with the difficulties presented by certain age groups.

RECOMMENDATION 14: Storage of multiple frontal images.

We recommend to allow for the storage of multiple frontal face images for the same person in order for a SIS ABIS-Face to support a multiple comparison strategy. As long as it is clearly established that the images belong to the same person, having multiple samples can increase the accuracy of the comparison process. Allowing to update alerts with the most recent images of a subject is especially relevant in order to minimise the ageing effect (see recommendation 15).

RECOMMENDATION 15: Corrective measures for the ageing effect.

We recommend to update, whenever possible, old alerts with the most recent face images available in order to reduce as much as possible the ageing effect (reduction in the accuracy of the system due the time separation between the two compared images). This is especially relevant for the case of children where a substantial difference may be observed in their face appearance even for short periods of time. This dimension might be more particularly relevant for alerts related to missing persons.

RECOMMENDATION 16: Storage of additional off-angle (yaw) images.

Off-angle images are unlikely to be used to search by the ABIS-Face in a border context, However, in addition to the frontal face images, we recommend to store as well, whenever possible (e.g., access to subject at a police station), face images at +90, +45, -45 and -90 degrees of the yaw angle. Therefore, the system should allow to label each image with the yaw angle at which it was captured. These images can be useful for:

- The manual verification of a match at the second line of inspection.
- For future potential uses of the ABIS-Face, like for example consultation using images acquired in unconstrained environments (e.g., coming from video surveillance footage), where faces may be seen off-angle.

It should be taken into account that, in order to perform reliable automatic recognition of off-angle images with a large yaw angle (e.g., profile pictures with 90º yaw), it would very likely be necessary to integrate a specific algorithm to operate on those images.

RECOMMENDATION 17: Presentation attack detection measures.

In case an Automatic Border Control (ABC) gate is used at the border crossing instead of a border guard, we highly recommend to put in place the necessary safeguards in the ABC gate in order to minimise the impact of potential prestentation attacks (e.g., ABC gates with integrated presentation attack detection measures). The most likely presentation attacks foreseen are the evasion attacks (i.e., attacks in which the subject tries to hide his identity not to be recognised).

In the case of the presence of a human supervisor, known presentation attacks (e.g., printed pictures, masks) should be easily detected after a brief training of the guard.

An evaluation of presentation attacks and of presentation attack detection methods should follow to the largest extent possible the guidelines and metrics given in the standard "ISO/IEC 30107, Biometric presentation attack detection".

RECOMMENDATION 18: Use of NIR FR technology within CS-SIS.

We recommend that the main ABIS-Face in CS-SIS should remain based primarily on facial images captured in the visual spectrum.

We recommend that images captured in NIR could complement those captured in the visual domain for the case of bad illumination conditions during consultation, or during enrollment with the aim to cope with consultation requests from NIR domain.

In case that NIR images were eventually stored in CS-SIS (and labbeled as specific NIR image), we recommend to have:

- The primary search engine ABIS-Face VIS to perform VIS-VIS comparison.
- A secondary search engine ABIS-Face NIR-VIS to perform NIR-VIS comparison in the case of a consultation using NIR image to be compared with the visible domain images stored in the SIS.

A third search engine ABIS-Face NIR to perform NIR-NIR comparison.

RECOMMENDATION 19: Use of 3D FR technology within CS-SIS.

For the near future we do not recommend the inclusion of 3D technology in CS-SIS since it does not adapt well to the use-cases of this system.

3D technology, however, can be useful in unsupervised capture points (e.g., ABC gates at airports) in order to perform Presentation Attack Detection.

Introduction

The Schengen Information System (SIS) is the most widely used and largest information sharing system for **security** (law-enforcement) and **border management** in Europe. It is important that the reader bears in mind these two dimensions of SIS throughout this report: law-enforcement and border management. While the system is unique, it has to deal with the reality of these two contexts that, in some cases, present different challenges and constraints. Just to give a simple example, at the first line of check of a border, the guard has very limited time to take a decision on a traveller, while in a police station the officer has almost no limitations time-wise to do the necessary checks on a subject. This dual nature of SIS leads to differences in the use cases of the system that will be later highlighted in the report.

The main purpose of SIS is to make Europe safer. The system assists the competent authorities in Europe to preserve internal security in the absence of internal border checks. To reach this objective, SIS enables competent national authorities, such as the police and border guards, to enter and consult **alerts** on **persons** or **objects**. A SIS alert always consists of three parts:

- A set of data for identifying the person or object, subject of the alert,
- A statement why the person or object is sought and
- An instruction on the action to be taken when the person or object has been found.

The quality, accuracy and completeness of the data elements enabling identification are the key conditions for the success of SIS. For alerts on persons, the minimum data set is name, year of birth, a reference to the decision giving rise to the alert and the action to be taken. When available, facial images and fingerprints must be added in order to facilitate identification and to avoid misidentification.

SIS consists of three major components:

- A Central System, CS-SIS;
- The national systems, N-SIS;
- A communication infrastructure (network) between the systems.

An alert entered in SIS in one Member State is transferred in real time to the central system. It then becomes available in all the other Member States so that authorised users can search the alert on the basis of the entered data-elements. Specialised national SIRENE bureaus located in each Member State (MS) serve as single points of contact for the exchange of supplementary information and coordination of activities related to SIS alerts. The responsibility of SIS management is divided as follows:

- Each **Member State** using SIS is responsible for setting up, operating and maintaining its national system and its national SIRENE bureau.
- The EU Agency for large-scale IT systems (**eu-LISA**) is responsible for the operational management of the Central System and the communication infrastructure.
- The **European Commission** is responsible for the general supervision and evaluation of the system and for the adoption of implementing measures where uniform conditions for implementation are needed, such as the rules for entering and searching data.

At the end of 2018[7], SIS contained approximately 82.2 million records (i.e., alerts), out of which, 940K were related to persons and the rest to objects. From the person alerts, around 25% contained at least one fingerprint image and around 30% contained at least one facial image.

In 2018, SIS processed a total 6.2 billion queries (including queries related to both object and person alerts), out of which around 0.005% were processed by the current AFIS.

According to the SIS regulation, alerts on persons shall be kept only for the time required to achieve the purposes for which they were entered. Initially, a Member State may enter an alert on a person for a period of five years. Then, within that five-year period, the Member State has the right to review the need to retain the alert for longer. This way, SIS is a quite dynamic system were alerts are constantly inserted and deleted. Following this process, it seems that the size of the database has reached a quite stationary size that is not expected to change significantly in the near future independently of the growing number of consultations.

For further details on the current functionality of CS-SIS we refer the reader to PART II of the present study.

i. Policy, technical and legal context of CS-SIS

From its inception, CS-SIS offered the possibility to store biometric data in alerts related to persons. As mentioned above, in addition to alphanumeric data, alerts related to persons should contain as well the fingerprints and facial image of the subject of the alert, whenever they are available.

However, while the storage of fingerprints and facial images of persons was allowed in the CS-SIS Database, in the original version of SIS, these could not be used to search the DB in order to identify a person. All searches were performed based on alphanumeric data. Then, fingerprints and facial images would be used to *verify* a given identity, in case the search based on alphanumeric data resulted in a positive identification of a subject. However, for a person it is more difficult to modify his biometric identifiers (e.g., fingerprints, face) than his alphanumeric data (e.g., name, surname, date of birth). As such, the use of only alphanumeric data to perform searches in the database resulted in the introduction in the system of duplicated identities belonging to the same subject who would provide false alphanumeric data at the time of the creation of the alerts.

This situation would change with Articles 22.c of CS-SIS Decision[8] and Regulation[9] from 2007, which stated that the CS-SIS could also be used to *identify* a person on the basis of his/her *fingerprints*. This option required the implementation of an Automatic Fingerprint Identification System (AFIS) "*once it becomes technically possible*" and when the Commission had presented "*a report on the availability and readiness of the required technology on which the European Parliament is consulted*". In October 2015 DG JRC provided such a report supporting the final decision of integrating 10-prints fingerprint

[7] https://www.eulisa.europa.eu/Publications/Reports/SIS%202018%20statistics.pdf

[8] http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007D0533&from=EN

[9] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:381:0004:0004:EN:PDF

identification technology within the functionalities of CS-SIS[10]. The CS-SIS AFIS went into production in March 2018.

In December 2016, following the decision to integrate 10-print fingerprint identification technology in CS-SIS, a revision of the regulation was proposed which was finally approved on the 28th of November 2018, for **police use**[11], for **border use**[12] and for the **return**[13] of illegally staying third country nationals. In article 33 of the new SIS-Border regulation and article 43 of the new SIS-Police regulation, it is defined the new use that can be given to face related data stored in alerts:

- Article 33.4 Border and Article 43.4 Police

  "As soon as it becomes technically possible, and while ensuring a high degree of reliability of identification, photographs and facial images may be used to identify a person in the context of regular border crossing points.

  Before this functionality is implemented in SIS, the Commission shall present a report on the availability, readiness and reliability of the required technology. The European Parliament shall be consulted on the report.

  After the start of the use of the functionality at regular border crossing points, the Commission shall be empowered to adopt delegated acts in accordance with Article 75 to supplement this Regulation concerning the determination of other circumstances in which photographs and facial images may be used to identify persons."

ii. Objectives of the study

In support of the new 2018 Regulation presented in Section i, the objectives of the present DG JRC study are to:

OBJECTIVE 1: Determine the readiness of facial recognition technology, to be integrated in CS-SIS for the identification of a person.

OBJECTIVE 2: Provide recommendations on the best way to integrate face recognition technology in CS-SIS based on: 1) the current state of the art of this technology; 2) the particularities and constraints of CS-SIS and its dual use for law-enforcement and border management.

As will be further explained in Section v, in order to address these two objectives, the present report describes first in PART I the current state of the art in ABIS-Face technology and clearly states the challenges faced by this type of systems. Then, in PART II, it contextualises face recognition technology given the specificities of CS-SIS, providing a

[10] https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/fingerprint-identification-technology-its-implementation-schengen-information-system-ii-sis

[11] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1862&from=EN

[12] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1861&qid=1544694006055&from=EN

[13] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R1860

series of recommendations on how to best address those challenges so that the outcome of the eventual integration of ABIS-Face technology in CS-SIS is successful.

iii. Technology: Readiness and Availability

According to the Horizon 2020 EU Research and Innovation Framework Programme the readiness and availability of a given technology is assessed using nine different levels (Technology Readiness Levels, TRL):

- TRL 1 – basic principles observed,
- TRL 2 – technology concept formulated,
- TRL 3 – experimental proof of concept,
- TRL 4 – technology validated in laboratory,
- TRL 5 – technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies),
- TRL 6 – technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies),
- TRL 7 – system prototype demonstration in operational environment,
- TRL 8 – system complete and qualified,
- TRL 9 – actual system proven in operational environment.

As will be explained throghout the report, although ABIS-face technology has reached TRL 9, with multiple large-scale systems already deployed and working worldwide, each operational scenario has its own specificities. As such, the successful application of a certain technology to a given specific use-case and environment, does not necessarily guarantee the same level of success when those operational conditions are changed.

In particular, for ABIS-face technology to achieve the expected level of performance, there are certain parameters that have to be considered. Probably, the most important of these features is the **accuracy** that can be expected from ABIS-face systems. Unfortunately, the answer to the question of how accurate current systems are is not straightforward, as it largely depends on the data (i.e. facial images and photographs) a system will have to deal with and, more specificaly, with the **quality** of that data. Furthermore, depending on the **use-cases** defined for an ABIS-face system, a different level of accuracy may be acceptable and/or expected for different operational conditions/scenarios.

iv. Methodology followed

In order to reach the objectives set forth in Section ii, the study was conducted in three steps with some slight overlap between them:

- STEP 1: Wide collection of information regarding ABIS-Face technology.
- STEP 2: Synthesis of the information obtained from multiple sources.
- STEP 3: Production of the report.

STEP 1 was the most important and, as such, the most time and resource consuming. This step provided all the necessary information for the JRC analysis and eventually led to the current report and the different recommendations contained in it. This information was collected over five phases, each of them involving different sources. These phases are detailed in the next sections.

Phase 1: Analysis of the state of the art in ABIS-Face technology

Relevant bibliography and scientific literature were extensively reviewed in order to consolidate and complement JRC knowledge and obtain an initial solid overview of the main features and challenges of ABIS-Face systems. The study oriented on the two areas (Police and Border) was necessary in order to prepare the set of visits and consultations carried out in the subsequent phases.

Phase 2: Consultation with national ABIS-Face operators

The end-users of a future ABIS-Face in CS-SIS will be the competent authorities of the different Member States (MS), such as law-enforcement and border-control authorities. It is therefore extremely important to know the operational contexts in which MS are using their national ABIS-Face systems, the similarities and differences between them, as well as to understand their operational needs.

Following the rationale described above and in order to address the objective of assessing "*the availability and readiness of the required technology*" for the inclusion of an ABIS-Face in CS-SIS, the JRC first contacted and visited seven EU Schengen Member States' and associated state law-enforcement and border-control entities (Norway, Germany, Netherlands, France, Poland, Sweden and Estonia). The choice of the MS participating to this study was based on the availability and operational status of a national ABIS. The objectives of these exchanges were threefold:

- Obtain knowledge regarding the technical aspects of the ABIS-Face solution implemented in these countries;
- Identify the operational constraints and best-practice followed;
- Describe the challenges they face.

The visits have led to better understanding and subsequently to a definition of the use-cases, in which the MS were using their national ABIS-Face system. The visits have also provided an opportunity for the JRC to collect the possible expectations and recommendations these authorities had regarding the introduction of ABIS-Face functionality in CS-SIS.

The visits to the seven EU Schengen MS were complemented with a visit to the United States of America (US). In particular the visit included: 1) the National Institute of Standards and Technology (NIST) which is renowned worldwide for its contributions to the field of biometric quality and 2) the Federal Bureaus of Investigation (FBI) which is managing some of the biggest and most advanced ABIS-Face system, presenting broad similarities with the objectives and expected use of the CS-SIS ABIS-Face technology.

Visit to the US was followed by the visit to Israel, country with very rich operational experience in using ABIS-Face system.

In addition to the visit to Eu-LISA (see next section), two EU and international organizations – Frontex and Interpol, both having operational experience with the use of this technology were also visited.

These visits were facilitated by the permanent support of DG HOME colleagues during the SISVIS committee meetings in 2018. Prior to each visit, An outline of the envisaged technical exchange was sent to the selected countries prior to the visit. This outline aimed to inform them by providing a list of preliminary questions regarding the different technical fields the JRC wished to explore during the visit. Each visit focused on the following subjects:

- Identification of the use-cases in which face images are processed;
- A technical description of the national ABIS-Face;
- The management of the life cycle of face images in their system;
- The possibility to have a live demonstration of the use of the ABIS-Face system.

The JRC visits targeted national ABIS-Face used in the context of criminal-investigation and managed by national police forces. However, for each visit, authorities in charge of border-control (also using the national ABIS-Face) were invited to participate in the presentations and discussion.

At the beginning of each visit, the JRC gave an introductory presentation and proposed an agenda divided into three main steps:

- The National ABIS-Face system;
- Current and future uses of CS-SIS;
- Use of other EU/international system such as Prüm, INTERPOL, etc.

The visits were conducted by two JRC scientific officers. The team of two was necessary to cope with the rich and intensive discussion offered by the visited countries.

At the end of each visit, the JRC provided the timescale of the study and invited participants to review the final draft of the present report. The timeline and most relevant information concerning the visits are summarized in **Table 1**.

Table 1. Summary of the key information concerning the visits to the institutions managing the national AFIS in different countries.

VISIT	DATE	INSTITUTION
eu-LISA	05/02/2018	Strasbourg, France
Norway	01/03/2018	National Criminal Investigation Service (Oslo)
Netherlands	06/03/2018	National Police Corps (Zoetemeer)
Germany	11/03/2018	Federal Criminal Police Office -BKA (Wiesbaden)
USA	26/03/2018	US National Institute for Standards and Technology (NIST) Federal Bureau of Investigation (FBI) – Criminal Justice Information Services (CJIS) Federal Bureau of Investigation (FBI) – Criminal Justice Information Services (CJIS) – Latent Print Support Unit
Israel	16/04/2018	Department of Identification and Forensic Science - DIFS (Jerusalem)
France	20/04/2018	Institut de Recherche Criminelle de la Gendarmerie Nationale - IRCGN (Paris)
Poland	25/04/2018	National Police Forensic Service (Warsaw)
FRONTEX	26/04/2018	FRONTEX (Warsaw)
Sweden	10/09/2018	National Forensic Centre - NFC (Linköping)
INTERPOL	26/09/2018	Lyon, France
Estonia	15/11/2018	Intelligence Management and Investigation Department (Tallinn)

Phase 3: Consultation with eu-LISA

The visit on 5th of February 2018 to the European Agency for the Operational Management of large-scale IT Systems in the Area of Freedom, Security and Justice (eu-LISA), allowed the JRC to obtain an accurate picture of the central part of the CS-SIS AFIS currently in operation, EURODAC and EU AFIS systems currently in production, such as the Visa Information System (VIS), European Criminal Records Identification System (ECRIS) and the EU Entry Exit System (EES). It also provided a detailed description of the CS-SIS central system and its reporting capability. This visit was followed by a series of exchanges and conference calls with the officers respectively in charge of CS-SIS, VIS and EURODAC until the end of the study, providing the latest up-to-date statistics of those systems when available.

Phase 4: Consultation with ABIS technology vendors

The information collected from authorities already using ABIS-Face was completed by discussions with the vendors of such technology. This also allowed the JRC to have a better understanding of the deployment challenges faced by the actual designers of such systems.

Although numerous companies offer ABIS-Face in multiple domains, most of them are integrators and do not themselves develop ABIS-Face solutions. Six vendors, suppliers of automatic face recognition technology responded to the JRC meeting requests (Idemia, Cognitec, Anyvision, Gemalto, NEC and Microsoft).

Phase 5: Consultation with external review board of experts

In order to review the results and conclusions established in this report, an External Review Board has been established. This expert review board is composed of five internationally renowned researchers in the field of face recognition. Their main objective was to review the report in order to: 1) give comments to complement/improve it, 2) correct possible content mistakes and 3) suggest missing pieces of relevant information. The final draft version of the report was submitted to them at the end of December 2018. The five experts presented their reviews, comments and suggestions at the beginning of 2019. The experts also had the opportunity to discuss among them and to further detail their point of view where necessary. The five experts were:

- Prof. Christoph Busch (Hochschule Darmstadt, GER)
- Dr. Patrick Grother (US National Institute for Standards and Technology, US)
- Prof. Josef Kittler (University of Surrey, UK)
- Dr. Sébastien Marcel (IDIAP Research Institute, CH)
- Prof. Raymond Veldhuis (University of Twente, ND)

v. Structure of the report

The approach adopted by the JRC for the initial analysis had as main objective to explore and assess the main characteristics and challenges of ABIS-Face technology from a general perspective and then apply these identified elements to the specific context of SIS and suggest recommendations to appropriately address them. Accordingly, the JRC report contains two main parts:

- **PART I** sets the scene on the current status of ABIS-Face technology. It introduces the key parameters and concepts of ABIS-Face systems, such as its feature extraction process, its comparison algorithms, its performance evaluation, existing databases, quality of face data or face related standards.
  As a wrap-up, PART I finishes with a section dedicated to the main challenges faced by ABIS-Face designers when putting in place such new large-scale systems. All these challenges have been extracted from the large amount of information provided by the different sources consulted during the preparatory stages of the report (i.e. bibliography, Member States, vendors, eu-LISA, and external experts board).

- **PART II** focuses first on SIS as it is implemented today, presenting some facts related to the current architecture, expected use-cases for an ABIS-Face or relevant legislation.
  After this initial presentation of the system, PART II builds on the initial scene, concepts and key features for the ABIS-Face technology introduced in PART I and on the specificities of SIS, to give a series of recommendations, suggestions and options on how each of the challenges presented at the end of PART I could be potentially dealt with in the use-cases identified for SIS in order to successfully implement an ABIS-Face functionality in the most effective way possible.
  PART II finishes with a more prospective look into the future giving some possible actions (still not contemplated by the current legislation) that could be undertaken in the years to come in order to further improve the accuracy, flexibility and ultimately the added-value offered by SIS to the Member States.

vi. Audience of the report

Even though some general aspects of SIS are presented in the introduction of the report, the document has been thought for readers who are knowledgeable in SIS and have some basic understanding of biometric technology.

As such, in PART I the study focuses on describing the state of the art of face recognition, while in PART II it describes how this technology can be applied to SIS. The reader should bear in mind that many details about the functioning and purpose of SIS are not described here as they are assumed to be known.

Regarding the biometric content, although the report has been conceived as a self-contained document to be read by a wide audience, many specific aspects related to biometric technology are discussed in the different sections (especially in PART I). Therefore, for those readers who are laymen in biometrics, it is strongly recommended to first read some general introduction to biometrics such as the research overview articles [1], [2], or the standardised biometric tutorial "ISO/IEC TR 24741 Biometrics – Overview and Application". In these documents basic concepts related to biometric technology are described. This initial reading can facilitate a better grasp of the implications and findings of the report.

vii. A note on terminology

The present document tries to adhere, to the largest extent possible, to the standardised biometric vocabulary that can be consulted in the international standard[14] "ISO/IEC 2382-37 – Information Technology – Vocabulary – Part 37: Biometrics".

The reader should be aware that the definitions given in the SIS Regulation adopted in November 2018 and the standardised vocabulary may not be exactly equivalent for some terms. In these cases, priority has been given to the Regulation. As such, unless specified otherwise, the next definitions taken from Article 3 of the Regulation are used:

Match. A match means the occurrence of the following steps:

(a) a search has been conducted in SIS by an end-user;
(b) that search has revealed an alert entered into SIS by another Member State; and
(c) data concerning the alert in SIS match the search data;

Hit. A hit means any match which fulfils the following criteria:

(a) it has been confirmed by:
(i) the end-user; or
(ii) the competent authority in accordance with national procedures, where the match concerned was based on the comparison of biometric data;
and
(b) further actions are requested;

Biometric data, means personal data resulting from specific technical processing relating to the physical or physiological characteristics of a natural person, which allow or confirm the unique identification of that natural person, namely photographs, facial images, dactyloscopic data and DNA profile.

[14] Freely available at: http://standards.iso.org/ittf/PubliclyAvailableStandards/c066693_ISO_IEC_2382-37_2017.zip

Dactyloscopic data, means data on fingerprints and palmprints which due to their unique character and the reference points contained therein enable accurate and conclusive comparisons on a person's identity.

Facial image, means digital images of the face with sufficient image resolution and quality to be used in automated biometric matching.

We would also like to highlight here that in the report we will refer as **facial portraits**, or simply portraits, to high-quality frontal images of the face acquired under controlled conditions (e.g., illumination, pose, background). These are the typical images that may be found in identity documents such as passports. These images are also referred to in the forensic/law-enforcement field as **mugshots**. However, given the dual use of SIS for law-enforcement and border management, we have given preference to the term portrait.

Part I
Overview of ABIS-Face technology

As already mentioned in Section v, the main objective of this PART I is to summarise the state of the art in ABIS-Face technology and to introduce the main concepts and key parameters that will play a key role for its integration in CS-SIS.

In order to help the reader to clearly identify these key concepts, each section finishes with a summary of the most important points, which will later have an impact in the recommendations made for SIS in PART II.

As a wrap-up, in the final section of the present PART I, we list and summarise the main challenges faced by current ABIS-Face technology which have to be taken into account when considering the introduction of this functionality in a new large-scale system like SIS. These challenges shape the structure of PART II and guide the recommendations given there.

Nowadays, the domain of face recognition technology is too wide to be covered in one single document. Therefore, in the present report we will focus in a review of those specific areas that are more relevant for the CS-SIS system. Although other topics can be mentioned or partially covered, the scope of the present overview is restricted to:

- face identification systems,
- working in the open set scenario,
- on 2D face images captured in the visual spectrum,
- under controlled or unconstrained conditions.

With that scope in mind, this PART I is structured as follows:

- Section 1 presents a review of the state-of-the-art of face recognition systems with special focus on current technology (i.e., deep learning-based systems).
- Section 2 presents a review of the evaluation of face recognition technology including datasets, evaluation scenarios and metrics, and finally results obtained in official and non-official evaluation campaigns.
- Section 3 presents a review of the state of the art in face quality estimation.
- Section 4 presents a review of existing standards for the interchange of face data.
- Section 6 finally presents the challenges that have to be addressed when deploying this type of technology.

1. Face Recognition Systems

Face Recognition (FR) can be defined as the task to recognise a subject based on his facial image. The first automated face recognition system was developed by Takeo Kanade in his Ph.D. thesis work in 1973 [3]. There was a dormant period in automatic face recognition until the work by Sirovich and Kirby [4] [5] on a low dimensional face representation, derived using the Karhunen–Loeve transform or Principal Component Analysis (PCA). In the early 1990s, the study of FR became increasingly popular following the introduction of the historical Eigenface approach by Turk and Pentland [6].

The milestones of feature-based FR over the past years are presented in **Figure 1**, in which the times of four major technical streams are highlighted. The holistic approaches derive the low-dimensional representation through certain distribution assumptions, such as linear subspace [7] [8], manifold [9] [10], and sparse representation [11] [12] [13] [14]. This idea dominated the FR community in the 1990s and 2000s. However, a well-known problem is that these theoretically plausible holistic methods fail to address the uncontrolled facial changes that deviate from their prior assumptions.

In the early 2000s, this problem gave rise to local-feature-based FR. Gabor [15] and Linear Binary Pattern (LBP) [16], as well as their extensions multilevel Gabor LBP (GLBP) and high-dimensional LBP (HD-LBP) [17] [18], achieved robust performance through some invariant properties of local filtering. Unfortunately, handcrafted features suffered from a lack of distinctiveness and compactness.

In the early 2010s, learning-based local descriptors were introduced to the FR community [19] [20] [21], in which local filters are learned for better distinctiveness, and the encoding codebook is learned for better compactness. However, these shallow representations still have an inevitable limitation on robustness against the complex nonlinear facial appearance variations.

As shown in **Figure 2**, Traditional methods attempt to solve FR problem by one or two-layer representation, such as filtering responses or histogram of the feature codes. In contrast, deep learning methods use a cascade of multiple layers of processing units for feature extraction and transformation.

In 2014, DeepFace [22] and DeepID [23] achieved state-of-the-art accuracy on the famous Labelled Faces in the Wild (LFW) benchmark [24], surpassing human performance in the unconstrained scenario for the first time. Since then, research focus has shifted to deep-learning-based approaches. FR is a different task from generic object classification tasks [25] because of the particularity of faces: a massive number of classes with small inter-class difference (differences between different faces), and large intra-personal variations (i.e., variations of different samples of the same face) due to different poses, illuminations, expressions, ages, and occlusions. These challenges have inspired many novel architectures and loss functions to promote the discrimination and generalization capability of deep models. Larger scale face databases and advanced face processing techniques are also developed to facilitate deep FR.

Enforced by the developed Graphics Processing Units (GPUs) and big training data, deep FR techniques have kept refreshing the record of performance on academic benchmark databases and fostered numerous successful real-world applications in recent five years. Over the past few years, there have been several surveys on FR [26] [27] [28] [29] and its subdomains, including illumination-invariant FR [30], 3D FR [31] or pose-invariant FR [32]. The aforementioned surveys mainly cover the methodologies on shallow FR.

In the present overview of the FR technology we will focus almost exclusively in deep-learning techniques as, for the last 4-5 years, they have fully taken the FR field, clearly outperforming previous systems in all the known public benchmarks and in independent evaluations. It is safe to say that technology based on holistic learning, local handcraft and shallow learning, are a thing of the past, and that deep learning represents both the present and future of face recognition. Even though the review may be understood as is, not all concepts related to deep learning are defined, therefore, for readers totally new to this field, some previous general reading on deep learning techniques is recommended [33], [34].

It should be noticed that several parts of this section have been directly adapted from the excellent survey in deep FR presented in [35], which is a very recommended read for any person interested in getting a more detailed insight into this field.

Figure 1. Milestones of feature representation for FR. As the representation pipeline becomes deeper and deeper, the performance on the LFW DB (Labelled Face in-the-Wild) steadily improves from around 60% to above 97%. Source: Adpated from (Wang & Deng, 2018)

References to the algorithms that appear in **Figure 1** are: Eigenfaces [6], Fisherface [7], Bayes [36], Laplacianface [9], 2-dimensional Principal Component Analysis (2DPCA) [37], Sparse Representation-Based Classifier (SRC) [13], Collaborative Representation-Based Classifier (CRC) [14], metric learning [38], Elastic Bunch Graph Matching (EBGM) [39], Linear Binary Pattern (LBP) [16], Local Gabor Linear Binary Pattern (LGLBP) [17], High-Dimensional Local Binary Pattern (HD-LBP) [18], Learning Based descriptors (LE) [19], Discriminant Face Descriptor (DFD) [20], Fischer Vector (FV) [40], PCAnet [21], Deepfaces [22], DeepID [23], Visual Geometry Group Network (VGG) [41], Facenet [42].

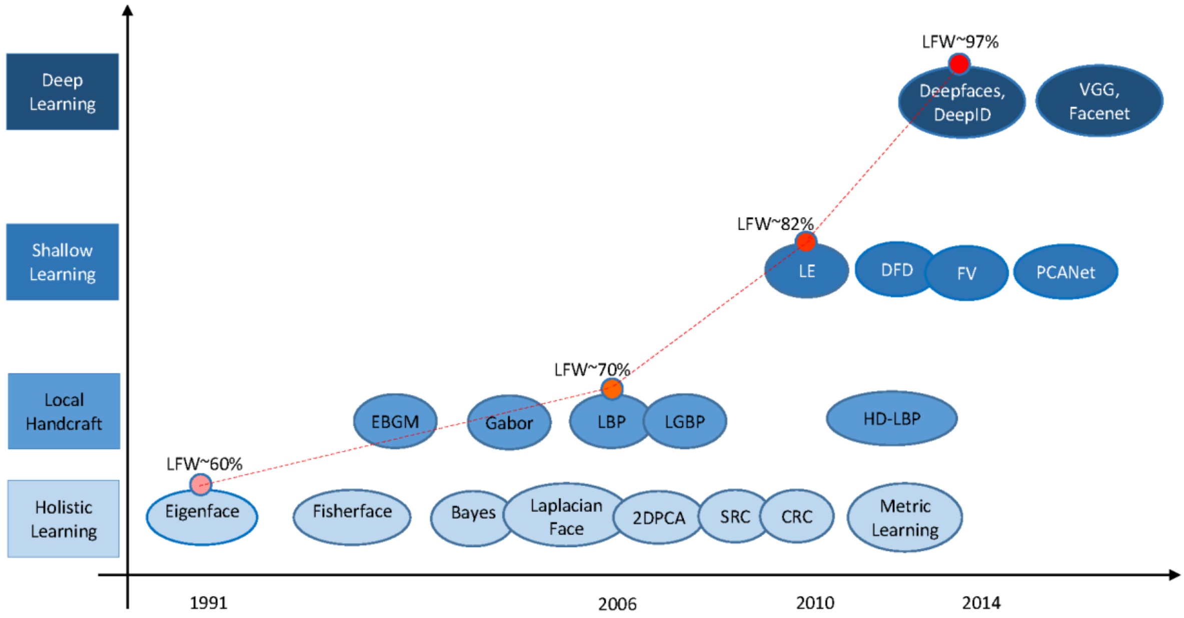
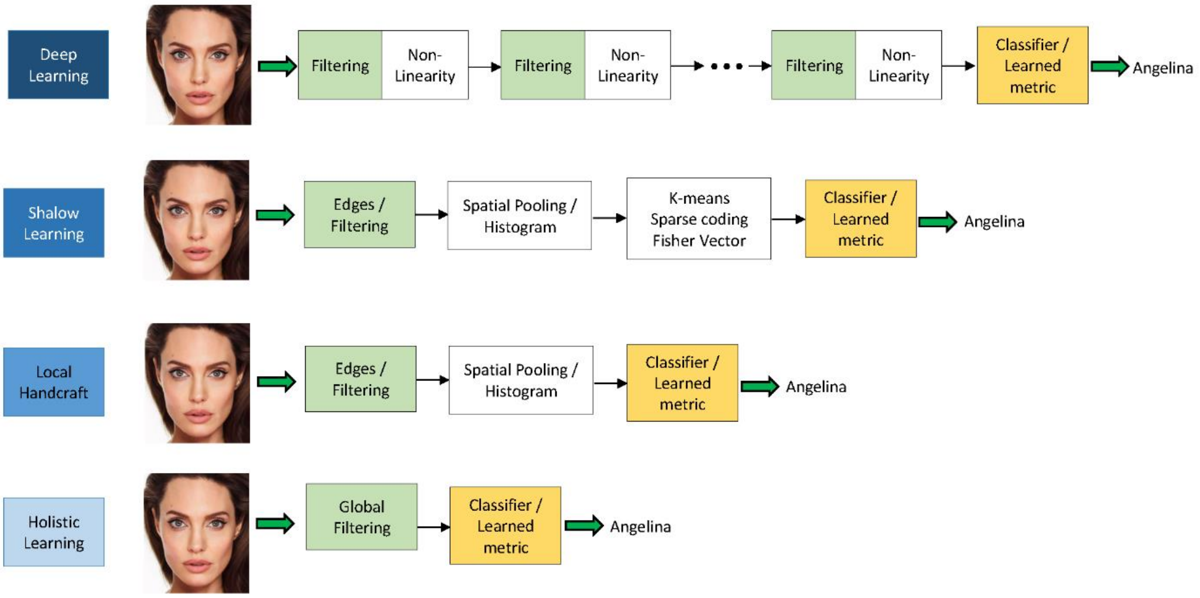


Figure 2. Diagram showing the evolution of the typical workflow for the four main types of face recognition technology considered in this work: holistic approaches, systems working on local handcrafted features, shallow learning systems and deep learning systems. (Source: Adpated from (Wang & Deng, 2018)



1.1. Workflow of face recognition systems

Face recognition is a visual pattern recognition problem, where the face, a 3D object that is subject to varying illumination, pose, expression, and other factors, needs to be identified based on acquired 2D images. While two-dimensional face images are commonly used in most applications, certain applications use three-dimensional (depth or range) images or optical images beyond the visual spectrum. As already mentioned, these latter systems lie beyond the scope of the present state of the art (please see Section 9.2 for a brief discussion of the possible future use of such technology within SIS).

A traditional face recognition system, generally consists of five modules: face detection, normalization, face processing, feature extraction, and matching. These modules are explained below.

Face detection segments the face area from the background. Face detection provides a coarse estimate of the location and scale of the face. A finer location is achieved through the face alignment module.

Face alignment/normalization is performed to normalize the face geometrically to canonical coordinates. This is usually achieved through face landmarking which localizes facial landmarks such as eyes, nose, mouth and facial outline.

Face processing is necessary because recognition methods are expected to recognize face images with varying pose and illumination and face samples in different conditions are not always available. This module is especially relevant in deep-learning systems, while it is not always present in traditional systems.

Face feature extraction is performed on the normalized face to extract salient information that is useful for distinguishing faces of different individuals and is preferably robust with respect to the geometric and photometric variations. The extracted face features are used for face comparison.

In the **face comparison** module, the extracted features from the input face are compared against one (verification) or many (identification) face references. Face verification computes one-to-one similarity between the reference and the probe to determine whether the two images are of the same subject, whereas face identification computes one-to-many similarity to determine the specific identity of a probe face. When the probe is certain to belong to one of the reference identities, this is referred to as closed-set identification; when the probes may or not belong to one of the reference identities, this is open-set identification. Further details about FR system evaluation under verification and identification operation modes are given in Section 2.1.

The main difference between traditional face recognition systems and deep-based approaches lies in the feature extraction algorithm:

- The features extracted in traditional systems are hand-crafted, that is, they are *defined* by the human expert designing the algorithm.
- On the other hand, the features extracted by the deep-based approaches are *learned* by the neural network based on a pool of data subjects which is used to train a network based on a specific loss function. The most common and successful deep neural networks architectures contain millions of parameters that need to be learned and, as such, the amount of data required for their accurate training is immense (in the range of tens of millions of images) compared to traditional systems where databases of tens of thousands of images are enough.

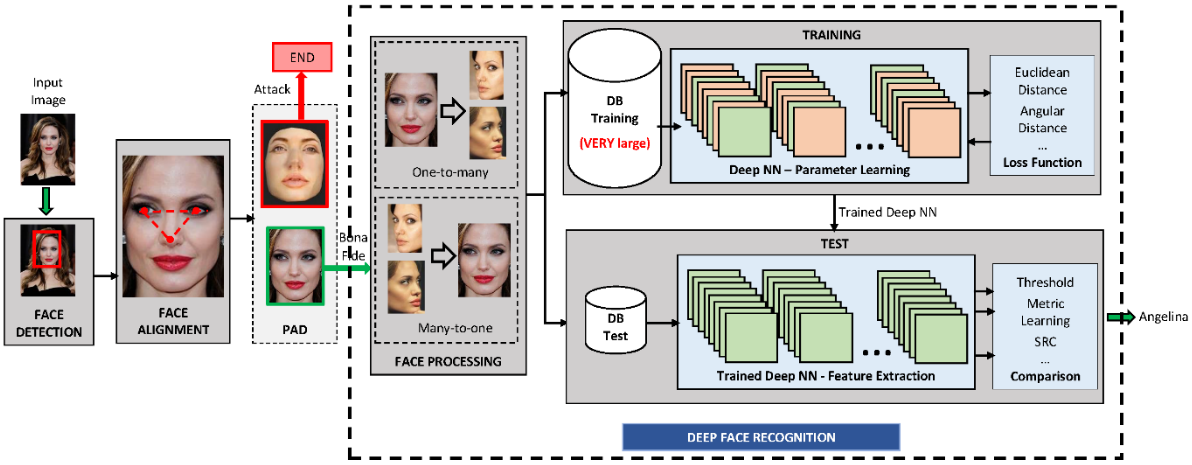
In **Figure 3** we show the typical pipe-line of a deep-learning based face recognition system. It should be highlighted that a fifth optional module has been added to this figure: the **presentation attack detection (PAD) and morphing attack detection module** (also referred to in the literature as anti-spoofing module). While such algorithm is not strictly required to perform face recognition, it is required in order to enhance the security of the system against presentation/morphing attacks. Its objective is to differentiate between real live faces presented to the sensor (bona-fide scenario) and replicated characteristics artefacts designed to deceive the system like photographs of a different user, morphed photographs or masks (presentation attack scenario). It should be noted that in **Figure 3** we have represented this optional module after the alignment algorithm, although it could be integrated in other parts of the system (e.g., within the sensor acquiring the image). Although somewhat out of the scope of the present report, for completeness, a brief review of potential threats to FR systems is given in Section 1.6.

In this state-of-the-art overview, we will mainly focus on up-to-date deep-feature-learning-based FR, as well as its closely related database development, face processing, and face comparison methods.

The first two modules of the workflow, face detection and alignment are beyond our consideration, and one can refer to Ranjan et al. [43], who provided a brief review of a full FR pipeline.

In the next subsections we will present a summary of the current state of the art for the three main modules described above for the case of deep-based systems: face processing, deep feature extraction and face comparison.

Figure 3. Deep FR system with face detector and alignment. First, a face detector is used to localize faces. Second, the faces are aligned to normalized canonical coordinates. Third, the FR module is implemented. The presentation attack detection module determines whether the presentation is an attack presentation or a bona fide presentation; face processing is used to handle recognition difficulty before training and testing; different architectures and loss functions are used to extract discriminative deep features when training; face comparison methods are used to do feature classification when the deep features of testing data are extracted. Source: Adpated from (Wang & Deng, 2018)



1.2. Face processing

Although deep-learning-based approaches have been widely used due to their powerful representation, [44] proved that various conditions, such as poses, illuminations, expressions and occlusions, still affect the performance of deep FR and that face processing is beneficial, particularly for poses. Since pose variation is widely regarded as a major challenge in automatic FR applications, we mainly summarize the deep methods of face processing for poses in this paper. Other variations can be solved by similar methods.

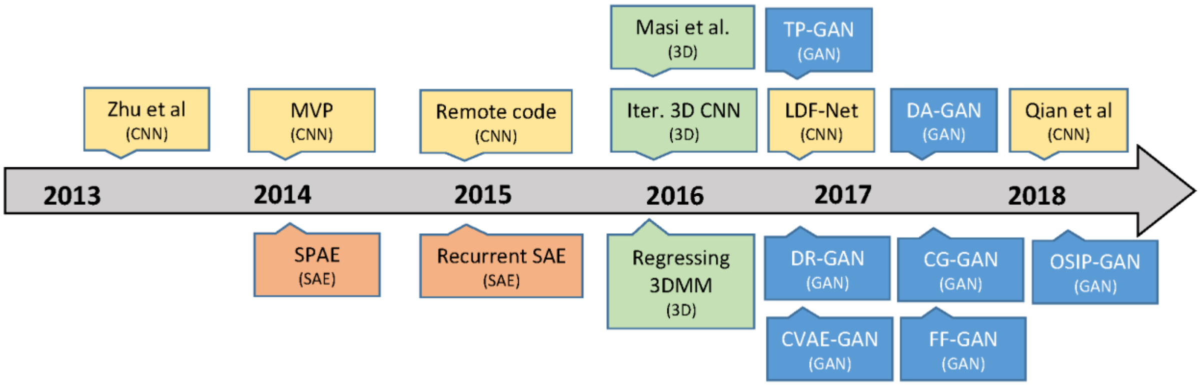
The face processing methods are categorized as "one-to-many augmentation" and "many-to-one normalization", as shown in **Table 2**. The timeline evolution of the different algorithms presented in **Table 2** is depicted in **Figure 4**.

- "One-to-many augmentation": generating many patches or images of the pose variability from a single image to enable deep networks to learn pose-invariant representations.
- "Many-to-one normalization": recovering the canonical view of face images from one or many images of a non-frontal view; then, FR can be performed as if it were under controlled conditions.

Table 2. Classification of different data pre-processing approaches. References are not comprehensive and are given as representative examples. See **Figure 4** for a timeline of the different algorithms presented here (colours in the table correspond to colours in the figure).

TYPE	BRIEF DESCRIPTION	ALGORITHMS
One to many	Generates many patches or images of the face variability from a single image (training data augmentation)	3D model [45] [46]
		2D deep model [47] [48]
		Data augmentation [49] [50]
Many to one	Recovers the canonical view of face images from many different images	SAE (Stacked AutoEncoders) [51] [52]
		CNN (Conv. Neural Nets.) [48] [53]
		GAN (Graph. Adversarial Nets.) [54] [55]

Figure 4. Timeline of the evolution of face processing algorithms for deep face recognition systems. The main types of algorithms given in **Table 2** are highlighted in different colours (corresponding to the colours in the cells of the table). Source: Adpated from (Wang & Deng, 2018)



1.3. Deep feature extraction

As shown in **Figure 3**, the feature extraction process in deep-based systems depends on two main factors: the architecture of the network used for FR and the loss function used for its training (i.e., learn the parameters that define the network).

Real-world FR can be regarded as an extremely fine-grained object classification task. For most applications, it is difficult to include the candidate faces during the training stage, which makes FR become a "zero-shot" learning task. Fortunately, since all human faces share a similar shape and texture, the representation learned from a small proportion of faces can generalize well to the rest.

The idea is to use the discriminative features learned from the training set to recognise faces of unseen subjects. To reach this objective it is key to include as many IDs as possible in the training set. For example, Internet giants such as Facebook and Google have reported their deep FR system trained by 10^6-10^7 data subjects [42] [22]. Unfortunately, these personal datasets, as well as prerequisite GPU clusters for distributed model training, are not usually accessible for academic community.

Currently, publicly available training databases for academic research consist of only 10^3-10^5 data subjects. Instead, the academic community is making the effort to design effective loss functions to make deep features more discriminative using the relatively small available training data sets.

Another initiative which is gaining momentum among small research laboratories and academia, in order to tackle the scarcity of training data for their models, is the development of algorithms capable of generating synthetic face images which present an appearance and variability similar to that of real faces. For this purpose, several works have shown the great potential of Generative Adversarial Networks (GANs).

1.3.1. Network architecture

Deep FR networks are in most cases based on a hierarchical architecture. Algorithms consist of multiple layers of simulated neurons that convolute and pool input, during which the effective receptive field size of simulated neurons are continually enlarged to integrate the low-level primary elements into multifarious facial attributes, finally feeding the data forward to one or more fully connected layer at the top of the network. The output is a compressed feature vector that represents the face. Such deep representation is widely considered the state-of-the-art technique for face recognition.

The architectures can be categorized as backbone and multiple networks, as shown in **Table 3**. Inspired by the extraordinary success on the ImageNet [56] challenge, the typical CNN architectures, such as AlexNet [25], VGGNet [57], GoogleNet [58], ResNet [59] and SENet [60], are introduced and widely used as the baseline model in FR (directly or slightly modified), as can be seen in the timeline evolution given in **Figure 5**. In addition to the mainstream, there are still some novel architectures designed for FR to improve efficiency.

Moreover, when adopting backbone networks as basic blocks, FR methods often train multiple networks with multiple inputs or multiple tasks. One network is for one type of input or one type of task. Hu et al. [61] shows that it provides an increase in performance after accumulating the results of multiple networks.

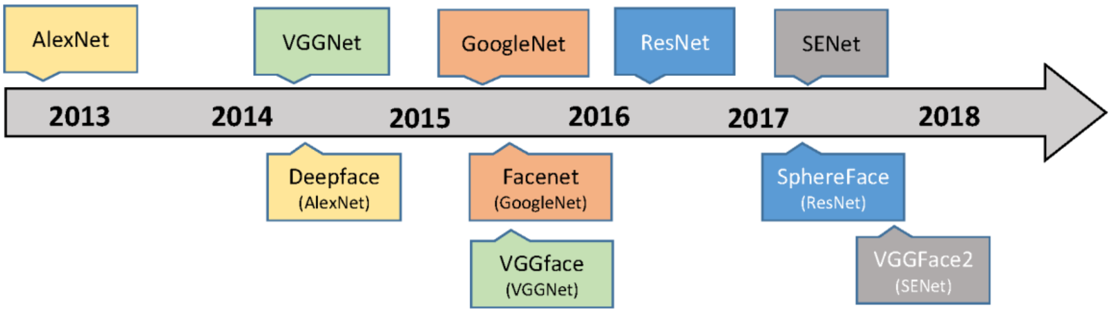
The main issue of these networks is the amount of data required for their proper training (i.e., the number of parameters needed to be learned are in the range of tens or hundreds of millions). A straightforward way to do this training is to include as many IDs as possible in the training set. For example, Internet giants such as Facebook and Google have reported their deep FR system trained by 10^6-10^7 data subjects [42] [22] (with multiple samples per data subject).

Unfortunately, these personal datasets, as well as prerequisite GPU clusters for distributed model training, are not accessible for academic community. Fortunately, since all human faces share a similar shape and texture, the representation learned from a small proportion of faces can generalize well to the rest. In any case the academic community is making the effort to design effective loss functions (the functions used to learn the parameters in the network) to make deep features more discriminative using the relatively small available training data sets.

Table 3. Classification of different network architectures used in deep FR. References are not comprehensive and are given as representative examples. See **Figure 5** for a timeline of the mainstream architectures. Source: Adpated from (Wang & Deng, 2018)

TYPE OF ARCHITECTURES	NETWORKS/REFERENCES
Backbone network	Mainstream architectures: AlexNet [42] [62], VGGNet [57] [63], GoogleNet [58] [64], ResNet [59] [65], SENet [60] [66]
	Special architectures [50] [67] [68]
	Joint alignment-representation architectures [69] [70] [71]
Multiple networks	Multipose [72] [73], multipatch [49] [74] [75], multitask [76]

Figure 5. The top row presents the typical network architectures in object classification, and the bottom row describes the well-known algorithms of deep FR that use the typical architectures and achieve good performance. The same colour rectangles mean the same architecture.
It is easy to find that the architectures of deep FR have always followed those of deep object classification and evolved from AlexNet to SENet rapidly. Source: Adpated from (Wang & Deng, 2018)



1.3.2. Loss function

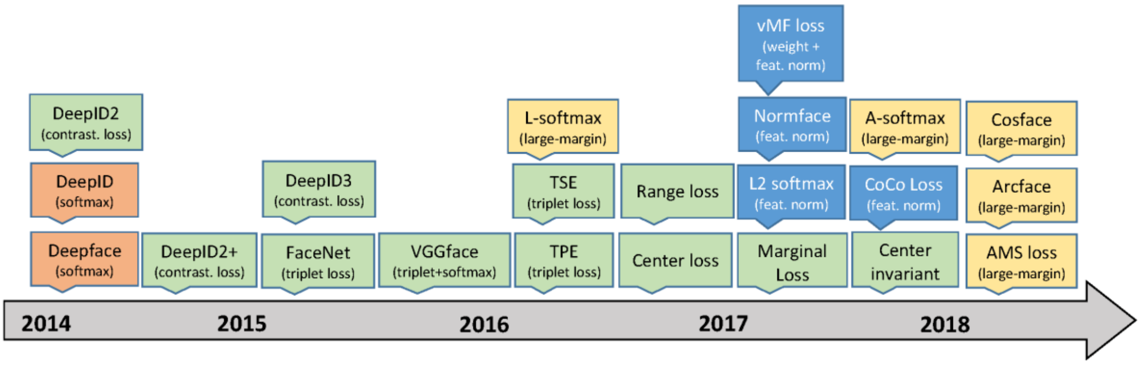
As mentioned above, the loss function is used to learn the parameters of the deep neural network based on a training pool of data subjects (as large as possible). The softmax loss is commonly used as the supervision signal in object recognition, as it tends to improve the separability of features. This function is defined on the output (decision layer) of deep networks. However, for FR, when intra-variations could be larger than inter-differences, the softmax loss is not sufficiently effective. For this reason, softmax is in many cases used together with other loss functions that operate on the feature vectors before the final fully connected layers of deep networks. Many works focus on creating novel loss functions to make features more separable, as shown in **Table 4**. A timeline of the evolution of loss functions is given in **Figure 6**.

- Euclidean-distance-based loss: compressing intravariance and enlarging inter-variance based on Euclidean distance.
- Angular/cosine-margin-based loss: learning discriminative face features in terms of angular similarity, leading to potentially larger angular/cosine separability between learned features.
- Softmax loss and its variations: directly using softmax loss or modifying it to improve performance, e.g., L2 normalization on features or weights as well as noise injection.

Table 4. Classification of loss functions used in deep FR. References are not comprehensive and are given as representative examples. See **Figure 6** for a timeline of the evolution of the loss functions where colours correspond to the rows in this table.

LOSS FUNCTIONS	**DESCRIPTION/REFERENCES**
Euclidean Distance-based	Compressing intra-variance and enlarging inter-variance based on Euclidean distance [62] [23] [42] [77] [78] [79].
Angular/cosine, large-margin, based	Making learned features potentially separable with larger angular/cosine distance [65] [80] [81] [82] [83].
Softmax	Modifying the softmax loss to improve performance [84] [85] [86].

Figure 6. The development of loss functions. It marks the beginning of deep FR that Deepface [22] and DeepID [74] were introduced in 2014. After that, Euclidean-distance-based loss always played the important role in loss function, such as contractive loss, triplet loss and center loss. In 2016 and 2017, L-softmax [83] and A-softmax [65] further promoted the development of the large-margin feature learning. In 2017, feature and weight normalization also begun to show excellent performance, which leads to the study on variations of softmax. Red, green, blue and yellow rectangles represent deep methods with softmax, Euclidean-distance-based loss, angular/cosine-margin-based loss and variations of softmax, respectively. Source: Adpated from (Wang & Deng, 2018)



1.4. Face comparison

After the deep networks are trained with massive data and an appropriate loss function, each of the test images is passed through the networks to obtain a deep feature representation. Once the deep feature vectors are extracted, most methods directly calculate the similarity between two feature sets using cosine distance or L2 distance. In addition to these, other methods are introduced to post-process the deep features and perform the face comparison efficiently and accurately, such as metric learning or sparse-representation-based classifier (SRC).

Please see Section 2.1 for further details on face comparison under the verification and identification operation modes and on the standard metrics used for the performance assessment under both tasks.

1.5. Other attributes related to face

While the present report is focused on the analysis of the very specific task of Face Recognition, human faces reveal also other attributes in addition to identity. This is the case for instance of age, gender or race. These other attributes are, in a way, associated to the identity and, therefore, can potentially help to improve the accuracy of FR systems. However, analysing the specific algorithms developed to perform each of these tasks (e.g., age estimation, gender or race detection), fall out of the scope of the report and the interested reader is referred elsewhere to obtain more specific details on any of them.

We would like to highlight here, however, that the revolution brought to FR by deep learning technology, has also influenced the recognition of these other face-related attributes. In fact, at the moment, in many cases, the same deep networks used for face recognition are being employed as well to detect age, gender or race, changing only the very last layers of the networks to adapt to each specific modality. This is one of the big advantages of deep learning, the features automatically learned by the networks embed all the relevant information regarding the face including identity, gender, age or race. This way, changing only the final classifier, the same features can be used to obtain any of these attributes [76].

1.6. Threats to FR systems

In addition to the threats that have to be taken into account in any security system (e.g., cyberattacks), Face Recognition systems are subjected to two types of technology specific attacks that have been largely studied in the literature: presentation attacks (also referred to in the literature as spoofing) and morphing attacks. Even though their detailed analysis falls out of the scope of the present report, for the sake of completeness, the next two subsections present a brief review of each of these two potential security weaknesses. This way these sections can also help to raise awareness of the existence of such threats in case that they need to be addressed in certain future use-cases of SIS.

1.6.1. Face presentation attacks

Biometric presentation attacks, are in general understood as attacks directed against the acquisition sensor of the system, in which the attacker may have one of two main purposes:

- **Impersonation attacks**. Where the objective is to trick the system to think that he is a different person, in this case the attacker uses an artificial physical artefact modelling the biometric characteristic of a legitimate user and presents it to the sensor to gain illegitimate access.
- **Evasion attacks**. Where the attacker disguises his identity in order not to be recognised by the system.

Independently of the purpose of the attacks (i.e., impersonation or evasion), in the case of face biometrics, the vast majority of presentation attacks reported in the literature may be classified in four main types of attacks:

- **Make-up Attacks**. In this case the attacker does not use any type of artefact but simply wears some specific make up that may deceive the recognition system. Although the success rate of these attacks is still under evaluation, they can pose a very important threat to the system as they may be difficult to detect even for a human supervisor.
- **Photo Attacks**. These fraudulent access attempts are carried out presenting to the recognition system a photograph of the genuine user. The photograph may have been taken by the attacker using a digital camera, or even retrieved from the internet after the user himself uploaded it to one of the very popular online social networks available today. The image can then be printed on a paper (i.e., print attacks, which were the first to be systematically studied in the literature) or may be displayed on the screen of a digital device such as a mobile phone or a tablet (i.e., digital-photo attacks).
- **Video Attacks**. Also referred in some cases as replay attacks. They represent a more sophisticated version of the simple photo spoofs. In this case, the attacker does not use a still image, but replays a video of the genuine client using a digital device (e.g., mobile phone, tablet or laptop). Such attacks appeared as a further step in the evolution of face presentation attacks and are more difficult to detect, as not only the face 2D texture is copied but also its dynamics.
- **Mask Attacks**. In these cases, the presentation artefact is a 3D mask of the genuine client's face, increasing the difficulty to find accurate countermeasures against them. Since the complete 3D structure of the face is imitated, the use of depth cues which could be a solution to prevent the previous two types of attacks (carried out with flat surfaces), becomes inefficient against this particular threat. Although the possibility to bypass a biometric system wearing a mask imitating the face of a different user is an idea that has been circulating for some time, these attacks are less common than the previous two categories in research publications.

All previous attacks have a number of variants depending on the resolution of the device presented to the sensor (e.g., tablet, printed paper), the type of support used to present the fake copy (e.g., handheld or fixed support), or the external variability allowed (e.g., illumination or background conditions).

It is worth highlighting that face recognition systems may also be subjected to attacks from identical twins claiming to be the same person. Strictly speaking these are not presentation attacks (as there is no physical artifact involved) but rather zero-effort impostor attempts in which Bob presents his own biometric trait while trying to access the system as John. We refer the interested reader to some specific works on this topic to understand the implications and performance of face recognition systems in the presence of twins [87].

Regarding the specific case of **SIS**, given the nature of the system, the attacks that could be expected are the ones in which an individual who is already registered in the system (i.e., there is an alert related to that person), tries to hide his identity in order not to be recognized, that is, **evasion attacks**. This is a situation that could potentially happen in a border crossing using automatic kiosks or ABC gates, while it is an unlikely attack to be performed if there is any human supervision (e.g., border crossing with a border guard or in a police station).

In order to prevent these attacks, extensive research has been carried out to develop Presentation Attack Detection (PAD) methods that should be integrated in those systems with a significant risk of facing this type of threat such as ABC gates. Further reading on face PAD technology can be found in [88], [89], [90].

The presentation attacks and the presentation attack detection technology are standardised in: ISO/IEC 30107:2017, Biometric presentation attack detection.

1.6.2. Face morphing attacks

Morphing algorithms are a special type of transformation methods which aim at merging a real face image belonging to one subject (i.e., source) with a real face image belonging to a second subject (i.e., target), usually with the intention that the resulting synthetic sample can be positively matched to both of the previous identities. These algorithms can pose a threat to certain applications as one unique "synthetic" face is not linked to one identity but to two different persons.

It was shown recently, that the issuing protocol of the ePass presents a security issue with respect to morphing algorithms [91]. The key deficiency in the passport issuance process lies in the way the facial picture of an applicant is processed. In many countries, the applicant provides a printed facial image which is scanned and then digitally transferred to the passport production site. As the facial image is provided by the applicant, it can be manipulated prior to the disposal at the federal offices. As a consequence, a specific attack scenario are morphed face attacks. An artificial facial image, which is referred to as morph, is created by blending the facial images of two or more different data subjects into one. If the newly generated facial image is enrolled to a Face Recognition system, the subjects contributing to the morphed image are positively verified against the morphed face attack reference, as the resulting morphed image resembles the constituent faces, both in visual and feature representation. Exploiting this fact, a black-listed subject (criminal) could be potentially able to obtain a legitimate ePass, by morphing his facial image with that of a non-listed subject (accomplice), which the accomplice utilizes to apply for a passport. Due to the infiltration during the issuance process, the accomplice, as well as the criminal, are able to verify successfully against the reference stored in the ePass. The feasibility of such morphed face attacks has been empirically confirmed in [91].

The application process of the ePass in many countries (e.g. most of the European Schengen states) still requires a printed face image that will be handed over to the public authority office during the application process. A way to protect systems against morphing attacks is to use live enrolment at the issuing of the document. In case of live-enrolment under supervision of an officer, the morphed face attack is not relevant. Unfortunately, only very few countries to date have decided to establish live-enrolment and many that have live-enrolment operate kiosks in an unsupervised manner. However, in the meantime, in the majority of countries the face image is still captured by a photographer or private person, printed and handed to the public authority office, where it is scanned again, a process compliant to the ICAO-standard. Thus, morphed face attack detection is not limited to the digital domain, but extended to the detection of morphing attacks after printing and scanning of the morphed face attack sample, which is even more challenging, as during the scanning process of the face images information is lost and noise and granularity are added.

For the specific case of SIS, morphing attacks are very unlikely to happen since the main objective of this threat is to have two or more subjects positively matched to the same enrolled picture. Given the nature of the alerts in SIS (see PART II of the present report), the expected attacks will come from subjects in the system trying to avoid recognition, rather than from persons who are not in the system trying to be recognised as someone with an alert in the database. Furthermore, as mentioned above, in order for a morphing attack to be successful, the attacker has to provide his own "morphed" facial image. This will very rarely be the case in SIS, since in all the foreseen use-cases (see PART II of the present report), both the enrolled image and the image used to query the database will be taken live under human supervision.

Section 1. Summary of key concepts:

1. Since 2014 Face Recognition (FR) based on deep-learning technique has clearly superseded classical techniques based on holistic learning, local handcraft features or shallow learning.
2. The biggest challenge of deep-learning technology is the large amount of training data which is required to obtain good accuracy. These training data should be representative of the population that the system will operate on, in order to avoid biased results (e.g., if a system is trained on faces of Asian people, it will obtain better accuracy on this race than on black or white people)
3. Currently there are a number of pre-trained deep-learning networks publicly available to perform high-accuracy FR. These networks can in general be downloaded and applied directly to the task of FR with quite good results.
4. Just retraining the last layers, these same deep networks have been successfully used to detect other face-related attributes such as age, gender or race, which, in turn, can help to boost the comparison accuracy in FR.
5. FR systems are subjected to a number of potential threats that should be considered on a case by case basis depending on the specific context of each application. From these vulnerabilities, presentation attacks and morphing attacks have shown to be especially dangerous in certain contexts. Their applicability in the context of SIS is however quite unlikely.
6. An evaluation of presentation attacks and of presentation attack detection methods should follow to the largest extent possible the guidelines given in the standard "ISO/IEC 30107-1, Biometric presentation attack detection".

2. Face Recognition Accuracy Evaluation

Considered as the main pillar for determining the readiness and availability of FR technology, accuracy constitutes the main focus of the present section. This dimension is addressed in the present section through an analysis of: 1) a review of the most common evaluation scenarios and metrics; 2) the available training and testing databases; 3) the most significant independent evaluation campaigns conducted so far.

Although the most important, accuracy represents only one of the parameters which determine the performance of an FR system. Other performance parameters that are not considered in this section (or only from a very general perspective) are for example: transaction times, computational efficiency and template size.

Objective accuracy evaluation of biometric systems is not a straightforward task. In an ideal situation, one would like to assess the application-independent accuracy of a recognition system and be able to predict its real operational accuracy in any context. In this ideal scenario, rigorous and realistic modelling techniques simulating data acquisition and comparison processes are the only way to obtain and extrapolate the accuracy evaluation results. More research effort is still required to further address this problem.

In the meantime, performing comparative evaluations on specific scenarios is the norm. Even if aspects of biometric recognition algorithms and application requirements are clearly understood, comparative and empirical application-dependent evaluation techniques will be predominant and the evaluation results obtained using these techniques will be meaningful mainly for a specific database in a specific test environment and a specific application.

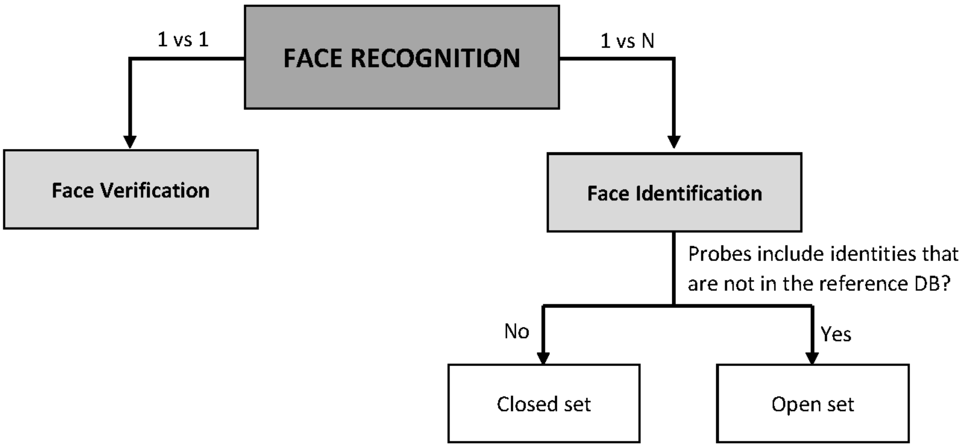
Another disadvantage of empirical evaluation is that it is usually expensive to collect the data for each evaluation, complement them with the ground-truth metadata and implement appropriate data protection measures so as to fulfil the obligations related to this purpose. It has to be highlighted that objectively comparing the evaluation results of two different systems, tested under different conditions, presents clear limitations regarding the relevance of the benchmark.

Depending upon the data collection protocol, the accuracy results can vary significantly from one benchmark to another. Within the biometric recognition context, a benchmark is defined by a database and an associated testing protocol. Generally, the protocol defines (at least) the subsets of images that can be used for training and testing, the pair of images that have to be compared, the performance metrics to be used and how they must be computed.

2.1. Evaluation scenarios and metrics

As in any other biometric related technology, face recognition systems can be evaluated in a different number of tasks as shown in **Figure 7**. Each of these tasks has associated specific metrics in order to assess the accuracy of the systems. Although it falls out of the scope of the present document to review how each of these metrics is computed, for completion, we include here a brief description of both tasks and metrics. For further details on this topic the reader is referred elsewhere [92].

Figure 7. Classification of different evaluation tasks. The performance of recognition model can be evaluated under face verification, close-set face identification, open-set face identification settings. Source: EC 2018



As explained in the previous section, deep learning-based FR algorithms are trained on specific sets of data (usually very large). In terms of training protocol, FR model should always be evaluated under subject-independent protocol:

- *Subject-independent protocol*: the testing identities are disjoint from the training set. That is, no data subjects are shared between the training and the test sets. This is the closest evaluation protocol to real world applications. In this setting it is not possible to train specific models for a given data subject in the training set, since it will not be present in in the test sets. Therefore, it is essential to achieve a subject-independent (generalized) representation of the human face. Due to the fact that human faces exhibit similar intra-subject variations, deep models are able to generalize well when training with a sufficiently large set of subjects, where the key is to learn discriminative large-margin deep features. All major face-recognition benchmarks, such as LFW, IJB-A/B/C and Megaface, require the tested models to be trained under subject-independent protocol. Also, all relevant independent FR evaluations are performed under the subject independent protocol.

Once the model has been trained, FR systems can be tested under the verification or identification scenarios (see **Figure 7**). In the identification scenario, the evaluation can be closed set or open set. Each of these evaluations tasks has its own metrics.

For further details on the evaluation of biometric systems and on the metrics used in each of the scenarios we refer the reader to the standard ISO/IEC 19795-1 2006 "Information Technology – Biometric Performance and reporting – Part 1: principles and framework". Although some parts of this standard are of particular interest for the report, it is beyond its scope to describe in detail the different technical aspects that go into accuracy evaluation of biometrics. The interested reader is therefore referred to the ISO online library.

FACE VERIFICATION: refers to a one-to-one comparison. That is, the input to the system are two face images and the system has to decide whether or not they come from the same person. This scenario is relevant to access control systems, re-identification, and application independent evaluations of FR algorithms. This scenario would be the case for instance, of a border guard that compares a live image captured at the border with the image contained in the passport of a traveller. The result of such comparison will tell the border guard whether or not the document belongs to the traveller (i.e., the person enrolled in the document and the person holding the document are the same person).

METRICS. As defined in the standard ISO/IEC 19795-1 2006, the fundamental performance metrics that should be used to measure accuracy in the verification scenario are:

- False Match Rate (FMR): is the proportion of non-mated samples (i.e., not belonging to the same subject), that are falsely declared as match. The True Match Rate (TMR) can then be easily computed as 1-FMR.
- False Non-Match Rate (FNMR): is the proportion of mated samples (i.e., belonging to the same subject), that are falsely declare to non-match. The True Non-Match Rate (TNMR) can then be easily computed as 1-FNMR.

It should be noted that there are other possible metrics to evaluate performance in the face verification scenario, however, in most cases they can be derived from these two. The most common way to report the verification accuracy of a system is to give the FNMR for a fixed FMR (e.g., FNMR@FMR=0.01%).

CLOSED-SET IDENTIFICATION: in this case the input to the system is a face image (probe) corresponding to a subject which is known to be already inside the reference database. The system has to find the person within the database (i.e., the output of the system is the entry of the database corresponding to the subject of the input image). Therefore, in this scenario there is a one-to-N comparison, where N is the size of the reference database.

METRICS. As defined in the standard ISO/IEC 19795-1 2006, the fundamental performance metrics that should be used to measure accuracy in the closed-set identification scenario are:

- Identification rate at rank r: is the probability that a transaction by a user enrolled in the system includes that user's true identifier within the top r matches returned. When a single point identification rank is reported, it should be referenced directly to the database size. Example: "The identification rate at rank 1 was 95 % against a database of 250 entries".

The primary measure of closed-set identification performance is normally shown as a cumulative match characteristic (CMC) curve in which the (true-positive) identification rate at rank r is plotted as a function of r.

One drawback of the CMC is its dependence on the number of people enrolled in the system. For this reason, a graph plotting the identification rate at rank 1 (i.e., r=1) as a function of the number of enrolments should be included with the results.

OPEN-SET IDENTIFICATION: in this case the input to the system is a face image (probe) corresponding to a subject which may or not be contained in a database (gallery). This scenario is relevant to high throughput face search systems (e.g., de-duplication, watch list), where the recognition system should reject unknown/unseen subjects (probes who are not present in the gallery) at test time. In this case the output of the system is either the identity of the search subject within the database (if, in fact, the subject is present in the database), or a notification that the person has not been found in the database.

This is the operation mode of CS-SIS, where the subjects of queries may or not be present in the database.

METRICS. As defined in the standard ISO/IEC 19795-1 2006, the fundamental performance metrics that should be used to measure accuracy in the open-set identification scenario are:

- (True positive) identification rate at rank r: is the probability that a transaction by a user enrolled in the system includes that user's true identifier within the top r matches returned. When a single point identification rank is reported, it should be referenced directly to the database size. Example: "The identification rate at rank 1 was 95 % against a database of 250 entries".
- False-negative identification-error rate (FNIR): is the proportion of identification transactions by users enrolled in the system, for which the user's correct identifier is not included in the candidate list returned.
- False-positive identification-error rate (FPIR): is the proportion of identification transactions by users not enrolled in the system, for which a non-empty list of candidate identifiers is returned. NOTE: The false-positive identification-error rate increases with the number of people enrolled in the system.

The overall identification performance of an open-set system, as the enrolment database grows may be shown as a plot of identification rate (at rank 1) against size of enrolment database, for a constant value of the false-positive identification-error rate (requiring the threshold to be adjusted as the database grows).

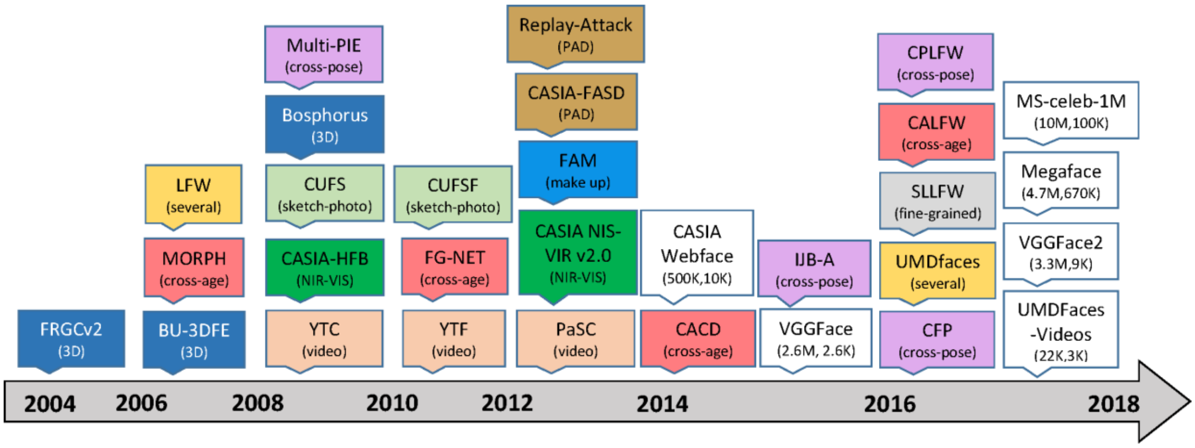
A common way to report the accuracy of a system in the open-set identification scenario is to give, for a given rank, the FNIR for a fixed FPIR (e.g., FNIR@FPIR=0.01% for r=1).

It is important to highlight that, for most real-world applications conducting searches in a reference database, including SIS as already mentioned, the most relevant evaluation scenario is **identification in open-set**. Any performance evaluation of SIS should specifically focus on this scenario, following the guidelines given in the ISO/IEC 19795-1 standard.

2.2. Training and evaluation databases

In the past three decades, many face databases have been constructed with a clear tendency from small-scale to large-scale, from single-source to diverse-sources, and from lab-controlled to real-world unconstrained conditions, as shown in **Figure 8**. As the performance of some simple databases became saturated, more and more complex databases were continually developed to facilitate the FR research. It can be said that the development process of the face databases has largely led the direction of FR research. In this section, we review the development of major training and testing databases for deep FR.

Figure 8. The evolution of FR datasets. Before 2007, early work in FR focused on controlled and small-scale datasets. In 2007, LFW [24] dataset was introduced which marks the beginning of FR under unconstrained conditions (see Section 3 for further details regarding the differences between controlled and unconstrained conditions). Since then, more testing databases with different tasks and scenes are designed. In 2014, CASIA-Webface [93] provided the first widely-used public training dataset. White rectangles represent training datasets. Rectangles of other colours represent testing datasets designed for different tasks indicated in parenthesis (PAD stands for Presentation Attack Detection). Source: Adpated from (Wang & Deng, 2018)



2.2.1. Large-Scale Training Datasets

As already mentioned in the present document, the prerequisite of effective deep FR is a sufficiently large training dataset. Large amounts of training data combined with deep learning improve the performance of FR. Also, it should be highlighted again here that the variability observed in the training data should be similar to that observed in the test data in order to obtain unbiased results. For example, a system trained on a very large database containing only male face images, will perform very poorly on a database of female face images.

The results of Megaface Challenge revealed that premier deep FR methods were typically trained on data larger than 0.5M images and 20K subjects. The early works of deep FR were usually trained on private training datasets. Facebook's Deepface [22] model was trained on 4M images of 4K subjects; Google's FaceNet [42] was trained on 200M images of 3M subjects; DeepID serial models [74] [23] [79] were trained on 0.2M images of 10K subjects. Although they reported ground-breaking performance at this stage, researchers cannot accurately reproduce or compare their models without public training datasets.

To address this issue, CASIA-Webface [93] provided the first widely-used public training dataset for the purpose of training deep models. This database consists of 0.5M images of 10K celebrities, collected from the web. Given its moderate size and easy usage, it has become a great resource for fair benchmarks for academic deep models. However, its relatively small data and ID size may not be sufficient to reflect the power of many advanced deep learning methods. Currently, more databases have been made available, providing public large-scale training datasets, which are summarized in **Table 5**. Especially relevant are three databases with over 1M images, namely MS-Celeb-1M [94], VGGface2 [66], and Megaface [95]. These large training sets are expanded from depth or breadth:

- VGGface2 provides a large-scale training dataset in depth, that is, it contains a limited number of subjects but many images for each subject. The depth of a dataset enforces the trained model to address a wide range of intra-class variations, such as lighting, age, and pose.
- In contrast, MS-Celeb-1M and Megaface (Challenge 2) offer large-scale training datasets in breadth, that is, they contain many subjects but limited images for each subject. The breadth of a dataset ensures the trained model to cover the variability observed among different people.

Cao et al. [66] conducted a systematic study on model training using VGGface2 and MS-Celeb-1M, and found an optimal model by first training on MS-Celeb-1M (breadth, inter-variability) and then fine-tuning on VGGface2 (depth, intra-variability).

Table 5. Publicly available large-scale datasets to train deep FR models. These databases correspond to the red databases shown in **Figure 8**.

DATASET	YEAR	#IMAGES	#SUBJECTS
CASIA WebFace [93]	2014	500K	10K
VGGFace [41]	2015	2.6M	2.5K
MS-celeb-1M (Challenge 1) [94]	2016	10M 3.8M (clean)	100K 85K
MS-celeb-1M (Challenge 2) [94]	2016	1.5M (base set) 1K (novel set)	20K 1K
Megaface [95]	2016	4.7M	650K
VGGFace2 [66]	2017	3.31M	10K
UMDFaces-Videos [96]	2017	22K	3K
MS-celeb-1M (Challenge 3) [97]	2018	4M (MSv1c) 2.8M (Asian-celeb)	80K 100K

2.2.2. Test Datasets

There are many testing datasets with different scenarios to mimic FR in real life as shown in **Table 6**. According to their characteristics, these scenarios can be divided into four categories:

- **Cross-factor FR**. This accounts for intra-class variability such as cross-pose, cross-age or make-up.
- **Heterogeneous FR**. It refers to the problem of comparing faces across different visual domains. The domain gap is mainly caused by sensor devices and cameras

settings, e.g. visual light vs. near-infrared and photo vs. sketch (some further details about this field are given in Section 9.2).

- **Multiple media FR**. Ideally, deep models are trained with massive images per subject and are tested with one/few image/s per subject, but the situation may be different in reality. Sometimes, the number of images per subject in the training set could be very small, referred to as low-shot FR; or each subject face in the test set is often enrolled with a set of data coming from multiple media sources (e.g, images and videos).
- **FR in industry**. Deep FR has achieved higher performance than humans on some standard benchmarks. However, other performance factors different than accuracy should be paid attention to when deep FR is adopted in industry, e.g. presentation attack detection.

In **Table 6** we present a review of the most popular test databases for FR, together with the metrics used in the evaluation (see Section 2.1 for further details on accuracy metrics) and the best performing algorithms.

Table 6. Summary of some popular test databases for FR and the top accuracy obtained so far together with the references to these best performing deep-based algorithms. See section 2.1 for a brief explanation of the accuracy metrics. TAR=True Acceptance Rate; FAR=False Acceptance Rate; TPIR=True Positive Identification Rate; FPIR=False Positive Identification Rate

DATASET	DATE	#IMAGES	#SUBJECTS	METRICS	ACCURACY
MORPH [98]	2006	55K	13.6K	1:N Rank-1	94.4% [99]
LFW [24]	2007	13K	5K	1:1 TMR vs FMR	99.78% [86]
				1:N Rank-1	99.63% [42]
FG-NET [100]	2010	1K	82	1:N Rank-1	88.1% [101]
IJB-A [102]	2015	25K	500	1:1 TMR@FMR=10-3	92.10% [66]
				1:N Rank-1	98.20% [66]
CFP [103]	2016	7K	500	1:1 TMR vs FMR	Fr-Fr: 98.67% [104] Fr-Pr: 94.39% [105]
UMDFaces [96]	2016	350K	8.5K	1:1 IR@FPIR=10-2	69.30% [25]
MegaFace [95]	2016	1M	690K	1:1 TMR@FMR=10-6	86.47% [42]
				1:N Rank-1	70.50% [42]
IJB-B [106]	2017	12K images	1.8K	1:1 TMR@FMR=10-5	70.50% [66]
		7K videos		1:N Rank-1	90.20% [66]
CFPLW [107]	2017	11.5K	4K	1:1 TMR vs FMR	77.90% [41]
MS-Celeb-1M Challenge 3 [97]	2018	274K (ELFW)	5.7K	1:1 TMR@FMR=10-9	46.15% [80]
		1M (DELFW)	1.6K	1:N IR@FPIR=10-3	43.88% [80]

2.3. Large-scale evaluation campaigns

Below, a summary of the most important independent evaluation campaigns that have been performed in face recognition will be presented. It is important to note that, as pointed out before, any effort to assess the general performance of any biometric system in verification or identification transactions must be undertaken with care. The accuracy and overall performance of any method or system will depend on multiple factors which are difficult to model or to objectively quantify, including the quality of data input (and hence the sensor and feature extraction algorithms), the specific comparison algorithms used, the population being assessed and in the case of identification from a database, the number of entries to be searched. Thus, the results presented in this section should be assessed with these caveats as they are valid only for the use-cases and situations in which the testing described was carried out. Performance expectations in other scenarios based on these results cannot simply be assumed.

Bearing in mind the previous caution, the results obtained in a performance evaluation can be useful for environments similar to the one envisaged by the data used and, if properly analysed, they can reveal important factors to consider in other environments (and most importantly: how to consider such effects). What is important to emphasize is that, in those cases where the scenario under study is not identical to the one defined in a benchmark, the results previously obtained in that benchmark should be carefully interpreted and probably adapted, e.g. previous NIST Face Recognition Vendor Test (FRVT) evaluations (see Section 2.3.1.2) disclose very important information to be considered as a basis for the proposed development of an Automatic Face Recognition System in SIS (a basis that can be fine-tuned with specific and more targeted benchmarks using real SIS data).

Therefore, the series of evaluations presented below, can help to provide an overall picture not only of the evolution of the state of the art over the last 15 years but also on the performance capabilities of face recognition systems today. In the following we will distinguish between two types of evaluations:

- **Governmental evaluations**. These evaluations are competitions characterized by: 1) organised by an institution; 2) with a specific time framework and deadlines; 3) with a very specific and clear protocol defining tests and tasks to be evaluated; 4) where test data is only available to the organising institution, not to the participants; 5) the FR algorithms are made available from participants to the organising institution so that the latter performs all the tests under the same conditions for all systems.
- **Academic evaluations**. With this term we refer to very popular benchmarks, defined by a dataset, testing protocol and metrics. Some of these benchmarks, due to their extended use by the community, have become over the years a reliable testbed to track the evolution of the state of the art in FR. This is the case, for instance, of the largely used LFW dataset where the authors have kept over the years an up-to-date website with all the results obtained by the different algorithms that have been assessed on the dataset following the official protocol. However, in these cases, both the training and test data are available to participants, who obtain the results in their own platform. Therefore, since there is no organising trustworthy third-party, comparative results should be taken with more care than in governmental evaluations.

All these initiatives will be briefly described in the following sections. In these evaluations FR algorithms are in general assessed both under the verification and identification scenarios. The most relevant results for SIS are those related to identification in open set, which will be the operation mode of SIS (see Section 2.1 for a description of testing scenarios and associated metrics).

In particular, the closest evaluation to SIS, both in terms of type of face data and size of the database, is the FRVT 2017 (described in Section 2.3.1.2). Even if results **cannot be directly extrapolated**, this evaluation can be considered as a good estimation of the accuracy that can be expected from FR technology in the SIS operational environment at the present moment.

The reader should also bear in mind that all these evaluations present a **snapshot** at a given moment in time of the state of the art of FR systems. As has been already described in Section 1, this technology is evolving very rapidly and, with the arrival of the new deep-learning algorithms and the increase in computation power, systems are getting faster and more accurate every year. Therefore, the results obtained for instance in FRVT 2017 will surely be improved in the near future.

Please see Section 2.1 for a description of the metrics used in these evaluations to assess the accuracy of the systems.

2.3.1. Governmental evaluations

The key to the organisation of a biometric technology evaluation is the access by the hosting institution to a set of test data not previously seen by the participants. This test dataset has to be designed to suit the type of experiments and tasks being assessed (e.g., cross-pose face recognition, cross-age face recognition, face recognition in the wild, etc.) To date, the US National Institute of Standards and Technology (US-NIST), with the support of IARPA (Intelligence Advanced Research Project Activity) and other US institutions like the FBI (Federal Bureau of Investigation) or DHS (Department of Homeland Security), which have provided access to the necessary data, has led all the most relevant official FR evaluations.

In this regard, Europe is still lacking behind as, to date, there has not been organised any similar initiative within the EU.

The most significant governmental evaluation campaigns that have been carried out so far in face recognition are specified here below. Please be aware that reference to quite old evaluations is given here for completeness and also as a way to show the huge evolution that FR has accomplished throughout the last three decades.

However, regarding SIS, the most relevant results are those that have been achieved in the most recent evaluations, where the submitted FR systems are up-to-date applications based on deep-learning algorithms such as the ones that can be expected to run in SIS. In particular, further details are given in Sections 2.3.1.1 and 2.3.1.2 of the last two evaluations carried out by NIST in 2017, the Face Recognition Prize Challenge (FRPC) and the Face Recognition Vendor Test 2017 (FRVT).

The **NIST Face Recognition Technology (FERET)** campaign which was organised in 1993 and set the beginning of official face recognition technology assessment.
NIST Website:
https://www.nist.gov/programs-projects/face-recognition-technology-feret

The **NIST Face Recognition Grand Challenge (FRGC)**, which was organised in 2004.
NIST Website:

https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc

The **NIST Face In Video Evaluation (FIVE)**, organised in 2014.
NIST Website: https://www.nist.gov/programs-projects/face-video-evaluation-five

The series of **NIST/IARPA Janus Benchmark Face Challenges** A, B and C (IJB-A, IJB-B and IJB-C) organised between 2015-2017.
NIST Website: https://www.nist.gov/programs-projects/face-challenges

The **NIST/IARPA Face Recognition Prize Challenge (FRPC)**, organised in 2017 (see further details in Section 2.3.1.1).
NIST Website: https://www.nist.gov/programs-projects/face-recognition-prize-challenge
IARPA Website: https://www.challenge.gov/challenge/face-recognition-prize-challenge/

The series of **NIST Face Recognition Vendor Test evaluations**: FRVT 2000, 2002, 2006, 2010, 2013 and 2017 (see further details for the 2017 edition in Section 2.3.1.2).
NIST Website: https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

The reader should bear in mind that this last evaluation, the FRVT 2017, is the one that better represents the capabilities that FR technology may present in a scenario such as SIS considering that the results are from the end of 2018, the size of the reference database used in the experiments, the type of data used in the experiments.

As mentioned above, in the following we give further details about the campaigns that are the most related to SIS and its use-cases, given the scenarios considered in the tests: the FRPC 2017 and the FRVT 2017.

2.3.1.1. Face Recognition Prize Challenge (FRPC - 2017)

Year: from June 2017

Organized by: NIST (U.S. Government)

Goal: to assess capability of the latest algorithms operating on unconstrained images.

Report: [108]Patrick Grother, Mei Ngan, Kayee Hanaoka, Chris Boehnen, Lars Ericson, "The 2017 IARPA Face Recognition Prize Challenge (FRPC)", NIST Interim Report NISTIR 8197, NIST, 2017.

Overview

In conjunction with Intelligence Advanced Research Project Activity (IARPA), NIST ran its first Face Recognition Prize Challenge (FRPC) to assess capability of the latest algorithms operating on unconstrained images. The FRPC was aimed at the measurement of automated face recognition accuracy when sub-optimal images are identified and verified either against other such images, or against standard-compliant portraits. Sub-optimal images are characterized by non-frontal head pose, low resolution, poor and uneven illumination, non-neutral facial expression and occlusion. In almost all cases, the input image contained only one face.

The Challenge came in two parts: 1) face identification open set, and 2) face verification. Both tasks involved "non-cooperative" images where subjects were unaware of the camera or, at least, did not engage with, or did not pose for the camera.

Dataset

From June to September 2017, NIST evaluated 41 face recognition algorithms from 16 developers. The algorithms processed the datasets of 2D still photographs (each one containing one face only) in two ways: verification of "wild" photojournalism and social media images, and identification of faces from surveillance videos against portrait galleries of size up to 691,000. The photojournalism set contains 141,331 faces images of 3,548 adults. The composition of the dataset is presented in **Table 7**.

Table 7. FRPC dataset description.

Context	**Photojournalism**	**Video surveillance**	
Task	**Verification**	**Identification**	
Protocol	*1-vs-1 comparison*	*Templates* *(Mugshot-like)*	*Target* *(Video frames)*
N° images	141,331	691,282	- + 79,403 non-mate
N° subjects	3,548	691,282	825 + - non-mate

Results

The results of the competition are summarised in **Table 8**. For the verification prize challenge, the best performance is obtained by the NTechLab algorithm, which gives a False Non-Match Rate (FNMR) of 22% with a False Matching Rate (FMR) of 0.1%. This FNMR would be unacceptably high for an access control application, but is achieved with images of non-cooperative subjects that have very few of the image quality constraints that are engineered into, for example, border crossing gates.

For the identification prize challenge, Yitu algorithm achieved the best performance. Using probes from the travel concourse dataset to search in a gallery of N = 691,282 portraits, the Yitu algorithm gives False Negative Identification Rate (FNIR) of 20.4% with a False Positive Identification Rate of FPIR = 0.1%.

Table 8. Upper bound performance reached in FRPC for verification and identification tasks. FMR stands for False Match Rate; FPIR stands for False Positive Identification Rate.

	Performance	**Algorithm**
False Non-Match Rate	22% @ 0.1% FMR	1st NTechLab algorithm
False Negative Identification at rank-1	20.4% @ 0.1% FPIR	Yitu algorithm
Identification speed (size ≅ 690k)	592 ± 51 μs	2nd NTechlab algorithm

References

Report: Patrick Grother, Mei Ngan, Kayee Hanaoka, Chris Boehnen, Lars Ericson, "The 2017 IARPA Face Recognition Prize Challenge (FRPC)", NIST Interim Report NISTIR 8197, NIST, 2017.

NIST Website: https://www.nist.gov/programs-projects/face-recognition-prize-challenge

IARPA Website: https://www.challenge.gov/challenge/face-recognition-prize-challenge/

2.3.1.2. Face Recognition Vendor Test (FRVT – 2017/2018)

As already mentioned in the introduction of the present Section 2.3, this is the closest evaluation to the SIS system, both in terms of type of face data and size of the database. Even if results **cannot be directly extrapolated**, this evaluation can be considered as a good estimation of the accuracy that can be expected from FR technology in the SIS operational environment.

Year: From February 2017

Organized by: NIST (U.S. Government)

Goal: Face Recognition Vendor Tests (FRVT) provide independent government evaluations of commercially available and prototype for face recognition technologies. These evaluations are designed to provide U.S. Government and law enforcement agencies with information to assist them in determining where and how facial recognition technology can best be deployed. In addition, FRVT results help identify future research directions for the face recognition community.

Reports: [109] Patrick Grother, Mei Ngan, Kayee Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) - Part 1: Verification", NIST Interim Report NISTIR, NIST, 2018.

[110]Patrick Grother, Mei Ngan, Kayee Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) - Part 2: Identification", NIST Interim Report NISTIR 8238, NIST, 2018.

In the following we summarise the most relevant information found in the two NIST reports referenced above [109], [110]. Note that parts of this section have been directed extracted from those reports. We refer the reader to them for further details.

Overview

The competition was aimed at measuring the performance of automated face recognition technologies applied to a wide range of civil, law enforcement and homeland security applications including verification of visa images, de-duplication of passports, recognition across photojournalism images, and identification of child exploitation victims. In all the considered cases, the input image contained one face only.

Performance reports include measurements of accuracy, speed, storage and memory consumption, and resilience. NIST reports the dependence of performance on the properties of the images and the subjects. In its initial form over 2017, FRVT had one assessment track, for face verification (1:1 verification). Then, open-set identification (1:N identification) task has been addressed in a successive dedicated report in 2018.

FRVT 2017 Verification Dataset (1:1)

The dataset of FRVT 2017 – Verification is composed of different types of images, coming from several contexts. The main features of the dataset are summarised in **Table 12**.

- *Visa Images*: The images have geometry in reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type. Pose is generally excellent. Subjects are from different countries and of different ages (including children). Many of the images are live capture. A substantial number of the images are acquisitions of paper photographs. The number of images and the number of subjects are $O(10^5)$ magnitude.

- *Mugshot Images*: The images have geometry in reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type. The images are of variable sizes. The mean Inter Ocular Distance (IOD) is 123 pixels. The images are of adult subjects from the United States. The images are all live capture. The number of images is $O(10^5)$, as well as the number of subjects.
- *Selfie Images*: The portrait images are in reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type. The selfie images are taken with camera above and below eye level, with one hand or two hands. Pitch angles vary more than yaw angles, which are frontal. Some perspective distortion is evident. The images have mean IOD of 140 pixels. The images are of adult subjects from the United States. The images are all live capture. The number of images is less than 500, as well as the number of subjects.
- *Webcam Images*: The portrait images are in reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type. The webcam images are taken with camera at a typical head height, with mild pitch angles, low yaw angles, but some variation in range, such that low perspective distortion is sometimes evident. The images have mean IOD of 68 pixels. The images are of adult subjects from the United States. The images are all live capture. The number of images is less than 1500, as well as the number of subjects.
- *Wild Images*: The images include many photojournalism-style images. Images are given to the algorithm using a variable but generally tight crop of the head. Resolution varies very widely. The images are very unconstrained, with wide yaw and pitch pose variation. Faces can be occluded, including hair and hands. The images are of adults. All of the images are live capture, none are scanned. The number of images is $O(10^5)$, while the number of subjects is $O(10^3)$.
- *Child exploitation images*: from real investigative case, from different countries. The ages of the subjects vary from infancy to late adolescence. The images are arbitrarily unconstrained in terms of quality, face poses, occlusions, resolution, face size, lighting. The number of images is $O(10^4)$ while the number of subjects is $O(10^3)$.

Table 9. FRVT 2017 dataset description for verification task.

VERIFICATION DATASET				
Image Set	**N° of images**	**N° subjects**	**Standard compliance**	**Mean IOD**
VISA	$O(10^5)$	$O(10^5)$	ISO/IEC 19794-5 Full Frontal	-
Mugshot	$O(10^5)$	$O(10^5)$	ISO/IEC 19794-5 Full Frontal	123
Selfies	< 500	< 500	ISO/IEC 19794-5 Full Frontal	140
Webcam	< 1500	< 1500	ISO/IEC 19794-5 Full Frontal	68
Wild	$O(10^5)$	$O(10^3)$	-	-
Child Exploitation	$O(10^4)$	$O(10^3)$	-	-

FRVT 2018 Identification Dataset (1:N)

For the identification challenge, the FRVT dataset is composed of four types of images, namely mugshots, webcam, wild and surveillance, containing in excess of 30.2 million still photographs of 14.4 million people. The dataset is summarised in **Table 10**.

The primary dataset is composed of 26.6 million controlled live portrait photos of 12.3 million individuals. The other three smaller datasets contain more unconstrained photos such as 3.2 million webcam images; 2.5 million photojournalism and amateur photographer photos; 90 thousand faces cropped from surveillance-style video clips. To the best of our knowledge, this constitutes the largest public and independent dataset to date.

All four types of images are used in the evaluation. The primary dataset is a set of law enforcement mugshot images, which are enrolled and searched with respect to 3 datasets: other mugshots (i.e. within domain); poor quality webcam images collected in similar detention operations (i.e. cross domain); frames from surveillance videos. Additionally, wild images are searched against other wild images.

Table 10. Overview of FRVT 2018 Identification Dataset. If available, number of mated comparison trials (m) and non-mated comparisons trials (n-m) is shown. Note that Webcam and Surveillance pictures are used only for searching.

IDENTIFICATION DATASET					
		Mugshots	**Webcam**	**Surveillance**	**Wild**
Enrolment	N° pictures	Up to $\sim 26\ M$	-	-	$\sim 1.1\ M$
	N° subjects	Up to $\sim 12\ M$	-	-	$\sim 1.1\ M$
Search	N° pictures	$\sim 150\ k$ m	$\sim 80\ k$ m	N.A.	$\sim 330\ k$
		$\sim 330\ k$ n-m	$\sim 330\ k$ n-m	N.A.	
	N° subjects	$\sim 150\ k$ m	$\sim 80\ k$ m	N.A.	$\sim 330\ k$
		$\sim 330\ k$ n-m	$\sim 330\ k$ n-m	N.A.	

Results – Verification (1:1)

Here, we provide an overview of the obtained results, giving an insight of the accuracies achieved in face verification, in different settings (constrained/unconstrained – cooperative/less cooperative subjects).

The following **Table 11** summarizes the performance reached in this challenge of the top ranked algorithms out of the total 53 evaluated, in terms of False Non-Match Rate (FNMR, i.e. miss rate) at a given False Match Rate (FMR). For more details about the performance of each algorithm, we refer to the NIST report (see the references in this same section).

Table 11. Top FRVT performer for each dataset category. Results are shown in terms of False Non-Match Rate (FNMR) at a given False Match Rate FMR.

	CONSTRAINED/COOPERATIVE			**LESS CONSTRAINED/NON-COOPERATIVE**			
Category ***@ FMR***	**VISA** ***0.0001%***	**VISA** ***0.01%***	**MUGSHOT** ***0.01%***	**WEBCAM** ***0.01%***	**SELFIE** ***0.01%***	**WILD** ***0.01%***	**CHILD EXP** ***1%***
Neurotechnology-003	8.8 %	1.8 %	2.3%	**0.00%**	1.7%	6.5%	69.1%
Ntechlab-003	3.9%	1.1%	1.6%	0.3%	1.4%	4.5%	43.3%
Ntechlab-004	1.3%	**0.3%**	**1.3%**	0.2%	-	4.3%	-
Tevian-001	12.7 %	3.3%	1.8%	0.2%	**0.6%**	8.4%	59.8%
Visionlabs-0004	4.6%	0.7%	**1.3%**	-	-	**4.2%**	-
Yitu-001	**0.7%**	**0.3%**	**1.3%**	-	-	5.4%	-

Results – Identification (1:N)

For the identification task, 127 algorithms from the research laboratories of 39 commercial developers and one university have been benchmarked. The major result is that massive gains in accuracy have been achieved in the last five years (see **Table 13**). With good quality pictures, the most accurate prototype finds mated entries, when present, in a gallery containing 12 millions of individuals, with error rates less than 0.2% (see **Table 14**); the remaining errors are due mainly to long-term aging and injuries. However, for at least 10% of facial images, those with significant aging and low quality, identification often succeeds, but recognition confidence becomes smaller, such that true matches become indistinguishable from false positives, and human adjudication becomes necessary. Such accuracy gain is mostly due to the adoption of approaches based on (deep) convolutional neural network (CNN).

The results in the tables below can be interpreted as typical searches conducted into an enrolled population of **N** identities, and for the algorithm to be configured to return the closest **L** candidate identities. These candidates are ranked by their score, in descending order. A human analyst might examine either all L candidates, or just the top R<L identities, or only those with score greater than threshold **T** defined in general in terms of the maximum FPIR allowed.

More in detail, this challenge explored several aspects of face identification systems evaluation. Hereafter, we list the most important ones.

***Absolute accuracy in 2018*:** For the most accurate algorithms, the proportion of searches that do not yield the correct mate in the top 50 hypothesized identities is very close to zero (**Table 14**). Moreover, the correct response is almost always at the top rank. Considering Microsoft_4 algorithm, the proportion of mated comparison trials that does not yield to the correct mate at rank 1 is 0.45%, over a gallery of 12 millions of identities. An outstanding achievement, close to perfect recognition, that must be put in context: first, most of the algorithm are not close to achieving such an accuracy; second, it only applies to mugshot facial images searched in mugshot galleries (**Table 12**); third, in many cases, the correct response is at rank 1, but the similarity score is below typical operational thresholds (**Table 12**); fourth, as the number of enrolled subjects grows, some mates are displaced from rank 1 because of lookalike subjects (**Table 13**). Several further considerations are stemmed from the challenge concerning accuracy.

- **Accuracy across commercial providers:** Recognition accuracy is very strongly dependent on the algorithm, and more generally on the developer of the algorithm. Recognition error rates in a particular scenario range from a few less than 1% up to beyond 50%. It implies that technological diversity remains in face recognition, and algorithms from some developers are quite far from being able to be used in an operational workflow.
- **Error rates at high threshold:** A threshold is usually adopted to limit the rate at which non-mate searches produce false positives. The counterpart is that mated searches may report the mate below the threshold, even if it is at rank 1. The utility of this is that many non-mated searches will usually not return any candidate identities at all. As shown in **Table 14** and **Table 12**, miss rates become higher when a stringent threshold is imposed. This occurs for three main reasons: poor image quality, ageing, and presence of lookalikes.
- **Image Quality:** Poor quality images badly affects recognition, either because the imaging system is poor (lighting, camera, etc.) or because the subject wrongly presents himself to the camera (head orientation, non-neutral expression, occlusion, etc.). In some cases, i.e. when the person is at disposal, imaging problem can be eliminated by design – i.e. by ensuring adherence to face capture standards. However, presentation problems must be detected at capture time and a re-capture

has to be performed. The most accurate algorithms in FRVT 2018 are indeed tolerant with respect to image quality. This derives from the invariance advantages provided by CNN-based algorithms, and this is the main reason why the accuracy has improved since 2013. For example, Microsoft algorithms are highly tolerant to non-frontal pose, to the point that profile-views images were correctly classify when searched against frontal mugshot.

- **Ageing:** The change in appearance, due to ageing, causes a reduction in the recognition capabilities over a long-time lapse. This behaviour is unavoidable, but it can be mitigated by scheduled re-captures, for instance. To quantify ageing effects, the more accurate algorithms were used to enrol the earliest image of 3.1 million adults and then search against 10.3 million newer photos taken up to 18 years after the initial enrolment photo; as summarized in **Table 14**, accuracy degrades progressively with time. Algorithms that are more accurate tend to be less sensitive to ageing, although accuracy alone does not predict ageing tolerance perfectly. In some cases, the most accurate algorithms provided less errors with 18 years older pictures, than middle tier algorithms after 4 years.
- **Accuracy in a large population:** Prior NIST tests had run on reference databases of enrolled populations of maximum 1.6 million. In this challenge, the number of enrolled people climbs up to 12 million data subjects. Nevertheless, identification miss rates grows very slowly as population size increases. For the most accurate algorithm when searching a database of size 640 000, about 0.27% of searches fail to detect the correct mate. In a database of 12 million this rises to 0.45% (see **Table 14**). This growth in miss rates justifies the utility of face recognition testing in large-scale one-to-many search applications. The reason is that the more identities are enrolled into the dataset, the higher is the possibility of a false positive due to lookalike faces. However, in this challenge rank-one identification miss rates scale very favourably with population size, N, growing slowly, approximately as a power law aN^b with b << 1. Depending on the algorithm, the exponent b for mugshot searches is low, around 0.06 for the Cogent algorithms with up to 12 million identities. The most accurate algorithms have somewhat larger values b = 0.17 (Microsoft-4) and 0.08 (Yitu-2).
- **Twins:** One error component is the incorrect associations of twins. Even the most accurate face recognition algorithms are essentially incapable of distinguishing twins, not just identical (monozygotic), but also same-sex fraternal (dizygotic) twins. A twin, when present in an enrolment database will invariably produce a false positive if the twin is searched.

***Accuracy within commercial providers*:** It is worth noting that the intra-provider accuracy variations are usually much smaller than the inter-provider variations. However, from phase 1 (February 2018) to phase 2 (June 2018), some developers attained up to a five-fold reduction in misses. Such rapid gains imply that the revolution is not yet over, and the chances are that further gains will be realized in a near future.

***Utility of adjudicating long candidate lists*:** When a system is configured with a zero threshold, such that human adjudication is always needed (i.e. non-mated comparison trials), the reviewer will find some mates on candidates at ranks far above one. This usually occurs when either the probe image or its corresponding enrolled mate image have poor quality or large time-lapse. The benefit of traversing 50 candidates is that the miss rate is reduced up to 50% (see **Table 14**). However, the accuracy achieved now is so high such that reviewers can expect to review substantially fewer candidates.

***Utility of enrolling multiple images per subject*:** Three kinds of enrolment have been adopted in this competition:

- Enrolling the most recent image;
- Creating a single template from a data subject's full lifetime history of images;
- Enrolling multiple images of a data subject separately (treated as different identities).

The overall effect is that the enrolment of multiple images lowers miss-rates by almost half (**Table 14**). This occurs because the most recent image may sometimes be of poorer quality than historical images.

***Reduced template size*:** The trend is to produce reduced template sizes. The most accurate algorithm uses a template of size 1 KB. Close competitors produce templates in range from 256 bytes to 4.442 bytes. In 2014, the leading competitors had templates of size 4KB to 8KB.

***Template generation time*:** Measured on a single circa-2016 server processor core, the times vary from 50 milliseconds up to nearly 1 second. Such a wide variation may be relevant for end –users who deal with high-volume workflow.

***Search times*:** They vary massively across competitors. For a database of one million subjects, the more accurate algorithms employ durations from 4 to 500 milliseconds, with other less accurate algorithms going much slower. Several algorithms show sublinear search time to the size of the database.

Table 12. Left, miss rate with no threshold; right, with threshold set to target FPIR=1%. Values in bold indicate most accurate algorithm. N denotes the size of the database where the search is carried out. T denotes the threshold set at a given FPIR. L denotes the size of the returned list of candidates (if it is not defined by T).

ACCURACY BY DATASET

	Rank 1 miss rate, FNIR(N,0,1)				Rank 1 miss rate FPIR=1%, FNIR(N,T,L)			
Gallery	N=1.6M	N=1.6M	N=0.7M	N=1.1M	N=1.6M	N=1.6M	N=0.7M	N=1.1M
Algorithm	FRVT18	Webcam	FRPC	Wild	FRVT18	Webcam	FRPC	Wild
Microsoft-4	**0.3%**	1.2%	1.5%	3.9%	1.3%	5.3%	5.5%	4.3%
Siat-1	0.4%	33,3%	**0.9%**	**4.0%**	**0.9%**	34.8%	**3.3%**	**4.1%**
Yitu-2	0.4%	**1.0 %**	1.9%	4.6%	1.1%	**2.8%**	5.5%	5.1%

Table 13. Values are FNIR with N=1.6 million with thresholds and T set to produce FPIR=0.001, 0.01 in non-mate searches, for best algorithms.

THRESHOLD BASE ACCURACY

FNIR(N,T>0,R>L)	ENROL MOST RECENT MUGSHOT, N= 1.6M								
	FRVT 2014 Mugshots			FRVT 2018 Mugshots			Webcam probes		
Algorithm	FPIR= 0.1%	FPIR= 1%	FPIR= 10%	FPIR= 0.1%	FPIR= 1%	FPIR= 10%	FPIR= 0.1%	FPIR= 1%	FPIR= 10%
Microsoft-4	1.7%	**0.7%**	**0.4%**	2.9%	1.3%	**0.5%**	8.7%	5.3%	2.6%
SIAT-1	1.8%	0.7%	0.5%	2.0%	0.9%	0.5%	36.5%	34.8%	33.7%
SIAT-2	9.3%	8.4%	8.2%	2.4%	0.9%	0.5%	47.8%	46.0%	45.1%
YITU2	**1.6%**	0.7%	0.5%	**2.0%**	**1.1%**	0.6%	**4.9%**	**2.8%**	**1.6%**

Table 14. Values are threshold-based FNIR, at FPIR=0.1% for five enrolment population sizes, N. The left six columns apply for enrolment of a variable number of images per subject. The right six columns apply for enrolment of the most recent image.

ACCURACY BY ENROLLED POPULATION

	Enrol Lifetime					Enrol Most Recent				
Algorithm	0.64 M	1.6 M	3 M	6 M	12 M	0.64 M	1.6 M	3 M	6 M	12 M
MISSES NOT AT RANK 1 FNIR(N,T=0, R>1)										
Microsoft-4	0.08%	0.10%	0.13%	0.15%	0.2%	0.27%	0.31%	0.34%	0.38%	0.45%
MISSES NOT AT RANK 50 FNIR(N,T=0, R>50)										
Microsoft-4	0.04%	0.04%	0.05%	0.05%	0.06%	0.18%	0.19%	0.20%	0.21%	0.22%
MISSES BELOW THRESHOLD, T FNIR(N, T>0, R>L)										
SIAT-1	26.95%	27.27%	27.58%	-	-	1.60%	2.01%	2.60%	3.80%	10.69%
YITU-2	0.96%	1.33%	1.74%	2.74%	11.8%	1.56%	2.04%	2.58%	3.82%	12.41%

References

NIST Website: https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

Reports: Patrick Grother, Mei Ngan, Kayee Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) - Part 1: Verification", NIST Interim Report NISTIR, NIST, 2018.

Patrick Grother, Mei Ngan, Kayee Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) - Part 2: Identification", NIST Interim Report NISTIR 8238, NIST, 2018.

2.3.2. Academic evaluations

As mentioned above, these are not evaluations that take place in one moment in time organized by a governmental institution, but on-going challenges usually updated by academic institutions on known sets of data with well-defined testing protocols that have become over the years de-facto benchmarks to assess the capabilities of FR technology. Here below we give details of the two most significant ones: Labelled Faces in the Wild and MegaFace.

It should be highlighted here that, since each participating institution carries out the evaluation of his own system, it cannot be guaranteed that the results have been achieved following exactly the same protocol on exactly the same set of data (e.g., in some cases, for instance, an institution may correct labelling errors in the data).

2.3.2.1. Labelled Faces in the Wild (Ongoing)

Year: From 2007.

Organized by: University of Massachusetts.

Goal: To provide a labelled database of image for studying face recognition in unconstrained environment.

Overview

Organizers provide a database of face photographs designed for studying the problem of unconstrained face recognition. The database has been widely used by academic communities around the world. Results are published in a dedicated web page, allowing a comparison between the different subscribed methods.

LFW provides information for supervised learning under two different training paradigms: image-restricted and unrestricted. Under the image-restricted setting, only binary "matched" or "mismatched" labels are given, for each pair of images. Under the unrestricted setting, the identity information of the person appearing in each image is also available, allowing one to potentially form additional image pairs.

An algorithm designed for LFW can also choose to abstain from using this supervised information, or supplement this with additional, outside training data, which may be labelled (matched/mismatched labelling or identity labelling) or label-free.

Dataset

The data set is described in **Table 15**. It contains more than 13,000 images of faces collected from the web, representing 5749 different people. Moreover, 1680 people have two or more pictures. Each face has been labelled with the name of the person pictured. The only constraint on these faces is that they were detected by the Viola-Jones face detector. There are now four different sets of LFW images including the original and three different types of aligned images.

The aligned images include: funneled images [111], LFW-a, which uses an unpublished method of alignment, and deep funneled images [112]. Among these, LFW-a and the deep funneled images produce superior results for most face verification algorithms over the original images and over the funneled images.

Table 15. LFW dataset description.

DATASET			
	N° of pictures	N° of subjects	N° of subjects with at least 2 pictures
Original images	> 13,000	5749	1680
Funneled images	> 13,000	5749	1680
LFW-a	> 13,000	5749	1680
Deep funneled	> 13,000	5749	1680

Results

Its popularity among several academic and industrial entities provides an interesting and fair comparison among a large amount of algorithms. So that, it can be considered not far from a continuously on-going challenge. The main results obtained on this dataset are presented in **Table 16** and **Table 17**.

LFW provides information for supervised learning under two different training paradigms: image-restricted and unrestricted. Under the image-restricted setting, only binary "mated" or "non-mated" labels are given, for pairs of images. Under the unrestricted setting, the identity information of the data subject appearing in each image is also available. An algorithm can also choose to abstain from using this supervised information, or supplement this with additional, outside training data, which may be labelled (mated/non-mated labelling or identity labelling) or label-free. The results given in **Table 16** correspond to algorithms using outside training data (unrestricted scenario) and the identity labelling.

Table 16. Top results in LFW obtained by deep-based FR systems. The verification accuracy is reported as the point in which FMR=FNMR. The confidence of that accuracy measure is also given.

METHOD	TRAIN DATA	#MODELS	VERIF. ACCURACY
DeepFace	4M	4	97.35%±0.25
Canonical View	203K	60	96.45%±0.25
DeepID	203K	60	97.45%±0.26
DeepID2	203K	25	99.15%±0.13
DeepID2+	290K	25	99.47%±0.12
DeepID3	290K	25	99.53%±0.10
Face++	5M	1	99.50%±0.36
FaceNet	260M	1	99.60%±0.09
Tencent	1M	20	99.65%±0.25

Table 17. Human performance of identification accuracy in three scenarios. In the experiments, no control for whether subjects had prior exposure to the people pictured has been made.

HUMAN PERFORMANCE	VERIF. ACCURACY
Human, funneled	99.20%
Human, cropped	97.53%
Human, inverse mask	94.27%

References

Website: http://vis-www.cs.umass.edu/lfw/

Technical reports: http://vis-www.cs.umass.edu/lfw/lfw_update.pdf

2.3.2.2. Megaface (Ongoing)

Year: From 2016

Organized by: Department of Science and Engineering, University of Washington. Funded by Samsung, Google's Faculty Research Award and by NSF/Intel grant.

Goal: assess face recognition techniques performance with up to a million of distractors. i.e. a million of people who are not in the test set, in an unconstrained setting.

Overview

The MegaFace challenge evaluates how face recognition algorithms perform with a very large number of "distractors," i.e., individuals that are not in the probe set. MegaFace is used as the reference database; the two probe sets used in the challenges are FaceScrub and FG-NET. The challenges (2) address fundamental questions and introduce the following key findings:

- How well do current face recognition algorithms scale?
- Is the size of training data important?
- How does age affect recognition performance?
- How does pose affect recognition performance?

The first challenge deal with unrestricted recognition with varying number of distractors, both for identification and verification scenarios. The second challenge deals with a training on large scale of identities (672.000 identities). Also, in this case, verification and identification scenarios are explored.

Dataset

The MegaFace dataset includes 1 Million photos of more than 690,000 unique subjects, free of licensing restriction. The main characteristics of the dataset are presented in **Table 18** (challenge 1) and **Table 19** (challenge 2). A detailed description of the dataset composition can be found in [95].

Table 18. MegaFace Challenge 1 datasets description.

CHALLENGE 1 DATASET

	Templates	Probes	
Image set	*MegaFace*	*FaceScrub*	*FG-NET*
N° of pictures	> 1.0 M	> 100k	975
N° of subjects	> 690k	530	82

Table 19. MegaFace Challenge 2 datasets descritption.

CHALLENGE 2 DATASET

	Train	Templates	Probes	
Image set	*MegaFace*	*MegaFace*	*FaceScrub*	*FG-NET*
N° of pictures	4.2 M	> 1 M	> 100k	975
N° of subjects	672k	> 690k	530	82

Results

One limitation of the Megaface evaluation is that it only reports results in the identification **closed-set** scenario.

This scenario is less relevant for many real-world applications than the open-set case (please see Section 2.1 for a definition of both scenarios). Therefore, the Megaface evaluation does not report False Positive Identification Rates (FPIR), which are essential for a system like SIS.

Challenge 1 discovered that:

- algorithms’ performance degrades given a large gallery even though the probe set stays fixed,
- testing at scale allows to uncover the differences across algorithms (which at smaller scale appear to perform similarly),
- age differences across probe and gallery are still more challenging for recognition.

A first conclusion from the challenge 2 is that providing good training data to the public allows to better evaluate face recognition algorithm.

Hereafter an overview of the results obtained in both challenges. **Table 20** and **Table 21** show the identification rates at the first rank obtained by the three best algorithms, for type of data set.

Table 20. Identification rate at first rank.

CHALLENGE 1

Algorithm	FaceScrub (Celebrity)			FGNet (Age-invariant)		
	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
iBUG_DeepInsight	98.063%	98.058%	98.053%	-	-	-
Intellivision (Gagan Gupta)	93.125%	93.123%	93.136%	-	-	-
Vocord - deepVo V3	91.763%	91.711%	91.704%	-	-	-
THU CV-AI Lab	-	-	-	77.977%	77.995%	77.968%
Google - FaceNet v8	-	-	-	74.594%	74.585%	74.558%
SIATMMLAB TencentVision	-	-	-	71.247%	71.283%	71.256%

Table 21. Identification rate at first rank.

CHALLENGE 2

Algorithm	FaceScrub (Celebrity)			FGNet (Age-invariant)		
	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
TencentAILab FaceCNN v1	77.068%	77.068%	77.068%	-	-	-
Yang Sun	75.786%	75.786%	75.786%	-	-	-
GRCCV - GRCCV	75.772%	75.772%	75.772%	-	-	-
TencentAILab FaceCNN v1	-	-	-	61.179%	61.179%	61.179%
Yang Sun	-	-	-	53.067%	53.067%	53.067%
Team 2009-(GT-CMU-SYSU)	-	-	-	38.208%	38.208%	38.208%

References

Website: http://megaface.cs.washington.edu/index.html

2.4. Overall Trend

In order to provide a comprehensive overview of the results of the main competitions presented above, we depict in

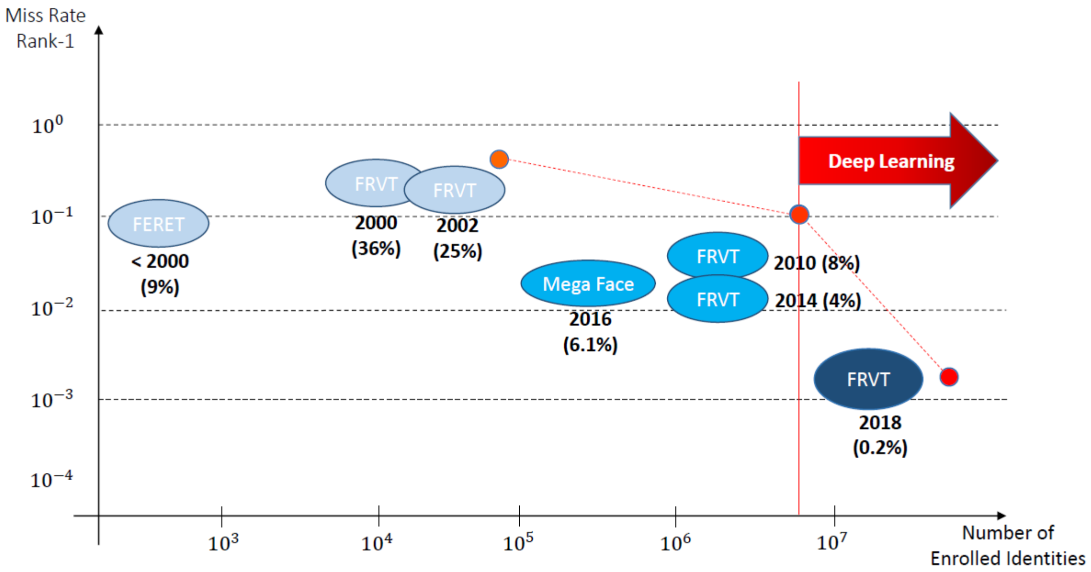
Figure 9 the overall trend in terms of identification accuracy and database size over the past 20 years.

Following the time line, the natural evolution has been to increase the size of enrolled images to explore the effects of database size at large-scale in order to get closer to real operational conditions. In fact, at large scale, the accuracy is expected to decrease due to the higher probability of presence of lookalike subjects in the enrolled image set.

At the same time, the progress in the design of face algorithms implied a continuous improvement of recognition rates over time. However, we have witnessed the most significant breakthrough in the last 5 years.

Thanks to (either full or partial) adoption of Deep Convolutional Neural Networks, we have assisted to almost twenty-fold reduction in miss rate, even though benchmark databases have reached as much as twelve million facial images.

Figure 9. Evolution of accuracy in face identification, in function of enrolled gallery size, across different competitions. Portrait-vs-portrait case is presented. Source: EC 2018.



Section 2. Summary of key concepts:

1. The accuracy of FR systems depends on the data used for its training and for its evaluation. The training data has to model the variability found on the evaluation data, otherwise, the results obtained will not be representative of the real system accuracy. For example, if a system is trained only on data coming from black people, but the test population is a mixture of black, white and Asian people, the system will perform poorly.
2. Even if the train and test data present a similar variability, the accuracy of the system will be dependent on the quality of the evaluation data (see Section 3 for further details on biometric quality).
3. Other factors that affect the accuracy of FR systems are: size of the reference database, age and ageing effects, number of faces used for the search/enrolment, expected response time and size of the ranked list of identities returned by the system.
4. The accuracy of FR systems has improved by two orders of magnitude since year 2000, with the largest leap reported after 2014 with the introduction of new technology based on deep-learning networks.

5. Current state of the art systems based on deep-learning technology, have reached or even surpassed human performance on *verification* tasks. In the case of identification (i.e., search in large databases), human operation is simply not feasible due to speed.
6. Under controlled conditions, very high true positive identification accuracy has been reported: 0.2% on the open-set scenario over a DB of 12.2 million identities, with no threshold. However, a wide range in accuracy can still be observed depending on the algorithm being evaluated.
7. Under controlled conditions, most of the errors are due to long-term ageing and presence of twins.
8. Under unconstrained acquisition conditions and for restrictive thresholds in the open set identification scenario, accuracy is still clearly lower: over a DB of 1.6 million identities, miss rates (for individual search trials) at rank-1 drops from 0.3% (with high quality mugshots and no threshold) to 2.8% (with webcam mugshot with high threshold).
9. Most of the technology shows a low increase in the miss rates when the database size grows.
10. The template size ranges between 256 Byte to 4.442 Kbyte, with an extraction time varying from 50 milliseconds to nearly 1 second.
11. Any evaluation of FR systems should follow as close as possible the directives given in: ISO/IEC 19795-1 2006 "Information Technology – Biometric Performance and reporting – Part 1: principles and framework".

3. Face biometric quality

3.1. Introduction to biometric quality

Many studies and benchmarks have shown that the accuracy of biometric systems heavily depends on the quality of the acquired input samples [113], [114], [115]. If quality can be improved, either by sensor design, user interface design or by standards compliance better accuracy will be obtained. For those aspects of quality that cannot be controlled by design, the ability to analyse the quality of the captured samples can be of great utility. This is useful, for instance, at the time of capture to decide whether or not a subject should be reacquired due to low quality, but also for the real-time selection of the best sample, or the selective invocation of different processing methods. That is why quality measurement algorithms are increasingly deployed in operational biometric systems.

Biometric quality measurement plays vital roles in improving biometric system accuracy and efficiency during the capture process (as a control-loop variable to initiate reacquisition), in database maintenance (sample update), in enterprise wide quality-assurance surveying, in invocation of quality-directed processing of samples and even in security-related tasks [115], [114]. Neglecting quality measurement will adversely impact the accuracy and efficiency of biometric recognition systems (e.g. verification and identification of individuals). Accordingly, biometric quality measurement algorithms are increasingly deployed in operational systems. These elements motivated the need for biometric quality standardization efforts.

This section, summarizes some of the main issues that should be considered regarding the estimation of biometric quality and how it can be used to enhance the performance of biometric systems, giving an overall framework of the challenges involved. The section starts with some general concepts regarding biometric quality and then focuses on specific factors that concern face quality.

3.1.1. Signal quality and system accuracy

One of the main challenges faced by biometric technologies is accuracy degradation in less controlled environments such as, for instance, portable handheld devices or forensic scenarios like for instances images coming from Video Surveillance Systems (VSS). These environments will require robust recognition algorithms that can handle a range of changing characteristics. In such uncontrolled situations there are intrinsic operational factors that further degrade recognition performance and that are not generally replicated in controlled studies.

Conditions that are progressively more difficult significantly decrease performance, despite improvements in technology. This can be clearly seen in the results of the FRVT 2018 competition presented in Section 2.3.1, where the accuracy of the top-ranked systems decreases by over 10 when comparing the “mugshot vs mugshot” scenario with the “mugshot vs face in the wild” scenario.

3.1.2. What is biometric sample quality?

Broadly, a biometric sample is of good quality if it is suitable for personal recognition. Recent standardization efforts (ISO/IEC 29794-1) have established three components of biometric-sample quality:

- *Character* indicates the source’s inherent discriminative capability.

- *Fidelity* is the degree of similarity between the sample and its source, attributable to each step through which the sample is processed.
- *Utility* is a sample's impact on the biometric system's overall performance, where the concept of sample quality is a scalar quantity (usually derived from a vector of quality features) that is related monotonically to the performance of the system.
- In general, in the specialised literature, when speaking about biometric quality experts refer to their *utility* component. This will be the case in the present document.

3.1.3. What is a biometric quality metric?

Essentially, a quality metric is a function or algorithm that takes as input a biometric sample and outputs a value or score defining the quality of the sample. It is important to note that automatic quality metrics do not necessarily measure quality in the same way that humans perceive it, therefore, their results are not always aligned with the subjective quality estimation of experts [116], [117].

Researchers have developed quality assessment algorithms mainly for fingerprint, iris, voice, face, and signature. Unfortunately, almost all of the many algorithms have been tested under limited, heterogeneous frameworks. This diversity of test conditions is due primarily to the fact that the biometrics community has only recently formalized the concept of sample quality and developed evaluation methodologies.

One of the biggest challenges to be addressed by biometric quality metrics is the fact that although biometric comparison involves at least two samples, these are not acquired at the same time. Reference samples are stored in the system database and are later compared with new samples provided during system operation. So, a quality assessment algorithm should be able to work with individual samples, even though it ultimately aims to improve recognition performance when comparing two samples.

One of the main characteristics a quality metric is expected to present is to mirror the sample's *utility* so that higher-quality samples lead to more accurate recognition of individuals. Accordingly, quality should be predictive of recognition accuracy. This concept was formalized in [114], where the authors presented a framework for evaluating and comparing quality measures in terms of the capability of predicting system performance. Broadly, they defined biometric sample quality as a scalar quantity monotonically related to the biometric recognition accuracy when that biometric sample is used for recognition.

3.2. Face recognition scenarios: controlled VS unconstrained

The performance of a face biometric system largely depends on a variety of factors that affect how good an image is for recognition purposes. Based on these factors, two operation scenarios can be distinguished: 1) controlled scenario and 2) unconstrained scenario. Although these two problems share common characteristics, their differences are quite significant and, in general, a face recognition system that presents high accuracy under the controlled scenario, will not necessarily perform well under the unconstrained scenario and vice versa.

The main actors that defined the controlled and the unconstrained scenarios are: quality of the sensor/camera, distance/angle to the subject, level of focus/sharpness, pose, illumination, background, and occlusions. Based on these parameters, both scenarios may be defined as follows:

- The **controlled scenario** implies that the recognition task will be performed on *portrait*-like images (also referred to as *mugshots* in forensic environments). That is, the acquisition of the face is carried out under full control of the environment: cooperative subject, high-quality camera, close range, high resolution, on-focus, frontal-neutral pose (e.g., eyes opened, no smile), controlled homogeneous illumination, constant background, no occlusions (e.g., hair, glasses). These conditions can be found in applications such as ID documents (national ID cards, passports) or law-enforcement criminal records where the suspect is brought into custody and enrolled at the police station.
- At the other side of the spectrum we find the fully **unconstrained scenario**, which is typical for instance in surveillance applications. In this case, the subject is non-cooperative and there is no control over the environmental conditions. Therefore, typical images used for recognition in this context would present features such as: low resolution, long-range images (i.e., face of small size), face captured off-angle, possibly out of focus, predominant light source that produces marked shadows, heterogeneous background (possibly even with other faces), occlusions (e.g., cap, hair, shadows).

The above two examples, represent the two most extreme cases for face recognition: the optimal and the worst situations. There are scenarios that may lie in between. For example, in the case of a face recognition system to unlock a smart phone the most typical situation would be to have a cooperative user, frontal pose, close range, sufficiently good sensor but with uncontrolled illumination and background.

In general, in the specialised literature, the term-controlled scenario is only used to refer to the case with close-to-optimal conditions, that is, where all the main parameters defining the quality of the image are those of a mugshot. In case that any of these parameters differs significantly from the optimal case, the image is already considered as unconstrained.

Therefore, there is just one type of controlled images (facial portraits), but there are many levels or degrees of unconstrained images. This way, in evaluations of face recognition systems where unconstrained scenarios are considered, it must be very well defined the type of unconstrained images being used. For instance, the pose of the subject varies among the different samples but all the other factors remain constant.

It is also worth mentioning that in the case that scanned face images previously printed on paper are stored in a system or are used as probes, the dual printing-scanning process also introduces degradation in the quality of the image. For this reason, it is recommended to avoid to the largest extent possible the use of scanned pictures in applications related to FR (e.g., the issuing of passports), and favour in all cases the use of live acquired images.

In the literature there are multiple works describing FR systems designed to increase the accuracy on very specific types of unconstrained or non-portrait images. This is the case for instance of systems designed for: off-angle pictures, bad illumination pictures, profile pictures or low-resolution pictures. It should be noted that, while those systems can increase the accuracy for the specific type of images for which they were designed, their performance very often drops significantly for all the other type of images. For this reason, in order to avoid this over-fitting effect to a very unique subset of low-quality images, in many cases it is more efficient to have a system capable to generalise well to all type of images, even if for a specific type there may be other algorithms slightly more accurate.

In fact, one of the advantages of the new deep learning-based systems is that they have shown a great ability to generalise and perform well under quite variable conditions, as long as they have been trained on sufficient data modelling those conditions.

3.3. Factors affecting face quality

Quality factors may be classified on the basis of their relationship with the system's different parts. In face recognition we propose to distinguish four classes:

- Related to the biometric characteristic (i.e., face) of the captured subject,
- Interaction of the captured subject with the capturing device,
- Biometric capture device,
- Factors related to the biometric processing subsystem.

Each of these factors is briefly analysed in the following sections.

3.3.1. Factors related to the biometric characteristic of the captured subject

These factors include physical/physiological and behavioural factors. As they are entirely related to the capture subject — the biometric characteristic (i.e., the face) of an individual is difficult or impossible to modify — they are the most difficult to control.

Physical/physiological. These include, for instance, age or gender — subjects cannot alter their biometric characteristic depending on the biometric system being used. Therefore, recognition algorithms must account for data variability in these categories. Also, diseases or injuries can alter features such as the face, sometimes irreversibly, possibly making them impractical for recognition.

- The *age* of the data subject can affect recognition according to the age and ageing effects [118], [119]. The age-effect accounts for the different performance of recognition systems for different age-groups (e.g., children, adults, elderly). The ageing effect accounts for the degradation in the accuracy of biometric systems when the probe and reference samples drift apart in time. The face is less stable over time than other biometric characteristics such as the fingerprint or the iris, therefore it is expected that these two effects will have a deeper impact in its performance.
- It has been shown that g*ender* causes differences in face recognition algorithms. In general, female faces present less inter-class variability which entails that the error rates are higher on female faces than on male faces [120].

Behavioural. Sometimes, people can modify their behaviours or habits. It is possible to alleviate many behavioural factors by taking corrective actions. However, this is not always possible, such as in forensic or surveillance applications. On the other hand, depending on the application, such corrective actions could be counterproductive, resulting in subjects being reluctant to use the system. In general, the supervision of the acquisition process by a well-trained human operator can reduce, to a large extent, the influence of these factors. Some of these factors include:

- Tiredness, distraction, cooperativity, motivation, nervousness.
- Pose (yaw and pitch).

- Make-up.
- Facial hair.
- Glasses, caps, hair.

3.3.2. Factors related to the interaction of the captured subject with the capturing device

In principle, these factors, which include environmental and operational factors, are easier to control than factors related to the captured subject, provided that it can be possible to supervise the interaction between the captured subject and the capture device — for example, in controllable premises such as a police station. In general, these factors also become less relevant as individuals get habituated to use the systems and learn how to interact with them. As in the previous case, the supervision of the capture process by a well-trained human operator can reduce, to a large extent, the influence of the following parameters:

- *Outdoor operation* is especially problematic because control of other environmental factors such as illumination can be lost. It also demands additional actions regarding sensor condition and maintenance.
- *Height/distance* the difference between the sensor height and the face height is key to acquiring good frontal images. Also, the distance to the capture device and embedded sensors can affect the resolution of the face and the level of focus/sharpness.
- *Feedback* to the captured subject regarding the acquired data has been demonstrated to lead to better acquired samples, which can lead to habituation of the capture subject with the system.
- *Automatic acquisition guidance* given by the sensor at the time of acquisition for example providing de user some ques on where to place the face. This can also increase the friendliness of the environment and the overall predisposition of the individual to use it.

3.3.3. Factors related to the biometric capture device

The sensor (i.e., camera) is responsible for reliably translating the physical biometric trait (i.e., subject’s face) in the digital domain. Therefore, its fidelity in reproducing the original face is crucial for the recognition system’s accuracy. The diffusion of low-cost sensors and portable devices is rapidly growing in the context of widening access to information and services. This represents a new scenario for automatic face recognition systems.

Unfortunately, these low-cost, portable devices produce data that is of lower quality from that obtained by more expensive sensors (e.g., reflex cameras). This is primarily due to smaller light sensitive sensors, worse quality optics, and the possibility of user mobility. Additional problems arise when data from different devices coexist in a face system—something common in multi-vendor markets. Algorithms must account for data variability in this scenario of sensor interoperability [121].

3.3.4. Factors related to the biometric processing subsystem

These factors relate to how a biometric sample is processed after it has been acquired. In principle, they are the easiest to control. Constraints on storage or exchange speed might impose data compression techniques — for example in the case of smart cards. Also, governments or regulatory bodies might specify that biometric data must be kept both in raw form as well as in post-processed templates, as templates might depend on proprietary algorithms and that can lead to a vendor lock.

3.4. Incorporating quality in face recognition systems

Quality measurement algorithms can be used to modify and improve the processing and final performance of biometric systems. Such influence in the general working flow of the system includes:

Quality-based processing. An identification system might apply image restoration algorithms or invoke different feature extraction algorithms for samples with some discernible quality problem.

- Quality-specific enhancement algorithms.
- Conditional execution of processing chains, including specialized processing for poor-quality data.
- Extraction of features robust to the signal's degradation.
- Extraction of features from useful regions only.
- Ranking of extracted features based on the local regions' quality.

Template updating (updating of the enrolment data and database maintenance). A quality measurement may be used to determine whether a newly-acquired sample should replace the already enrolled sample. Some systems combine old and new sample features. Quality can be used in both processes.

- Storing multiple samples representing the variability associated with the user (for example, different poses, illumination conditions, images at different ages).
- Updating the stored samples with better-quality samples captured during system operation.

Quality-based comparison, decision, and fusion. Certain systems may invoke a slower but more powerful comparison algorithm when low-quality samples are compared. Also, the logic that provides acceptance or rejection decisions may depend on the measured quality of the original samples. This might involve changing a verification system's operating threshold for poor quality samples. For example, in multimodal biometrics, the relative qualities of samples of the separate modes may be used to augment a fusion process by:

- Using different comparison or fusion algorithms,
- Adjusting those algorithms' sensitivity,
- Quantitative indication of acceptance or rejection reliability,
- Quality-driven selection of data sources to be used for comparison or fusion — for example, weighting schemes for quality-based ranked features or data sources.

Monitoring and reporting across the different parts of the system help to identify problems leading to poor-quality signals and initiate corrective actions. This process can assess signal quality according to these factors:

- *Application*. Different applications might require different scanners, environment set-ups, and so on, which might have different effects on the acquired signals' overall quality.
- *Site or terminal*. Such assessment identifies sites or terminals that are abnormal owing to operator training, operational and environmental conditions etc.
- *Capture device*. Such assessment identifies the impact due to different acquisition principles, mechanical designs etc. It also determines whether a specific scanner must be substituted if it doesn't provide signals that satisfy the quality criteria.
- *Subject*. Such assessment identifies interaction learning curves, which can help better train new users and alleviate the "first-time user" syndrome.
- *Stored template*. Such assessment detects how the database's quality varies when new templates are stored or old ones are updated.
- *Biometric input*. If the system uses multiple biometric traits, such assessment improves how they're combined.

Monitoring and reporting can also support trend analysis by providing statistics from all applications, sites etc. This will let analysts identify trends in signal quality or sudden changes that need further investigation.

3.5. Face quality metrics

Not all the variability factors stated in section 3.3 affect face recognition performance in the same way and this impact also varies depending on the recognition system under consideration. Therefore, generating a metric for the objective evaluation of face quality is a challenging task. To date there is no standard face quality metric, such as NFIQ2 in fingerprints [122] which has even been included in the last version of the standard "ISO/IEC 29794-4:2017 Information Technology – Biometric Sample Quality – Part 4: Finger image data". In spite of this lack of a standard metric, several works have addressed this topic in the literature. These algorithms can be classified according to: 1) their input, that is, the type of features used to estimate quality and 2) their output, that is, the type of information they generate to express the quality level.

3.5.1. Face quality metrics: input features

According to the type of features used to estimate the quality of face images, algorithms can be classified into "hand-crafted" or "deep-learning based".

HAND CRAFTED FEATURES. First studies in face quality assessment employed features that were designed and developed by the researchers using both either their knowledge about the Human Visual System or the knowledge about the internal logic of the employed face recognition system. This type of approach is referred to as "hand-crafted features" in the literature. This type of features extracted from the images can be further divided into several categories regarding if they are related to general digital image processing, face recognition or sensor-related.

- *General Image Quality features*: these features are related to general image processing techniques. Some of them are, for instance, sharpness, contrast or compression artefacts. They have the potential of being applied to a huge number of fields not only to face recognition (though they also have a great impact in face sample quality).
- *Specific Face Quality features*: these are features closely related to properties from the human faces and the face recognition systems, e.g. face geometry, pose, eye location, face detection confidence, face illumination, or orientation. These features are normally related to human perceived quality.
- *Sensor Quality features*: these are features related to the devices and technology employed in the sample acquisition process such as the cameras. Some examples are: lens distortion, thermal noise or histogram equalization functions.

DEEP LEARNING FEATURES. As already mentioned in Section 1, in the last five years, the vast majority of works regarding computer vision (e.g., object detection, classification, face recognition, scene understanding, video segmentation, etc.), rely on Deep Neural Networks, and more specifically on Convolutional Neural Networks (CNNs). These systems have shown a high potential for automatically learning features from the input data. Thank to this autonomous learning process, the problem of designing and developing hand-crafted features has slowly lost a lot of strength. DNNs have outperformed state-of-the-art results based on previous approaches and almost all current face recognition systems employ them as their back-ends.

In face quality related systems based on deep-learning, the quality features are autonomously learned by the DNN based on the data with which they are fed. These networks, e.g. CNNs, learn the best features for achieving a specified target with the highest possible accuracy. There are methods to try to control the type of learnt features or at least to visualise them in order to try to understand the learning process and the physical meaning of the features, if any. However, normally the process is uncontrolled and the features may not have a clear match with the features of the previous categories and can be really abstract or diffused.

3.5.2. Face quality metrics: Output

The output of the face quality assessment algorithms can vary significantly. According to this output face quality metrics can be classified into:

- *Raw decision*. Some techniques just output a decision on whether an image is or not compliant with a given set of quality conditions (e.g., frontal face, no glasses, on focus, eyes detected). While this is an advance with respect to the absence of a quality metric, the amount of information provided is limited and it is difficult to assess how much a given factor affects the overall quality of the image.

- *Qualitative label*. Some metrics produce a qualitative label for each image in order to classify them into a few quality levels.

- *Probability Density Functions (PDFs).* More complex works try to estimate the PDFs of the different variability factors present in the images. These PDFs will estimate the grade of presence of this good/bad quality factor in each sample.

- *Numerical value*. More advanced outputs consist on computing a numerical score for each input image that can predict the expected performance of the image in a face recognition task. This is the most desirable output by a quality metric and the way in which progress has been made in other characteristics such as fingerprints with the new standard quality metric NFIQ2 [122] which has been included in the last version of the standard "ISO/IEC 29794-4:2017 Information Technology – Biometric Sample Quality – Part 4: Finger image data".

3.5.3. Face quality metrics: Evaluation

A very challenging task in the development of face quality metric is the evaluation of their goodness, that is, how to assess if a quality metric is giving a reliable output. The evaluation should be as objective as possible in order to allow for the comparison of different metrics. The main issue is that, in order to assess a face quality metric, you need to generate the ground-truth for a given number of images with which to compare the values given by the automatic metric. How to generate this ground-truth? Two main approaches are followed in the literature:

- *Human perceived quality*: the Human Visual System (HVS) has a great potential performing the face recognition task, even when the face samples present high levels of variability. The HVS can recognize faces with occlusions, small size, extreme poses, etc. HVS is also capable of learning face models with only a few gallery samples. However, generating databases with human perceived quality labels is difficult, as it is a subjective and expensive task. Also, an image that represents bad quality for a human observer may be of sufficient quality for an automatic FR system, and vice versa.
- *Performance-based quality*: as far as the face recognition systems do not work in the same manner than the HVS, using human generated labels may not be the best approach to define the biometric quality ground-truth. In this type of approaches, the final target consists in obtaining a quality score than can be a predictor of the recognition performance when employing one specific sample as an input. The quality score represents a correlation between the input quality features (see Section 3.5.1) and the expected performance. The two main challenges of this type of approaches are: 1) in order to generate a comparison score you need to compare two images, while for a quality score you only need one. Therefore, it is not easy to determine whether a low comparison score is due to low quality of image 1 or of image 2, and as such is difficult to establish the ground-truth quality for each of the images. 2) The expected recognition performance highly depends on the employed recognition system, as some variability factors can affect more to some systems than to others. A quite extended way of measuring the reliability of biometric quality metrics from an accuracy perspective is to use the Error versus Reject Curve (ERC), as introduced by Grother and Tabassi in [114]. This evaluation method has been instrumental in the development of biometric quality metrics and for the ISO/IEC 29794 standard.

3.6. Face quality metrics: Existing works

Face biometrics is nowadays one of the two top-ranked biometric characteristics in terms of amount of funding investment and scientific efforts. However, the amount of research work carried out in face quality estimation is, to date, quite scarce compared to other widely used characteristics such as fingerprints. This has led to a situation where, unlike the fingerprint characteristic with NIFQ2 [122], in face biometrics there still does not exist a standard quality metric that is extensively used. This lack of a general face quality metric is currently being addressed with the on-going revision of the Technical Report (not a standard) “ISO/IEC 29794-5:2010 Information Technology – Biometric Sample Quality – Part 5: Face image data”.

After the huge recognition accuracy leap witnessed in the last five years with the introduction of deep-based technology, the development of new reliable quality metrics is one of the gaps that should be addressed in the short coming future in face recognition technology.

In **Table 22** we include a compilation of the most relevant works carried out so far in face quality estimation according to: their input features (see Section 3.5.1), their output (see Section 3.5.2) and the methodology used for their evaluation (see Section 3.5.3). A brief summary of each of these works is given in the next paragraphs.

Table 22. Summary of face quality assessment works classified by: 1) type of input features (see Section 3.5.1); 2) type of output produced (see Section 3.5.2); 3) evaluation method used (see Section 3.5.3).

YEAR	REF.	INPUT FEATURES	OUTPUT RESULT	EVALUATION METHOD
2007	[123]	ICAO/ISO compliance testing	Lighting+symmetry	Comparison of images different conditions
2009	[124]	ICAO/ISO compliance testing	Lighting+symmetry	Comparison of images different conditions
2013	[125]	Image features, sensor features	Low/High	FRR at fixed FAR
2012	[126]	5 image/face features: Contrast, brightness, focus, sharpness, illumination	FQI (Face Quality Index): 0 to 1	Rank-1
2012	[127]	20 features ICAO/ISO compliance testing	Score 0-100 for each individual test	% of rejection
2014	[128]	12 texture features	FQI: 0 to 1	EER
2011	[129]	Face features, image features	Score reflecting presence of each factor	% of images with minimum Q for each feature
2015	[130]	2 face features: pose, illumination	Predicted FMR/FNMR	FMR/FNMR
2016	[131]	Image gradient features	FQI: 0 to 1	Correlation
2017	[132]	FAR, yaw and confidence	Predicted VR	FAR/RMSE
2017	[133]	Image-based features	Numerical quality metric	CMC
2017	[134]	DNN features	Similarities between test and reference image	Average subjective Q score
2018	[135]	DNN features	MQV (Machine Quality Value) and HQV (Human Quality Value)	FMR/FNMR

Several face quality standards have been proposed so far, being the most relevant and extended ones the ICAO 9303 and the ISO/IEC 19794-5 (see Section 3.7 for further details). These standards are composed of a series of guidelines for the acquisition of good quality (i.e., mugshot-like) images, usually for their inclusion in ID documents (e.g., ID cards or passports). A number of vendors and academic works have developed tools to automatically check if an image complies with the guidelines given in these standards. In general these works provide as output a binary vector where each feature defines whether or not a specific guideline was passed/not-passed by the image [127] [123] [124].

In [125] the authors presented one of the first works related to quality assessment in face recognition. They employed 12 quality features divided into three categories: 1) The first class consists in image processing and face recognition related features, e.g. edge density, eye distance, face saturation, pose, etc. 2) The second category is composed by sensor-related features like the ones that can be encountered in the EXIF headers of the images. 3) The last class consists of features related with the comparison algorithms they employed, i.e. SVM. They extracted conclusions about which features are more relevant to the specific dataset they employed (PaSC) regarding to the overall recognition performance. They used that knowledge for splitting the whole dataset into two categories regarding their quality: low and high.

In [126] the authors proposed a, performance based, Face Quality Index combining individual quality factors extracted from five image processing features: contrast, brightness, focus, sharpness and illumination. They employed the CASPEAL database adding synthetic effects to the images (data augmentation), being able to emulate different real-world variations. After computing a numerical value of quality for each feature, they defined the Face Quality Index normalizing each quality measure, modelling the distribution of quality scores as Gaussian PDFs. Values close to the mean of each PDF means good quality, while scores far to the mean represents bad quality. The good quality reference PDFs were obtained using a good quality subset from the FOCS database. Finally, they performed an average of all individual quality scores to compute the FQI.

The work presented in [130] establishes a relationship between two image features: pose and illumination, and the final face recognition performance. They developed individual score metrics using PDFs in a similar way to [126]. However, the main difference between both works is the fact that in [130] the individual scores are employed to finally estimate expected performance values, i.e. FMR and FNMR. The authors used six different face recognition systems in order to extract performance values from the databases: two of them were Commercial Off-The-Shelf software (COTS) and four open-source algorithms, and they applied them to three different datasets: MultiPIE, FRGC and CASPEAL.

In [135] the authors predict both quality values (scores) related to machine performance (they called it Machine Quality Values - MQV) and other related to human perceived quality (Human Quality Values- HQV). They annotated the LFW database with human perceived quality using the Amazon Mechanical Turk platform where the task consisted in comparing pairs of images from LFW and determine which one had the best perceived quality. Differently to [130], where they predicted a value for recognition performance, in this case they employed FMR and FNMR as accuracy measurements in the training stage, but their final output/target was a predicted value for MQV or HQV. Other differential point of this work is the fact that they employed a pre-trained CNN (VGG Face) to extract features from the images. Then they used those features to train their own classifier, which means that they successfully transferred learning from face recognition to the quality prediction task. They extracted interesting conclusions such as that both measures (MQV and HQV) are highly correlated with recognition performance, even for cross-database prediction. However, they also concluded that automatic HQV is a more accurate predictor of performance than automatic MQV.

3.7. Face quality standards

Even though there is still no standard tool/metric to estimate facial quality level, given its relevance, several standards and official recommendations have been published describing some best practices for the acquisition of good quality face images specially in the context of ID documents.

3.7.1. ISO/IEC 19794-5

ISO/IEC 19794-5 defines a standard scheme for codifying data describing human faces within a CBEFF-compliant data structure, for use in facial recognition systems. Modern biometric passport photos should comply with this standard. Many organizations have already started enforcing its directives, and several software applications have been created to automatically test compliance to the specifications.

In order to enable applications that run on a variety of devices, including those with limited resources (such as embedded systems), and to improve face recognition accuracy, the specification describes not only the data format, but also additional quality requirements, namely: scene constraints (lighting, pose, expression, hair style, eye glasses, head coverings, children); photographic properties (positioning, camera focus); and digital image attributes (image resolution, image size, color saturation).

3.7.2. ISO/IEC 29794-5

The purpose of ISO/IEC 29794-5 is to define and specify methodologies for computation of objective, quantitative quality scores for facial images. It also discusses on the purpose, intent and interpretation of face quality scores.

It should be noted that the current version of this document, from 2010, is a Technical Report and not a standard. Given its publication date, it contains reference to algorithms and techniques which have been clearly superseded by new technology. As such, at the moment, ISO has started the process to review this document in order to generate a standard equivalent for instance to the one existing on biometric quality for fingerprint images (ISO/IEC 29794-4).

This standard complements the information given in ISO/IEC 19794-5 by giving samples of a classification scheme of facial quality and also by defining approaches for the determination of certain aspects of quality. In particular, it further defines the next features for face image quality analysis:

- Dynamic subject characteristics (e.g., glasses, hair)
- Subject’s behaviour (e.g., pose, expression)
- Analysis based on statistical differences of the left and right half of the face.
- Static characteristics of the acquisition process.
- Image resolution and size.
- Image noise.

- Characteristics of the image acquisition sensor.
- Image properties.
- Image appearance.
- Illumination intensity.
- Image brightness.
- Image contrast.
- Exposure.
- Focus, blur and sharpness.
- Colour.
- Subject camera distance.

3.7.3. ICAO 9303 – Part 3: Specifications common to all MRTDs

The ICAO 9303 standard defines the technical specifications for Machine Readable Travel Documents (MRTDs), such as passports. Among its recommendations, it defines a number of guidelines for the facial images that have to be included in this type of documents. These requirements are largely based on the ISO/IEC 19794-5 standard. In particular, the standard states that:

- The photograph shall show a close up of the head and shoulders with the subject facing square on and looking directly at the camera with both eyes visible and with a neutral expression with the mouth closed.
- The pose should be such that an imaginary horizontal line drawn between the centres of the eyes is parallel to the top and bottom edges of the rectangular image.
- The facial image shall be in focus from the crown (top of the head ignoring any hair) to the chin and from the nose to the ears.
- If the additional detail of one ear is required (sometimes referred to as "half-on profile"), the face shall be at such an angle to the imaginary plane as to reveal the detail of the ear while maintaining full-face frontal details on that side of the face opposite to the exposed ear.
- Both edges of the face must be clearly visible. The subject shall not be looking, portrait-style, over one shoulder.
- The face shall be in sharp focus and clear with no blemishes such as ink marks or creases.
- The eyes must be open, and there must be no hair obscuring them.
- If the subject wears glasses, the photograph must show the eyes clearly with no lights reflected in the glasses. The glasses shall not have tinted lenses. Avoid heavy frames if possible and ensure that the frames do not cover any part of the eyes. Glasses should appear only if permanently worn.
- Head coverings shall not be accepted except in circumstances that the issuing State specifically approves. Such circumstances may be religious, medical or cultural. The face must be visible from the hairline to the chin and forward of the ears.

- Coverings, hair, headdress or facial ornamentation which obscure the face are not permitted.
- The issuing State shall use its discretion as to the extent to which facial ornaments (e.g. nose rings, studs), not obscuring the face, may appear in the portrait. A facial ornament should appear only if it is permanently worn.
- A facial image of a baby should conform to the same specifications as for adults. Ideally, the baby should be in an upright position but it is acceptable to capture the facial image with the baby lying on a white or plain light-coloured blanket. Alternatively, the baby may be placed in a baby seat but there shall be white or plain light-coloured background behind the head. The baby's eyes shall be open and no supporting hands visible.
- There must be no other subjects or objects in the photograph.
- Adequate and uniform illumination shall be used to capture the facial image ensuring there are no shadows or reflections on the face or in the background.
- The subject's eyes must not show red eye.
- The photograph must have appropriate brightness and contrast.
- The displayed portrait shall be monochrome greyscale [or black and white] or a true-colour representation of the holder. Where the picture is in colour, the lighting and photographic process must be colour balanced to render skin tones faithfully.
- A uniform light-coloured background shall be used to provide a contrast to the face and hair. For colour portraits, light blue, beige, light brown, pale grey or white are recommended for the background.

3.7.4. ICAO Technical Report, 2018 - Portrait Quality (Reference Facial Images for MRTD), 2018

Although this is not in itself a standard yet, in 2018 ICAO published a technical report to further detail the recommendations given in ICAO 9303 with regard to facial images stored in MRTDs.

The technical report is based on ISO/IEC 19794-5:2005 and ISO/IEC 19794-5:2011 as well as on Doc 9303 statements on portraits. The content of those documents is rearranged, consolidated, enriched, and improved in the technical report.

The scope of technical report is to describe the requirements and best practice recommendations to be applied for portrait capturing in the application case of enrolment of biometric reference data for electronic MRTD. In this sense, the document is an application profile. The document:

- shares the lessons learned using the stored and displayed portrait in an MRTD,
- describes how the portraits should be captured that serve as the content of ISO/IEC 19794-5 and ISO/IEC 39794-5 data structures,
- provides the experiences made applying facial recognition technology in ABC gates, manual border control, identity screening, and other applications based on the portraits provided by electronic MRTD's. It also gives guidance on the requirements

for capturing and processing portraits contained in MRTD's to support the inspection process,

- provides comprehensive recommendations for portrait capturing including scene, photographic and digital requirements,
- provides requirements for image printing and scanning as well as on digital image processing,
- provides requirements for portraits printed on MRTD's to ensure good visibility for inspection, and
- gives guidance for reader system manufacturers on the use of unified reflection free illumination and view angles.

3.7.5. DHS Technical Report, 2009 - Facial image quality improvement and face recognition study final report

The Facial Image Quality Improvement and Face Recognition Study (FIQIFRS) project was initiated in February 2007 to investigate technology for improving the quality of face images captured at United States (U.S.) ports of entry (POEs). The project's goal was to bring US-VISIT face images into compliance with standards defined in the Registry of U.S. Government Recommended Biometric Standards and to improve quality sufficiently to ensure accurate recognition by both humans and computer systems while minimizing operational impacts and allowing for technology maturation.

The technical report includes a series of findings and recommendations including:

- Camera hardware specifications, physical infrastructure of POEs,
- client-side software specifications,
- client-side Graphical User Interface (GUI) specifications,
- standards compliance
- or image specifications.

Section 3. Summary of key concepts:

1. The most determinant factor that defines the accuracy of a Face Recognition system is the quality of the data it runs on.
2. There are systems specifically designed to increase the accuracy of certain very particular type of low-quality images (e.g., profile, bad illumination, low-resolution). However, in general, these systems present a significant drop in accuracy for the rest of image types for which they have not been trained. Depending on the context, it can be more efficient to have a system that generalises well and performs at a good level for all image types.
3. Today, the most widely used standard to obtain good quality images is the ICAO standard. Many of the requirements established by the ICAO standard to obtain good quality pictures are not defined according to accuracy principles of ABIS-Face technology, but according to guidelines based on the subjective cognitive capacities of border guards to distinguish individuals. That is, in addition to trying to obtain images that work well for recognition purposes using ABIS-Face technology, the ICAO standard also tries to produce portraits that can be usable by border guards. This way, some of its specification may not be relevant for current FR systems.
4. By following the ICAO standard, perfect quality images from an accuracy perspective are obtained. However, not all the parameters defined in the ICAO standard affect the performance of ABIS-Face systems equally.
5. The picture taken for the purpose of issuing of a biometric travel document should be compliant with the ICAO standard. However, the level of compression in the chip of the passport changes greatly among countries. In some cases, the level of compression is so big that the picture is not of sufficient resolution to be used for face recognition. It should also be noted that in some countries it is allowed to provide a printed facial image of the individual in order to issue a biometric travel document. This picture is scanned and converted to digital format. The printing-scanning process adds degradation to the quality of the face image. For this reason, at borders, it is more reliable to take a live picture of the traveller than to use the picture of the travel document in order to query a database such as the one in CS-SIS.
6. The main factors, identified in the scientific literature presented above, which affect the quality of face images from an accuracy perspective are:
    a. Resolution of the image and size of the face (i.e., number of pixels in between the eyes)
    b. Illumination conditions (e.g., one predominant lighting source from the right or the left can totally change the physiognomy of the face). It is preferable to have a homogeneous and controlled artificial lighting.
    c. Sharpness of the image (on focus).
    d. Pose (frontal, neutral face)

7. From a quality point of view, the main type of face images are: portrait-like and unconstrained images. There is no standard way to differentiate between the two types of images. The best way to set a threshold to separate both categories is to determine a minimum accuracy that needs to be met, and then define the quality requirements for the images in order to meet that accuracy.
8. The automatic estimation of the quality of face images can be a very useful tool to ensure a certain level of performance of the system, depending of the context where it is being used (e.g., border control, law enforcement). However, to date, there is no standard metric to estimate quality.
9. There is the need to develop a reliable vendor-independent face quality metric. The development process should contribute and get feedback from the currently under-review standard: "ISO/IEC TR 29794-5 Information Technology – Biometric sample quality – Part 5: Face image data".

4. Standards for face biometrics

Biometric data interchange standards are needed to allow the recipient of a data record to successfully process data from an arbitrary producer. In other words, biometric interoperability means that biometric data, in whatever form (i.e., raw samples, templates, scores) can be accurately exchanged and interpreted by different applications. This can only be achieved if the data record is both syntactically and semantically in compliance with a published standard.

Following advances in biometric technologies as a reliable identity authentication technique, more large-scale deployments (e.g. e-passport) involving multiple organizations and suppliers are being rolled out. Therefore, in response to a need for interoperability, biometric standards have been developed.

Without interoperable biometric data standards, exchange of such data among different applications coming from different vendors is not possible. Seamless data sharing is essential to identity management applications when enrolment, capture, searching and screening are done by different agencies, at different times, using different equipment in different environments and/or locations. Interoperability allows modular integration of products without compromising architectural scope, and facilitates the upgrade process and thereby mitigates risk of obsolescence.

Table 23. Main organizations working on the development of Biometric standards

BIOMETRIC STANDARD ORGANIZATIONS
International Standards Organizations: - IEC: International Electrotechnical Commission (www.iec.ch) - ISO SC37: International Organization for Standardization, Subcommittee 37 for Biometrics (www.iso.org/iso/jtc1_sc37_home) - CEN: European Committee for Standardization, Technical Committee 224, Working Group 18 – Biometrics (https://www.cen.eu)
National standards bodies: - ANSI: American National Standards Institute (www.ansi.org)
Standards-developing organizations: - ICAO: International Civil Aviation Organization (www.icao.int) - INCITS M1: International Committee for Information Technology Standards, Technical Committee M1 on Biometrics (http://standards.incits.org/a/public/group/m1) - NIST-ITL: American National Institute of Standards and Technology, Information Technology Laboratory (www.nist.gov/itl)
Other organizations: - BC: Biometric Consortium (www.biometrics.org) - BCOE: Biometric Center of Excellence (www.biometriccoe.gov) - BIMA: Biometrics Identity Management Agency (www.biometrics.dod.mil) - EAB: European Association for Biometrics (https://www.eab.org/) - IBG: International Biometric Group (www.ibgweb.com) - IBIA: International Biometrics and Identification Association (www.ibia.org)

This section focuses on face recognition standardization. Broadly, face recognition standards serve the same purpose as many other standards, which is to establish an interoperable definition, interpretation and exchange of face data. Like other standards, face-related standards create grounds for a marketplace of off-the-shelf products and are also a necessary condition in order to achieve supplier independence and to avoid vendor lock-in.

Table 23 lists the main standards organizations and other bodies working on the development of biometric standards. Current development focuses on acquisition practices, sensor specifications, data formats and technical interfaces. Also, a registry of US-government-recommended biometric standards (www.biometrics.gov/standards) offers high-level guidance for their implementation. The two main entities working in biometrics standards are the ISO/IEC JTC 1/SC 37 and the ANSI/NIST.

4.1. Most relevant face data exchange standards

Concerning the specific exchange of face data, the most relevant efforts are:

- The ANSI/NIST ITL 1-2011 Update 2015. Data format for the interchange of fingerprint, facial and other biometric information.
- The ISO/IEC 19794-5:2005 Biometric data interchange formats – Face image data.
- The ISO/IEC 19785-1:2015 Common biometric Exchange Formats - Framework.
- The ISO/IEC 39794-1:2019 Extensible biometric data interchange formats – Framework.

4.1.1. ANSI/NIST ITL 1-2011 Data format for the interchange of fingerprint, facial and other biometric information

This standard is the one on which the current SIS face image exchange is based on.

It defines the content, format, and the units of measurement for electronic exchange of fingerprint, palmprint, plantar, facial/mugshot, scar, mark and tattoo, iris, deoxyribonucleic acid (DNA), and other biometric sample and forensic information that may be used in the identification or verification process of a subject.

The information consists of variety of mandatory and optional items. This information is primarily intended for interchange among criminal justice administrations or organizations that rely on automated identification systems for use other biometric and image data for identification purposes.

4.1.1.1. ISO/IEC 19794-5:2005 Biometric data interchange formats – Face image data

ISO/IEC 19794-5:2005 specifies scene, photographic, digitization and format requirements for images of faces to be used in the context of both human verification and computer automated recognition. The approach to specifying scene and photographic requirements in this format is to carefully describe constraints on how a photograph should appear rather than to dictate how the photograph should be taken. The format is designed to allow for the specification of visible information discernible by an observer pertaining to the face, such as gender, pose and eye colour. The digital image format can be either ISO standard JPEG or JPEG2000. Finally, the 'best practice' appendices provide guidance on photo capture for travel documents and face recognition performance versus digital compression.

4.1.2. ISO/IEC 19785-1:2015 Common Biometric Exchange Formats Framework

This standard defines structures and data elements for Biometric Information Records (BIRs). It defines also the concept of a domain of use to establish the applicability of a standard or specification that complies with CBEFF requirements. It defines the concept of a CBEFF patron format, which is a published BIR format specification that complies with CBEFF requirements, specified by a CBEFF patron. This standard defines also the abstract values (and associated semantics) of a set of CBEFF data elements to be used in the definition of CBEFF patron formats. It specifies the use of CBEFF data elements by a CBEFF patron to define the content and encoding of a standard biometric header (SBH) to be included in a biometric information record.

The ISO/IEC 19785-1:2015 provides the means for identification of the format of the BDBs in a BIR. It also provides a means (the security block) for BIRs to carry information about the encryption of a BDB in the BIR and about integrity mechanisms applied to the BIR as a whole; the structure and content of security blocks are not in the scope of this part of the standard, as well as the specification of encryption and integrity mechanisms for BIRs.

This standard specifies transformations form one of CBEFF patron format to a different CBEFF patron format. The encoding of the abstract values of CBEFF data elements to be used in the specification of CBEFF patron formats is not in the scope of this part of standard.

The standard specifies several patron format specifications for which ISO/IEC JTC 1 SC 37 is the CEBFF patron. It also specifies several security block format specifications for which ISO/IEC JTC 1 SC 37 is the CBEFF patron.

4.1.3. ISO/IEC 39794-1:2019 Extensible biometric data interchange formats – Framework

The ISO/IEC is currently working on a new standard of extensible biometric data interchange formats that will eventually supersede the previous two standards (ISO/IEC 19794-5 and ISO/IEC 19785-1). Both parts 1 (Framework) and part 5 (face image data) are currently under development and are expected to be published by the end of 2019. This will be, at the time of publishing, the most advanced and up-to-date standard in biometric technology and specifically in face recognition.

Its most relevant and innovative feature with respect to previous standards is that it is "extensible", that is, it is not only backward compatible but also forward compatible so that new features can be added to the containers in order to keep up with the fast-moving biometric technology.

4.2. Other face standards

Regarding the context of SIS, the following document is also relevant:

ICAO, Doc 9303, Part 3: Machine Readable Travel Documents Machine Readable Official Travel Documents

Technically, this document is not a standard but a specification that refers to ISO/IEC standards (mostly the ISO/IEC 19794 specified in Section 4.1.1.1). The ICAO vision for the application of biometrics technology encompasses:

- Specification of a primary interoperable form of biometrics technology for use at border control (verification, watch lists) as well as by carriers and document issuers, and specification of agreed supplementary biometric technologies;
- Specification of the biometric technologies for use by document issuers (identification, verification and watch lists);
- Capability of data retrieval for 10 years, the maximum recommended validity for a travel document;
- Having no proprietary element thus ensuring that any States investing in biometrics are protected against changing infrastructure or changing suppliers.

Doc 9303 considers only three types of biometric identification systems. With respect to the storage of these three biometric features in the contactless IC of an eMRTD, the issuing State or organization shall conform to the relevant international standard.

The types of biometrics are:

- Facial recognition – MANDATORY. MUST comply to ISO/IEC 19794-5;
- Fingerprint recognition – OPTIONAL. If used, MUST comply to ISO/IEC19794-4;
- Iris recognition –OPTIONAL. If used MUST comply to ISO/IEC 19794-6

Section 4. Summary Of Key Concepts

1. Standards are essential in order to ensure the correct communication between parties in large IT systems. This is of special relevance for the case of systems where the enrolment, capture, searching and screening are done by different agencies, at different times, using different equipment in different environments and/or locations.
2. The two most widely used standards for Face Recognition are the ones developed by ISO/IEC and ANSI/NIST.
3. While the SIS is already running in an adapted version of the ANSI/NIST type 10 containers, at some point in time it could be worth considering the change to the most modern and up-to-date standards developed by ISO/IEC, in particular the ISO/IEC 39794 standard.

5. Lessons learned: challenges to be addressed by face recognition technology

All the previous sections, that conform PART I of the report, contain in a structured way, all the information regarding automatic face identification technology gathered during: 1) the review of the state of the art, 2) the visits to the MSs, 3) the interviews with the different vendors and 4) the exchange with the board of five external experts.

The information presented has allowed us to identify the different challenges currently faced by this type of technology. Such challenges, which in many cases are connected, are summarised here below.

The aim of the next part of the report (PART II), starting right after this section, will be to detail and address these challenges, whenever possible, in the context of CS-SIS in order to give some recommendations on the best possible way to integrate the ABIS-Face functionality in the system.

1. **Use-cases**: probably the most critical parameter in the design of an ABIS-Face system is the definition of the use-cases, scenarios and operational context in which the system will be deployed. These use-cases will determine, to a large extent: the size of the database, the number of consultations that the system will have to support, the population it will run on, the quality of the data used for enrolment/query, the possible threats that the system will be subjected to, and the speed required to receive an answer from the system, etc. Therefore, all the other parameters listed below are somewhat linked to this one.
2. **Performance**: this feature refers to the accuracy of the system, that is, its ability to find in a given database the queried identity. Very high performance becomes especially critical in the case of ABIS-Face that have to cope with large databases. Given current FR deep-based technology, it is key for a correct accuracy and to avoid biased results that the data used to train the system in representative of the population on which it will operate.
3. **Quality**: this feature refers to the biometric quality of the face images that are processed in the system. Ensuring high quality, especially of the samples enrolled in the database, is a critical parameter in order to achieve a high level of performance. In the case of face recognition, high quality refers to high resolution portrait-type images captured under controlled conditions, while low quality refers to images acquired under uncontrolled conditions usually including bad illumination, low resolution and non-frontal pose.
4. **Integrity of the database**: this feature refers to the correctness of the data stored in the ABIS-Face database. Typical errors that are usually observed in ABIS-Face databases include: face images not corresponding to the right person, missing faces, inconsistency between alphanumeric data and face data, duplicate enrolments under different identities, etc. It is critical for the correct functioning of an ABIS-Face to mitigate, as much as possible, this risk.
5. **Type of data being processed**: with regard to the use-cases, it is important to define the type of face data that the system will have to work with at enrolment and consultation. This feature is very tightly related to quality. For instance, possible types of face images are: portraits, profile, in the wild. The quality of the different

type of data differs significantly and therefore it also has an impact on the final performance of the system.

6. **Face images in the wild of unknown subject**: these images typically come from video surveillance cameras and constitute probably the biggest challenge faced by current ABIS-Face due to their typical very low quality. Defining a specific processing strategy for this particular type of data (i.e. fully automatic, partially assisted etc.) is usually required to obtain the required performance.
7. **Speed**: this feature refers to the response time of the system when a query (i.e. consultation) is launched. The response time can be a critical parameter for certain use-cases where the time constraints are very strict (e.g., first line border control).
8. **Size of the database**: this refers to the number of data subjects enrolled in the system database and which will be used to perform the searches. This parameter is one of the key design features and should be carefully estimated before putting in place any ABIS-Face. The size of the database will have a big impact on the response time of the system an on the False Positive Identification Rate (FPIR). This is one of the features to be taken into account when defining the minimum accuracy expected for the system.
9. **Number of transactions at peak hours**: together with the database size and the expected response time, this feature is also a key design feature to size the ABIS-Face (in terms of the necessary processing power). It refers to the number of consultations that the system will have to process and, as in the case of the database size, it should be carefully estimated in the design phase.
10. **Comparison capacity**: this feature is totally linked to the previous one (i.e. number of transactions). It refers to the number of comparisons between individual face samples that the system should be able to perform at peak hours.
11. **Strategy to handle the queries**: although this may be considered a secondary feature, it may play a very important role in the transaction response time and therefore in the resources needed by the ABIS-Face. For instance, in many cases, it is useful to assign a priority to each transaction depending, for instance, on the expected response time. This has been discussed in the biometric literature as workload reduction and has been shown that can have a positive impact in hardware costs, operational costs, transaction times and accuracy error rates.
12. **Exchange formats**: it is essential to commit to a unique, standardized exchange format for the different type of data handled by the system (e.g., face images, face templates, comparison scores etc.)
13. **Multiple face records**: the possibility to store multiple face records could offer the opportunity to apply an improved quality strategy such as using the best record or produce a composite face record with the best available face images. The strategy may vary according to the face submitted for consultation (portraits or other quality). Having different face records of the same data subject over time can also help to reduce the effect of ageing as the latest enrolled images are expected to provide better comparison scores (template update strategy).
14. **Operational procedures**: in some ABIS-Face, captured subjects follow different operational procedures to interact with the system (e.g., face acquisition methodology). Although such diversity is not crucial for the successful integration of an ABIS-Face, it can have a negative impact on its accuracy. Therefore, it is

preferable to work towards the harmonization of such methodologies and their best practices in order to achieve the maximum possible performance of the system.

15. **Human intervention**: although representing a decreasing part of the overall process, thanks to the accuracy improvement of face technology over the last five years with the use of deep-based technology, manual pre-processing of images before submitting them to the system can be an important step in some use-cases and should therefore be considered. This would be the case for instance of low-quality face images acquired from surveillance cameras in a forensic context.
    Note that, according to best practices in forensics [136], not every type of processing on the images should be allowed, as certain manipulations could be considered tampering of the forensic evidence. In general, only linear transformations (rotation, translation) and landmark annotation are allowed.

16. **Maintenance and performance evaluation**: benchmarking the performance of an ABIS-Face is a healthy and important task to be conducted during the life-cycle of the system. This task not only provides important information on the performance of the system in production (with real data) but can also be a useful tool for fine-tuning the system and eventually improving its performance.

17. **System architecture**: all the previous technical features, as well as other parameters derived from the specific context in which an ABIS-Face will be deployed, should be taken into account during the design phase in order to select the most suitable architecture (e.g. distributed, centralised, hybrid). In the case of SIS this has a lesser important since the centralised nature of the system already in place defines and restricts the possibilities of future modifications.

18. **Threats**: The fact that the face can be captured in a non-intrusive way at a distance makes it a public mean of identification but also an easier target of attacks such as presentation or morphing attacks that could be potentially performed against it depending on the specific context of each application. The risk posed by these threats should be evaluated for each specific use-case.

Part II
ABIS-Face within CS-SIS

PART II of this report is focused on CS-SIS and its future ABIS-Face functionality. This part refers to and builds upon many of the concepts, terms and general aspects of ABIS-Face technology already described in PART I. PART II is based on the following rationale:

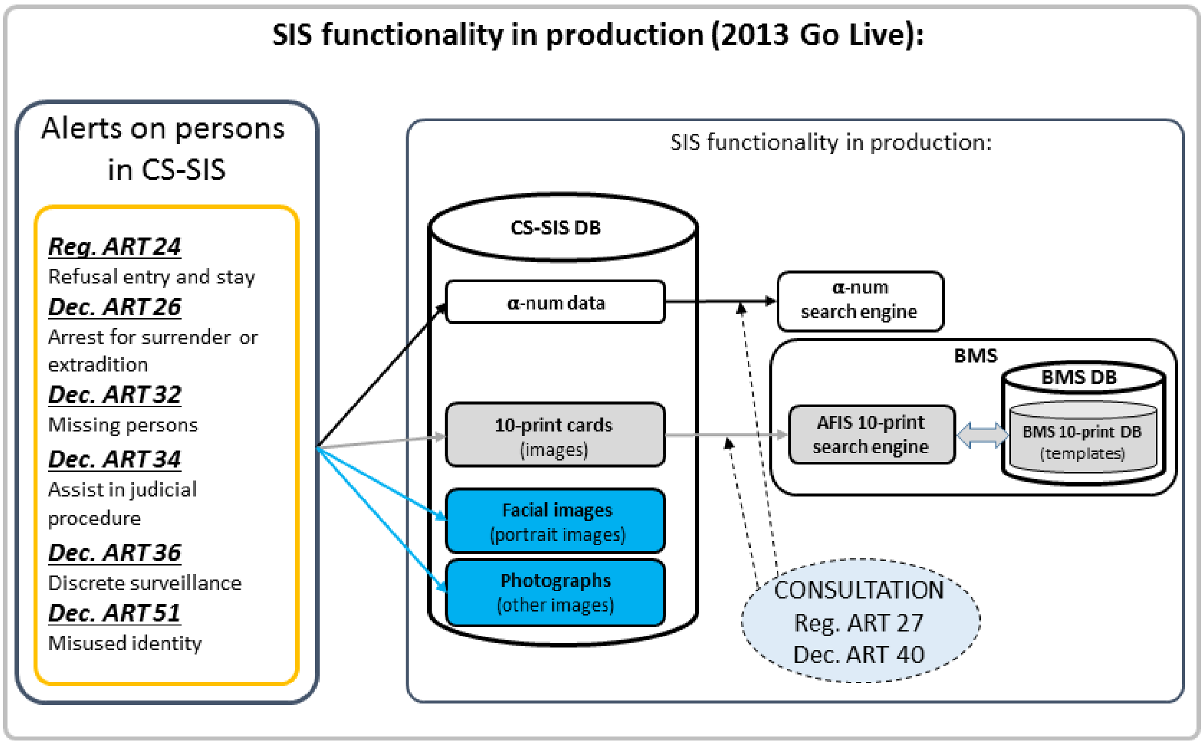
- First, taking into account its regulatory framework, a description of the key aspects concerning CS-SIS today is presented.
- Second, according to the new legislation that has been adopted in November 2018, we present the main changes that will be performed in CS-SIS in order to integrate the new functionalities defined in this legislation.
- Third, according to the challenges exposed in Section 5 (PART I) and to the specificities of the SIS described below in Section 7 (PART II), a series of recommendations are set out on how to tackle such challenges in order to implement an ABIS-Face in CS-SIS in the most effective manner.
- Fourth, adopting a more prospective view which goes beyond today's regulatory framework, we describe some possible functionalities that could be further added to SIS in order to improve its utility and accuracy and provide consolidated services.
- Fifth, we present the final conclusions of the report.

6. Current use of biometrics in CS-SIS since 2013

The present section is mainly focused on the current use of biometric technology within CS-SIS. For a wide and general introduction to SIS, please refer to the introduction of the present report.

At present, the SIS works under the first Regulation and Decision and the system put in production since 2013. According to that regulation, 6 articles allow the end user (i.e., law-enforcers and border guards) to create person-related alerts and consult the CS-SIS (see **Figure 10** below). For further details on these articles, we refer the reader to the legislation, or to the 2015 DG JRC study on the AFIS for SIS where a summary of the regulation can be found. It should be noticed that in Article 51, related to miused identities, only the fingerprint images are stored in the CS-SIS DB, but the searchable templates are not extracted nor stored in the BMS DB.

Figure 10: present functionalities of CS-SIS in production since 2013. Source: EC 2018.


SIS functionality in production (2013 Go Live):
Alerts on persons in CS-SIS
Reg. ART 24
Refusal entry and stay
Dec. ART 26
Arrest for surrender or extradition
Dec. ART 32
Missing persons
Dec. ART 34
Assist in judicial procedure
Dec. ART 36
Discrete surveillance
Dec. ART 51
Misused identity
SIS functionality in production:
CS-SIS DB
α-num data
10-print cards (images)
Facial images (portrait images)
Photographs (other images)
α-num search engine
BMS
BMS DB
AFIS 10-print search engine
BMS 10-print DB (templates)
CONSULTATION
Reg. ART 27
Dec. ART 40

The CS-SIS database stores person alerts which can contain:

- Alphanumeric data.
- 10-print cards (images).
- Photographs and facial images.

Even though photographs and facial images can be stored as part of person-related alerts, the only data that can be used to identify a subject in CS-SIS (i.e., peform a consultation) are:

- Alphanumeric data.
- 10-print cards.

The use of 10-print cards for consultation of SIS implied the integration into its functionality of a **Biometric Matching System (BMS)** which, at the moment, consists only of an **Automatic Fingerprint Identification System (AFIS)**. The AFIS started its roll-out phase in March 2018 with 8 Member States and one associated State connected to it. The BMS database contains only the extracted templates from the 10-print cards stored in the CS-SIS DB. These templates are used by the AFIS search engine to consult the database.

Therefore, as shown in **Figure 10**, the SIS contains two different databases with biometric data:

- **CS-SIS DB**. This database stores the alerts with the original biometric data sent by the MSs to the central SIS (CS-SIS). That is, the CS-SIS DB contains the fingerprint and facial images present in SIS together with the alphanumeric data of the alerts.
- **BMS DB**. The BMS database is associated to the Biometric Matching System and stores **searchable templates** extracted from the biometric samples stored in the CS-SIS DB (i.e., 10 print cards). At the moment, only templates coming from 10 print cards are extracted and stored in this database for the purpose of biometric **10-print comparison**. That is, currently no face templates are extracted or stored since the previous regulation didn't allow for face images to be used for identification.
  It should be noted that, if at any point there is a major update of the BMS (e.g., a new provider is selected), the new BMS would have to go back to the CS-SIS DB in order to extract again the new searchable templates from the original images (biometric samples) submitted by the MSs at the time of the creation of the alerts.

As already mentioned in the introduction of the report, the SIS is basically used in two main contexts which present different operational requirements:

- Law enforcement use.
- Border management use.

In the next subsections we describe the functioning of the current CS-SIS in each of these scenarios, with respect to the identification of subjects using biometric data, that is, their 10-print information (which is the only biometric information that can be used at the moment for identification purposes).

In the next section the terms "match" and "hit" are used according to the definitions given in Article 3 in the Police and Border new SIS regulation from November 2018 (see Section 7 for further details on this new regulation):

- **Article 3 Border, Police; Definition (7)**: a '**match**' means the occurrence of the following steps:
  (a) a search has been conducted in SIS by an end-user;
  (b) that search has revealed an alert entered into SIS by another Member State;
  (c) data concerning the alert in SIS match the search data;

- **Article 3 Border, Police; Definition (8)**: a '**hit**' means any match which fulfils the following criteria:

  (a) it has been confirmed by:

    (i) the end-user; or

(ii) the competent authority in accordance with national procedures, where the match concerned was based on the comparison of biometric data;

(b) further actions are requested;

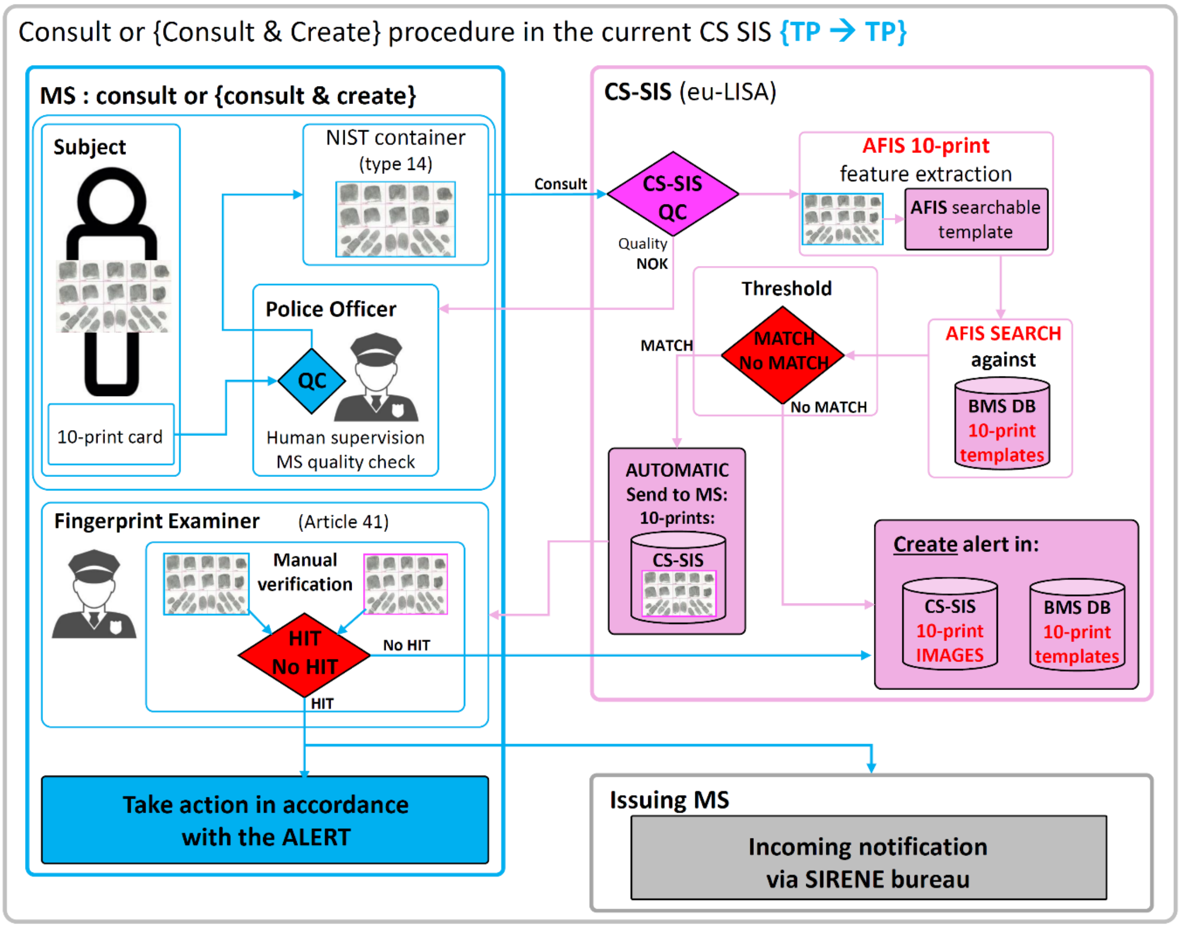
6.1. Current use of the CS-SIS AFIS in a law-enforcement context

According to the current legislation, the creation of alerts in CS-SIS is strictly in competence of national law-enforcement agencies. Every MS connected to CS-SIS is allowed to create alerts in the system following a "Consult and create" procedure. This means that, before creating a new alert related to a subject, the system conducts a search in order to verify if there is already an existing alert associated to that same subject. The full "biometric" operation of SIS in this scenario is depicted in **Figure 11**. As mentioned above, at the moment the consultation process with regard to biometrics involves only 10-prints and can be described as follows:

- STEP 1. The subject of the alert is booked at the police station. His 10-print card is acquired using live-scanners (usual case) or else the ink-and-paper process (always less common). Alternatively, if the subject of the alert is not available at the police station (e.g., alert regarding a missing person) the 10-print card may be taken from the national registry (if available).
- STEP 2. The quality of the fingerprint images is usually verified at the level of the MS. Fingerprints which are not of sufficient quality can be re-enrolled (if the subject is present at the police station, which is the most usual case).
- STEP 3. Once the 10-print card has been created at the MS, it is sent using a NIST container type 14 to CS-SIS.
- STEP 4. At CS-SIS it is checked that the NIST container is compliant with the specifications of central system. After that, there is a biometric quality check in order to ensure that the quality of the fingerprint images is sufficient in order for the AFIS to extract a searchable template. In case that any of these two checks fails (NIST container compliancy or minimum quality to extract a template), CS-SIS notifies the MS.
- STEP 5. If the templates are extracted from the 10-print card, the AFIS searches the BMS DB containing the 10-print templates of person-related alerts existing in the system.
- STEP 6. The AFIS technology based on 10-print has shown to be accurate enough in order to pre-define a threshold, based on which the system can produce a match (if there is a comparison score above the threshold between the searched fingerprints and any of the templates in the BMS DB) or a no-match (if there is not a comparison score above the threshold).
- STEP 7. In **case of a no-match**, it means that there is no alert containing 10-prints in CS SIS related to the searched subject. Following this outcome, the MS conducting the consultation has the option to create a new entry in SIS, where the 10-print card is stored in the CS-SIS 10-print image DB and the searchable templates are stored in the BMS 10-print template DB.
- STEP 8. In **case of a match**, the 10-print card corresponding to the alert producing the match is sent to the MS consulting the SIS. At the MS, a forensic expert verifies the match manually, comparing the existing 10-print card in SIS, to the 10-print

card of the subject of the consultation. If the manual verification results in a **no-hit** (i.e., error of the AFIS system), a new alert is optionally created in CS-SIS in an analogous way to the no-match case (described above). In case of a hit (i.e., there is already an alert in SIS related to the subject of the consultation), the MS will take action according to the alert and the MS owner will be informed of the hit through the Sirene Bureau.

Figure 11: Police – CS-SIS Consultation and Alert Creation procedure. Source: EC 2018.



6.2. Current use of the CS-SIS AFIS in a border context

In the case of checks at regular border crossings, the main differences to be taken into account with respect to the use of SIS in a law-enforcement context are:

- In a border crossing, the subject of the consultation is a standard traveller (i.e., a citizen or visitor of the Schengen area), while in the context of law-enforcement the data subject is in general the suspect of some crime.
- In a border crossing, there is a strict limitation in the time that the border guard can spend with every person. The border crossing should be as fast as possible. In the usual law-enforcement scenario, the officers have some good cause to stop the

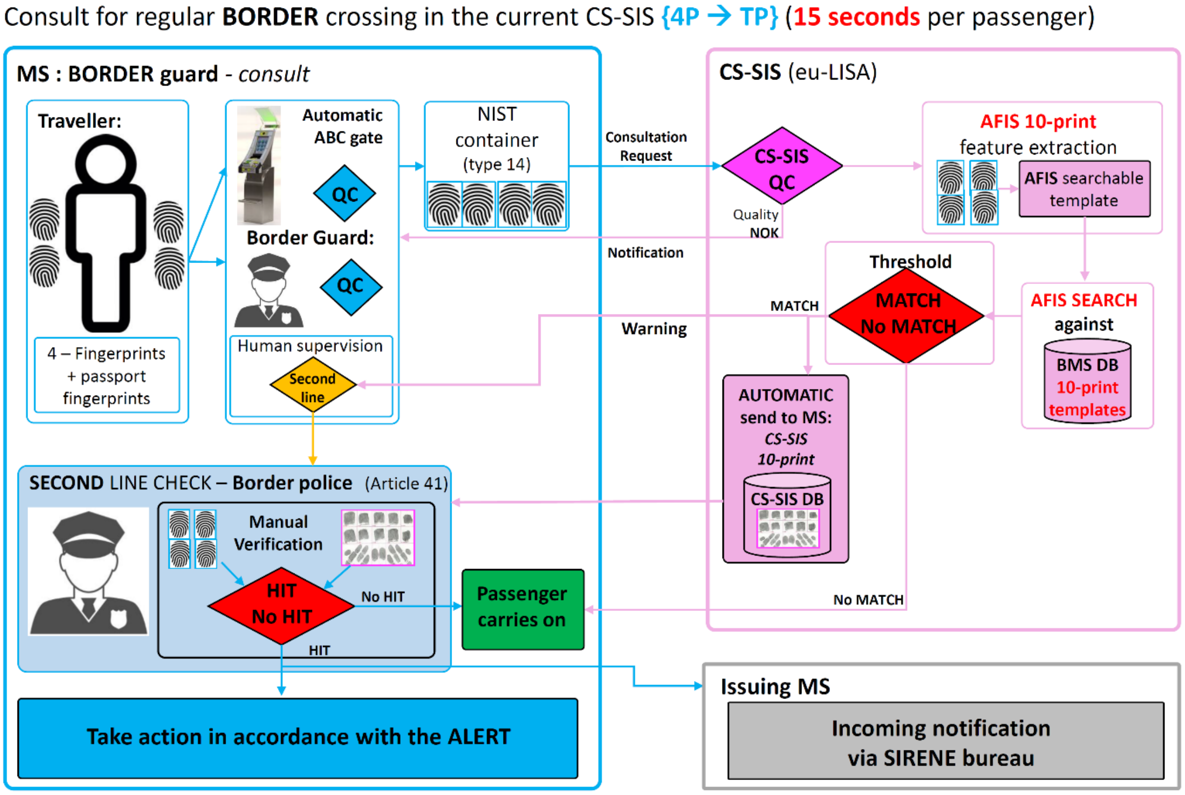
subject and time is usually not a limitation to perform the necessary checks on the suspect.

- In a border crossing, only consultations of the CS-SIS using the BMS AFIS are performed, but no new alerts are created in the system.

The full "biometric" operation of SIS in this scenario is depicted in **Figure 12**. As mentioned above, at the moment the consultation process with regard to biometrics involves only 10-prints and can be described as follows:

- STEP 1. In this scenario, a subject who wants to enter the Schengen space arrives in front of a Border Guard or at the Automatic Border Checking (ABC) gate. Typically, his index and middle fingers of both hands are acquired using a live-scanner and used to consult CS-SIS after a sufficient image quality level has been reached. It should be noticed that the number of fingerprints acquired for the consultation can vary. It should also be noticed that although speed plays a significant role in this scenario (15s turnover to be guaranteed), quality of the fingerprints consulted from the border post against CS-SIS database should be verified and only "good quality fingerprints" should be submitted for consultation.
- STEP 2. The acquired fingerprints are embedded in a NIST container type 14 and sent to the CS-SIS where two checks are performed: 1) if the NIST container is compatible with the CS-SIS; 2) if the fingerprint images are of sufficient quality for a searchable template to be extracted from them. In case either of these two checks is not successful, a notification is sent to the border guard.
- STEP 3. In case both of the previous checks are successful, a searchable template is extracted from the fingerprints and a search is conducted by the SIS AFIS on the BMS DB conatining 10-print templates.
- STEP 4. The AFIS technology based on 10-print has shown to be accurate enough in order to pre-define a threshold, based on which the system can produce a match (if there is a comparison score above the threshold between the searched fingerprints and any of the templates in the BMS DB) or a no-match (if there is not comparison score above the threshold).
- STEP 5. In **case of a no-match**, it means that there is no alert containing 10-prints in CS SIS related to the searched subject. Following this outcome, the borderguard is informed and, if all other checks performed by the guard are also satisfactory, the person is allowed to enter the SCHENGEN space.
- STEP 6. In **case of a match**, the 10-print card corresponding to the alert producing the match is sent to the MS consulting the SIS. The border guard sends the person to a second line of check where police officers can take more time to examine the case. At this second line of check, a forensic expert verifies the match manually, comparing the existing 10-print card in SIS, to the fingerprints of the subjec used to perform the consultation. If the manual verification results in a **no-hit** (i.e., error of the AFIS system), the person will eventually be allowed to carry on. In case of a hit (i.e., there is already an alert in SIS related to the subject of the consultation), the MS will take action according to the alert and the MS owner will be informed of the hit through the Sirene Bureau.

Figure 12: Border – CS-SIS Consultation procedure. Source: EC 2018.



According to the current legislation, "*… Dactyloscopic data may be searched in all cases to identify a person. However, dactyloscopic data shall be searched to identify a person where the identity pf the person cannot be ascertained by other means…*" (current legislation on Border, Article 33, paragraph 2). In practice, this implies that the search using biometric fingerprint data is preceeded either by:

- Detection of inconsistency in the biometric travel document;

or following:

- Failed alphanumeric search;
- Failed automatic fingerprint verification (1:1 comparison by ABC gate or border officer);
- Failed automatic face verification (1:1 comparison by ABC gate) or visual inspection by border officer.

It is important to take into account that in order to ensure a smooth and fast processing of passengers at the borders, the feedback given to the border guard at the **first line** of check should be as concise and informative as possible, for instance:

- No passenger-related information in CS-SIS (green light – passenger proceeds)
- Information present in CS-SIS (amber light – passenger goes to the **second line** of check)

7. New Functionality of CS-SIS: 2018 Regulation

On the 28th November 2018 a new legislation for SIS was adopted by the European Union. This legislation is divided into three documents, one related to POLICE[15] context, one directed to BORDER[16] context and one for the RETURN[17] of illegally staying third country nationals. In the following we summarise the articles related to person alerts:

- Border ARTICLE 24: Refusal of entry and stay.
- Police ARTICLE 26: Alerts on persons wanted for arrest for surrender or extradition purposes.
- Police ARTICLE 32: Alerts on missing persons or vulnerable persons who need to be prevented from travelling.
- Police ARTICLE 34: Alerts on persons sought to assis with a judicial procedure.
- Police ARTICLE 36: Alerts on persons and objects for discreet checks, inquiry checks or specific checks.
- Police ARTICLE 62: Additional data for the purpose of dealing with misused identities. It is important to notice that, as was the case in the 2013 regulation, this article allows to store in the CS-SIS DB fingerprint and face images of subject whose idenity has been misused, but templates are not extracted nor stored in the BMS DB to be searched.
- Police ARTICLE 40: Alerts on unknown wanted persons for the purposes of identification under national law.
- Return ARTICLE 3: Alerts on the return of illegally staying third country nationals.

In light of the previous articles, the CS-SIS is to be enhanced by new functionalities (see **Figure 13**). According to the new legislation, the main biometric-related novelties with respect to the 2013 regulation may be summarised as follows:

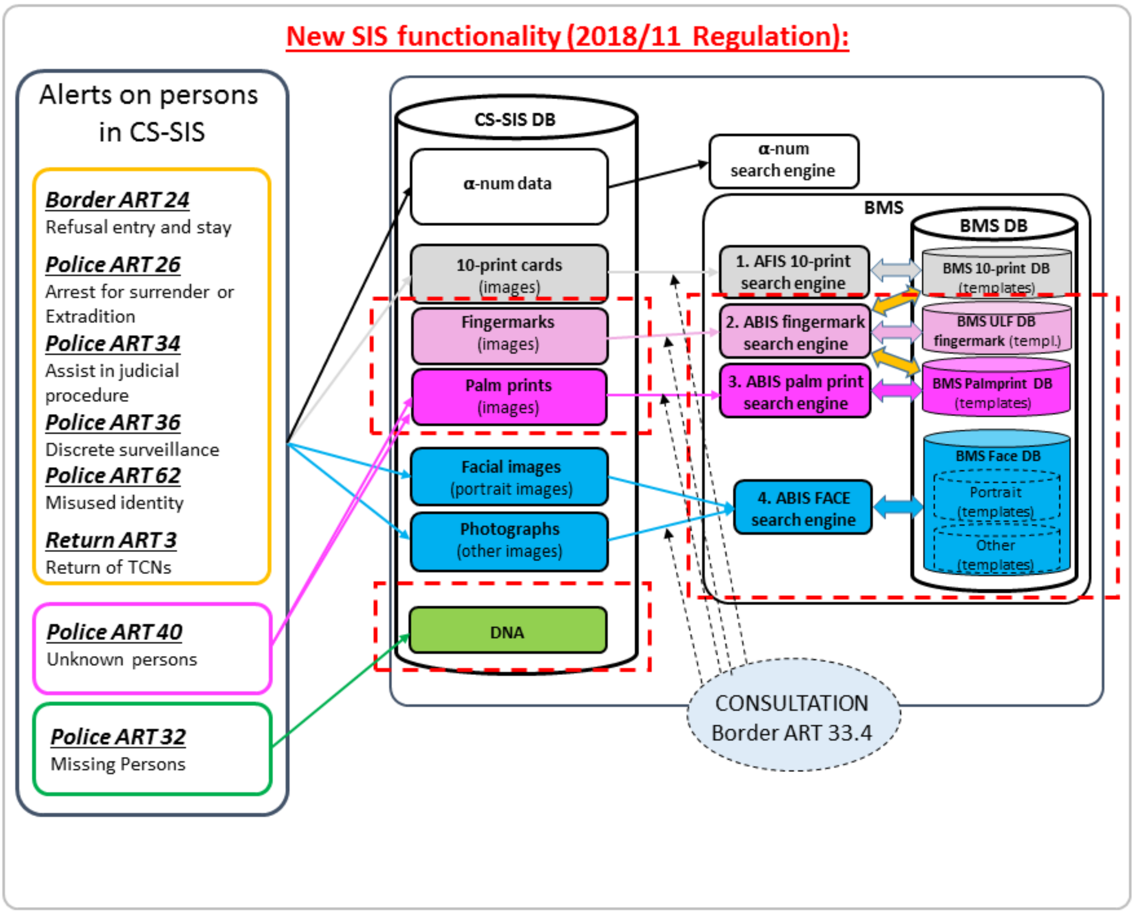
- A new type of alert related to "unknown persons" is introduced in article 40 (POLICE document). As a result, two new biometric characteristics are introduced: fingermark and palmmarks (to allow searching for unknown persons). Therefore, while in the 2013 legislation there were 6 categories of alerts related to persons, in the 2018 legislation there are 7 alerts related to persons.
- A search engine is to be added for automatic face recognition whenever the technology becomes ready (the storage of facial images and photographs was already allowed under previous legislation).
- In the case of missing persons, if fingerprints or facial image(s) are not available, the storage of DNA profiles is to be allowed (subject to an independent readiness and availability assessment).

[15] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1862&from=EN

[16] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1861&qid=1544694006055&from=EN

[17] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R1860

Figure 13. New functionalities of CS-SIS according to the 2018 Regulation. Source: EC 2018.


New SIS functionality (2018/11 Regulation):
Alerts on persons in CS-SIS
Border ART 24
Refusal entry and stay
Police ART 26
Arrest for surrender or Extradition
Police ART 34
Assist in judicial procedure
Police ART 36
Discrete surveillance
Police ART 62
Misused identity
Return ART 3
Return of TCNs
Police ART 40
Unknown persons
Police ART 32
Missing Persons
CS-SIS DB
α-num data
10-print cards (images)
Fingermarks (images)
Palm prints (images)
Facial images (portrait images)
Photographs (other images)
DNA
α-num search engine
BMS
BMS DB
1. AFIS 10-print search engine
2. ABIS fingermark search engine
3. ABIS palm print search engine
4. ABIS FACE search engine
BMS 10-print DB (templates)
BMS ULF DB fingermark (templ.)
BMS Palmprint DB (templates)
BMS Face DB
Portrait (templates)
Other (templates)
CONSULTATION Border ART 33.4

Following the current architecture of the SIS, as depicted in **Figure 13**, there are two blocks at the CS-SIS level, the CS-SIS DB and the Biometric Matching System (BMS). However, based on the new 2018 legislation, these two blocks will be now substantially updated:

- The Central System Database (CS-SIS DB), will store alphanumeric data, facial images, photographs, 10-print cards, palmprints, fingermarks, palmmarks and DNA profiles.
- The Biometric Matching System (BMS) part is formed by:
  - A BMS database which will contain searchable templates which have been extracted from the different biometric modalities: 10-print (including palmprints), Face (portraits and other type of images), Fingermark (Unsolved Latent Files, ULF) or Palmmarks.
  - A number of biometric search engines that will perform the consultations on the BMS DB: AFIS 10-print, Automatic Biometric Identification System (ABIS) fingermark, ABIS-Palmmark and ABIS-face.

The previous databases and search engines will interact in the following way:

- ABIS 10-print search engine interacts with BMS 10-print templates

- ABIS-Fingermark search engine interacts with:
- BMS Unsolved Latent Files (ULF) fingermark and palmmark template database.
- BMS 10-print template database (all of the 10-print fingerprint images from the CS-SIS have to be reprocessed in order to be compatible with ABIS-Fingermark search).
- BMS Palmprint template database (similar as in the previous case, all the palmprint images stored in CS-SIS have to be processed to be ABIS-Fingermark search compatible).

- ABIS-Palmprint search engine interacts with:
  - BMS ULF fingermark and palmmark template database
  - In principle BMS palmprint template database although fingerprint are always favored

- ABIS-Face search engine interacts with:
  - BMS face image database of portraits (high quality, high resolution, full frontal, controlled conditions acquired photographs)
  - BMS face image database of other faces of lower quality, produced in uncontrolled conditions, though still showing "sufficient resolution" to produce an ABIS-searchable template (e.g. webcam, CCTV, photograph, etc). Please see below for further details

Please note that, according to the definitions introduced in Article 3 of the 2018 regulation:

<u>DEFINITION (Reg. ART 3): Facial Image.</u> *'Facial image' means digital images of the face with sufficient image resolution and quality to be used in automated biometric matching*.

From a technical biometric perspective, this definition is not a closed one and is up for interpretations of what constitutes or not a facial image. Furthermore, no definition is given in the Regulation with regard to what is understood by a 'photograph' and the difference that exists with respect to a 'facial image'. Therefore, in the present JRC report we will consider and adopt the next definitions for those two concepts:

<u>DEFINITION 'Facial Image' (JRC report).</u> A portrait image, of very high quality, as close as possible to the ICAO standard.

<u>DEFINITION 'Photograph' (JRC report).</u> Any other type of image containing a face that does not comply with the definition of a 'facial image'. The quality range in this case can be very broad, from relatively high quality images to very poor images from which a searchable template cannot be extracted.

It should be noticed that for technical reasons related to the current limitations of the accuracy of face recognition technology, it is the assumption of the present study that templates extracted from "facial images" (i.e., portraits) and "photographs" (i.e., other face images) will be flagged on different ways within the BMS Face DB. The rationale behind keeping this logical separation in the searchable database comes from two observations:

- Observation 1. The need to have different answers from the system in different contexts. In the border context, at the first line, the border guard should only get as answer from the system a match or no-match, as the time constraints in this scenario do not allow him to check a rank list of candidates for every consultation.

However, in the second line of check or in a police scenario, a rank list of candidates is the preferred output from the system.

- Observation 2. The results of the last NIST FRVT evaluation (see Section 2.3.1.2). These results have shown that under good quality conditions and for databases of up to 10 million entries, current face recognition technology is capable of returning with a high degree of reliability a match/no-match answer. However, on lower quality images, the reply shall still be in all cases a rank list of candidates to be manually verified by an expert.

Based on these two observations, a good compromise is to flag the images stored in the database with their quality (i.e., "portrait" or "other") in order to be able to select the type of images the search should be performed on depending on the scenario: for the first line of check only portratis would be used (match or no-match answer), while in the second line of check (or other law-enforcement contexts) all images would be searched (the output would be a list of candidates).

Recommendation 1:

Searchable database logical separation.

We recommend that the searchable database containing the templates extracted from face images of the eventual SIS ABIS-Face is logically separated into two quality types: 1) portraits; 2) other type of images.

This will enable to perform the searchers on a given subset of images depending on the context. This will allow the system to provide as output a "match/no-match" response (portrait images) or a "list of candidates" (all images) depending on the quality of the images that the search has been performed on.

Recommendation 2

Single ABIS-Face search engine.

Even though we recommend to have two logically separated databases according to image quality (i.e., "portraits" and "other", see Recommendation 1), we recommend to have only one unique ABIS-Face search engine to perform the automatic consultations on either of the two face image quality types.

Having two dedicated search engines for "portraits" and "other", presents the high risk of overfitting the systems to a specific type of images. In this case, if the images used in the comparison differ slightly from the images expected by the system, the accuracy will drastically drop. Therefore, such systems will not be able perform well in cases where it is needed a comparison between images belonging to different quality classes (i.e., portraits VS other)

On the other hand, current deep-learning based technology has shown that, if trained on a sufficiently large quantity of data, it is able to generalise well to different types of images. Therefore, one single system trained on a significant large quantity of variable data will perform similarly to systems tuned to a specific quality type, without presenting the accuracy drop in case the compared images do not comply with the expected quality.

While the recommendation is to have one single ABIS-FACE, this unique search engine may include several comparison algorithms inside, in order to optimise the accuracy and the respond time. For large biometric IT systems, it is typical that a system includes several comparison algorithms in cascade, for instance:

1) a first algorithm very fast but with not perfect accuracy that does an initial coarse filtering of the complete reference database in order to discard obvious non-matches;

2) a second comparison algorithm, slower than the previous one but more accurate in order to do a second filtering;

3) a last algorithm slow but very accurate, which only runs on a very reduced pre-filtered database (by the other two algorithms).

How this internal comparison strategy is defined is part of the vendor design and will not be discussed here. However, independently of how it is accomplished internally, it is important to evaluate that the final complete system complies with the accuracy and speed requirements.

The key point to fulfill recommendation 2, that is, to separate templates between "portraits" and "other" is to determine what constitutes an image that can be flagged as "portrait" quality. This can be done in two ways, which are correlated to some extent:

- Measurable characteristics such as resolution, pixels in between eyes, off-angle allowed, sharpness etc. In this case the guidelines would set minimum requirements for each of these characteristics.
- A single overall quality metric that estimates the expected accuracy of a given image for recognition purposes. In this case the guidelines would simply establish the miminum quality required to qualify as "portrait".

Although both approaches are feasible, the second one is preferred as it has shown in other biometric modalities to give more consistent results (e.g., the use of NFIQ2 in fingerprints to define minimum quality requirements). The main drawback is that, at the moment, there is no standard face quality metric to be relied on (as has been explained in Section 3).

Recommendation 3:

Definition of parameters for portrait images.

We recommend to clearly set the parameter(s) defining what consitutes an image that can be flagged as "portrait" quality.

These parameters should be set together with the supplier of the ABIS-Face system, as they will depend on the system accuracy. The requirements should be set to guarantee a certain accuracy on the subset of "portrait" images. For example, False Positive Identification Rate, FPIR=0.001% over a DB of 1 million entries.

Recommendation 4:

Development of an overall face quality metric.

Linked to recommendation 3, we recommend to promote the development of a vendor-independent, robust and reliable, face quality metric to be integrated in the ABIS-Face as soon as it becomes available.

This quality metric could be the result of: 1) the combination of a number of individual values estimating human-defined features such as illumination, sharpness, pose, background, etc. 2) deep-learning derived features; or 3) a combination of both hand-crafted and deep-based features.

The development of a face quality metric should contribute and get feedback from the currently under review standard: "ISO/IEC TR 29794-5 Information Technology – Biometric sample quality – Part 5: Face image data".

Whenever such a quality metric, ideally subjected to its incorporation into an international standard becomes available, we recommend the following actions:

- Integration in CS-SIS. We recommend to include in the CS-SIS ABIS-Face the quality metric algorithm. The quality metric at central level can be of great utility to: 1) as monitoring tool of the face images stored in CS-SIS; 2) to automatically classify images between the two quality types 'portrait' and 'other'; 3) to give feedback to the MSs regarding the quality of the face images submitted to CS-SIS.
- Integration at MS level. We recommend to implement the quality metric also at the level of the MS. In this case the quality metric can be useful to incorporate in an acquisition loop/recapture procedure to be carried out until satisfactory quality face images have been obtained both at the time of enrolment and of querying the system. This procedure should contemplate alternative acquisition processes, according to the sample quality, and should include human intervention, where appropriate.

As was clearly shown in the description of the evaluation of face recognition systems presented in Section 2, the accuracy of this technology is totally dependent on the data that it is tested on. Therefore, in order to define an accuracy to set the requirements to establish what consitutes a "portrait" image, it is necessary to perform an evaluation of the system on the type of data it will be running on.

Recommendation 5:

Evaluation of the ABIS-Face on operational data.

We recommend to perform an evaluation of the ABIS-Face on the operational data already present in CS-SIS in order to determine:

- The accuracy that should be expected for portrait images (or, if available, the minimum quality required to categorise an image as portrait).
- Decision thresholds to produce match/no-match responses (see recommendation 12).

We recommend that, in addition to the initial evaluation on operational data to determine certain parameters of the system (e.g., minimum quality for portrait images, threshold to determine match/no-match), a similar evaluation is performed on a regular basis in order to adapt the parameters to possible changes in the accuracy of the system due to an increase/decrease of the enrolled data or to an update of the system.

Any evaluation of FR systems should follow as close as possible the directives given in: ISO/IEC 19795-1 2006 “Information Technology – Biometric Performance and reporting – Part 1: principles and framework”.

8. Integration of an ABIS-Face in CS-SIS

According to the article 33 of the new 2018 SIS Border Regulation: "*As soon as it becomes technically possible, and while ensuring a high degree of reliability of identification, photographs and facial images may be used to identify a person in the context of regular border crossing points.*"

This means that, while it is in the law-enforcement context where SIS alerts are created (including photographs and facial images), the SIS for the time being can only be consulted using face data in the context of regular border crossings.

This way, law-enforcers will consult the SIS based on alphanumeric data and/or dactyloscopic data, but not face data. In the case that a consultation with alphanumeric data and/or dactyloscopic data results in a no hit, then law-enforcers have the opportunity to create a new alert, including in it face data. This face data will later be consulted at regular border crossings. Therefore, although not directly consulting it, the role of law-enforcers is key in order to obtain a high accuracy from the ABIS-Face in CS-SIS, since they are responsible for the quality of the data being enrolled to the CS-SIS DB. The data should not only be of high quality but should also be as consistent as possible in terms of acquisition parameters: type of sensors used, illumination conditions, resolution, size of the images, etc.

RECOMMENDATION 6:

High-quality enrolment process.

We recommend that, whenever a cooperative data subject is available at the enrolment process, that is, in most of the cases, the enrolment phase should favor the use of high quality cameras, in fully controlled conditions, to adhere as much as possible to the ICAO Standard specifications or to the ISO/IEC 19794-5 specifications, under the supervision of experienced operators, as is usually the case in a law enforcement context. This should result in the production of high-quality portrait-like face images which are to be stored in the CS-SIS database.

In order to promote this high-quality enrolment process, we recommend that best practices for face acquisition are compiled and distributed to the Member States in order to obtain a central database as homogeneous as possible.

As such, in section 8.1 the border use-case will be described, as is the only one considered in the current legislation. Other possible future use-cases of the face modality will be treated in section 9.

8.1. ABIS-Face: Border use-case

The main boundary conditions of the border use-case for the new SIS ABIS-Face system are as follows:

- **Time constraints**: Each border crossing should be processed fast by border guards. There is a very strict time limitation that they have to comply with (around

30s per person). All necessary system based checks should be completed within this time frame.

- **Number of consultations**: The SIS might be consulted every time a person wants to enter the Schengen space as well as in the course of a VISA request procedure. This results in a potentiall very large number of consultations that the system should be able to handle in real time. The exact number of consultations related to persons at border crossings may be roughly estimated as the total number of annual crossings of the Schengen borders.
- **Database size**: The size of the database being consulted is relatively small (compared for instance to the 12 million data subjects in the last FRVT 2017, see Section 2.3.1.2). At the moment it contains around 1 million alerts related to persons. Should all alerts contain facial images, this would entail that the ABIS-Face would have to perform searches on a database of 1 million images.

Recommendation 7:

Need for complementary statistics.

We recommend that eu-LISA identifies the best possible ways to include in its annual report the statistics of CS-SIS:

- The number of consultations per year related to persons at border checks. In order to complement this assessment at central level, we also recommend that Member States report annually on the number of consultations related to persons that have been carried out on their national copies.
- Once the ABIS-Face is running, the number of consultations performed based on the ABIS-Face.
- The number of person related alerts that contain face images.
- The number of hits obtained based on ABIS-Face.
- The number of duplicated alerts detected based on ABIS-Face.
- The quality of the enrolled face images in CS-SIS.
- The quality of the live images submitted to perform queries in CS-SIS.

The previous parameters, together with the state of the art of current face recognition technology (presented in PART I of the present study), determine the workflow for the new ABIS-Face functionality at regular border crossings, which is summarised in the flow chart of **Figure 14**. The workflow might be as follows:

- STEP 1. When the person who wants to enter into the Schengen area arrives at the border, the border guard takes a live picture of him/her and performs a first quality check (preferably assisted by a ISO/IEC standardised quality metric). If the quality is not satisfactory he/she should have the opportunity to recapture the image.
  Typically the border guard could also use the picture in the traveller’s passport to perform the consultation. However, the level of compression of pictures in passports

varies greatly among countries. In some cases, this compression is so big that the picture is not usable for automatic face recognition systems. It has been proven that, in terms of accuracy, it is more reliable to take a live picture of the traveller at the time of the border crossing.

Recommendation 8:

Use of live captured images.

For consultation of CS-SIS at border crossings we recommend to use in all cases a live picture of the traveller, carefully designing the set-up of the capture points (see recommendation 10). The additional use of the face image stored in the passport chip can be optional although it is not recommended as this image:

- Is in general of lower resolution than the images captured live.
- Face Recognition systems have shown to obtain worse accuracy using passport images than live-captured images.
- Increases the vulnerability of the system, since it cannot be guaranteed that the image in the passport belongs to the traveller (e.g., morphing attacks).

Recommendation 9:

Quality of capture points.

Supervision by an operator. Adequate operator training is recommended, in order to:

- Train the operator to capture good quality face images (e.g., indicate him the best position for the capture subject, pose, face expression, presence of glasses).
- As supervision of biometric acquisition is a repetitive task and requires additional attention in the case of centralised enrolment stations. The aim is to avoid tiredness and boredom adversely affecting the process.
- Train the operator to detect Presentation Attacks.

In case of automatic ABC gates, they should be thoroughly tested in each location where they will be deployed to ensure their ability to capture good quality face images. ABC gates should in all cases be equipped with Presentation Attack Detection measures.

Adequate sensor. We recommend to use performant cameras (e.g. in speed, imaging sensor and resolution), offering also enhanced capabilities to acquire good quality images in sub-optimal environments.

Enhanced graphic user interface (GUI). We recommend that capture points have large displays and provide real-time feedback regarding the quality of the acquired data.

Proper user interaction. The enrolment process should be user-friendly with clear step-by-step procedures properly explained. The use of good ergonomics should be considered to support better acquisition practices. The user should receive some feedback from the system as where to locate himself.

Adequate environment. The acquisition environment should be appropriate in terms of illumination, temperature and backgrounds both for the subject and the operator. These elements are recommended mainly for fixed stations but similar considerations are instrumental as well for mobile stations. It is especially relevant to pay attention to the illumination factor, as it is key to the acquisition of good quality face images.

Sensor maintenance. There should be regular and systematic maintenance of the enrolment stations to avoid a decrease in performance, especially in the case of consultation processes taking place in heavily used check points (e.g., high-traffic airports).

- STEP 2. The face image is sent in a NIST container type 10 to the CS-SIS. The CS-SIS makes two initial checks at this point: 1) check of the compliance of the NIST container with the SIS requirements; 2) check that the face image is of sufficient quality to extract from it a searchable template. If either of this checks results negative, the border guard gets a notification in order to take action (e.g., recapture the face).

Recommendation 10:

Common exchange standard.

At the moment, the exchange of face data in the SIS system is done on slightly modified version of the ANSI/NIST ITL 1-2011 containers type 10, as required by the SIRENE manual. These containers seem to provide an appropriate basis regarding the exchange of face data. We recommend that an automatic check for their mandatory and complete implementation should be developed in order to appropriately support the deployment of the SIS ABIS-Face functionality.

A transition between the NIST container to the ISO/IEC 39794-5 standard (which will soon be available) could be explored. Two main reasons for this possibility:

- The ISO/IEC 39794 standard is an extensible data format that guarantees both backward and forward compatibility (in case that future versions of the standard require further data fields to be included in the containers).
- The ISO/IEC 39794 standard allows for human annotated points to be encoded in facial images. These points can help to enhance the accuracy of FR systems under certain contexts.
- The ANSI/NIST ITL 1-2011 is mainly a forensic-based standard. This could be seen as user-unfriendly in order to process the data of travellers.

- STEP 3. Since the face image sent for consultation is taken under relatively controlled conditions at the border (usually assisted by a border guard), it is assumed to be of quite high quality: frontal pose, sufficiently good illumination, good resolution, no occlusions. Therefore, it is assumed that in the vast majority of cases the system will extract a searchable template from it. This template should be searched ONLY against the portrait templates contained in the BMS Face DB (NOT against the templates coming from "other" type of images). Please see Recommendation 1 above, for an explanation of this design recommendation.
- STEP 4. The system will initially return a list of N candidates. A decision threshold will be applied to this list of candidates in order to reduce the output of the system to a "match/no-match" reply. The inclusion of such a threshold is necessary in order for the border guard at the first line of check to take a fast decision on the data subject as, in a border scenario, the time limitations do not allow him/her to check a list of candidates for every border crossing.

Recommendation 11:

Computation of the match/no-match threshold.

We recommend to set the threshold that defines the match/no-match output of the system based on the acceptable number of false matches to be produced by the system.

This rate is defined by the False Positive Identification Rate (FPIR) of the system and determines the number of subjects that will be sent to the second line of inspection due to a mistake of the system. Therefore, the FPIR is a determinant factor to set the amount of workload and manpower that will be needed for the second line of inspection (based only on face consultations).

We recommend to perform an evaluation of the ABIS-Face on the real operational data where it will be used in order to set the threshold for the match/no-match reply according to the FPIR predefined (see recommendation 6).

While the FPIR may be the determinant factor to determine the accuracy of the system, a lower FPIR necessarily implies a higher FNIR, that is, the number of non-detected subjects in SIS will increase.

Recommendation 12:

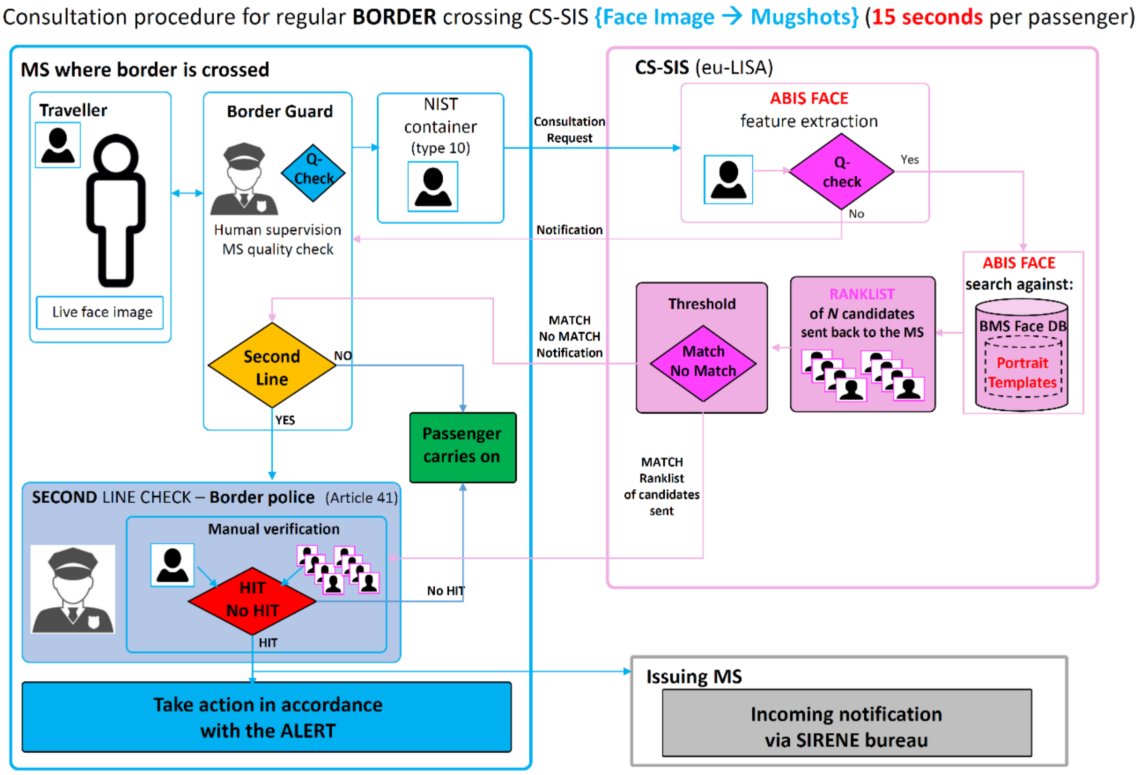
Accuracy evaluation across ethnicities and gender.

We recommend to perform an evaluation of the ABIS-Face on data coming from different ethnicities (e.g., caucasian, black, asian) and also gender (i.e., male, female).

Current face recognition technology based on deep learning has shown that, if not properly trained on data that represents the variability and nature of the data that the system will operate on, it can be biased towards a specific ethnicity or gender. It is important to test that the system has been trained on data that models to the largest extent possible, the variability present in human faces. The ABIS-Face in SIS will be used (initially) in border control, therefore it should be able to perform consistently on all ethnicities and genders.

- STEP 5. The match/no-match reply from the system is returned to the border guard in the first line of inspection. In **case of a no-match**, if all other checks are also satisfactory, the subject is allowed to carry on and access the Schengen space.
- STEP 6. In **case of a match** the subject is sent to the second line of check, where the border police does not have the time restriction of the border guard, and can perform further checks. The border police will receive the initial ranklist of N candidates and manually verify the match against the live picture of the subject. They may even perform a second consultation on CS-SIS using a better quality portrait type of image acquired from the subject at the second line. If the result of the manual verification is a **no hit** (i.e., the system made an error and there is no alert in SIS related to the subject), the passenger is allowed to carry on.
- STEP 7. In case of **a hit** (i.e., there is an alert already in SIS concerning the subject), action will be taken according to the alert in SIS and the MS owner of the alert will be notified through the SIRENE bureau.

Figure 14. Consultation procedure of CS-SIS ABIS-Face for regular border crossing. Source: DG JRC 2018.



In addition to the recommendations specified above, some other general recommendations for the integration of an ABIS-Face in SIS that should be taken into account are:

Recommendation 13:

Need to study the age effect.

We recommend to analyse the difference in accuracy of face recognition technology between different age groups (e.g., children, adults, elderly). Some initial studies have shown that accuracy drops drastically for children below 13 years of age, although these results need further confirmation. There may be an age limit to be set for the accurate use of face recognition technology. Alternatively, specific algorithms may have to be developed to cope with the difficulties presented by certain age groups.

Recommendation 14:

Storage of multiple frontal images.

We recommend to allow for the storage of multiple frontal face images for the same person in order for a SIS ABIS-Face to support a multiple comparison strategy. As long as it is clearly established that the images belong to the same person, having multiple samples can increase the accuracy of the comparison process. Allowing to update alerts with the most recent images of a subject is especially relevant in order to minimise the ageing effect (see recommendation 15).

Recommendation 15:

Corrective measures for the ageing effect.

We recommend to update, whenever possible, old alerts with the most recent face images available in order to reduce as much as possible the ageing effect (reduction in the accuracy of the system due the time separation between the two compared images). This is especially relevant for the case of children where a substantial difference may be observed in their face appearance even for short periods of time. This dimension might be more particularly relevant for alerts related to missing persons.

Recommendation 16:

Storage of additional off-angle (yaw) images.

Off-angle images are unlikely to be used to search by the ABIS-Face in a border context, However, in addition to the frontal face images, we recommend to store as well, whenever possible (e.g., access to subject at a police station), face images at +90, +45, -45 and -90 degrees of the yaw angle. Therefore, the system should allow to label each image with the yaw angle at which it was captured. These images can be useful for:

- The manual verification of a match at the second line of inspection.
- For future potential uses of the ABIS-Face, like for example consultation using images acquired in unconstrained environments (e.g., coming from video surveillance footage), where faces may be seen off-angle.

It should be taken into account that, in order to perform reliable automatic recognition of off-angle images with a large yaw angle (e.g., profile pictures with 90º yaw), it would very likely be necessary to integrate a specific algorithm to operate on those images.

Recommendation 17:

Presentation attack detection measures.

In case an Automatic Border Control (ABC) gate is used at the border crossing instead of a border guard, we highly recommend to put in place the necessary safeguards in the ABC gate in order to minimise the impact of potential prestentation attacks (e.g., ABC gates with integrated presentation attack detection measures). The most likely presentation attacks foreseen are the evasion attacks (i.e., attacks in which the subject tries to hide his identity not to be recognised).

In the case of the presence of a human supervisor, known presentation attacks (e.g., printed pictures, masks) should be easily detected after a brief training of the guard.

An evaluation of presentation attacks and of presentation attack detection methods should follow to the largest extent possible the guidelines and metrics given in the standard “ISO/IEC 30107, Biometric presentation attack detection”.

9. Beyond the current CS-SIS regulatory framework

In this section we suggest a series of possible new operational tools and functionalities which might be interesting to consider in the context of a potential further developments of the SIS Regulatory framework according to article 43.4 or future review of the legislation. These suggestions are largely based on observations and findings obtained during the visits to MSs carried out for the fulfilment of this study.

It has to be underlined that, at this stage, these suggestions are of a prospective nature and will require additional targeted analysis in order to be possibly included in further development or a revision of the SIS legislative framework.

9.1. ABIS-Face consultation in police context

As clearly mentioned in the new legislation adopted in 2018 (see Sections 7 and 8.1), the foreseen use of an ABIS-Face in SIS for consultation is only at regular border crossings. However, Article 43.4 of the Police regulation allows for future uses of this functionality:

> **Article 43.4 Police Regulation. Specific rules for verification or search with photographs, facial images, *dactyloscopic* data and DNA profiles**: *"After the start of the use of the functionality at regular border crossing points, the Commission shall be empowered to adopt delegated acts in accordance with Article 75 to supplement this Regulation concerning the determination of other circumstances in which photographs and facial images may be used to identify persons."*

As explained in Section 8.1, the police, although not allowed to use ABIS-Face for consultation, play a vital role in its performance as law enforcers are in charge of creating the alerts and, therefore, of introducing the face images against which consultations at border crossings will be performed. Furthermore, based on the current use of the SIS AFIS in a police context, it seems reasonable to assume that a possible future use-case for an ABIS-Face in SIS would be to allow for consultations coming from the police environment.

The following subsections describe the workflow for such use-case both for known and unknown subjects.

9.1.1. ABIS-Face consultation in police context: known subject

The main differences between the police use case where the subject of the consultation is at the police station (i.e., known subject) and the border crossing use case are:

- The quality of the face image used for consultation is expected to be higher in the police context since the officers do not usually have time restrictions or equipment restrictions for the acquisition of the face image.
- There are no strict restrictions regarding the time required to obtain an answer on whether the subject is present or not in the CS-SIS database.

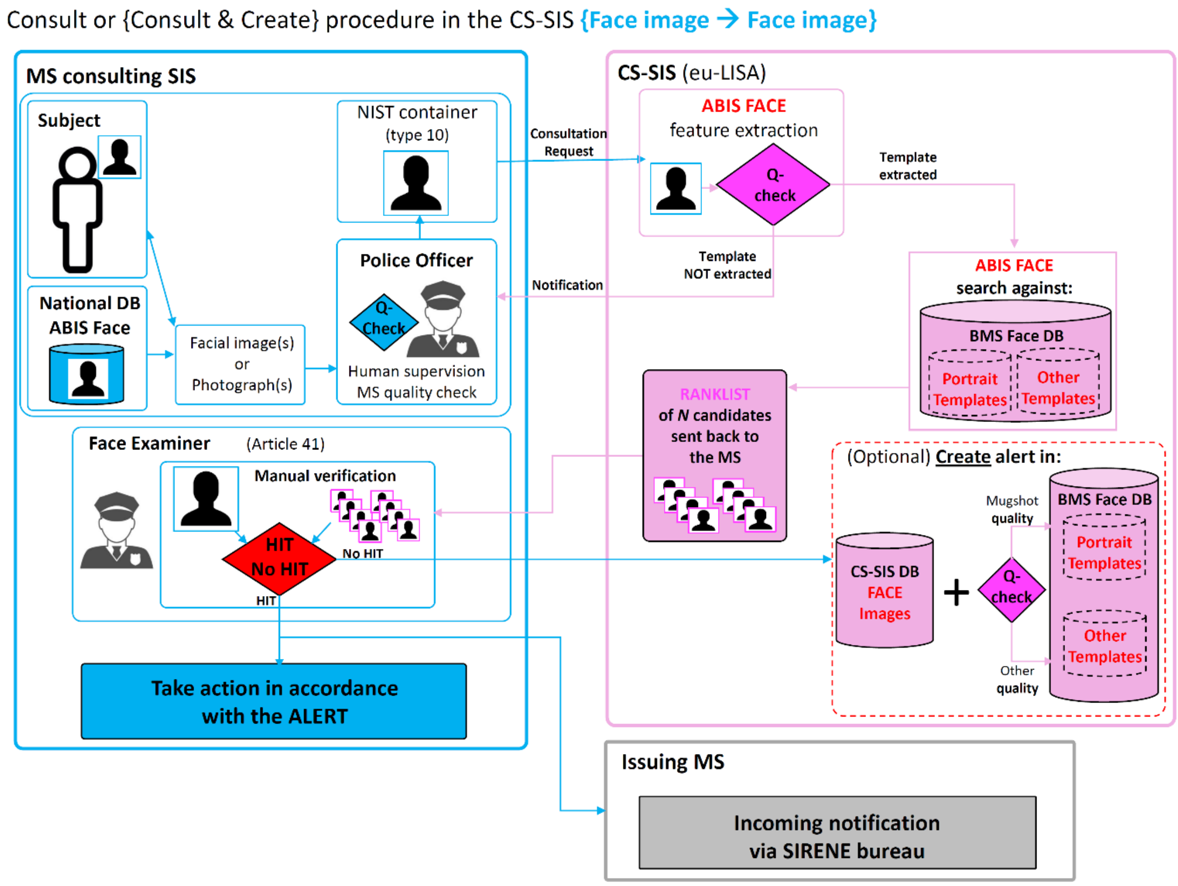
Based on these two differences, the typical workflow of a SIS ABIS-Face is depicted in **Figure 15** and can be described as follows:

- STEP 1. In the usual scenario, the subject of the SIS consultation is at a police station and the police officer takes a live image of him. Since the officer does not have time constraints to take the picture, this image is expected to be a portrait like image, captured in perfectly controlled conditions with top quality photograph equipment. If the quality is not satisfactory he has the opportunity to recapture the image.

In a less usual scenario, the officer may retrieve the facial image from the national ABIS-Face registry (e.g., missing person). In this case, the officer has to accept whatever quality he obtains as he is not in the position to recapture the face image.

- STEP 2. The image is sent in a NIST container type 10 to CS-SIS where it is checked for: 1) the NIST container complies with the CS-SIS especifications; 2) the image quality is high enough to extract a searchable template from it. If either of this two checks is negative, the officer is notified.
- STEP 3. The template extracted from the image used for consultation is submitted to the ABIS-Face to search the complete BMS Face DB, with both the "portrait" and the "other quality" templates. A ranklist of N candidates is produced and sent back to the MS performing the consultation. With respect to the border case, in this scenario it is not needed to set a threshold in order to produce a match/no-match reply from the system, since the police officers: 1) do not have the time limitations of border guards; and 2) the number of consultations would be far lower than in the case of border crossings. This way, police officers can manually check a rank-list for each consultation.
- STEP 4. At the MS a manual verification of the ranklist is performed. If the result of the verification is a no-hit (i.e., there is no alert in CS-SIS associated to the individual), the officer would have the possibility to create a new alert, storing the face image(s) in the CS-SIS Face DB (images) and the templates in the BMS Face DB, flagging the template either as "portrait" or "other" depending on the quality.
- STEP 5. In case of hit after the manual verification (i.e., there is already an alert in CS-SIS associated to the subjet), the MS performing the consultation will take action according to the alert an the MS owner of the alert will be informed through the SIRENE Bureau.

Figure 15. Consult or {consult and create} procedure of CS-SIS ABIS-Face in a police context for a **known** suspect Source: EC 2018.



Regarding this use case, it could be argued that, if the subject of the consultation is present at the police station, his ten-prints are also available. As of today, AFIS 10-print technology is significantly more accurate than ABIS-Face. Therefore, if no match is found using fingerprints, it is highly unlikely that ABIS-Face will find a positive match.

However, CS-SIS DB contains around 1 million alerts related to persons, and a part of them contain facial images and not fingerprints. This is a situation that will very likely continue in the future as it is allowed in CS-SIS to create new alerts containing only face images and not fingerprints. As such, a search using 10-prints only considers a limited part of the CS-SIS DB. In order to consider a wider part of the DB it is also necessary to perform a search based on face.

9.1.2. ABIS-Face consultation in police context: Unknown subject

In this use-case the subject of the consultation is not present at the police station and his/her identity is unknown. This would be the usual case of a crime that has been committed and there is video surveillance footage of it, but the perpetrators have not yet been identified. An illustrative specific example would be the Boston Marathon Bombings, where video surveillance footage of the criminals was available and was used to apprehend them. This scenario would be analogous to the case of fingermarks considered in Article 40 of the new 2018 Police Legislation.

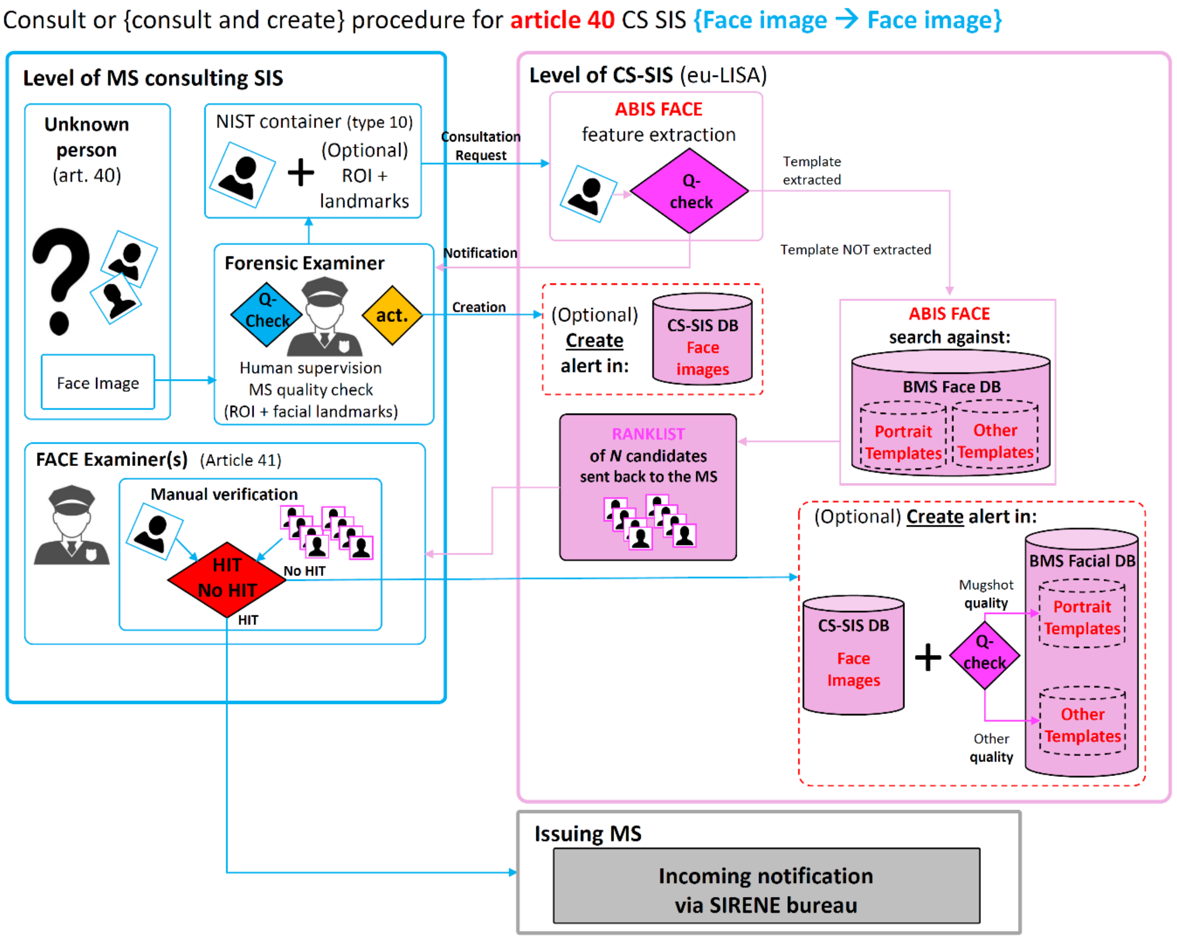
The most significant difference between this scenario and the one with a known subject is (see previous Section 9.1.1):

- The police officer has no control upon the quality of the face image used for consultation. This image will be in general of quite low quality since, in the typical case, it will come from surveillance cameras.

Given this difference, the typical workflow for this use-case is depicted in **Figure 16**, and may be described as follows:

- <u>STEP 1.</u> The police officer receives images with an unknown person. Depending on the quality of the image, some human interventions may be required from a forensic examiner such as: delimitation of the face (i.e., delimitation of the Region of Interest, ROI) or mark-up of specific landmarks such as eyes or sides of the mouth. The need for human intervention can be especially necessary in the case for instance of low-quality face images coming from surveillance cameras. In this type of images some pre-processing by a human examiner can help to increase the accuracy of the system.
  Note that, according to best practices in forensics [136], not every type of processing on the images should be allowed, as certain manipulations could be considered tampering of the forensic evidence. In general, only linear transformations (rotation, translation) and landmark annotation should be performed.
- <u>STEP 2.</u> Both the image and any additional information provided by the forensic examiner are sent in a NIST Container type 10 to the CS-SIS, where two initial tests are performed: 1) compliance of the NIST container with the specification of the CS-SIS; 2) the image is of sufficient quality to extract from it a searchable template.
- <u>STEP 3.</u> In case the image is not of sufficient quality to use it for automatic search, the officer is notified and he/she has the possibility to create a new alert in SIS associated to an unknown person. Only the face image will be stored in the CS-SIS Face DB, but no template will be stored in the BMS Face DB (since the system is not able to extract it). This way, the alert is created but it will not be able to consult it in the future based on the ABIS-Face search engine only. Therefore, the face image is stored in the system in order to be used for manual verification of a hit obtained through some other means (e.g., alphanumeric data, fingermarks).
- <u>STEP 4.</u> In case a template can be extracted from the image, it is used to search the complete BMS Face DB (both portraits). A rank list of N candidates is produced from this search and sent to the MS performing the consultation.
- <u>STEP 5.</u> At the MS a manual verification of the ranklist is performed. If the result of the verification is a no-hit (i.e., there is no alert in CS-SIS associated to the individual), the officer would have the possibility to create a new alert, storing the face image(s) in the CS-SIS Face DB (images) and the templates in the BMS Face DB, flagging the template either as "portrait" or "other" depending on the quality.
- <u>STEP 6.</u> In case of hit after the manual verification (i.e., there is already an alert in CS-SIS associated to the subjet), the MS performing the consultation will take action according to the alert and the MS owner of the alert will be informed through the SIRENE Bureau.

Figure 16. Consult or {consult and create} procedure of CS-SIS ABIS-Face in a police context for an **unknown** suspect. Source: EC 2018.



9.2. Face recognition technology beyond the 2D visible spectrum

As mentioned in Sections 1 and 3, automatic FR under unconstrained conditions is still a difficult task due to the wide range of variations in human faces caused for instance by illumination, pose, facial expressions or occlusions. Hence, in the unconstrained scenario, it is important to find methods that result in highly accurate feature extraction with high robustness to variations.

Traditionally, most research in the field of automatic Face Recognition has been focused on the visible domain (VIS) as this is the most common spectrum for the vast majority of face data available nowadays (see Sections 1, 2 and 3 for a review of the state of the art of this field). One of the main difficulties encountered by FR systems working on visual data is the variability on illumination conditions. In order to tackle this issue two kinds of methods have been proposed:

- Passive methods work with images captured in the visual spectrum. These methods focus on developing specific algorithms robust to illumination changes combined with illumination-invariant sets of features. One drawback of this type of approaches, however, is the loss of useful information about facial images during the illumination compensation which results in a certain loss of accuracy.

- Active methods use active imaging techniques to overcome illumination variation. That is, the sensors used to capture the images are different from those used in the visual spectrum. These methods are used to obtain facial images of illumination-invariant modalities or to acquire facial images taken in consistent illumination conditions. Active methods can be divided into: those that use 3D information and those based on infrared. Infrared methods can be divided into thermal infrared (TIR) and near infrared (NIR).

Active methods based on different sensing technologies have shown promising results under highly varying illumination conditions or under poor illumination, even outperforming the results obtained using traditional 2D visual spectrum images. However, these relatively new technologies still present a number of challenges that have prevented their wide deployment:

- In general, the acquisition equipment (sensors embedded in the capture device) used to capture this type of images is significantly more expensive than the one needed for images in the visual spectrum. This is especially true for the case of good quality 3D sensors. Therefore, deploying a FR system based on NIR or 3D images has a higher financial cost.
- Most existing databases, as is the case for SIS, contain images in the visual spectrum. This means that, if the new technology NIR/3D is introduced, it has to be made compatible with the existing data.
  That is, a new image captured in the NIR should be used to search in a database of visible domain images (i.e., NIR-VIS comparison). This is a problem known as Heterogeneous Face Recognition (HTR) as it tries to compare samples coming from different acquisition sources. For the time being, such a challenge is far from being solved and presents large accuracy decrease with respect to the cases where the comparison is performed in the same domain (i.e., NIR-NIR, VIS-VIS).
  A similar challenge is faced when comparing 3D-VIS, where the type of data is in general intrinsically different: while VIS images are a pixel-based matrix, 3D models are in general a point-cloud where each point is defined by its x, y and z coordinates.
- There are scenarios in which VIS images are the only ones available. This is the case for instance of data coming from surveillance cameras working in the visual domain, or the use of face images obtained for instance from social networks. In these very common scenarios it is not possible to select the type of images to be used. As in the back-compatibility case, it would be necessary to perform a NIR-VIS comparison process which would entail a significant decrease in accuracy.
- The use of 3D technology has the extra challenge of producing very large data samples. This results in the need for significantly more computing power than in the case of 2D images. The processing of such samples and templates is also slower, which may derive in the impossibility to perform searches in real time if the volume of consultations to a database is large, or can even pose storage issues.

In view of the previous advantages and challenges of alternative sensing technologies to the traditional visual spectrum, the recommendations of DG JRC for a potential future inclusion of this type of systems in CS-SIS are:

Recommendation 18:

Use of NIR FR technology within CS-SIS.

We recommend that the main ABIS-Face in CS-SIS should remain based primarily on facial images captured in the visual spectrum.

We recommend that images captured in NIR could complement those captured in the visual domain for the case of bad illumination conditions during consultation, or during enrollment with the aim to cope with consultation requests from NIR domain.

In case that NIR images were eventually stored in CS-SIS (and labbeled as specific NIR image), we recommend to have:

- The primary search engine ABIS-Face VIS to perform VIS-VIS comparison.
- A secondary search engine ABIS-Face NIR-VIS to perform NIR-VIS comparison in the case of a consultation using NIR image to be compared with the visible domain images stored in the SIS.
- A third search engine ABIS-Face NIR to perform NIR-NIR comparison.

Since NIR images only bring improved performance under varying or poor illumination conditions, the NIR images would not apply to the case of portrait searches where illumination is considered to be stable and controlled (i.e., portrait vs portrait consultations). This is the most typical case foreseen for the SIS system. Therefore, the final added value of NIR images to SIS would be somewhat limited at first to the situation during which a consultation is conducted under degraded conditions.

Recommendation 19:

Use of 3D FR technology within CS-SIS.

For the near future we do not recommend the inclusion of 3D technology in CS-SIS since it does not adapt well to the use-cases of this system.

3D technology, however, can be useful in unsupervised capture points (e.g., ABC gates at airports) in order to perform Presentation Attack Detection.

9.3. Multimodal biometrics

When face recognition technology will eventually integrated in SIS, the system will already operate two unimodal biometric search engines based on two different biometric characteristics: fingerprints and face. It has been shown in the literature that usually the combination of several biometric traits gives in general better results than their independent accuracy. This approach is known as multimodal or multibiometric systems [137].

Some of the limitations imposed by unimodal biometric systems can be overcome by including multiple sources of information for establishing identity. Such systems, known as multimodal biometric systems, are expected to be more reliable due to the presence of multiple, (fairly) independent pieces of evidence. These systems are able to meet the stringent performance requirements imposed by various applications. They address the problem of non-universality, since multiple traits ensure sufficient population coverage. They also deter presentation attacks since it would be difficult for an impostor to spoof multiple biometric traits of a genuine user simultaneously.

A multimodal system can operate in one of two different modes:

- Serial operation mode: the output of one modality is typically used to narrow down the number of possible identities before the next modality is used. Therefore, multiple sources of information (e.g., multiple traits) do not have to be acquired simultaneously. Further, a decision could be made before acquiring all the traits. This can reduce the overall recognition time.
- Parallel operation mode: the information from multiple modalities are used simultaneously in order to perform recognition.

The strategy adopted for integration depends on the level at which fusion is performed:

- Feature level fusion: it can be accomplished by concatenating two compatible feature sets. Feature selection/reduction techniques may be employed to handle the curse-of-dimensionality problem.
- Score level fusion: has been well studied in the literature. Robust and efficient normalization techniques are necessary to transform the scores of multiple comparison algorithms into a common domain prior to consolidating them. In the context of verification, two distinct strategies exist for fusion at this level. In the first approach the fusion is viewed as a classification problem where a feature vector is constructed using the comparison scores output by the individual algorithms; this feature vector is then classified into one of two classes: Accept (genuine user) or Reject (impostor). In the second approach the fusion is viewed as a combination problem where the individual comparison scores are combined to generate a single scalar score which is then used to make the final decision. General strategies for combining multiple classifiers have been suggested in [138]. It has been shown that combining different scores according to the quality of the samples that generated them gives very positive results compared to other combination strategies.

9.4. Iris biometrics

Iris recognition systems have proven over the years to have an accuracy comparable to that of fingerprint recognition systems and higher than that of face. The main challenge of iris recognition is the acquisition of good quality iris images which, in general, requires stringent cooperation from the data subject. However, with the improvement of imaging devices, it has been shown in some works that it is possible to perform iris recognition from frontal facial images of sufficient resolution captured under good illumination conditions.

Potentially, this could be a way of enhancing the biometric capabilities of SIS, without needing to include specific iris acquisition devices: the same face image that is used for face recognition is also the input to the iris recognition system (ABIS-Iris). Of course, this would entail putting in place an ABIS-Iris system and keeping a new BMS database of iris templates.

The advantages would be: 1) Increased accuracy in the searches that could be done at the same time using face and iris (multimodal biometrics, see Section 9.3); 2) The vast majority of iris recognition systems are based on standard templates known as iris-codes. This makes iris recognition systems highly interoperable.

The drawback at the moment is that the requirements of the face images in order to be used also for iris recognition are very restrictive, therefore, many of the images expected to be enrolled in SIS or used to query SIS would most likely not be eligible to be used in the ABIS-Iris.

9.5. Tattoo recognition

The ANSI/NIST container type 10 which is the basis for the transmission of face image data in SIS, also allows for the inclusion of tattoo images. For this reason, we include in the present report an Annex dedicated to image-based tattoo recognition (see Annex 1). The objective of the Annex is not to be exhaustive in the review of the state of the art in this area, since it constitutes in itself a full field of research which falls out of the scope of the present report. However, tattoo recognition does share some common elements with Face Recognition and, given that it is transmitted in the same container, it can be a future functionality to be considered in SIS. The aim of the Annex therefore, is to present a first general overview of the different technologies that are being developed today for the use of tattoo images contributing to the identification of individuals.

Conclusions

The present study was conducted in support of the new SIS regulatory framework published in November 2018. In article 33 of the new SIS-Border regulation and article 43 of the new SIS-Police regulation, it is defined the new use that can be given to face related data stored in SIS alerts:

- Article 33.4 Border and Article 43.4 Police

  "As soon as it becomes technically possible, and while ensuring a high degree of reliability of identification, photographs and facial images may be used to identify a person in the context of regular border crossing points.

  Before this functionality is implemented in SIS, the Commission shall present a report on the availability, readiness and reliability of the required technology. The European Parliament shall be consulted on the report.

  After the start of the use of the functionality at regular border crossing points, the Commission shall be empowered to adopt delegated acts in accordance with Article 75 to supplement this Regulation concerning the determination of other circumstances in which photographs and facial images may be used to identify persons."

As a direct consequence of this new Regulation, the objectives defined in Section ii for the study were:

OBJECTIVE 1: Determine the readiness of facial recognition technology, to be integrated in CS-SIS for the identification of a person.

OBJECTIVE 2: Provide recommendations on the best way to integrate facial recognition technology in CS-SIS based on: 1) the current state of the art of this technology; 2) the particularities and constraints of CS-SIS and its dual use for law-enforcement and border management.

Given all the information presented in the study, the conclusion reached in the study with respect to objectives 1 and 2 is that:

CONCLUSION:

Given the great boost in accuracy that face recognition technology has experimented since 2014 with the advent of deep learning-based systems, it is the conclusion of the present study that: ABIS-Face systems have reached a sufficient level of readiness and availability for its integration into CS-SIS, provided that the recommendations listed in the present report are implemented and respected, to the largest extent possible, during the rollout and utilisation of this new functionality.

References

[1] A. Jain, K. Nandakumar and A. Ross, “50 years of biometric research: Accomplishments, challenges, and opportunities,” *Pattern Recognition Letters,* vol. 79, pp. 80-105, 2016.

[2] A. Jain, A. Ross and S. Pankanti, “Biometric: a tool for information security,” *IEEE Trans. on Information Forensics and Security,* vol. 1, pp. 125-143, 2006.

[3] T. Kanade, “Picture processing system by computer complex and recognition of human faces,” 1973.

[4] L. Sirovich and M. Kirby, “Low-dimensional procedure for the characterization of human faces,” *Journal of the Optical Society of America A,* vol. 4, pp. 519-524, 1987.

[5] M. Kirby and L. Sirovich, “Application of the Karhunen-Loeve procedure for the characterization of human faces,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, pp. 103-108, 1990.

[6] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of cognitive neuroscience,* vol. 3, pp. 71-86, 1991.

[7] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 19, pp. 711-720, 1997.

[8] B. Moghaddam, W. Wahid and A. Pentland, “Beyond eigenfaces: probabilistic matching for face recognition,” in *Proc. ICFGR*, 1998.

[9] X. He, S. Yan, Y. Hu, P. Niyogi and H.-J. Zhang, “Face recognition using laplacianfaces,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, pp. 328-340, 2005.

[10] S. Yan, D. Xu, B. Zhang and H.-J. Zhang, “Graph embedding: A general framework for dimensionality reduction,” in *Proc. IEEE Computer Society*, 2005.

[11] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma, “Robust Face Recognition via Sparse Representation,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, pp. 210-227, 2009.

[12] L. Zhang, M. Yang and X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?,” in *Proc. ICCV*, 2011.

[13] W. Deng, J. Hu and J. Guo, “Extended SRC: Undersampled face recognition via intraclass variant dictionary,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 34, pp. 1864-1870, 2012.

[14] W. Deng, J. Hu and J. Guo, “Face recognition via collaborative representation: Its discriminant nature and superposed representation,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 99, pp. 1-1, 2018.

[15] C. Liu and H. Wechsler, “Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition,” *IEEE Trans. on Image Processing,* vol. 11, pp. 467-476, 2002.

[16] T. Ahonen, A. Hadid and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 2037-2041, 2006.

[17] W. Zhang, S. Shan, W. Gao, X. Chen and H. Zhang, “Local gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition,” in *Proc. ICCV*, 2005.

[18] D. Chen, X. Cao, F. Wen and J. Sun, “Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification,” in *Proc. CVPR*, 2013.

[19] Z. Cao, Q. Yin, X. Tang and J. Sun, “Face recognition with learning-based descriptor,” in *Proc. CVPR*, 2707-2714.

[20] Z. Lei, M. Pietikainen and S. Z. Li, “Learning discriminant face descriptor,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 36, pp. 289-302, 2014.

[21] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng and Y. Ma, “PCAnet: A simple deep learning baseline for image classification?,” *IEEE Trans. on Image Processing,* vol. 24, pp. 5017-5032, 2015.

[22] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proc. CVPR*, 2014.

[23] Y. Sun, Y. Chen, X. Wang and X. Tang, “Deep learning face representation by joint identification-verification,” in *Proc. NIPS*, 2014.

[24] G. B. Huang, M. Ramesh, T. Berg and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” 2007.

[25] A. Krizhevsky, I. Sutskever and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. NIPS*, 2012.

[26] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, “Face recognition: A literature survey,” *ACM Computing Surveys,* vol. 35, pp. 399-458, 2003.

[27] K. W. Bowyer, K. Chang and P. Flynn, “A survey of approaches and challenges in 3{D} and multi-modal 3{D}+2{D} face recognition,” *Computer vision and image understanding,* vol. 101, pp. 1-15, 2006.

[28] A. F. Abate, M. Nappi, D. Riccio and G. Sabatino, “2D and 3D face recognition: A survey,” *Pattern recognition letters,* vol. 28, pp. 1885-1906, 2007.

[29] R. Jafri and H. R. Arabnia, “A survey of face recognition techniques,” *Journal of Information Processing Systems,* vol. 5, pp. 41-68, 2009.

[30] X. Zou, J. Kittler and K. Messer, “Illumination invariant face recognition: A survey,” in *Proc. BTAS*, 2007.

[31] A. Scheenstra, A. Ruifrok and R. C. Veltkamp, “A survey of 3D face recognition methods,” in *Proc. ICAVBPA*, 2005.

[32] X. Zhang and Y. Gao, “Face recognition across pose: A review,” *Pattern Recognition,* vol. 42, pp. 2876-2896, 2009.

[33] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen and S. S. Iyengar, “A Survey on Deep Learning: Algorithms, Techniques, and Applications,” *ACM Computing Surveys,* vol. 51, 2019.

[34] S. Skansi, Introduction to Deep Learning, Springer, 2018.

[35] M. Wang and W. Deng, “Deep face recognition: a survey,” *arXiv:1804.06655v5,* 2018.

[36] B. Moghaddam, T. Jebara and A. Pentland, “Bayesian face recognition,” *Pattern Recognition,* vol. 33, pp. 1771-1782, 2000.

[37] A. Xu, X. Jin, Y. Jiang and P. Guo, “Complete Two-Dimensional PCA for Face Recognition,” in *Proc. ICPR*, 2006.

[38] M. Guillaumin, J. Verbeek and C. Schmid, “Is that you? Metric learning approaches for face identification,” in *Proc. ICCV*, 2009.

[39] L. Wiskott, J.-M. Fellous, N. Kruger and C. von der Malsburg, “Face recognition by elastic bunch graph matching,” in *Proc. ICIP*, 1997.

[40] K. Simonyan, O. M. Parkhi, A. Vedaldi and A. Zisserman, “Fisher Vector Faces in the Wild,” 2013.

[41] O. M. Parkhi, A. Vedaldi and A. Zisserman, “Deep Face Recognition,” 2015.

[42] F. Schroff, D. Kalenichenko and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proc. CVPR*, 2015.

[43] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. C. Chen, V. M. Patel, C. D. Castillo and R. Chellappa, “Deep learning for understanding faces: Machines may be just as good, or better, than humans,” *IEEE Signal Processing Letters,* vol. 35, pp. 66-83, 2018.

[44] M. M. Ghazi and H. K. Ekenel, “A comprehensive analysis of deep learning based representation for face recognition,” in *Proc. CVPR Workshops*, 2016.

[45] A. T. Tran, T. Hassner, I. Masi and G. Medioni, “Regressing robust and discriminative 3D morphable models with a very deep neural network.,” in *Proc. Intl. Conf on Computer Vision*, 2017.

[46] I. Masi, T. Hassner, A. T. Tran and G. Medioni, “Rapid synthesis of massive face sets for improved face recognition,” in *Proc. FG*, 2017.

[47] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan and J. Feng, “Dual-agent {GANS} for photorealistic and identity preserving profile face synthesis,” in *Proc. NIPS*, 2017.

[48] Z. Zhu, P. Luo, X. Wang and X. Tang, “Multi-view perceptron: a deep model for learning face identity and view representations,” in *Proc. NIPS*, 2014.

[49] C. Ding and D. Tao, “Robust face recognition via multimodal deep face representation,” *IEEE Trans. on Multimedia,* vol. 17, pp. 2049-2058, 2015.

[50] Y. Sun, X. Wang and X. Tang, “Sparsifying neural network connections for face recognition,” in *Proc. CVPR*, 4856-4864.

[51] M. Kan, S. Shan, H. Chang and X. Chen, “Stacked progressive autoencoders (SPAE) for face recognition across poses,” in *Proc. CVPR*, 2014.

[52] Y. Zhang, M. Shao, E. K. Wong and Y. Fu, “Random faces guided sparse many-to-one encoder for pose-invariant face recognition,” in *Proc. ICCV*, 2013.

[53] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park and J. Kim, “Rotating your face using multi-task deep neural network,” in *Proc. CVPR*, 2015.

[54] L. Tran, X. Yin and X. Liu, “Disentangled representation learning GAN for pose-invariant face recognition,” in *Proc. CVPR*, 2017.

[55] R. Huang, S. Zhang, T. Li, R. He and others, “Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis,” *arXiv preprint arXiv:1704.04086,* 2017.

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, Z. H. S. Ma, A. Karpathy, A. Khosla, M. Bernstein and others, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision,* vol. 115, pp. 211-252, 2015.

[57] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556,* 2014.

[58] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich and others, “Going deeper with convolutions,” in *Proc. CVPR*, 2015.

[59] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016.

[60] J. Hu, L. Shen and G. Sun, “Squeeze-and-excitation networks,” *arXiv preprint arXiv:1709.01507,* 2017.

[61] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li and T. Hospedales, “When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition,” in *Proc. ICCV Workshops*, 2015.

[62] S. Sankaranarayanan, A. Alavi, C. D. Castillo and R. Chellappa, “Triplet probabilistic embedding for face verification and clustering,” in *Proc. BTAS*, 2016.

[63] I. Masi, A. T. Trin, T. Hassner, J. T. Leksut and G. Medioni, “Do we really need to collect millions of faces for effective face recognition?,” in *Proc. ECCV*, 2016.

[64] J. Yang, P. Ren, D. Chen, F. Wen, H. Li and G. Hua, “Neural aggregation network for video face recognition,” *arXiv preprint arXiv:1603.05474,* 2016.

[65] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proc. CVPR*, 2017.

[66] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, “VGGface2: A dataset for recognising faces across pose and age,” *arXiv preprint arXiv:1710.08092,* 2017.

[67] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan and T.-K. Kim, “Conditional convolutional neural network for modality-aware face recognition,” in *Proc. ICCV*, 2015.

[68] X. Wu, R. He and Z. Sun, “A lightened CNN for deep face representation,” in *Proc. CVPR*, 2015.

[69] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel and R. Chellappa, “An end-to-end system for unconstrained face verification with deep convolutional neural networks,” in *Proc. ICCV Workshops*, 2015.

[70] Y. Zhong, J. Chen and B. Huang, “Toward end-to-end face recognition through alignment learning,” *IEEE Signal Processing Letters,* vol. 24, pp. 1213-1217, 2017.

[71] M. Hayat, S. H. Khan, N. Werghi and R. Goecke, “Joint registration and representation learning for unconstrained face identification,” in *Proc. CVPR*, 2017.

[72] M. Kan, S. Shan and X. Chen, “Multi-view deep network for cross-view classification,” in *Proc. CVPR*, 2016.

[73] I. Masi, S. Rawls, G. Medioni and P. Natarajan, “Pose-aware face recognition in the wild,” in *Proc. CVPR*, 2016.

[74] Y. Sun, X. Wang and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in *Proc. CVPR*, 2014.

[75] Y. Sun, X. Wang and X. Tang, “Hybrid deep learning for face verification,” in *Proc. ICCV*, 2013.

[76] R. Ranjan, S. Sankaranarayanan, C. D. Castillo and R. Chellappa, “An all-in-one convolutional neural network for face analysis,” in *Proc. FG*, 2017.

[77] Y. Wen, K. Zhang, Z. Li and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Proc. ECCV*, 2016.

[78] J.-C. Chen, V. M. Patel and R. Chellappa, “Unconstrained face verification using deep cnn features,” in *Proc. WACV*, 2016.

[79] Y. Sun, D. Liang, X. Wang and X. Tang, “DeepID3: Face recognition with very deep neural networks,” *arXiv preprint arXiv:1502.00873,* 2015.

[80] J. Deng, J. Guo and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” *arXiv preprint arXiv:1801.07698,* 2018.

[81] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao and L. Song, "Deep hyperspherical learning," in *Proc. NIPS*, 2017.

[82] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," *arXiv preprint arXiv:1801.09414,* 2018.

[83] W. Liu, Y. Wen, Z. Yu and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. ICML*, 2016.

[84] B. Chen, W. Deng and J. D. Noisy, "Softmax: improving the generalization ability of DCNN via postponing the early softmax saturation," *arXiv preprint arXiv:1708.03769,* 2017.

[85] A. Hasnat, J. Bohne, J. Milgram, S. Gentric and L. Chen, "Deepvisage: Making face recognition simple yet with powerful generalization skills," *arXiv preprint arXiv:1703.08388,* 2017.

[86] R. Ranjan, C. D. Castillo and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507,* 2017.

[87] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge and P. J. Grother, "Distinguishing identical twins by face recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, 2011.

[88] S. Marcel, M. Nixon, J. Fierrez and N. Evans, Eds., Handbook of biometric anti-poofing: Presentation Attack Detection, Springer, 2019.

[89] J. Galbally, S. Marcel and J. Fierrez, "Biometric Anti-spoofing Methods: A Survey in Face Recognition," *IEEE Access,* vol. 2, pp. 1530-1552, 2014.

[90] R. Raghavendra and C. Busch, "Presentation Attack Detection methods for Face Recognition System - A Comprehensive Survey," *ACM Computing Surveys,* vol. 50, 2017.

[91] M. Ferrara, A. Franco and D. Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*, 2014.

[92] A. J. Mansfield and J. L. Wayman, "Best practices in testing and reporting performance of biometric devices," 2002.

[93] D. Yi, Z. Lei, S. Liao and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923,* 2014.

[94] Y. Guo, L. Zhang, Y. Hu, X. He and J. Gao, "MS-celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. ECCV*, 2016.

[95] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proc. CVPR*, 2016.

[96] A. Bansal, C. Castillo, R. Ranjan and R. Chellappa., "The dos and donts for cnn-based face verification," *arXiv preprint arXiv:1705.07426,* 2017.

[97] Microsoft, *MS-celeb-1M challenge 3,* http://trillionpairs.deepglint.com, 2018.

[98] K. Ricanek and T. Tesafaye, “MORPH: A longitudinal image database of normal adult age-progression,” in *Proc. FGR*, 2006.

[99] L. Lin, G. Wang, W. Zuo, X. Feng and L. Zhang, “Cross-domain visual matching via generalized similarity measure and feature learning,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 39, pp. 1089-1102, 2016.

[100] G. Panis, A. Lanitis, N. Tsapatsoulis and T. F. Cootes, “Overview of research on facial ageing using the FG-NET ageing database,” *IET Biometrics,* vol. 5, pp. 37-46, 2016.

[101] Y. Wen, Z. Li and Y. Qiao, “Latent factor guided convolutional neural networks for age-invariant face recognition,” in *Proc. CVPR*, 2016.

[102] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah and A. K. Jain, “Pushing the frontiers of unconstrained face detection and recognition: {I}arpa {J}anus benchmark,” in *Proc. CVPR*, 2015.

[103] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa and D. W. Jacobs, “Frontal to profile face verification in the wild,” in *Proc. WACV*, 2016.

[104] X. Peng, X. Yu, K. Sohn, D. N. Metaxas and M. Chandraker, “Reconstruction-based disentanglement for pose-invariant face recognition,” *Intervals,* vol. 20, p. 12, 2017.

[105] X. Yin and X. Liu, “Multi-task convolutional neural network for poseinvariant face recognition,” in *Proc. TIP*, 2017.

[106] C. Whitelam, K. Allen, J. Cheney, P. Grother, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller and N. Kalka, “Iarpa Janus benchmark-B face dataset,” in *Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[107] T. Zheng and W. Deng, “Cross-pose LFW: A database for studying crosspose face recognition in unconstrained environments,” 2018.

[108] P. Grother, M. Ngan, K. Hanaoka, C. Boehnen and L. Ericson, “NISTIR 8197 - The 2017 IARPA Face Recognition Prize Challenge (FRPC),” 2017.

[109] P. Grother, M. Ngan and K. Hanaoka, “Ongoing Face Recognition Vendor Test (FRVT) - Part 1: Verification,” 2018.

[110] P. Grother, M. Ngan and K. Hanaoka, “NISTIR 8238 - Ongoing Face Recognition Vendor Test (FRVT) - Part 2: Identification,” 2018.

[111] E. Learned-Miller, “Data driven image models through continuous joint alignment,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 236-250, 2005.

[112] G. B. Huang, M. A. Mattar, H. Lee and E. Learned-Miller, “Learning to align from scratch,” in *Proc. of NIPS*, 2012.

[113] F. Alonso-Fernandez, J. Fierrez and J. Ortega-Garcia, “Quality Measures in Biometric Systems,” *IEEE Security & Privacy,* vol. 10, pp. 52-62, 2012.

[114] P. Grother and E. Tabassi, “Performance of Biometric Quality Measures,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, pp. 531-543, 2007.

[115] E. Tabassi, C. Wilson and C. Watson, “Fingerprint Image Quality, NFIQ - NISTIR 7151,” 2004.

[116] G. Langenburg, C. Champod and T. Genessay, “Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools,” *Forensic Science International,* vol. 219, pp. 183-198, 2012.

[117] M. A. Olsen, “Fingerprint image quality,” 2015.

[118] J. Galbally, R. Haraksim and L. Beslay, “A study of age and ageing in fingerprint biometrics,” *IEEE Trans. on Information Forensics and Security,* 2018.

[119] D. Michalski, S. Y. Yiu and C. Malec, “The Impact of Age and Threshold Variation on Facial Recognition Algorithm Performance Using Images of Children,” in *Proc. Intl. Conf. on Biometrics (ICB)*, 2018.

[120] P. Grother, M. Ngan and K. Hanaoka, “Ongoing Face Recognition Vendor Test (FRVT),” 2018.

[121] J. Galbally and R. Satta, “Biometric Sensor Interoperability: A Case Study in 3D Face Recognition,” in *Proc. Int. Conf. on Pattern Recognition Applications and Methods (ICPRAM)*, 2016.

[122] P. G. C. W. e. a. Elham Tabassi, “NFIQ2 NIST Fingerprint Image Quality,” 2016.

[123] X. Gao, S. Z. Li, R. Liu and P. Zhang, “Standardization of Face Image Sample Quality,” in *Proc. ICB*, 2007.

[124] J. Sang, Z. Lei and S. Z. Li, “Face Image Quality Evaluation for ISO/IEC Standards 19794-5 and 29794-5,” in *Proc. ICB*, 2009.

[125] P. J. Phillips, J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, S. Cheng, M. N. Teli and H. Zhang, “On the existence of face quality measures,” in *Proc. BTAS*, 2013.

[126] A. Abaza, M. A. Harrison and T. Bourlai, “Quality metrics for practical face recognition,” in *Proc. ICPR*, 2012.

[127] M. Ferrara, A. Franco, D. Maio and D. Maltoni, “Face Image Conformance to ISO/ICAO Standards in Machine Readable Travel Documents,” *IEEE Trans. on Information Forensics and Security,* vol. 7, pp. 1204-1213, 2012.

[128] R. Raghavendra, K. B. Raja, B. Yang and C. Busch, “Automatic face quality assessment from video using gray level co-occurrence matrix: An empirical study on Automatic Border Control system,” in *Proc. ICPR*, 2014.

[129] Y. Wong, S. Chen, S. Mau, C. Sanderson and B. C. Lovell, “Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition,” in *Proc. CVPRW*, 2011.

[130] A. Dutta, R. Veldhuis and L. Spreeuwers, “Predicting face recognition performance using image quality,” *arXiv preprint arXiv:1510.07119,* 2015.

[131] L. Liu, Y. Hua, Q. Zhao, H. Huang and A. C. Bovik, “Blind image quality assessment by relative gradient statistics and adaboosting neural network,” *Signal Processing: Image Communication,* vol. 40, p. 2016.

[132] P. J. Phillips, A. N. Yates, J. R. Beveridge and G. H. Givens, “Predicting face recognition performance in unconstrained environments,” in *Proc. CVPRW*, 2017.

[133] Y. Liu, J. Yan and W. Ouyang, “Quality aware network for set to set recognition,” in *Proc. CVPR*, 2017.

[134] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu and Y. Zhu, “Deepsim: Deep similarity for image quality assessment,” *Neurocomputing,* vol. 257, pp. 104-114, 2017.

[135] L. Best-Rowden and A. K. Jain, “Learning face image quality from human assessments,” *IEEE Trans. on Information Forensics and Security,* vol. 13, pp. 3064-3077, 2018.

[136] ENFSI, “Best practice manual for facial image comparison,” 2018.

[137] A. Ross, K. Nandakumar and A. Jain, Eds., Handbook of multibiometrics, Springer, 2006.

[138] J. Kittler, M. Hatef, R. Duin and J. Matas, “On combining classifiers,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 20, pp. 226-239, 1998.

[139] D. Lowe, “Distinctive Image Features from Scale Invariant Keypoints,” *Intl. Journal on Computer Vision,* vol. 60, pp. 91-110, 1999.

[140] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in *Proc. ICCV*, 2003.

[141] K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman, “The devil is in the details: an evaluation of recent feature encoding methods,” in *Proc. British Machine Vision Conference*, 2011.

[142] D. Manger, “Large-scale tattoo image retrieval,” in *Proc. ICCRV*, 2012.

[143] INTERPOL, *FASTID project: FAST and efficient international disaster victim IDentification,* 2018.

[144] A. Jain, J. Lee and N. Gregg, “Content-based image retrieval: An application to tattoo images,” in *Proc. ICIP*, 2009.

[145] J. Lee, R. Jin, A. Jain and W. Tong, “Image Retrieval in Forensics: Tattoo Image Database Application,” *IEEE Multimedia,* vol. 19, pp. 2-11, 2012.

[146] C. Szegedy, S. Ioffe and V. Vanhoucke, “Inception-v4: Inception-ResNet and the Impact of Residual Connections on Learning,” *arXiv:abs/1602.07261,* 2016.

[147] A. Karpathy and L. Fei-Fei, “Deep Visual-Semantic Alignments for Generating Image Descriptions,” *arXiv:1412.2306v2,* 2014.

[148] J. Johnson, A. Karpathy and L. Fei-Fei, “DenseCap: Fully Convolutional Localization Networks for Dense Captioning,” in *Proc. ICVPR*, 2016.

[149] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Jia-Li, D. Shamma, M. Bernstein and L. Fei-Fei, “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations,” *arxiv.org/1602.07332,* 2016.

[150] X. Di and V. Patel, “Deep Tattoo Recognition,” in *Proc. ICVPR*, 2016.

[151] H. Han, J. Li, A. Jain, S. Shan and X. Chen, “Tattoo Image Search at Scale: Joint Detection and Compact Representation Learning,” *arxiv.org/1811.00218,* 2018.

[152] M. Ngan and P. Grother, “Tattoo Recognition Technology - Challenge (Tatt-C): An Open Tattoo Database for Developing Tattoo Recognition Research,” in *Proc. ICISBA*, 2015.

[153] M. Ngan, G. Quinn and P. Grother, “Tattoo Recognition Technology - Challenge (Tatt-C) - Outcomes and Recommendations,” 2016.

[154] M. Ngan, P. Grother and K. Hanaoka, “Tattoo Recognition Technology -Evaluation (Tatt-E) Performance of Tattoo Identification Algorithms,” 2018.

[155] T. Hrkac, K. Brkic, S. Ribaric and D. Marcetic, “Deep Learning Architectures for Tattoo Detection and De-identification,” in *Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 2016.

[156] Z. Sun, J. Baumes, P. Tunison, M. Turek and A. Hoogs, “Tattoo Detection and Localization using Region-based Deep Learning,” in *Intl. Conf. on Pattern Recognition (ICPR)*, 2016.

List of abbreviations and definitions

ABIS	Automatic Biometric Identification System
AFIS	Automatic Fingerprint Identification System
AVC	Area Under the Curve
CCTV	Closed Circuit TV
CMC	Cumulative Match Characteristic
CNN	Convolutional Neural Network
CS-SIS	Central System SIS
DET	Detecion-Error Tradeoff
DG	Directorate General
DHS	Department of Homeland Security
EC	European Commission
ECRIS	European Criminal Records Information System
EES	Entry Exit System
EU	European Union
FAR	False Acceptance Rate
FBI	Federal Bureau of Investigation
FPIR	False Positive Identification Rate
FRVT	Face Recognition Vendor Test
FR	Face Recognition
GAN	Graphical Adversarial Networks
GPU	Graphics Processing Unit
IARPA	Intelligence Advanced Research Project Activity
ICAO	International Civial Aviation Organization
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
JRC	Joint Research Centre
LFW	Labelled Faces in the Wild
MRTD	Machine Readable Document
MS	Member State
NIST	National Institute for Standards and Technology
PAD	Presentation Attack Detection
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic

SIS	Schengen Information System
SRC	Sparse Representation-Based Classifier
SVM	Support Vector Machine
TAR	True Acceptance Rate
TPIR	True Positive Identification Rate
TRL	Technology Readiness Levels
VIS	VISA Information System
VSS	Video Surveillance System

List of figures

List of tables

Annexe(s)

Annex 1. Tattoo recognition for human identification

Using tattoos as soft biometrics is an additional key element which serves as complementary information for criminal identification. Content Based Tattoo Recognition and Identification System (CBTRIS) could be used by law-enforcement agencies, when included in the EU large-scale IT Schengen Information System (SIS).

Given an image of a tattoo, the objective is then to find one or several instances of the same tattoo from the same subject from a database. This use case has application in investigation supporting identification of an individual, for example, in the case of a criminal activity where the suspect is wearing a mask and gloves and the surveillance camera may be able to record a tattoo exposed on the neck or the arm from the suspect. The test data for this use case is composed of images of the same tattoo from the same subject collected during different encounters. For each probe image, there could be one or more correctly comparing tattoo image(s) in the database.

Such a system could be ideally use a tattoo query image, in a robust and practical manner, to identify the person carrying the tattoo. The tattoo in question will be compared against the SIS Database, which in addition to facial image will contain records with tattoo data, if exist, as well. The ultimate goal is to find the best match between a query tattoo and database tattoos, extracting a ranked estimate (usually the best 1 to 10 matches).

The classical approach

The standard technique already implemented in all law enforcement Authorities is to use keywords to describe tattoo classes such as those defined by US National Institute of Standards and Technology (NIST)[18]. A tattoo could be described by 3-4 words (at maximum) and this narrows down the search into the full tattoo database. The tattoo search and comparison are performed through comparison of the keywords produced from the tattoo at hand with those belonging to the database. The procedure for tattoo keyword includes a human intervention in the loop needed to assign by hand classes similar to those of NIST and therefore is a tedious, subjective and error-prone procedure.

In order to automate the process of tattoo comparison without human intervention, the classical approach is to extract robust features from the tattoo in question and compared them with similar features stored in a database. Most of the published papers make use of hand-crafted and carefully prepared features (key points) such as the robust SIFT descriptors [139], which consist of a vector of 128 dimensions. Additional information is the orientation and the scale for each key point.

However, the application of SIFT in large-scale CBTR systems with thousands or millions of tattoos, a pair-wise direct comparison of the query image key points with the key points for each database image, is impractical in terms of memory requirement and processing times.

Therefore, other techniques have been applied such as coding the SIFT key points allowing only quantised values for the descriptors. This leads to the adoption of the Bag-of-Words notion [140], [141] where each descriptor falls into one visual word from the total codebook (dictionary). Then each image is transformed into a set of words and querying

[18] ANSI/NIST-ITL 1-2011 https://www.nist.gov/programs-projects/ansinist-itl-standard

an image is equivalent to search for images in the database sharing the maximum number of common words with the query image.

Manger [142] presents a state-of-the-art Content Based Image Retrieval (CBIR) system which uses Bag-of-Words (BoW) based on SIFT features and enhancing the results using Hamming Embedding (HE) and Weak Geometric Consistency (WGC). The accuracy of identification is about 85 % and using a re-ranking scheme a further boost of 1st rank accuracy reaching 90.7%. The code has been developed within the framework of FASTID Project [143].

Jain [144] and Lee [145] presented a Tattoo-ID CBIR system based on SIFT features and reported 82% maximum rank-1 retrieval accuracy.

It should be noted that hand-crafted feature-based approaches has the benefit of being a mature technology with fast implementation, requires no training and a small testing dataset but accuracy is not as high as with that could be achieved with Deep Learning methods.

The Deep Learning approach

During the last decade, computing power in the form of parallel GPU processing and novel methods for computer vision, combined with novel machine learning techniques (deep learning) made feasible the automation of the image recognition, which could be applied to tattoo identification as well.

Deep Learning Platforms

Currently the most popular Deep Learning platforms used for computer vision are: TensorFlow[19] (Google), Keras[20] (Francois Chollet), Theano[21] (University of Montreal), Caffe[22] (University of Berkeley) and Torch[23] (Ronan, Clément, Koray and Soumith).

As a first step towards automating the tattoo identification procedure is the use of Deep Learning technology for assigning the relevant NIST classes to each tattoo in the database for each new tattoo input and to compare against the existing ones comparing their keywords (classes). Given a specific tattoo, the aim is to create automatically a set of keywords that verbally describe the tattoo content. These natural language labels could be used for searching similar tattoos in existing text-based descriptions for tattoo databases.

A second step and the most decisive one would be using tattoo search and comparison based solely on the image content, without any keyword comparisons. This is feasible by using Deep Learning methods for image comparison, however a large number of tattoos are needed for proper machine learning training.

Tattoo identification using keyword description

The current State-of-the-Art is the Google Inception v4 and Inception ResNet [146] using Very Deep Convolutional Neural Networks (VDCNN) for automatic classification and automatic image annotation and the end result is a labelled image.

[19] https://www.tensorflow.org

[20] https://github.com/keras-team/keras.git

[21] http://deeplearning.net/software/theano/

[22] https://github.com/BVLC/caffe

[23] https://github.com/torch/torch7

Karpathy [147] presented a model that generates natural language descriptions of images and their regions. Their model is based on a novel combination of Convolutional Neural Networks (CNN) over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding.

Johnson [148] reported a convolutional localisation network for dense captioning based on the Visual Genome dataset [149]. The system involves a convolutional neural network, a novel dense localisation layer and a recurrent neural network language model that generates the label sequence.

Tattoo identification using deep learning (no keywords)

Di and Patel [150] presented a novel deep CNN method for automatic comparison of tattoo images based on the AlexNet and Siamese networks. A triplet loss function utilized significantly improve the performance of a tattoo comparison system. Experiments over the Tatt-C dataset (see below) performed significantly better than many competitive tattoo recognition algorithms.

Han et. al. [151] propose an efficient tattoo search methodology based on tattoo compact representation in a single CNN via multi-task learning. A method of random image stitch and preceding feature buffering was used successfully, overcoming the problem of small tattoo dataset availability.

The US National Institute of Standards and Technology (NIST) has launched the *Tattoo Recognition Technology Program*[24] which features a family of activities designed with goals to evaluate and measure image-based tattoo recognition technology.

The Tattoo Recognition Technology – Challenge (Tatt-C)[25] [152], [153] is being conducted to challenge the commercial and academic communities in advancing research and development into automated image-based tattoo comparison technology. The goal of Tatt-C was to advance research and development into automated image-based tattoo recognition technology.

The challenge focused on tattoo comparison and retrieval from still images captured operationally by law enforcement agencies. Tatt-C activity culminated with a workshop and a final report with outcomes and recommendations. The Tatt-C dataset is available on an ongoing basis for interested researchers.

Tatt-BP[26] provides best practice recommendations and guidelines for the proper collection of tattoo images to support image-based tattoo recognition. Based on the outcomes of the Tatt-C, tattoo recognition algorithm accuracy is often influenced by the consistency and quality of the tattoo images.

The Tattoo Recognition Technology – Evaluation[27] (Tatt-E) [154] is being initiated by NIST to assess and measure the capability of systems to perform automated image-based tattoo recognition.

[24] https://www.nist.gov/programs-projects/tattoo-recognition-technology

[25] https://www.nist.gov/programs-projects/tattoo-recognition-technology-challenge-tatt-c

[26] http://www.nist.gov/itl/iad/ig/tatt-bp.cfm

[27] https://www.nist.gov/programs-projects/tattoo-recognition-technology-evaluation-tatt-e

Possible SIS Tattoo Identification System Application

Based upon the current state-of-the-art, a roadmap is proposed in this section by the JRC, in order for automatic tattoo identification process to be introduced into the SIS functionality, implemented by the competent authorities. JRC proposes a methodology and steps towards for the automation of tattoo-based criminal identification, in line with NIST best practices and guidelines.

JRC proposes to follow a three-steps procedure:

- Tattoo detection and localization (cropping)
- Tattoo description identification (keywords)
- Tattoo deep learning image identification (no keywords)

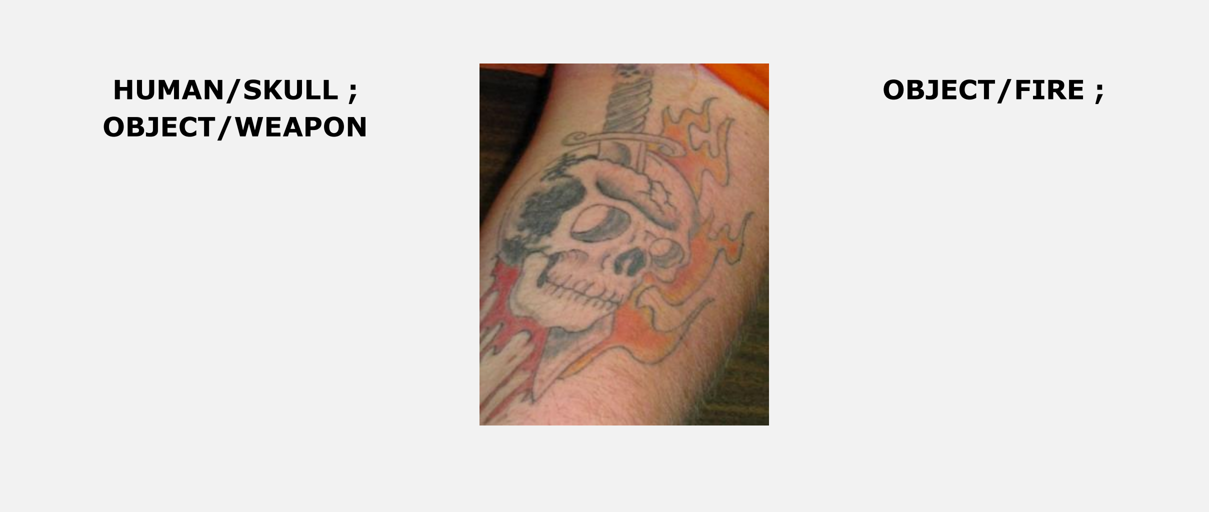
Tattoo Detection and Localisation

Using a pre-trained RCNN [155] and the TATT-C Database [consisting of 2349 total images (1349 tattoo and 1000 non-tattoo) together with ground truth provided for the locations of the tattoos (if any), a maximum performance of 97 % is obtained, which is very close with the one reported by Sun [156] (98%) using almost the same datasets. The localisation efficiency is proportional to the precision of the detected bounding boxes (cropping) with the ground truth ones. As a metric for the localisation efficiency, the percentage overlap between the detected and the ground truth boxes is used.

Tattoo Keyword Description Identification

An automatic tattoo description scheme is going to be studied, where a Deep Learning model [147] generates a set of natural language words describing the tattoo or the various discrete parts of it (e.g. a snake and a cross), an example is shown in **Figure 17**.

Figure 17. Tattoo keyword description example. Image Source: (Ngan, Quinn, & Grother, Tattoo Recognition Technology - Challenge (Tatt-C) - Outcomes and Recommendations, 2016)



A preliminary test is done using pre-trained CaffeNet and GoogleNet models and re-training for 5 NIST Classes of tattoos namely ‘WORDING’, ‘DRAGON’, ‘ROSE’, ‘SKULL’ and ‘BIRD’. The Tatt-C database and a limited set of Flickr tattoo images were used for training, validation and testing.

In order to evaluate the performance of tattoo identification, the Cumulative Matching Curve (CMC) will be employed, depicting how the identification precision varies with the number of ranked results. Precision-Recall will also be used as a metric.

The results are quite promising, providing tattoo description accuracies between 65% and 70%, and there is still plenty of room for further enhancements.

Tattoo Deep Learning Image Identification

A Siamese AlexNet similar to the one proposed by Di [150] will be used. Two identical AlexNet-based CNNs will be trained for learning visual similarity. In **Figure 18** the flowchart of the overall tattoo identification process is shown and it also includes the case of tattoo registration, which is entering the tattoo into a relevant database. In **Figure 19** a typical case for tattoo ranked identification is shown, indicating the correct and wrong matches, up to rank 5. The Cumulative Matching Curve (CMC) will be used as well, depicting how the identification precision varies with the number of ranked results.

Figure 18. Tattoo identification/registration flowchart using deep learning methods. Source: EC 2018

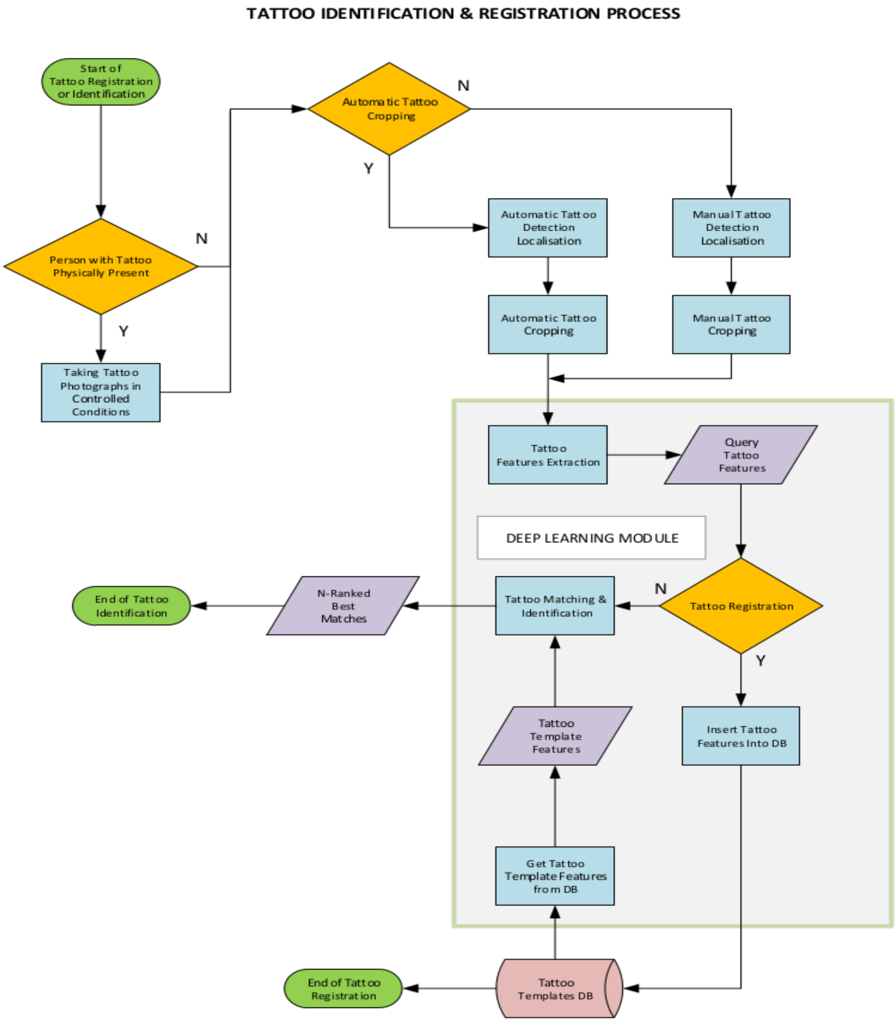
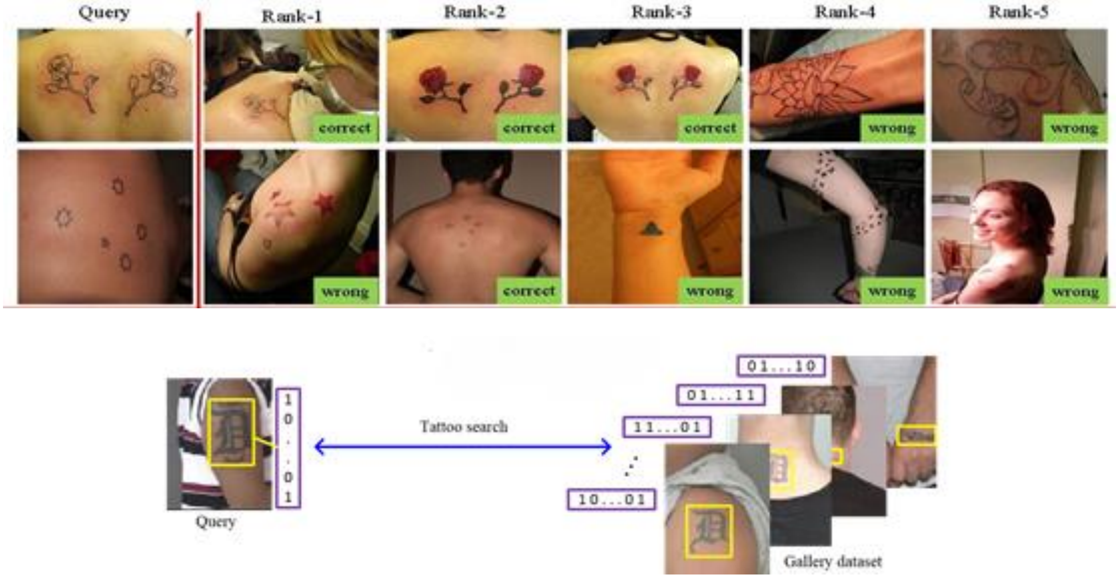

TATTOO IDENTIFICATION & REGISTRATION PROCESS
Start of Tattoo Registration or Identification
Person with Tattoo Physically Present
N
Y
Taking Tattoo Photographs in Controlled Conditions
Automatic Tattoo Cropping
N
Y
Automatic Tattoo Detection Localisation
Manual Tattoo Detection Localisation
Automatic Tattoo Cropping
Manual Tattoo Cropping
Tattoo Features Extraction
Query Tattoo Features
DEEP LEARNING MODULE
N
Tattoo Registration
Y
Tattoo Matching & Identification
N-Ranked Best Matches
End of Tattoo Identification
Tattoo Template Features
Insert Tattoo Features Into DB
Get Tattoo Template Features from DB
End of Tattoo Registration
Tattoo Templates DB

Figure 19. Tattoo identification example using deep learning methods. Source: (Han, Li, Jain, Shan, & Chen, 2018)



GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

KJ-NA-29808-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub

Publications Office

doi:10.2760/661464

ISBN: 978-92-76-08843-1