Year : 2021

Interprétation des scores de reconnaissance faciale automatique pour l'investigation et le tribunal

Jacquet Maëlig

Jacquet Maëlig, 2021, Interprétation des scores de reconnaissance faciale automatique pour l'investigation et le tribunal

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_F747FA4478CB9

Thèse de Doctorat

Interprétation des scores de reconnaissance faciale automatique pour l'investigation et le tribunal

Maëlig Jacquet

A. Bertillon, Identification anthropométrique, instructions signalétiques, 1893

Directeur de Thèse : Prof. Christophe Champod

Université de Lausanne
Faculté de Droit, des Sciences criminelles et d'Adminitration publique
Ecole des Sciences Criminelles

LAUSANNE - SUISSE
2021

UNIL | Université de Lausanne
Ecole des sciences criminelles
bâtiment Batochime
CH-1015 Lausanne

IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise l'impression de la thèse de Madame Maëlig Jacquet, candidate au doctorat en science forensique, intitulée :

« Interprétation des scores de reconnaissance faciale automatique pour l'investigation et le tribunal »

Le Président du Jury

Professeur Franco Taroni

Lausanne, le 13 décembre 2021

Remerciements

Ce projet est né en 2015 de presque rien et alors que je n'avais - à peu près - aucune connaissance dans ce domaine. Pendant les six dernières années, de nombreuses personnes ont participé à ce travail, directement ou non, en me soutenant, en m'aidant et en me secouant quand c'était nécessaire. Elles et ils se sont assuré.e.s que je mène à bien mon projet, de la manière dont je le voyais et je leur en suis profondément reconnaissante.

Au Professeur Christophe Champod, merci infiniment de m'avoir fait confiance en acceptant de diriger ce projet, quand je n'avais encore aucune idée de l'ampleur qu'il pourrait avoir. Tu m'as donné la chance de pouvoir conduire ma recherche avec énormément d'autonomie, en apportant toujours les connaissances et questionnements dont j'avais besoin pour avancer.

Je remercie également le Professeur Didier Meuwly et le Docteur Simon Baechler, tout d'abord pour avoir accepté, dès la rédaction de mon mémoire intermédiaire, de faire partie de mon jury, ainsi que pour l'intérêt et l'enthousiasme dont ils ont fait preuve tout au long de mon projet. Les échanges et opportunités qu'ils m'ont offerts m'ont donné le regain de motivation nécessaire lors de périodes de doutes, et m'ont éclairée quant à la réelle portée opérationnelle de ce projet.

Docteure Claude Bauzou, je vous remercie d'avoir accepté de partager votre avis, vos remarques et suggestions en faisant partie de mon comité de thèse. Je tiens également à remercier tous vos collègues avec qui j'ai pu échanger chez Idemia, notamment Alexandre Fabre, Christophe Varlet et Sébastien Tilly, de m'avoir aidée à adapter le MFE à l'utilisation académique que je voulais en faire et qui diffère de l'application opérationnelle pour laquelle il a été développé.

À toutes mes « POI », les personnes qui ont donné de leur temps pour participer à ma collecte de données et remplir les innombrables Doodles d'organisation. Ce projet n'aurait évidemment jamais abouti tel que je l'imaginais sans leur aide, mais je les remercie également et par-dessus tout pour leur enthousiasme et leur bienveillance.

Merci à toutes les personnes qui m'ont apporté leur l'aide dans les parties techniques de ce projet, en partageant leurs scripts et leur savoir, et m'aidant à améliorer les miens : Andrea Macarulla (*National Forensic Institute*, Pays-Bas), Lorenzo Gaborini (ESC), Marco De Donno (ESC) et Stackoverflow (*world-wide web*). Je remercie tout particulièrement Julien Furrer (ESC) pour son aide immense et sa patience face à ma relation compliquée avec Linux, notamment.

Je tiens à remercier la Professeure Manon Jendly, qui m'a apporté un regain de force et de confiance – aussi bien en moi qu'en mon projet – par l'intérêt et l'enthousiasme indéfectibles qu'elle a manifestés pour ma recherche. Je la remercie de m'avoir accordé sa confiance en me permettant d'intervenir dans le cadre de ses cours. Cette expérience a été le déclencheur de l'écriture d'un article collaboratif alliant les points de vues criminaliste et criminologique sur ce sujet. Merci, Manon, tu es une source d'inspiration pour les personnes (femmes, surtout, mais pas uniquement) qui veulent se frayer, tant bien que mal, un chemin dans le monde souvent hostile de la recherche académique.

Je remercie chaleureusement toutes les personnes – prédécesseur.e.s, assistant.e.s et professeur.e.s - qui m'ont aidée, soutenue et fait confiance dans mes activités d'enseignement. Lorsque l'on s'engage pour un doctorat, on s'engage également à dédier la moitié de notre temps à l'enseignement, et il n'aurait jamais été possible de mener à bien ma thèse sans un tel soutien en parallèle pendant ces cinq ans.

C'est là que ça devient plus compliqué à exprimer, car bien plus personnel. Ces remerciements-là vont être bien plus courts, mais si vous y êtes, vous savez pourquoi - et le comprenez. Samwise Gamgee a dit « *I cannot carry it for you, but I can carry you* »[1]. Vous avez été ma communauté de Sam.

Toutes les aides techniques et professionnelles du monde n'auraient été d'aucune utilité sans votre présence, votre soutien, vos tournées à Sat', vos comiques de répétitions, vos cinés, vos soirées geekage, votre humour douteux et énormément de nourriture. Ilaria, Natalia, Yionel, Tim, Marc, Giulia, Patteet, Hannes, Elénore, Francesco… Merci.

Virginie, Benjamin, Cécile, je ne serais pas la moitié de la personne que je suis, sans vous. J'aurais même probablement encore 11 ans. Alors merci, d'exister.

À mes parents, à Mané, à ma famille - proche ou éloignée, ceux qui sont partis et ceux juste arrivés. Merci.

À mes sœurs, bisounours.

[1] « Je ne peux le porter pour vous, mais je peux vous porter, vous. »

« Il semble que la perfection soit atteinte non quand il n'y a plus rien à ajouter, mais quand il n'y a plus rien à retrancher. »
Antoine de St-Exupéry (1939)

« It is done. »
Frodo Baggins (3019, 3ème Age)

Table des Matières

Définitions et abréviations

1vN / NvN	Comparaisons d'une image avec N images (1 *versus* N) / de N images contre N autres images (N *versus* N)
ATM	(*Automated teller machine*) Distributeur automatique de billets
BdD	Base de Données
CCTV	(*Closed-Circuit Television)* Caméra de surveillance en circuit fermé
FN	FaceNet
GEV-MLE	(*Generalized Extreme Value – Maximum Likelihood Estimation*) Méthode d'estimation de distribution de fonction de probabilité
KDE	*Kernel Density Estimation*
Intervariabilité	Ensemble des variations observables sur les photographies de visages de plusieurs personnes distinctes (syn. Variabilité intersources)
Intravariabilité	Ensemble des variations observables sur les photographies du visage d'une même personne (syn. Variabilité intrasource)
LFW	*(Labelled Face in the Wild*) Base de données de photographies faciales de célébrités prises dans des conditions non contrôlées (http://vis-www.cs.umass.edu/lfw/)
MFI / MFE	Systèmes automatiques de reconnaissance faciale Idemia MorphoFace Investigate / MorphoFace Expert
NIST	*National Institute of Standards and Technology* (États-Unis)
PAVA	*Pool Adjacent Violators Algorithm*
POI	(*Person of Interest*) Personne d'intérêt dans le cadre de l'enquête : victime, témoin, suspect, etc.
RegLog	Régression Logistique
RMEP / RMED	*(Rate of Misleading Evidence in Favor of Prosecution / Defense)* Taux de résultats soutenant à tort la proposition de l'accusation/de la défense
SLR	*Score-based Likelihood Ratio* (Rapport de vraisemblance basé sur un score)

Chapitre I. Objectifs et structure de la recherche

Cette section introductive décrit le contexte académique dans lequel s'inscrit ce projet, les objectifs et problématiques généraux de la recherche ainsi que la structure du manuscrit.

I.1. Objectifs de la thèse

L'objectif de ce projet de thèse est de développer et valider des modèles évaluatifs permettant d'interpréter les scores de comparaison issus de systèmes automatiques de reconnaissance faciale, en adéquation avec les contraintes de différents scénarios forensiques.

I.2. Plan du manuscrit

Les chapitres II et III introduisent les fondements académiques et opérationnels sur lesquels est basé ce projet. Le chapitre II dresse un portrait large des applications actuelles de la reconnaissance faciale en sciences criminelles et présente les enjeux liés aux images et aux méthodes selon les domaines d'application dans lesquels elles sont utilisées. Le chapitre III se focalise plus précisément sur l'exploitation d'images dans le contexte forensique, par comparaisons manuelle et automatique.

Dans le Chapitre IV, nous exposons les problématiques de travail ainsi que la méthodologie développée pour y répondre.

Dans les chapitres V, VI, VII et VIII nous analysons et discutons les résultats pour quatre grands axes de recherche : l'utilisation de la reconnaissance faciale en phase d'enquête, puis comme élément de preuve au tribunal, l'évolution du SLR de l'enquête à l'expertise, et enfin les variables impactant les performances des systèmes automatiques.

Le chapitre IX revient sur des points de discussion généraux pour répondre aux problématiques de travail, et ouvre les perspectives de recherches futures.

I.3. Place de la recherche dans les contextes académiques et opérationnels

Afin de mieux comprendre le but de ce travail de thèse il est important de le resituer dans le continuum académique. Pour ce faire, une comparaison avec le processus de recherche médical se révèle être un exemple des plus concrets.

Niveau de la technologie

Dans le domaine médical, la recherche fondamentale a pour but d'améliorer la compréhension de phénomènes naturels biologiques et physiologiques. Par exemple, la virologie vise à étudier la structure et les mécanismes des virus pour comprendre leurs modes de reproduction et d'action sur les cellules infectées.

En reconnaissance faciale, cette étape peut être transposée aux études fondamentales sur les outils statistiques, les algorithmes de comparaison de visages et le développement du *machine learning*.

Niveau des scénarios

La recherche clinique est l'étape d'application des principes soulevés par la recherche fondamentale. Elle fonctionne sur une méthodologie basée sur des tests successifs à différentes échelles : tests in vitro, tests in vivo et essais cliniques.

Les tests in vitro

En médecine, il s'agit de tests « sous verre » en laboratoire, dont l'environnement et les données sont totalement contrôlés par le chercheur afin de maitriser toutes les variables pouvant agir sur le résultat. Cela permet par exemple d'observer l'effet d'un vaccin contre certaines souches du virus de la grippe développé grâce aux connaissances acquises par la recherche fondamentale.

En reconnaissance faciale, cela correspond par exemple à l'évaluation des performances d'algorithmes sur des tâches d'identification (1vN). Cette étape nécessite l'utilisation de très larges bases de données, composées d'images prises aussi bien dans des conditions contrôlées que non contrôlées. Plus le jeu de données est grand, plus le résultat d'un algorithme est significatif. La base de données LFW est très exploitée dans cette étape car elle contient de multiples photographies de célébrités prises dans de bonnes conditions mais avec des distances et angles de prises de vue variées (Zeng *et al.*, 2020). L'équivalent des tests in vitro est, par exemple, les tests FRVT menés par le NIST : ils offrent un très large aperçu des performances et limitations des algorithmes existants (Grother *et al.*, 2019a ; Grother *et al.*, 2019b).

Les tests in vivo

Les tests « dans le vivant » ont pour but d'appliquer la même solution que celle utilisée dans les tests in vitro, mais dans un environnement moins contrôlé et plus proche de celui de l'objectif final du développement – le plus souvent sur des mammifères comme la souris ou les primates. Dans le domaine médical, cette étape permet d'observer l'efficacité d'un vaccin sur un être vivant auquel le virus ciblé a été inoculé. Cette étape permet de se rapprocher des conditions réelles visées sans soulever – en théorie - de préoccupations éthiques liées à l'implication de sujets humains.

En biométrie, le but de cette étape est de développer un modèle en sélectionnant des algorithmes et données grâce aux résultats des tests de la précédente étape et adaptés à l'objectif opérationnel.

C'est dans cette phase de tests que s'inscrit la majeure partie du présent projet. L'ensemble des connaissances engrangées par la recherche fondamentale (les outils statistiques, le développement d'algorithmes et de matériel de prise de vue, et les principes de l'interprétation de traces forensiques) et par la recherche appliquée telle que celle menée par le NIST, sont les piliers sur lesquels est fondé ce projet, dont l'objectif est de répondre aux contraintes opérationnelles rencontrées par les services de police. Une critique récurrente qui survient à ce stade concerne la structure, et plus spécifiquement la taille, des bases de données utilisées. Il est couramment admis que la recherche scientifique gagne en crédibilité et pertinence par l'augmentation de la taille de son échantillonnage. Un tel critère doit être pris en compte lors des étapes précédentes, mais peut mener à valoriser la quantité de données aux dépens de leur

qualité, et par conséquent aux dépens de leur pertinence, dans les dernières étapes de développement et d'implémentation d'une méthode forensique.

Niveau opérationnel

Le vaccin développé doit faire l'objet d'essais cliniques afin que son utilisation soit validée sur un échantillon significatif avant de pouvoir être mise à disposition à grande échelle. Cela nécessite une étude de faisabilité et de démonstration du bénéfice, avec comme critères principaux l'objectif visé, la méthodologie à appliquer, la sélection des patients, les risques éventuels, l'efficacité attendue, etc. Tout cela permet d'utiliser la solution développée au travers des tests in vitro et in vivo en la testant en conditions réelles, non contrôlées, sur un échantillon restreint de patients.

En reconnaissance faciale, le stade final du processus de recherche est de tester le modèle développé avec des données issues de cas réels. Dans ce projet, le modèle développé est appliqué à des données de cas rencontrés par la Police cantonale neuchâteloise, Suisse.

I.4. Publications et présentations associées à cette recherche

Articles

M. Jacquet, L. Grossrieder, « Enjeux et perspectives de la reconnaissance faciale en sciences criminelles », Criminologie, 54(1), pp. 135-170, 2021.

T. Bollé, E. Casey, M. Jacquet, « The Role of Evaluations in Reaching Decisions Using Automated Systems Supporting Forensic Analysis », Forensic Science International: Digital Investigation, 34, p. 301016, 2020.

M. Jacquet and C. Champod, « Automated face recognition in forensic science: Review and perspectives », Forensic Science International, 307, p. 110124, 2020.

Présentations

M. Jacquet, C. Champod, « Probabilistic evaluation of face recognition evidence from simulated forensic cases using CCTV images and ID documents », ENFSI Digital Imaging Working Group Annual Meeting, Madrid, Espagne, 2019.

M. Jacquet, C. Champod, « Comparison of generic and suspect-anchored score-based likelihood ratio assignation models using automatic face recognition systems », 16th International Summer School for Advanced Studies on Biometrics for Secure Authentication, Alghero, Italie, 2019.

M. Jacquet, C. Champod, « Développement d'une méthode évaluative en comparaison de visages », XVIe Colloque de l'AICLF, Lausanne, Suisse, 2018.

Chapitre II. Reconnaissance faciale : Applications, enjeux et perspectives

Cette section se base sur et complète les réflexions et la revue de littérature publiées dans l'article de Jacquet et Grossrieder (2021) intitulé « Enjeux et perspectives de la reconnaissance faciale en sciences criminelles » paru dans la Revue Criminologie.

II.1. Origine de la reconnaissance faciale et attentes sociétales

La première utilisation de la photographie dans le cadre judiciaire remonte à la fin du XIXe siècle. À cette époque, Alphonse Bertillon met au point des fiches signalétiques pour recenser les criminels. Celles-ci étaient basées sur des méthodes anthropométriques (mesures de certains éléments du corps tels que l'envergure, les dimensions du crâne, la longueur des membres et du visage, avec notamment les dimensions du front, du nez et de l'oreille) ainsi que sur des photographies et la description du visage, appelé « portrait parlé » (Bertillon, 1886 ; Bertillon, 1893).

Avec sa démocratisation grandissante, la reconnaissance faciale est accompagnée d'idées reçues véhiculées entre autres par les médias (fictionnels ou non) et les campagnes commerciales de systèmes de reconnaissance automatisée.

Depuis plus de dix ans, la reconnaissance faciale est présentée au grand public comme un outil « miracle » capable de retourner un résultat binaire irréfutable – « identifiée » ou « non-identifiée » - pour la reconnaissance de personne d'intérêt dans le cadre forensique. Dès les prémices du développement de cette technique, les séries télévisées, telles que « Les Experts », « NCIS » et « Bones », ont montré l'exploitation d'outils aux performances redoutables. Pour exemple, la Figure 1 est un extrait d'un épisode des Experts, datant de 2008, où un système automatique compare la photographie récente d'une personne âgée à une image d'archive représentant un jeune homme d'identité connue. L'algorithme de l'inspectrice forensique conclut qu'il s'agit de la même personne, survolant tous problèmes concernant le grand intervalle de temps entre les deux images, la différence d'angles et de conditions de prise de vue, ainsi que les distorsions multiples inhérentes aux appareils photographiques et aux expressions faciales. En réalité, chacun de ces éléments complique la comparaison de visages en ajoutant de nombreuses variations en termes de luminosité, dimensions, proportions, résolution de l'image et âge de l'individu. Cela a pour effet d'augmenter les dissemblances observées entre les deux images et de diminuer les performances aussi bien des systèmes de reconnaissance faciale que de l'humain (Peng, 2019 ; Phillips et O'Toole, 2018 ; Phillips *et al.*, 2018).

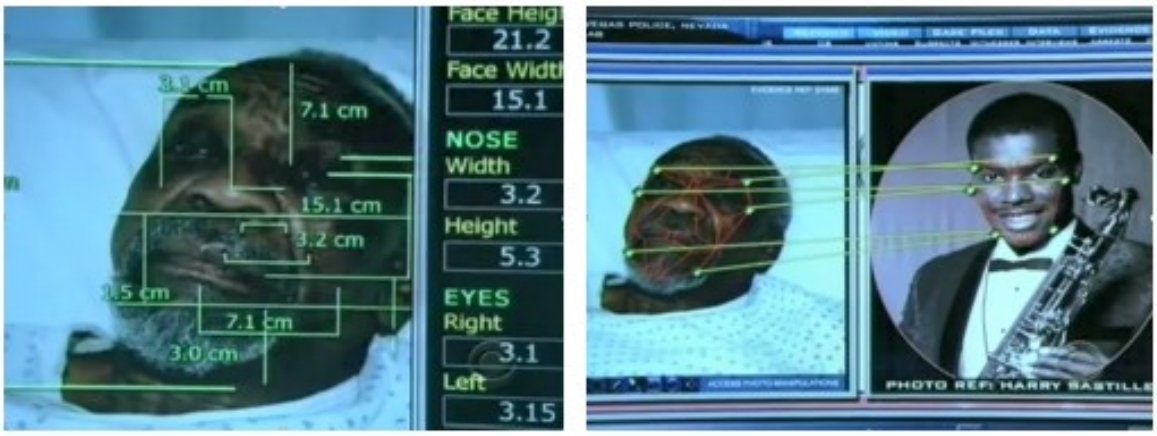


Figure 1 : Extrait du processus de reconnaissance faciale présenté dans « Les Experts » (saison 9 épisode 8, 2008).

Bien que l'effet de ces séries sur l'opinion publique, communément appelé « *CSI effect* », demeure mitigé (Cole et Dioso-Villa, 2007 ; Smith *et al.*, 2011), cet exemple (Figure 1) permet d'entrevoir un éventail représentatif de problématiques couramment ignorées dans les médias fictifs, pourtant toujours activement étudiées dans la littérature spécialisée. Les médias tels que la presse écrite et télévisée ne relatent que peu (voire pas) de détails sur l'utilisation judiciaire de ces technologies. En cause, la confidentialité des informations entourant des cas récents évidemment, mais également la faible utilisation des techniques de reconnaissance faciale dans ce domaine. En contrepartie, il existe et se développe une abondance d'articles, de débats, d'interviews, traitant de l'omniprésence et dangers liés aux caméras et systèmes de surveillance des populations. Cette technologie est présentée comme un outil de surveillance massive mettant en danger le respect de la vie privée, mais dont la fiabilité dans certaines conditions serait médiocre (Davies *et al.*, 2018 ; Fussey et Murray, 2019).

Un autre grand mythe entourant la reconnaissance faciale est la croyance selon laquelle le système agit à notre place et accomplit successivement une multitude de tâches variées pour accomplir un objectif final, tel qu'identifier un suspect, lier des délits et prouver la culpabilité ou non d'un auteur. Ici encore, les stratégies commerciales ainsi que les représentations médiatiques mettent en avant une simplicité et une autonomisation presque complète tout en occultant un élément indispensable du processus, l'être humain. Nombre de tâches essentielles dépendent encore presque exclusivement du facteur humain, comme par exemple, trier les images pertinentes, prétraiter les données pour les adapter à l'analyse, réviser et interpréter les résultats de comparaison. L'efficacité d'une tâche de reconnaissance faciale dépend donc d'un ensemble de choix et d'actions à différents niveaux et ne découle pas uniquement de l'utilisation d'un outil informatisé.

Finalement, le besoin d'un système à la pointe de la technologie et supposé onéreux est également une préconception erronée. Il n'est pas rare de faire un amalgame entre la complexité d'un logiciel, son prix et son efficacité. De la même manière que le nombre de pixels d'un appareil photo numérique est un puissant argument commercial pour le grand public, la mise en avant de la puissance de calcul d'un outil de reconnaissance faciale peut rapidement être assimilée à la qualité de ses résultats. La plupart des études sur les algorithmes automatiques (*cf.* III.3 p.32) sont effectuées sur des bases de données développées pour la recherche. Les outils qui en découlent présentent ainsi des performances élevées lorsqu'ils sont appliqués à ces données, mais lors d'une application en conditions réelles avec les contraintes et spécificités que cela implique, leur utilité peut se révéler décevante.

II.1.1 Sciences criminelles : de la scène de crime au tribunal

Une dichotomie récurrente, d'abord définie par Jackson et collègues (Jackson *et al.*, 2006), distinguait l'utilisation de méthodes forensiques à des fins d'investigation et de renseignement, d'une part, et le processus évaluatif, associé exclusivement à la présentation d'éléments de preuve au tribunal, d'autre part. Cependant, plusieurs publications récentes redéfinissent cette distinction (Baechler *et al.*, 2020 ; Jacquet et Champod, 2020 ; Ryser *et al.*, 2020a), et ne cantonnent plus le processus évaluatif à la seule présentation de résultats d'expertise au tribunal.

Ce processus est considéré comme l'objectif de l'expertise, alors qu'il s'agit d'une méthodologie scientifique utilisée pour produire un résultat aussi bien dans le cadre de l'investigation que de l'expertise.

L'utilisation d'images à des fins forensiques peut intervenir lors de 3 phases distinctes, communes à toutes les traces : lors de l'investigation, de la veille opérationnelle et au tribunal. À ces trois phases peut également s'ajouter une quatrième sous la forme de l'action de prévention. Les phases d'investigation et de veille peuvent se servir l'une l'autre, alors que la phase d'expertise clôture l'instruction.

Il est à noter que nous utilisons le terme « expertise », non pas au sens strict décrit dans le CPP, mais dans un souci de simplification pour décrire toute procédure destinée à fournir des éléments de preuve pertinents au tribunal lors d'un procès.

Une quatrième phase, la prévention, est également susceptible de tirer profit de manière proactive des éléments issus de l'investigation et du processus de veille de manière à réduire les activités délictueuses (Figure 2).

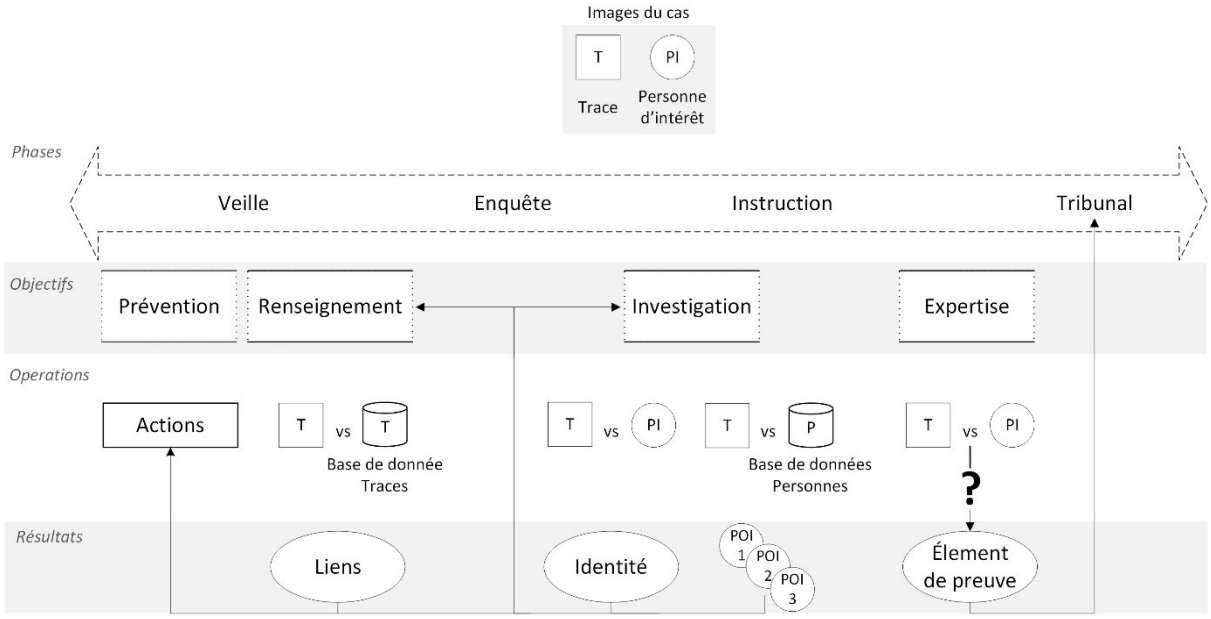


Figure 2 : Synthèse de l'utilisation forensique actuelle d'images dans l'exemple d'un cas comportant une trace et une personne d'intérêt (POI).

II.1.2 Investigation

Chaque cas porté à l'attention de la police débute par une enquête, plus spécifiquement nommée phase d'investigation dans cet article. L'objectif est de recueillir des informations (dont font partie les traces) et de les exploiter pour orienter l'enquête vers des personnes d'intérêt (POI) ou aider à reconstruire le déroulement de l'événement. Pour un cas donné, où la trace est une image faciale, deux opérations peuvent être effectuées (Figure 2). D'une part, un système automatique de reconnaissance faciale permet de comparer la trace à de larges bases de données d'images et de générer une liste de candidats potentiels. D'autre part, lorsque l'enquête s'oriente plus spécifiquement vers une personne d'intérêt (Figure 2) absente des bases de données à

disposition, l'opérateur peut comparer l'image trace avec les photographies de référence du POI. Les enquêteurs peuvent alors exploiter seuls les résultats de ces comparaisons - automatiques ou manuelles - ou recouper les informations fournies avec les éléments apportés par d'autres traces ou d'autres éléments d'enquête.

En pratique, les applications mises en place dans la phase d'enquête (dans laquelle collaborent renseignement et investigation) sont, par exemple, l'analyse rétroactive de vidéosurveillance à la suite d'un événement délictueux, ou l'analyse proactive d'enregistrements en direct afin de repérer des POI ou des personnes à risque. Ces systèmes de reconnaissance faciale en direct (*live face recognition/LFR*) sont implémentés dans différentes grandes métropoles et exploitent de vastes réseaux de CCTV[2] sur la voie publique. Des systèmes LFR ont été testés récemment par le *Metropolitan Police Service* à Londres (Fussey et Murray, 2019) et par le *South Wales Police* au Pays de Galle (Davies *et al.*, 2018). Là où certains pays et villes ont adopté massivement ces systèmes depuis plusieurs années, comme en Chine ou à Moscou[3], d'autres ont fait machine arrière en les retirant, par exemple à Orlando (États-Unis) en 2018 (Cagle et Ozer, 2018), ou les bannissant légalement comme la ville de Chicago (États-Unis, *Biometric Institute*, (2021)

II.1.3 Renseignement

Parallèlement à l'investigation, la phase de veille opérationnelle consiste à produire du renseignement pour contribuer non seulement à l'enquête, mais surtout au modèle d'action de sécurité (Hane, 2015 ; Ribaux, 2014 ; Ribaux *et al.*, 2011). Dans le cadre d'un processus de renseignement criminel opérationnel, des données sont recueillies sur chaque cas, intégrées dans une mémoire, puis analysées afin d'extraire des informations sur le modus operandi et les auteurs. En parallèle, la comparaison d'images-traces de différents cas entre elles aide à détecter des liens entre ces cas, qui étayent l'enquête en cours, et s'intègrent également au suivi systématique de la criminalité répétitive pour détecter, prolonger ou confirmer des séries (Rossy *et al.*, 2013). Dans Dessimoz et Champod (2015), les auteurs décrivent un exemple concret du traitement actuel des images à des fins d'enquête et de renseignement en Suisse, via une plateforme commune utilisée en Suisse latine[4] pour le suivi de la délinquance sérielle et itinérante. Cette étude a permis de constater que le nombre de liens détectés grâce aux images a fortement augmenté depuis 2009, de même que le nombre d'événements associés avec des images. En revanche, ceux associés avec d'autres types de traces (ADN, traces papillaires et traces de semelles) restent stables (Rossy *et al.*, 2013). Toutefois, les auteurs soulignent le fait qu'en 2013, au moment de la collecte des données, le pourcentage d'images adaptées à l'utilisation des systèmes automatiques était encore très faible (3,2 %). Cette proportion est susceptible d'être revue à la hausse actuellement, compte tenu notamment du fort développement de ces systèmes ces dernières années, tant en termes de performance (typiquement la résolution des capteurs) que de diffusion.

[2] Télévision en circuit fermé (*Closed-Circuit Television)*

[3] Source : https://www.letemps.ch/monde/moscou-opte-surveillance-totale-chinoise, consulté le 24.04.2020

[4] La Suisse « latine » regroupe tous les cantons de Suisse romande ainsi que le Tessin.

II.1.4 Expertise

Lorsque cela est requis par l'instruction judiciaire, la finalité de l'expertise forensique est l'interprétation des résultats de comparaisons entre les images-traces et les images de référence afin potentiellement de les présenter comme élément de preuve au tribunal (Figure 2). Le problème actuel en matière de comparaison de visages est qu'aucune méthode documentée ne permet de réaliser cette étape avec un système automatique, comme c'est le cas dans les domaines de la reconnaissance de locuteurs (Botti *et al.*, 2004 ; Meuwly, 2001) et des traces papillaires (Egli, 2009). La technique de comparaison, qu'elle soit manuelle ou automatisée, doit répondre à des exigences légales très variables selon le pays pour être utilisée devant les tribunaux. Nous proposons d'exposer un bref aperçu de l'éventail de critères légaux, du plus au moins exigent, respectivement en Amérique du Nord, au Royaume-Uni et en Europe de l'Ouest.

Aux États-Unis, la jurisprudence Daubert (1993) (révision de Frye (1923)) - toujours en vigueur dans plusieurs États) définit que le juge conserve le droit de juger si la méthode utilisée pour fournir la preuve est acceptée par la communauté scientifique (Federal Rules of Evidence 702 (2020)). Les révisions les plus récentes exigent également que la méthode soit fondée sur une base scientifique fiable et réfutable, qu'elle ait été vérifiée et que les taux d'erreur soient connus, et qu'elle soit disponible pour examen par les pairs et publication. Au Canada, la jurisprudence issue de R. v. Mohan (1992) met en place des règles semblables au texte Daubert.

Au Royaume-Uni, d'après R. v. Turner (1975) l'expertise scientifique doit aller au-delà du bon sens et être fiable[5] et l'expert.e doit posséder les connaissances suffisantes acquises par apprentissage et expérience, faire preuve d'impartialité. Contrairement aux jurisprudences d'Amérique du Nord, il est spécifié que le manque de validation, de publications scientifiques et de révision par les pairs n'est pas forcément rédhibitoire pour l'utilisation d'une méthodologie.

En Europe continentale, il n'y a pas de contrainte légale spécifique au sujet de l'admissibilité de la preuve scientifique. Il est seulement précisé que les juges sont garants de l'évaluation de la pertinence de la preuve en fonction de l'état des connaissances scientifiques. En Suisse, par exemple, le code de procédure pénale du 5 octobre 2007 (CPP (2007) art. 182 à 183), l'expertise est présentée comme « Une mesure d'instruction nécessitant des connaissances ou des investigations complexes confiées par le/la procureur·e ou le/la juge à un·e ou plusieurs spécialistes pour qu'il(s)/elle(s) l'informe(nt) sur des questions de fait excédant sa compétence technique ou scientifique » (Piquerez et Macaluso, 2011). Similairement, en France, l'expertise doit être justifiée et l'expert.e jugé.e compétent.e dans le domaine requis, puis le juge est seul à évaluer la valeur et la portée des preuves fournies par les parties (CPP – art. 427).

Pour conclure, il nous apparait que, malgré les disparités du cadre légal entourant l'expertise scientifique selon les pays, l'emploi de techniques nouvelles doit être soumis à des tests empiriques sur des données adéquates pour garantir de présenter au tribunal des résultats fiables et dont l'expert.e connaît les risques d'erreurs.

[5] Ici, « *reliability* » regroupe les notions de validité, reproductibilité, justesse et répétabilité.

Les systèmes de reconnaissance faciale sont de plus en plus performants et étudiés, mais les méthodes en développement manquent encore de validation. Cependant, comme expliqué précédemment, les méthodes manuelles manquent également toujours de standardisation et de validation (ENFSI, 2018 ; FISWG, 2012 ; Forensic Science Regulator, 2019) et ne constituent donc pas une solution suffisante.

En outre, un second frein à l'utilisation de ces systèmes est, selon Dessimoz et Champod (2015), la faible qualité des images CCTV. Néanmoins, les études récentes de Peng (2019) suggèrent que l'exécution de tâches de reconnaissance faciale dépend fortement de la qualité des données d'apprentissage sur lesquelles un système de *machine learning* a été entrainé. Par conséquent, la comparaison d'images de faible qualité nécessite l'utilisation d'un modèle adéquat entrainé sur des données de qualité similaire.

À notre avis, les barrières freinant pour l'instant l'utilisation de ces systèmes à des fins d'expertises ne sont pas infranchissables. Nous développons les limitations et perspectives liées à ces problématiques dans la section suivante.

II.1.5 Prévention

Au-delà du processus judiciaire, les tâches de reconnaissance faciale sont également utilisées de manière proactive, notamment eu égard à leur capacité à soutenir des mesures de prévention situationnelle. Issue des approches du même nom en criminologie, la prévention situationnelle réunit ainsi l'ensemble des mesures, qui ont pour objectif de prévenir les actes contraires aux normes en limitant les opportunités pour leurs auteurs de les commettre, par exemple en alourdissant le risque perçu par ces derniers d'être arrêtés et/ou en réduisant au minimum les avantages escomptés (Jendly, 2013). Parmi les techniques situationnelles synthétisées par Clarke et Eck (2005), la reconnaissance faciale est principalement utilisée pour augmenter les efforts et les risques perçus. Cela se traduit principalement par des contrôles d'accès, sous deux formes pouvant être combinées.

Dans la première forme, le visage des personnes souhaitant accéder à un lieu spécifique est comparé à une base de données de personnes autorisées. Dans le domaine pénitentiaire par exemple, il s'agit de contrôler les accès des visiteurs en vue de prévenir les différents trafics susceptibles de survenir au sein de l'établissement pénitentiaire (p. ex. stupéfiants, téléphones portables, cigarettes, etc.). Au Royaume-Uni, trois prisons ont testé un tel système en 2019[6]. Certaines sociétés proposent également des solutions de reconnaissance faciale permettant de localiser et suivre les détenus au sein de l'établissement pénitentiaire[7]. À notre connaissance, aucune institution n'a publiquement annoncé recourir à cette utilisation. Dans le domaine de la sécurité au sens large, de nombreux exemples récents montrent l'intérêt des institutions publiques et privées pour ces technologies. En 2019, deux lycées en France ont testé un tel système pour réguler l'accès à leurs établissements par les élèves. Néanmoins, la Commission

[6] Source : https://www.bbc.com/news/uk-47461035, consulté le 24.04.2020
[7] Source : https://www.facefirst.com/industry/correctional-facility-face-recognition/, consulté le 24.04.2020

nationale de l'informatique et des libertés (CNIL, 2019) a considéré le dispositif projeté comme disproportionné par rapport à l'objectif visé (contrôle d'accès d'un lycée). Dans le contexte des contrôles aux frontières, plusieurs aéroports aux États-Unis et en France ont mis en place un système permettant de déterminer si les passagers sont autorisés à pénétrer dans les zones d'embarquement[8].

Dans la seconde forme, les comparaisons sont effectuées avec une base de données de personnes non autorisées à accéder aux lieux. Par exemple, l'identification des hooligans à l'entrée des matchs de football permet de prévenir les violences et dégradations en les empêchant d'entrer (Woodward, 2001). L'objectif est de prévenir la survenue d'affrontement pendant et à l'issue du match en empêchant les personnes à risque d'entrer. Un principe similaire a été appliqué lors d'un concert de la chanteuse Taylor Swift en 2018 dans le but d'identifier des fans harceleurs (*stalkers*)[9].

Au-delà des contrôles d'accès, il est possible que la reconnaissance faciale produise des effets proactif et dissuasif liés au renforcement de la surveillance formelle et à la réduction de l'anonymat, par exemple par la présence de CCTV visible sur la voie publique. Cependant, si les méta-analyses sur l'effet des CCTV montrent que la présence de tels systèmes est à même de réduire la criminalité contre le patrimoine dans des contextes particuliers (Farrington *et al.*, 2007 ; Welsh et Farrington, 2009), rien ne permet à l'heure actuelle de démontrer un effet supplémentaire lié à l'ajout de reconnaissance faciale aux CCTV.

II.1.5.1 Perception des acteurs

L'usage croissant de ces technologies contraste fortement avec le faible nombre d'études empiriques destiné à évaluer leur efficacité et leur efficience sur leur application pratique. La plupart des sources disponibles concernant l'utilisation de la reconnaissance faciale en matière de sécurité se résume aux articles de presse ou à des rapports de travail institutionnels.

Peu d'études s'attardent sur la perception et l'acceptation de ces nouvelles technologies par les différents acteurs qui les emploient soit en tant qu'utilisateur direct (p. ex. services de police, tribunaux, etc.), soit en tant que citoyen susceptible de figurer dans une base de données. Une étude australienne portant sur l'utilisation et l'acceptation des mesures biométriques chez les victimes de vol et d'utilisation frauduleuse de données personnelles montre que plus d'un tiers des répondants seraient enclin à utiliser la reconnaissance faciale comme moyen d'identification, et l'adhésion à cette technologie serait encore plus forte chez les adultes plus âgées et les utilisateurs réguliers d'ordinateurs (Emami *et al.*, 2016). Dans le cadre d'une étude sur la perception des citoyens états-uniens sur l'utilisation de la reconnaissance faciale sur les caméras corporelles des policiers, Bromberg et ses collègues soulignent également l'importance de

[8] Sources : https://journalmetro.com/techno/2389332/la-reconnaissance-faciale-arrive-dans-les-aeroports/, consulté le 24.04.2020
https://www.washingtonpost.com/technology/2019/06/10/your-face-is-now-your-boarding-pass-thats-problem/, consulté le 24.04.2020

[9] Source : https://www.bbc.com/news/technology-49647244, consulté le 24.04.2020

l'influence sociale dans l'acceptation d'une nouvelle technologie (Bromberg *et al.*, 2020). En plus des caractéristiques sociodémographiques des acteurs, l'adhésion à la reconnaissance faciale est également tributaire de ses objectifs et domaines d'applications. Le soutien de la population est par exemple plus important dans le cadre de la protection des frontières (Unisys, 2014) que lors d'une utilisation en entreprise avec des caméras de vidéosurveillance (Rainie et Duggan, 2016).

Cet état des lieux permet de dépeindre le cadre opérationnel qui entoure l'utilisation actuelle de la reconnaissance faciale au sein des sciences criminelles et d'en identifier les principales caractéristiques et niveaux d'analyse. Le potentiel de ces utilisations s'observe principalement à travers des cas isolés car il est rarement évalué empiriquement. Parmi les exemples les plus récents se trouve celui d'un auteur d'une tentative de viol dans un métro new-yorkais qui a pu être arrêté grâce à la reconnaissance faciale (Miles, 2020). L'individu était connu des services de police, mais pas pour des délits d'ordre sexuels et se trouvait donc dans une base de données différentes, ce qui compliquait son identification. Le système automatique prend tout son sens dans ce type de situation, car il permet d'aider à identifier ou à exclure rapidement un auteur parmi de larges jeux de données que l'humain n'aurait pas pu parcourir entièrement et permet ainsi de prévenir de potentiels futurs délits. Malgré ces résultats prometteurs, à l'heure actuelle le manque d'études empiriques ne permet pas d'évaluer précisément les forces et les faiblesses des systèmes de reconnaissance faciale. D'où la nécessité de mieux situer leurs limitations et enjeux opérationnels ainsi que les perspectives de développement et de recherche que nous jugeons essentiels.

II.2. Enjeux et perspectives

Les enjeux de l'utilisation de la reconnaissance faciale forensique, ainsi que les obstacles rencontrés par son développement, diffèrent grandement selon le niveau d'analyse. On peut distinguer des enjeux à trois niveaux (Figure 3) : les enjeux liés aux images (1), aux méthodes (2) et aux objectifs (3).

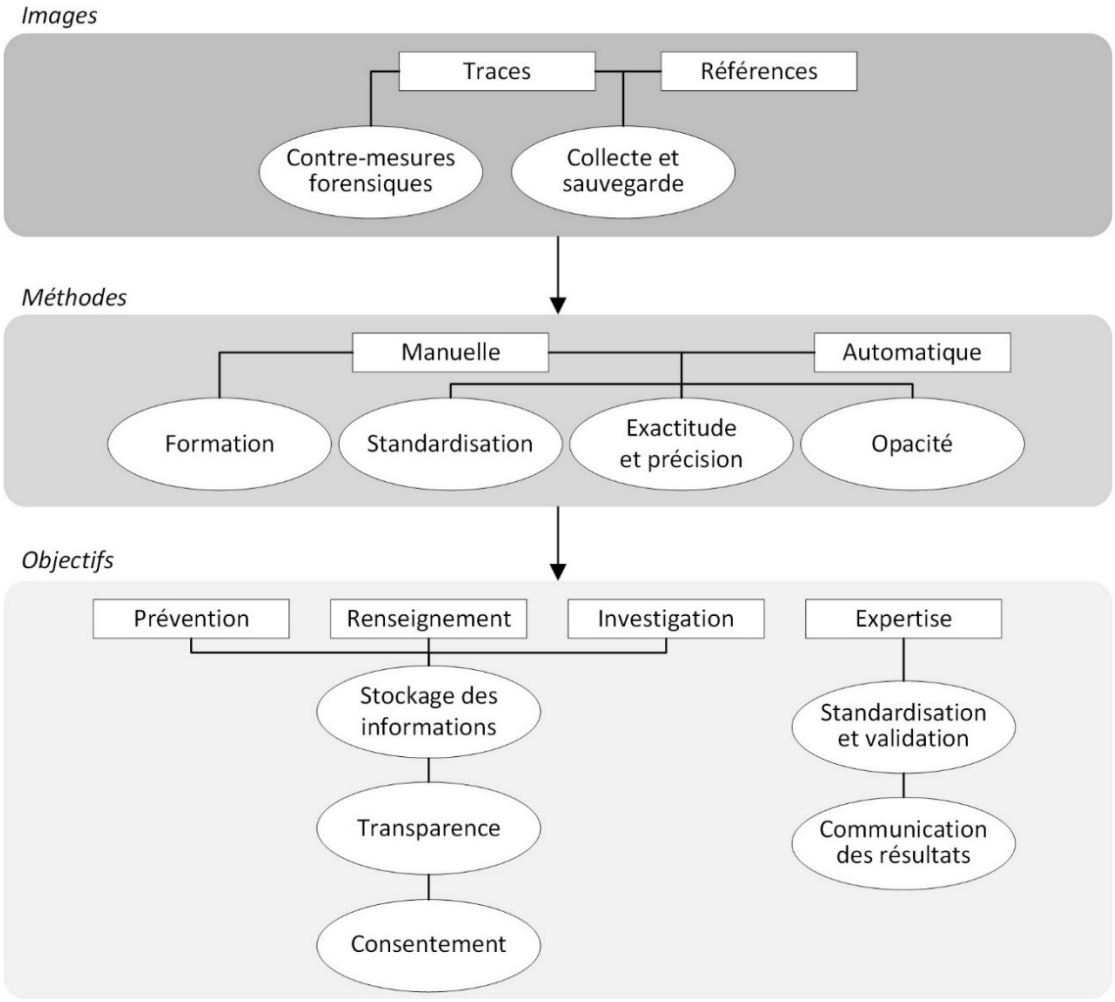


Figure 3 : Principaux enjeux de la reconnaissance faciale en sciences criminelles selon le niveau d'analyse.

II.2.1 Enjeux liés aux images

La reconnaissance faciale est présentée à travers les médias, surtout fictionnels, comme un outil capable de retourner un résultat irréfutable dans des conditions souvent irréalistes, omettant les facteurs qui influencent l'utilisation des images. En réalité, des facteurs à deux niveaux sont susceptibles d'affecter l'utilisation des images dans des tâches de reconnaissance faciale. Tout d'abord, au niveau de l'activité enregistrée par les outils de capture vidéo, les individus engagés dans des activités délinquantes peuvent masquer leur identité par différents stratagèmes, tels qu'un changement de tenue ou le port de cagoule, de masque, de capuche et de lunettes pour dissimuler leur visage. Ces contre-mesures constituent un des freins majeurs à l'utilisation des images de surveillance dans un but forensique, et il est plus difficile d'intervenir pour réduire cette problématique. Un autre problème récurrent se situe au niveau de la qualité de l'image fournie aux investigateur.trice.s. Dans un premier temps, cette qualité est régie par les caractéristiques intrinsèques de l'appareil d'enregistrement (résolution de l'image, colorimétrie, distorsions, etc.) ainsi que par les conditions de prise de vue (luminosité, angle, netteté, etc.). La qualité varie donc considérablement d'une source à l'autre. Par exemple, une photographie trouvée par le biais de réseaux sociaux, initialement prise avec un appareil à haute résolution,

constitue un matériel de meilleure qualité qu'un enregistrement vidéo d'une caméra de surveillance de résolution moyenne et située en hauteur par rapport au sujet.

Néanmoins, une image originale de très bonne qualité, enregistrée à un distributeur automatique de billets (ATM) équipé d'une caméra à haute résolution par exemple, peut être fournie aux investigateur.trice.s dans une qualité détériorée. En cause : le format et la manière dont les images sont récoltées, puis partagées.

Lors d'une investigation suite à un braquage de commerce par exemple, les investigateur.trice.s prennent connaissance de l'enregistrement de la caméra de surveillance et en demandent l'accès. Idéalement, le propriétaire de l'enregistrement - le commerce dans notre exemple - doit fournir le fichier vidéo original ou, à défaut, une copie de celui-ci sans modification de format ni traitement ou compression. Cependant, dans une grande majorité de cas, les images mises à disposition sont elles-mêmes des enregistrements de l'image originale. Par exemple, l'investigateur.trice photographie, à l'aide d'un téléphone portable, l'écran où est affichée l'image originale ou sauvegarde l'image sous un format causant la perte de qualité, telle que des fichiers au format PDF ou DOC. La détérioration de l'image due à ces pratiques diminue considérablement la quantité et la qualité de détails d'intérêt pour l'enquête. De plus, dans les cas où l'investigation débouche sur une procédure nécessitant une expertise, si des images de faible qualité sont fournies à l'expert.e, il/elle devra émettre une demande d'acquisition des images originales au propriétaire par l'intermédiaire du responsable de l'instruction du cas. En l'absence des originaux, l'expert.e peut être amené.e à refuser l'expertise à défaut de matériel de qualité suffisante, ou d'effectuer une expertise dont la fiabilité pourrait être discutée au tribunal. Contrairement à la présence de contre-mesures forensiques lors de l'activité, il est plus aisé d'agir pour améliorer le processus de récolte et de sauvegarde des images-traces. La Figure 4 résume le processus d'acquisition et de transmission des images d'intérêt forensique, et met en évidence les pratiques actuelles à l'origine de leur détérioration ainsi que nos recommandations d'alternatives.

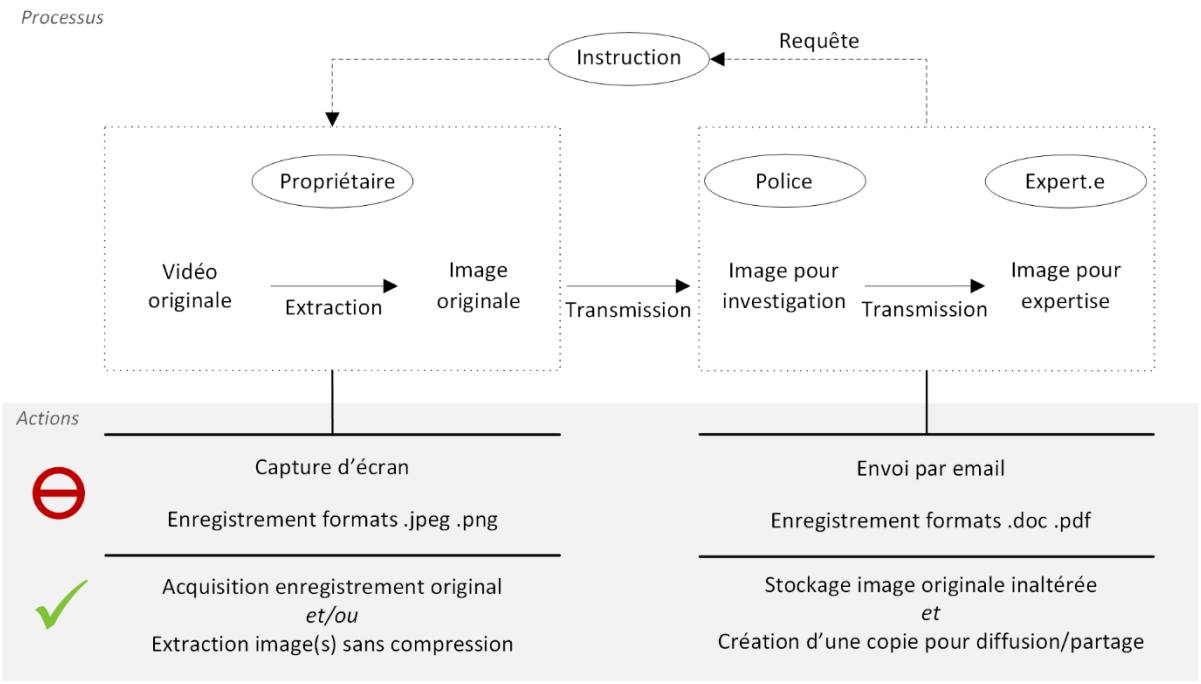


Figure 4 : Cheminement et manipulation des images dans le système judiciaire.

La Figure 5 illustre brièvement l'étendue des variations que provoquent différents modes et formats d'extraction au niveau de la qualité observable d'une image et de sa résolution. Pour cet exemple, le but était d'enregistrer une image où le visage de la POI est visible, à partir d'une vidéo enregistrée par une caméra de surveillance de haute résolution (1080p). Les deux images (a) ont été extraites à l'aide d'un script Python et du logiciel ouvert VirtualDub, respectivement aux formats TIFF et JPG. Dans les deux cas, la qualité observable de l'image semble similaire, mais la résolution de l'image au format JPG (72 ppp[10]) est en réalité inférieure. Cette différence, à peine décelable par l'humain, peut avoir en revanche un impact sur le processus d'analyse du système automatique. Les images (b) résultent de la capture de la même image, mais par l'intermédiaire de l'appareil photo d'un téléphone portable. Il s'agit alors de photographies au format JPG de l'image d'intérêt. Bien que la résolution soit comparable à celle de l'image JPG extraite (a), la qualité globale de l'image est très visiblement détériorée. De plus, ces deux captures ont été prises avec le même appareil, par le même opérateur, dans des conditions similaires et à quelques secondes d'intervalle l'une de l'autre. Malgré cela, la présence et l'intensité des distorsions sont indéniables et diminuent considérablement l'image d'intérêt. Il est également à noter qu'il ne s'agit ici que de variations survenant à l'étape d'extraction du matériel original, et que ces détériorations peuvent par la suite s'accumuler avec la multiplication d'étapes de transmission et de stockage de cette même image par les différents acteurs judiciaires (Figure 4). Malheureusement, comme exposé précédemment, les images-traces collectées par les forces de l'ordre sont très rarement extraites dans un format optimal, le plus souvent dans un souci de simplification et de gain de temps. Pour exemple, la plateforme intercantonale d'échange d'informations liées aux cas de Suisse latine (Dessimoz & Champod, 2015) contient

[10] Points Par Pouce

exclusivement des images stockées aux formats JPG et BMP, dont la taille, résolution et mode de capture sont très variables.

Ces pratiques courantes impactent par la suite tous les processus d'analyse et de comparaison, et peuvent nécessiter une requête pour l'acquisition du matériel original auprès du propriétaire, par le biais de l'autorité d'instruction chargée du cas (Figure 4). Cela peut être imputé en grande partie à la l'utilisation encore récente de détails tels que le visage dans les images récoltées, exacerbé par la perception souvent trop simpliste sur ces technologies, véhiculée par les médias, tels que discuté dans la section 2. En l'absence de sensibilisation adéquate de tous les acteurs intervenant sur l'ensemble des processus de collecte et transmission d'images, ces derniers sont enclins à baser leurs pratiques sur cette vision erronée de systèmes de reconnaissance faciale tout-puissants, capables de reconnaitre un individu à partir de vidéos pixellisées d'une POI au visage de profil, par exemple. De plus, dans un objectif évaluatif les exigences se multiplient, comme exposées précédemment, alors que les images fournies lors des demandes d'expertises sont souvent celles déjà collectées et transmises de manière inadéquate.

Figure 5 : Illustration des variations de qualité visuelle et de résolution d'images (a) extraites et (b) capturées à l'aide d'un téléphone portable[11] à partir d'une vidéo de caméra de surveillance[12].

[11] Samsung S10

[12] AXIS P3245-V (les résolutions indiquées sont celles de l'image enregistrée dont seule la partie recadrée sur le visage est présentée ici)

II.2.2 Enjeux liés aux méthodes

L'amalgame récurrent entre la complexité et l'efficacité d'un système se traduit auprès des décideurs par le besoin d'un système à la pointe de la technologie. Au contraire, cette impression de complexité est rarement synonyme d'efficacité et soulève plusieurs enjeux.

II.2.2.1 Opacité des systèmes automatiques

Bien que les algorithmes de *deep learning* utilisés initialement soient développés entièrement par l'humain, le processus d'apprentissage et la création des règles d'analyses par le système demeurent obscurs. Dès lors, l'opérateur ne peut pas expliquer le résultat d'une comparaison en toute transparence. C'est l'effet boîte noire (« *blackbox »)* (Pasquale, 2015).

Selon Burrell (2016), l'opacité des algorithmes utilisés peut être intentionnelle à des fins commerciales ou sécuritaires, mais il peut découler aussi du niveau de compétence à avoir pour les comprendre ou de la manière dont l'algorithme fonctionne au sein du système entier. L'opacité de ces systèmes et le manque de connaissances sur leur fonctionnement sous-jacent seraient alors un frein à leur optimisation et donc à leur performance. La période actuelle est une période charnière, durant laquelle les meilleurs algorithmes *rule-based* atteignent les limites de leur potentiel, alors que le *deep learning* laisse entrevoir des performances déjà remarquables et un potentiel d'amélioration exponentielle pour les années à venir –au prix d'une opacité accrue par la structure de ces algorithmes.

Les performances des systèmes de reconnaissance faciale les plus couramment utilisés ou récemment développés sont évaluées et publiées par le NIST (*National Institute of Science Technology*) dans les rapports « *Face Recognition Vendor Test* - FRVT ». Dans le dernier de ces rapports FRVT et ses derniers suppléments (Grother *et al.*, 2019a ; Grother *et al.*, 2021), 274 algorithmes sont soumis à deux tâches, identification et investigation, à partir d'images signalétiques (*mugshots*), de portraits « webcam » de qualité plus faible, de portraits de profils et de documents d'identité. Pour la tâche d'identification, les performances sont triées d'après le taux de faux négatifs atteint au seuil de 0.003% de faux positifs. Dans la tâche d'investigation, le classement est basé sur le taux des recherches où la POI ciblée n'apparait pas au premier rang de la liste de candidats. Par exemple, les algorithmes chinois de *Sensetime* et NEC ainsi que *cloudwalk* (Hengrui AI technology) présentent parmi les 10 meilleures performances et l'algorithme de Microsoft se classe parmi les 20 meilleurs. Le dernier système MorphoFace (Idemia), que nous utilisons dans le présent projet, se classe aux rangs 1, 2 et 3 respectivement pour les tâches d'identification, et entre les rangs 3 et 8 en investigation[13].

Les systèmes automatiques apportent de nombreux bénéfices, tant par leurs performances que par leur rapidité d'exécution. Cependant, il n'existe à ce jour aucun guide méthodologique dédié aux expertises en reconnaissance faciale assistées par un système automatique. Chaque système génère un résultat de comparaison selon une méthode de calcul qui lui est propre, le plus souvent sous la forme d'un score, et qui nécessitera une interprétation pour une utilisation devant un

[13] Source : https://pages.nist.gov/frvt/html/frvt1N.html, consulté le 13.05.2021

tribunal. Par exemple, l'algorithme FaceNet[14] génère des scores entre 0 et 3 représentant la distance – *c.-à-d.* le degré de dissimilarité - entre les deux images, alors que le système MorphoFace (Idemia) produit des scores de similarité entre 0 et 50'000. L'utilisation d'un système par un.e expert.e au tribunal peut être justifiée par les performances exposées dans les rapports FRVT du NIST (Grother et al., 2019). Cependant, ces tests évaluent les performances de systèmes à partir de très larges bases de données collectées dans un but de recherche, c'est-à-dire sous des conditions contrôlées qui ne représentent pas totalement des conditions réelles. Il est actuellement toujours nécessaire de conduire des études basées sur des données d'intérêt forensique reflétant au mieux les cas réels rencontrés dans le cadre judiciaire, et de produire un modèle permettant d'interpréter les différents scores de comparaison afin de présenter un résultat fiable de manière standardisée au tribunal.

Le manque de standardisation des méthodes de reconnaissance faciale n'est plus uniquement un débat de niche réservé aux spécialistes forensiques. En 2019, le tribunal de Lyon, en France, a été confronté au premier cas français dans lequel le seul élément de preuve disponible dans un procès pour vol était un enregistrement de vidéosurveillance. Le résultat de l'expertise en reconnaissance faciale, effectuée à l'aide d'un système automatique, soutenait la proposition de la partie accusatrice. L'avocat de la défense dénonce alors dans les médias l'utilisation de ce « robot accusateur »[15], en mettant en exergue le manque de bases scientifiques validées de cette méthode. Cet exemple résume à lui seul la précarité de la situation actuelle de la reconnaissance faciale automatique dans le domaine judiciaire et le besoin de fondations empiriques solides pour consolider la fiabilité de telles expertises.

II.2.2.2 Précision et exactitude

La précision se rapporte à la probabilité que deux images représentent effectivement la même personne lorsque la méthode produit un résultat positif, et l'exactitude est la probabilité de fournir un résultat positif en comparant deux images de la même personne. Malgré la complémentarité de ces deux notions, la performance d'une méthode est souvent réduite à sa précision. Or, bien qu'une méthode extrêmement précise produisant très peu de résultats positifs incorrects paraisse satisfaisante, le fait qu'elle puisse omettre un taux élevé de résultats positifs corrects peut avoir un impact négatif lors de son application.

Les taux d'erreurs, le plus souvent appelés taux de faux positifs (une personne reconnue à tort) et taux de faux négatifs (une personne non reconnue à tort), sont essentiels à l'évaluation de la robustesse d'une méthode (Dror, 2018). Ces taux sont interdépendants, de telle manière que la calibration d'une méthode visant à diminuer le taux de faux négatifs augmente inexorablement le taux de faux positifs en contrepartie, et inversement.

[14] Source : https://github.com/davidsandberg/facenet

[15] Source : https://www.rtl.fr/actu/justice-faits-divers/lyon-la-reconnaissance-faciale-utilisee-lors-d-un-proces-suscite-le-debat-7798350513, consulté le 24.04.2020

Tout l'enjeu se situe alors à trouver l'équilibre entre les taux de faux positifs et de faux négatifs servant le mieux l'objectif final du dispositif de reconnaissance faciale. Par exemple, dans le cadre d'une veille opérationnelle en renseignement visant à repérer un suspect potentiel, la priorité est de s'assurer de ne pas manquer un suspect potentiel (faux négatifs), au risque de considérer à tort d'autres personnes (faux positifs). En revanche, dans le cadre de l'évaluation d'éléments de preuve en vue de leur présentation au tribunal, il est essentiel de limiter le taux de faux positifs, car il est associé au risque de condamner une personne à tort.

II.2.3 Enjeux liés aux objectifs

L'idée de systèmes entièrement automatiques et généralisables, véhiculée par les stratégies commerciales et les représentations médiatiques, occulte les spécificités et enjeux propres aux différents objectifs poursuivis en matière de sécurité. Nous distinguons deux types d'enjeux : ceux liés au renseignement, à l'investigation et à la prévention, et ceux liés à l'évaluation.

II.2.3.1 Renseignement, investigation et prévention

Comme décrit précédemment, les applications de la reconnaissance faciale à des fins de renseignement, d'investigation et de prévention, sont principalement dédiés à l'identification de personnes ou à leur supervision dans une séquence vidéo. La poursuite de ces objectifs implique deux conséquences : la nécessité de disposer de bases données d'images et le déploiement de dispositifs sur l'espace public. Ces conséquences soulèvent alors un certain nombre d'enjeux qu'il s'agit de considérer.

Tout d'abord, la manière dont est construit et organisé le stockage des données utilisées par les systèmes de reconnaissance faciale a des impacts sur leur application (Marzouki, 2001). Parmi les caractéristiques liées au stockage, l'enjeu principal se situe au niveau des critères d'inclusion dans la base de données. Ceci est d'autant plus problématique dès lors que le dispositif de reconnaissance faciale est utilisé par les services de police. Ces critères peuvent être plus ou moins larges selon leur compréhension. Par exemple si l'un des critères est d'être une personne recherchée par la police et la justice pour figurer dans la base de données, la limite pourrait être fixée aux personnes sous mandat d'arrêt. Mais cette délimitation peut s'avérer plus large si cela concerne également les personnes suspectées d'un crime sous la seule décision d'un enquêteur (Fussey et Murray, 2019). En allant encore plus loin, des personnes considérées comme étant à risque sans implication dans un délit sont susceptibles d'être incluses. Les tests menés au Pays de Galle (Davies et al., 2018) incluent une base de données qui catégorise les personnes en trois niveaux selon le degré d'intérêt de la police et les risques perçus envers la société. En plus des critères d'inclusion, le croisement des images avec d'autres sources d'information, notamment étatiques, peut poser de sérieux dangers en matière de protection des données et de respect de la sphère privée.

Contrairement à d'autres mesures biométriques, le consentement nécessaire à la collecte de données est très difficile à obtenir lors de l'utilisation de la reconnaissance faciale dans les espaces publics (Castelluccia et Le Métayer, 2019). Le consentement peut intervenir à deux moments distincts : à l'étape de capture (enrôlement), c'est-à-dire au moment où l'image de la personne

est enregistrée et à l'étape de comparaison entre deux ou plusieurs images (reconnaissance) (Marzouki, 2001). Au moment de l'enrôlement, le dispositif nécessite la présence de la personne, mais pas de coopération de sa part pour enregistrer son visage, à l'inverse par exemple de dispositifs utilisant la reconnaissance par empreinte digitale. Ainsi, la capture du visage peut être prise à l'insu de la personne. En termes de protection des données et de la sphère privée, il s'agit là d'un enjeu majeur. L'information à l'aide d'affiches indiquant la présence de caméra et/ou de dispositifs de reconnaissance faciale ne peut en elle seule suffire comme mode de consentement. En effet, le consentement nécessite un choix libre et éclairé. La pose de panneaux est une information passive, c'est-à-dire que la personne est susceptible de ne pas avoir connaissance de sa présence. De plus, même en ayant connaissance de la présence du dispositif, il est difficile de connaître son étendue, le champ de vision des caméras étant inconnu. La notion de choix est également mise à mal dès lors que l'alternative est de ne pas se rendre dans la zone concernée. Du moment que le dispositif de reconnaissance faciale est actif dans un espace public, un consentement total devrait offrir aux citoyens des moyens alternatifs de continuer leur chemin sans emprunter la zone couverte par les caméras. Le refus de consentement en évitant les zones surveillées ou en se masquant le visage est également susceptible d'entrainer des réactions de suspicions de la part des forces de l'ordre dès lors que ces actions peuvent découler d'un effet préventif provoqué par le dispositif (Fussey et Murray, 2019). La question du consentement s'inscrit de manière plus large dans la transparence liée au dispositif. De manière similaire à l'effet boîte-noire qui entoure les algorithmes automatiques, une certaine opacité entoure les dispositifs qui font appel à la reconnaissance faciale, notamment en termes de processus. La transparence peut être alors comprise comme le degré d'information à disposition du public, notamment concernant l'utilisation du système de reconnaissance faciale, ses objectifs, ses localisations et son fonctionnement. Cette transparence joue un rôle important dans la perception de la légitimité du dispositif, et dans le contrôle (*accountability*) des acteurs qui le gèrent (services de police, sécurité privée, etc.) (Fussey et Murray, 2019). L'enjeu se situe alors à deux niveaux : (i) déterminer la quantité d'information à mettre à disposition des citoyens sans compromettre l'utilité opérationnelle du dispositif et (ii) impliquer activement la société civile tant dans la conception que la mise en œuvre des dispositifs de reconnaissance faciale dans les espaces publics. Ces différents enjeux liés à la reconnaissance faciale dans la prévention, le renseignement et l'investigation gravitent autour des libertés individuelles, notamment en termes de protection de données et du respect de la sphère privée. Malgré les risques générés, des recommandations à destination des services publics sont susceptibles de limiter ces risques et de guider la mise en œuvre de reconnaissance faciale dans le contexte d'action de sécurité (Castelluccia et Le Métayer, 2019 ; Dupont *et al.*, 2018) :

- Garantir l'accès au code source et aux algorithmes, afin d'éviter le phénomène de boîte-noire. Dans le cas d'algorithmes de *deep learning*, leur fonctionnement demeure opaque, il est alors nécessaire de passer par une validation empirique (analyse des résultats) à la place d'une validation analytique (analyse du système)

- Mettre en place des bases de données d'entrainement adaptées et représentatives de la diversité (notamment ethnique) et garantir leur accès pour éviter les biais de discrimination
- Vérifier les variables utilisées par les modèles pour calculer les scores (explicabilité) et mises en place de mesures de contrôle pour répondre à l'enjeu de transparence
- Questionner la légitimité du dispositif et sa proportionnalité par rapport à l'objectif poursuivi
- Évaluer la pertinence (adéquation), l'efficacité (effet) et l'efficience (coût-bénéfice) du dispositif
- Évaluer l'impact de l'adoption de systèmes de reconnaissance faciale tant sur les usagers que les professionnels pour faciliter leur intégration et acceptation au sein de la société civile

Ces recommandations certes idéalistes devraient néanmoins faire partie des fondations des recherches futures et des applications de systèmes biométriques automatiques, afin que les progrès technologiques – actuellement exponentiels – soient suivis de près par leur maitrise dans le cadre judiciaire.

II.3. Interprétation de la preuve scientifique

II.3.1 Approche probabiliste

Lors de l'évaluation forensique, le but du scientifique est d'évaluer le poids accordé à des propositions mutuellement exclusives en regard des éléments de preuve à disposition. Dans cette optique, l'utilisation d'une approche probabiliste, décrite par Aitken et Taroni (2004) et Robertson *et al.* (2016), est recommandée puisqu'elle permet d'actualiser les connaissances *a priori* par l'ajout d'éléments indiciaires. Cette approche est communément utilisée pour l'interprétation de nombreuses catégories de traces telles que les traces papillaires, les traces de semelles, la reconnaissance de locuteurs, l'ADN, les écritures et signatures ainsi que les armes à feu.

Le théorème de Bayes permet de calculer ces probabilités *a posteriori* en multipliant le LR associé à un élément de preuve par les probabilités d'observer chaque proposition alternative mutuellement exclusive, appelées *a priori*. Le théorème est formulé comme suit par Aitken et Taroni (2004) :

$$\underbrace{\frac{P(H_1|E,I)}{P(H_2|E,I)}}_{\textit{a posteriori}} = \underbrace{\frac{P(E|H_1,I)}{P(E|H_2,I)}}_{\text{LR}} \times \underbrace{\frac{P(H_1|I)}{P(H_2|I)}}_{\textit{a priori}} \qquad \text{(a)}$$

Avec :

- E (*Evidence*), les concordances et discordances relevées lors de la comparaison entre une trace et une référence (p. ex. le score issu d'une comparaison d'images faciales).

- I, les informations circonstancielles apportées par l'enquête et les connaissances de l'expert.e
- H_1 et H_2 les propositions alternatives mutuellement exclusives, comme :

  H_1 : « la référence et la trace sont de source commune »

  H_2 : « la référence et la trace sont de sources différentes »

De la valeur du LR, strictement supérieur à 0, dépend le poids que celui-ci apporte à l'une ou l'autre des propositions alternatives. Un LR inférieur à 1 soutient la proposition H_2 alors qu'un LR supérieur à 1 soutient la proposition H_1. Ce poids est exprimé verbalement selon une échelle de correspondance décrite, par exemple, par l'*European Network of Forensic Science Institute* (ENFSI, 2015 ; Robertson *et al.*, 2016). Pour exemple, un LR de 2000 soutient très fortement la proposition H_1 et indique qu'il est 2000 fois plus probable d'observer ce score (E) si les deux photographies comparées représentent la même personne plutôt que de deux personnes différentes de la population pertinente. En outre, un LR de 1 apporte un poids équivalent à l'une et l'autre des propositions, c'est-à-dire l'indice scientifique en question n'apporte aucune information permettant de discriminer les propositions.

II.3.2 Rapport de vraisemblance basé sur le score

L'utilisation du rapport de vraisemblance (*likelihood ratio*, LR) est préconisée dans de nombreux domaines forensiques, dont ceux des traces digitales, de la reconnaissance vocale, des armes à feu, etc. Plus spécifiquement, et comme déjà mentionné, le calcul de LR à partir des scores de similarités générés par des systèmes biométriques a déjà été décrit notamment pour la voix (Botti *et al.*, 2004 ; Meuwly, 2001) et les traces digitales, avec l'utilisation du système AFIS (Egli, 2009).

Pour la comparaison de visages, l'indice à considérer est le score de similarité qui traduit numériquement la distance qui sépare les deux images. Dans ce cas, le terme E dans l'équation (b) est exprimé par le score *s(S,T)*, similarité entre les vecteurs des images *S* et *T* (Bolck *et al.*, 2015).

Dès lors, le LR peut être décrit comme le ratio de ces densités de probabilités sous la forme suivante :

$$\mathrm{SLR(s)} = \frac{\mathrm{f(s(S,T)|H_1, I)}}{\mathrm{f(s(S,T)|H_2, I)}} \rightarrow \frac{\mathrm{(Densité\ de) Probabilité\ d'obtenir\ ce\ score\ si\ H_1\ est\ vraie}}{\mathrm{(Densité\ de) Probabilité\ d'obtenir\ ce\ score\ si\ H_2\ est\ vraie}} \quad \text{(b)}$$

Dans plusieurs domaines tels que la comparaison de traces papillaires ou d'enregistrements vocaux, les calculs de LR ont été étudiés à partir de scores de similarité générés par un système automatique, mais également à partir de la comparaison directe des caractéristiques observables sur les images. La méthode de calculs basés sur un score de similarité est la plus utilisée pour les cas de biométrie forensique (Meuwly *et al.*, 2017).

Nous présentons ici une formulation générale et simplifiée à titre d'exemple. En fait, la formulation d'une proposition d'alternatives et de ses informations conditionnelles I est l'une des

étapes les plus cruciales et elles doivent être adaptées à chaque cas en fonction des données et des informations dont on dispose. Le numérateur du SLR est conditionné par la variabilité intrasource (intravariabilité) et le dénominateur dépend de la variabilité inter-source (intervariabilité), qui à son tour dépend de I. L'intravariabilité représente les variations des valeurs de score lorsque l'on compare plusieurs images de la même personne, et l'intervariabilité est les variations de ces valeurs qui se produisent lorsque l'on compare des images de différents individus.

II.4. Synthèse

Avec le développement croissant des algorithmes automatiques et de l'intelligence artificielle, la reconnaissance faciale occupe une place toujours plus importante dans un contexte d'action de sécurité et dans la société en général. Cependant, la plupart des sources disponibles concernant son utilisation se résume aux articles de presse ou à des rapports de travail institutionnels. Il existe une réelle lacune en matière d'études empiriques sur les différentes utilisations de ces technologies en sciences criminelles et sur la perception de celles-ci par les acteurs qui les mettent en œuvre. Les effets bénéfiques apportés en matière de prévention, de renseignement, d'investigation et d'évaluation sont accompagnés d'enjeux et de limites qu'il s'agit de considérer, au risque de compromettre une mise en œuvre adéquate de dispositifs de reconnaissance faciale. Comme démontré à travers ce chapitre, il est crucial de comprendre et de délimiter les enjeux qui diffèrent fortement selon le niveau d'analyse considéré. Dans le cadre de notre recherche, nous nous focalisons sur les enjeux liés aux images ainsi qu'aux objectifs d'investigation et d'expertise.

Chapitre III. Exploitation d'images faciales : Pratiques forensiques

Ce chapitre dresse un état des lieux actuel des types d'images exploitées dans le cadre forensique, des méthodologies automatiques et « manuelles » ainsi que des apports et des limitations inhérents à chacune.

III.1. Les images faciales en science forensique

La science forensique, ou criminalistique, est la discipline qui étudie les traces laissées lors d'activités litigieuses (Ribaux, 2014). L'importance et la portée de l'utilisation des images en science forensique ont été décrites en détail par Milliet (2017). Dans le domaine de la reconnaissance faciale, la trace est l'image de question enregistrée au moment des faits investigués, sur laquelle peut apparaitre le visage de l'auteur des faits, ou de toute autre POI liée aux faits - victimes, suspects ou témoin. Le type et la qualité de ces images varient selon les scénarios de l'activité litigieuse, que nous décrivons ci-dessous.

III.1.1 Images témoins

Les « images témoins » sont définies par Milliet et collègues comme des images ambiguës liées d'une manière ou d'une autre à un événement, enregistrées par hasard ou délibérément par un individu ou un dispositif (Milliet *et al.*, 2014). Il s'agit donc, par exemple, d'images enregistrées sur des appareils mobiles par de véritables « témoins oculaires » lors d'un événement. Dans la plupart des cas, ces images sont de qualité moyenne ou médiocre, en raison des mouvements de l'opérateur et/ou des sujets et des conditions générales de prise de vue non contrôlées. Par exemple, ces enregistrements sont immédiatement recherchés par les autorités après un attentat à la bombe pour tenter d'identifier des personnes d'intérêt et reconstituer la chaîne des événements. Les images de témoins peuvent également faire référence à des images extraites de réseaux sociaux ainsi qu'à du matériel pédopornographique provenant de sources en ligne ou d'appareils appartenant à un POI. Ceux-ci peuvent aider à trouver une personne non seulement en utilisant la reconnaissance faciale, mais aussi n'importe quelle information de l'arrière-plan de l'image pour localiser où (et quand) la photo a été prise ou tout autre attribut du POI. Un autre type d'images témoins sont des séquences de surveillance, enregistrées à tout moment par des caméras fixes placées dans les points chauds.

III.1.2 Caméra de surveillance

Avec la constante hausse de l'utilisation de caméras de surveillance dans les secteurs privés (p. ex. magasins, banques) et publics (p. ex. transports, voies publiques), les images CCTV font désormais partie des traces les plus souvent disponibles dans une enquête comme nous l'avons déjà évoqué. Un problème récurrent concerne le fait que les images fournies (par le propriétaire aux enquêteurs, ou dans un second temps des enquêteurs à un.e expert.e) ne sont pas les séquences brutes enregistrées par l'appareil mais des captures d'écran ou des images dont le format a été modifié (compression due à un envoi par mail, à la copie de l'image dans un fichier texte ou dans un PDF, etc.). Cela complique la comparaison en diminuant la qualité de l'image (distorsion, compression, baisse de résolution, etc.). En outre, ces caméras sont le plus souvent installées en hauteur, sur un mur ou plafond, et enregistrent donc les images en plongée, ce qui peut entrainer une distorsion des proportions des éléments.

Figure 6 : Images CCTV d'un suspect de vol de carte bancaire et escroquerie (Belgian Federal Police[16]).

Un cas particulier d'images CCTV concerne les enregistrements aux distributeurs automatiques de billets (ATM). Celles-ci sont prises à très courte distance du sujet avec un objectif « *hypergon* », qui permet un grand angle de prise de vue malgré la proximité de l'appareil avec le visage mais déforme les images en les élargissant toutes au centre par rapport aux bords. Des images ATM peuvent être fournies dans des cas tels que les vols où la carte de crédit et/ou les billets de banque d'une victime sont volés lors d'un retrait ainsi que pour des utilisations frauduleuses de cartes de crédit. Un dernier type d'images dites de surveillance provient des investigateur.trice.s chargés de l'observation de personnes d'intérêt dans l'enquête, qui enregistrent des images à distance.

III.1.3 Documents officiels

Lorsqu'elles n'impliquent pas d'images de vidéosurveillance, les demandes de comparaisons de visages concernent majoritairement l'utilisation de faux documents d'identité, notamment lors de contrôles douaniers, de demandes de permis de séjour ou d'autres privilèges d'État (Dessimoz & Champod, 2015). Dans la plupart des cas, les documents de référence fournis sont des passeports ou des cartes d'identité authentifiés. Dans ce scénario, la comparaison est facilitée par le fait que les images de trace et de référence sont deux documents d'identité avec des photographies prises dans des conditions normalisées, mais parfois très espacées dans le temps.

[16] Disponible sur https://www.youtube.com/watch?v=wDLKPha24iY&list=PLC6D98801D730A8A4&index=66

III.2. Comparaison manuelle de visages

III.2.1 Les méthodes

La comparaison manuelle d'images est utilisée pour des tâches nécessitant d'analyser peu d'images simultanément, par exemple lors de contrôles d'identité (comparaison d'une personne avec un document d'identité) et lors d'investigations judiciaires (comparaison d'une image de vidéosurveillance avec la photographie signalétique d'un individu). La littérature spécialisée décrit quatre méthodes d'analyses pour la comparaison manuelle de visages par un.e expert.e : l'analyse holistique, morphologique, photo-anthropométrique et la superposition des deux images comparées (Ali *et al.*, 2012b).

L'analyse holistique est la description globale des caractéristiques du visage, sans aucune mesure. Cette méthode est notamment utilisée par un officier lors de contrôles d'identités à la douane, par exemple. Cela permet une comparaison rapide mais également très dépendante de la qualité des images et du degré de difficulté de la comparaison (forte ressemblance de personnes différentes). Dans un contexte judiciaire, il est recommandé de n'utiliser cette méthode qu'en première ligne, en la combinant aux approches suivantes, en raison de ces performances faibles et variables (FISWG, 2012).

L'analyse morphologique consiste à décrire les formes et proportions des éléments du visage (front, yeux, nez, bouche, sourcils, oreilles, joues, menton) de la manière la plus précise possible. Cela comprend également la description des marques faciales (c.-à-d. rides, cicatrices, tâches de rousseurs, grains de beauté, tatouages, etc.) ainsi que la couleur et la longueur des cheveux, l'implantation et la densité de la pilosité. Lors d'expertises forensiques, cette pratique est considérée comme la plus performante, applicable aux images de qualité variable et dont le processus est facilement explicable aux personnes non spécialisées. Elle souffre cependant d'un manque de standardisation, ainsi que d'une trop grande dépendance à la capacité innée de l'expert.e à reconnaitre les visages, et à la qualité de l'image (Noyes *et al.*, 2017 ; Peng, 2019 ; Phillips *et al.*, 2018). La capacité innée d'une personne à reconnaitre des visages prévaut de plus en plus face à des études comme celle de Towler et ses collègues, où les auteurs mettent en évidence l'inefficacité des programmes d'entrainement aux tâches de reconnaissance faciale dédiées aux expert.e.s (Towler *et al.*, 2019).

La troisième approche, l'analyse photo-anthropométrique, repose sur les mesures verticales et horizontales de distances entre certains points du visage tels que la base du menton, le centre de la bouche et des yeux, les commissures des lèvres, les bords des ailes du nez, etc. (Moreton et Morley, 2011). Enfin, la technique de superposition a pour but de visualiser plus directement les similarités et les discordances morphologiques et anthropométriques des visages sur les deux images.

La littérature recommande néanmoins de ne pas utiliser seules les approches anthropométriques et de superposition d'images à des fins de comparaison, mais de les combiner aux autres approches (Edmond *et al.*, 2009 ; ENFSI, 2018 ; FISWG, 2012). En cause, i) la nécessité d'utiliser des images prises dans des conditions optimales, ce qui est rarement le cas ii) le caractère très

chronophage de ces processus iii) l'entrainement avancé dont l'opérateur a besoin, si tant est qu'il soit efficace, et iv) le manque de fiabilité des résultats.

À titre d'exemple, il suffit de regarder les trois photographies de la figure 1, qui représentent le même mannequin pris sous trois distances et focales différentes.

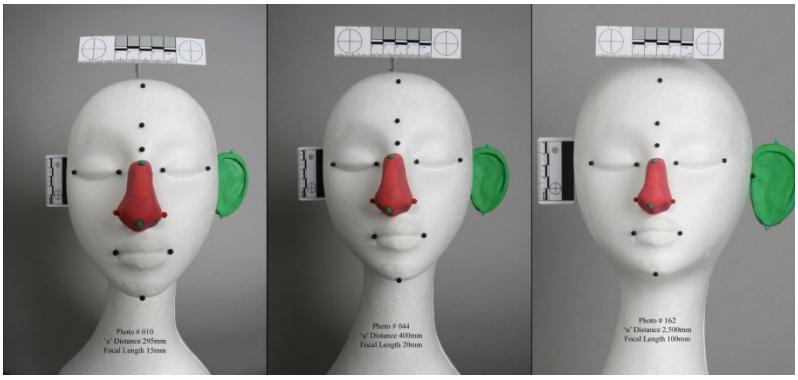


Figure 7 : Illustration des variations de morphologie faciale apparents sous trois perspectives différentes. Le sujet photographié ne change pas, seules la focale et distance de prise de vue varient (de gauche à droite, distance (mm) : 295, 400 et 2'500, focale (mm) : 15, 20 et 100) (Edmond et al. 2008).

Cela démontre très rapidement le problème lié à l'analyse de mesures morphologiques dans la comparaison de visages, même pour des photographies de face, de type images signalétiques. Des images de type CCTV, ATM, etc., sont non seulement de moindre qualité mais subissent également des distorsions dues à l'angle de prise de vue non perpendiculaire au sujet. C'est autant de paramètres qui vont déformer chacune des distances entre différents points fixes du visage.

Indépendamment de la méthode utilisée, les différents éléments comparés sont ensuite classés selon leur pouvoir discriminant (Ali *et al.*, 2012b). Des caractéristiques jugées très discriminantes par les spécialistes possèdent un poids plus important lors de l'évaluation des résultats, car il s'agit d'éléments plus rarement rencontrés dans une population cible ou très variables d'un individu à l'autre. Par exemple, une similarité dans la couleur des yeux renforce d'autant plus l'hypothèse selon laquelle les deux images représentent la même personne lorsqu'ils sont verts plutôt que marron, étant donné la différence de rareté de ces deux couleurs dans les populations.

Dans le cadre judiciaire, les expert.e.s basent leurs comparaisons sur les approches recommandées par les organismes compétents (ENFSI, 2018 ; FISWG, 2012). Ces derniers conduisent régulièrement des tests de performances afin d'évaluer les résultats des protocoles individuels, et adapter ses recommandations méthodologiques. Naturellement, comme évoqué précédemment, des limitations persistent, aussi bien en termes de standardisation des méthodes que dans leur application dans le cadre opérationnel.

III.2.2 Problématiques opérationnelles et académiques

III.2.2.1 Standardisation des méthodes

Le manque de standardisation est un frein à l'utilisation généralisée de la reconnaissance faciale aussi bien dans l'enquête qu'au procès. Cela concerne à la fois l'unification des méthodes d'analyse, de comparaison et d'évaluation, et les formations dispensées aux expert.e.s.

Aux États-Unis, la standardisation méthodologique est promue par le groupe de travail sur l'identification faciale (*Facial Identification Scientific Working Group*, FISWG) à travers la publication de glossaires, de guides pratiques et de recommandations destinées à l'usage des praticiens forensiques et judiciaires. L'objectif est le développement de standards, l'harmonisation des méthodes et l'échange d'informations entre praticiens et chercheurs.

En Europe, la rédaction de tels documents est assurée par le réseau européen d'instituts en science forensique (*European Network of Forensic Science Institute*, ENFSI). Le guide pratique publié en 2018, *Best Practice Manual for Facial Image Comparison*, couvre la pratique de comparaisons manuelles de visages, et non de l'utilisation de systèmes automatiques. De plus, il est spécifié qu'il ne s'agit pas d'une procédure opérationnelle standard, et que cela ne répond que de manière générale aux exigences des systèmes judiciaires.

Ces documents s'inscrivent dans la démarche d'unification et de standardisation des méthodes de travail et d'utilisation de la preuve scientifique. Cependant, il n'existe pas encore de guide décrivant une méthode d'interprétation de la trace en tant qu'élément de preuve suite à l'utilisation de systèmes automatiques, comme c'est le cas pour la comparaison de locuteurs (p. ex., Drygajlo *et al.*, 2016 ; Morrison *et al.*, 2021).

III.2.2.2 Limitations pratiques

Bien que l'être humain soit habitué à reconnaitre des visages tout au long de sa vie, il rencontre plus de difficultés lorsque la personne à reconnaitre n'est pas de la même ethnie (Furl *et al.*, 2002 ; Meissner *et al.*, 2005). Une difficulté supplémentaire se présente dans le cas où plusieurs années se seraient écoulées entre les deux photographies à comparer, d'autant plus pour les individus très jeunes dont la morphologie varie plus rapidement (Michalski et Snyder, 2019). De plus, les variations liées directement au matériel d'acquisition de l'image, tel que les caméras de surveillance, les smartphones, etc., ainsi qu'aux conditions dans lesquelles les images sont prises, sont autant de contraintes auxquelles l'expert.e doit faire face. Ainsi, les expressions du visage, l'angle de prise de la photographie, la luminosité et la qualité de l'image mais également les accessoires tels que des lunettes peuvent compliquer, voire rendre impossibles, les phases d'analyse et de comparaison.

Ces différents facteurs peuvent rendre non différentiables les photographies de deux personnes différentes, ou à l'inverse mettre en évidence des discordances entre deux photographies de la même personne.

La littérature scientifique ne décrit pas les cas où les conclusions d'expertises manuelles en comparaison de visages se sont avérées erronées par la suite. Cependant, le cas Talley décrit dans

« The Intercept »[17] met en exergue les problèmes liés à la fois aux fausses identifications des expertises manuelles en comparaison de visages et à la « toute-puissance » de la parole de l'expert.e aux États-Unis.

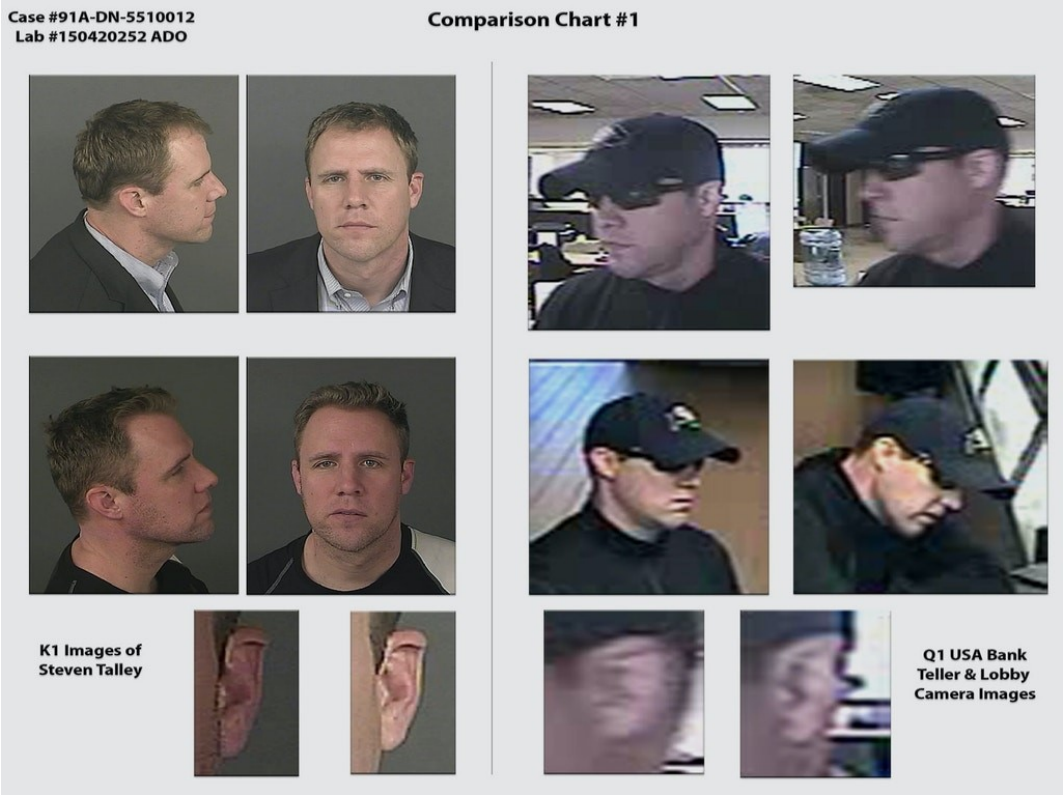


Figure 8 : Images traces de type CCTV (à droite) montrant le suspect d'un cambriolage, et mugshots de référence de la POI (à gauche) du Federal Bureau of Investigation. Source : The Intercept (13.10.2016).

En 2014, Steve Talley est accusé d'être l'auteur d'un cambriolage de banque sur la base de la comparaison de visage à partir d'images CCTV enregistrées lors des faits, bien que le témoignage de la personne ayant fait face à l'auteur l'innocentait sur la base de plusieurs discordances physiques importantes (carrure, taille, poids, signes distinctifs sur les mains et le visage).

Il est rare qu'un procès porte exclusivement sur l'exploitation d'images, mais le manque de standardisation des méthodes d'analyse et de comparaison ainsi que de données statistiques fragilise les conclusions d'expertises en comparaison de visages.

III.2.2.3 Formation des praticiens

L'analyse et la comparaison effectuées par l'humain peuvent manquer d'objectivité et de transparence, dans la mesure où les observations et les résultats d'une même expertise ainsi que la manière de les documenter peuvent varier d'un expert.e à un autre selon leur expérience. Cela est souvent mis en lien avec le manque de formation dédiée aux professionnels de la comparaison de visages (Spaun, 2009), mais les études menées sur l'impact de la formation et de l'expérience

[17] Source : https://theintercept.com/2016/10/13/how-a-facial-recognition-mismatch-can-ruin-your-life/

sur les performances des expert.e.s et des non spécialistes en reconnaissance faciale ne vont pas toutes dans ce sens.

D'une part, certaines publications mettent en évidence de meilleures performances des expert.e.s dans des tâches de reconnaissance de visage (1vN) et de comparaison de visages (1v1), ainsi que l'impact positif d'un entrainement basé sur le feedback sur les performances de personnes spécialisées ou non (Kemp *et al.*, 1997 ; Spaun, 2009 ; White *et al.*, 2014a). Il y est également démontré que les personnes non spécialisées dans ce domaine et non entrainées peinent à performer dans cet exercice. Kemp et collègues ont également montré que lorsque des hôtes de caisse de supermarchés devaient vérifier l'identité d'une personne à partir des photographies de cartes bancaires, le taux d'erreur oscillait entre 34 et 64% (Kemp *et al.*, 1997).

À l'inverse, certaines études démontrent que les performances de personnes entrainées et non entrainées sont comparables. Par exemple, White et collègues ont démontré que des officiers des douanes achevaient un exercice de vérification d'identité à partir de passeports avec un taux d'erreur d'environ 20% et que les années d'expérience ne semblaient pas avoir d'incidence sur ce résultat (White *et al.*, 2014b). De plus, Burton a mis en avant un taux d'erreur de 25% lorsque des individus devaient associer un visage à une image correspondante parmi 10 à partir d'images de faible qualité venant de vidéosurveillances (Burton *et al.*, 1999). Des résultats similaires sont exposés dans Lee *et al.* (2009). Récemment, Towler et collègues ont directement remis en question l'efficacité de plusieurs programmes d'entrainement courts (max. une journée) (Towler *et al.*, 2019 ; Towler *et al.*, 2014). De plus, plusieurs équipes de chercheurs - notamment de l'UNSW (Australie), de l'Université de Dallas et du NIST (US) - étudient plus en détail les variations de performances entre individus dans les tâches de comparaisons de visages. Parmi leurs nombreux résultats, les auteurs démontrent l'existence de « *super-recognizers* » aux performances innées très supérieures à la moyenne, et l'amélioration des résultats lors de comparaisons issues de la collaboration entre plusieurs expert.e.s, ou entre l'expert.e et un système biométrique automatique (Noyes et O'Toole, 2017 ; Phillips *et al.*, 2018 ; Towler *et al.*, 2017).

Enfin, lors de la comparaison de visages l'opérateur relève et évalue les concordances et discordances visibles, selon lui, entre les deux images. Le principal problème que cette démarche soulève réside dans le fait qu'aucune donnée statistique n'existe concernant les variations de ces différentes caractéristiques d'un homme à un autre, d'une population ethnique à une autre, etc.

III.3. Systèmes automatiques de reconnaissance faciale

Dans le chapitre précédent, nous avons largement abordé l'utilisation de systèmes automatiques de reconnaissance faciale dans différents domaines des domaines civils et judiciaires. Dans cette section, nous nous focalisons sur les aspects techniques et progrès algorithmiques actuels, et sur les performances et les limitations liées à l'utilisation de ces systèmes dans le cadre forensique.

Cette section reprend des aspects développés dans l'article de Bollé, Casey et Jacquet (2020), « *The role of evaluations in reaching decisions using automated systems supporting forensic analysis* », publié dans le journal *Forensic Science International : Digital Investigation*.

III.3.1 Les algorithmes

Les systèmes automatiques basés sur des techniques d'apprentissage (*machine learning*) sont développés pour aider les humains à trouver des informations de manière plus efficace et efficiente dans des quantités massives d'informations.

Le principe d'intelligence artificielle (IA) est décrit pour la première fois par Alan Turing en 1936. Le *machine learning*, et encore plus spécifiquement le *deep learning*, sont des sous-catégories d'IA dont le principe existe depuis des décennies mais dont la nomenclature a évolué à travers le temps. Depuis 2006, le terme *deep learning* ainsi que le regain d'intérêt des chercheurs pour ce processus sont revenus sur le devant de la scène technologique. À l'heure actuelle, les réseaux neuronaux profonds ont surpassé les systèmes d'IA concurrents basés sur d'autres technologies de *machine learning*. Cette troisième vague de popularité des réseaux neuronaux se poursuit jusqu'à l'heure où nous écrivons ces lignes, bien que l'objectif de la recherche sur l'apprentissage profond ait changé de façon spectaculaire au cours de cette vague. La troisième vague a commencé par se concentrer sur les nouvelles techniques d'apprentissage non supervisé et sur la capacité des modèles profonds à bien généraliser à partir de petits ensembles de données, mais aujourd'hui, on s'intéresse davantage aux algorithmes d'apprentissage supervisé beaucoup plus anciens et à la capacité des modèles profonds à exploiter de grands ensembles de données étiquetées. Ces dernières années, sa popularité et son utilité ont considérablement augmenté, en grande partie grâce à des ordinateurs plus puissants, des ensembles de données plus volumineux et des techniques permettant de former des réseaux plus profonds (Goodfellow *et al.*, 2016).

L'étape d'apprentissage peut être supervisée, ce qui signifie que les règles sont dérivées d'un ensemble de données labellisées considéré comme une vérité de base, ou non supervisé, c.-à-d. que les règles sont apprises par l'algorithme à partir de données non labellisées.

III.3.2 Utilisation forensique et limitations

Le guide publié par le groupe FISWG pose les bases de l'utilisation de systèmes automatiques dans un cadre judiciaire, mais il est indispensable de mettre à jour les guides de bonnes pratiques pour prendre en compte l'évolution exponentielle qu'ont subie ces systèmes dans la dernière décennie (FISWG, 2012). Indépendamment, plusieurs rapports étudient les performances actuelles et le potentiel opérationnel de la reconnaissance faciale. Par exemple, un récent rapport de la Commission européenne évalue la possibilité d'intégration de systèmes automatique dans le système d'information de l'espace Schengen (*Schengen Information System*), qu'il conclut assez performant pour une application aux contrôles des frontières et, à terme, dans le cadre judiciaire (Galbally *et al.*, 2019 ; TELEFI Project, 2021). En parallèle, les rapports d'utilisation de LFR (*live face* recognition) en Angleterre et aux Pays de Galles présentent cette technologie comme un danger pour le respect de la vie privée, à la fiabilité médiocre dans certaines conditions (Davies *et al.*, 2018 ; Fussey et Murray, 2019).

D'un point de vue tant scientifique que juridique, tout système automatisé qui aide un forensicien à tirer des conclusions à des fins forensiques doit être transparent et reproductible (Margagliotti et Bollé, 2019). En outre, les praticiens en science forensique doivent pouvoir comprendre et expliquer les résultats de tels systèmes automatiques de manière claire, complète, correcte et cohérente (Casey, 2020). De nombreux systèmes automatiques existants manquent de transparence et de reproductibilité suffisantes à des fins forensiques, et ne sont pas conçus de manière à aider les forensiciens à évaluer et à expliquer efficacement les résultats de ces systèmes. La solution à ce problème n'est pas seulement technique, mais implique un processus structuré d'évaluation qui respecte les principes du raisonnement et de l'interprétation scientifiques. Outre le fait que les résultats des systèmes automatiques doivent être transparents, reproductibles et compréhensibles, il est nécessaire de guider les utilisateurs à travers un processus structuré d'interprétation scientifique.

Il n'existe aucune méthode standardisée et validée dédiée à l'utilisation de systèmes automatiques dans le cadre forensique (Meuwly *et al.*, 2017). Dans un but investigatif, l'importance est moindre puisque le résultat de la recherche est discuté avec les enquêteurs avant d'orienter l'enquête alors que dans un but évaluatif le résultat peut devenir un élément de preuve soutenu par l'expert.e lors d'un procès. Dans ce cas, le résultat ainsi que toute la procédure scientifique suivie par l'expert.e doivent répondre à des exigences juridiques, où l'utilisation de systèmes automatique reste peu abordée à ce jour notamment à cause du manque de validation des méthodes exploitées.

À l'inverse des systèmes de reconnaissance faciale décrits précédemment, les logiciels utilisés en science forensique ne sont pas destinés à générer un résultat binaire basé sur la comparaison d'un score avec un seuil de tolérance fixé par l'utilisateur. C'est à l'opérateur d'interpréter le résultat traduit par le score de similarité sans que son jugement ne soit orienté par la connaissance d'informations d'enquête circonstancielles qui peuvent l'amener à manquer d'impartialité (Edmond *et al.*, 2009).

De plus, les systèmes commerciaux, aussi performants soient-ils, fonctionnent à partir d'algorithmes propriétaires. L'effet boîte-noire que cela impose représente un frein potentiel à l'utilisation de tels systèmes devant un tribunal où le besoin de transparence dans les démarches scientifiques est omniprésent. Cet effet devrait être endigué par la publication d'études empiriques attestant des hautes performances de ces systèmes afin d'en valider l'utilisation dans un cadre forensique. Malgré cela, l'utilisation et l'acceptation de systèmes boîte-noire dans la sphère juridique sont ralenties par le manque intrinsèque de transparence, comme par exemple dans les domaines de l'ADN (Imwinkelried, 2017).

D'autre part, bien que la structure exacte d'un tel algorithme soit inconnue, il est entendu que la majorité d'entre eux ne prend pas en compte les *soft biometrics*, informations primordiales pour l'expert.e telles que le genre perçu, l'ethnie, et la présence de signes distinctifs (cicatrices, taches de rousseur, grain de beauté, etc.). Néanmoins, l'algorithme prend en compte d'autres éléments que l'expert.e n'exploite pas et permet de traiter de larges bases de données. Par conséquent, dans le cadre investigatif de l'authentification d'un visage face à une base de données, il représente un outil complémentaire au travail de l'expert.e et pas une alternative à celui-ci dans

la mesure où l'opérateur étudie la liste des résultats générés par le système. Dans le cadre présent, où l'optique évaluative est de calculer un LR directement à partir de ces scores, l'approche et la finalité étant différentes, aucune étape intermédiaire manuelle n'est requise.

III.4. Synthèse

Dans ce chapitre, nous avons abordé, à partir de la littérature actuelle, l'état de l'art de la reconnaissance faciale à des fins forensiques par l'utilisation de systèmes automatiques ou de méthodologie de comparaison manuelle. Nous nous concentrons davantage sur les techniques automatiques telles que l'utilisation de systèmes biométriques que nous explorons dans la présente recherche. Pour plus de détails sur les méthodologies manuelles, Zeinstra et ses collègues ont offert en 2018 une étude détaillée sur la comparaison de visages par l'humain à des fins forensiques et les ensembles de données disponibles (Zeinstra *et al.*, 2018). Ce chapitre constitue le fondement de notre recherche, en identifiant les différents scénarios forensiques (vidéosurveillance, images témoins, documents officiels) et leur utilisation à des fins de renseignement, d'enquête ou d'expertise. Dans les chapitres suivants, nous exposons les scénarios (types d'images), les objectifs (cadre d'utilisation), les méthodes de comparaison (systèmes automatiques) sur lesquels nous avons choisi de focaliser notre recherche.

Chapitre IV. Problématiques et matériel

Cette section décrit les questions sur lesquelles est basée cette recherche ainsi que les images et systèmes de reconnaissance faciale que nous avons choisis pour y répondre.

IV.1. Problématiques de recherche

Les chapitres précédents ont permis de mettre en évidence de nombreuses problématiques sur lesquelles doit s'orienter la cherche pour mieux comprendre et optimiser l'application de la reconnaissance faciale automatique dans le cadre judiciaire. Le présent projet a été développé autour de trois problématiques principales (1-3), desquelles émergent des questions sous-jacentes (a-c) :

1) **Comment choisir une approche selon la quantité et la qualité de données à disposition, et quels sont les impacts sur les résultats ?**

Le calcul de SLR se fait idéalement à partir d'images multiples et de bonne qualité, mais ces conditions sont rarement rencontrées dans le cadre opérationnel. Notre but est de tester plusieurs approches en fonction du type et du nombre d'images à disposition, simulant ainsi les conditions rencontrées à différents moments de l'instruction judiciaire.

2) **Le potentiel des systèmes automatiques comme aide à l'investigation**

Les systèmes automatiques sont un moyen d'effectuer des tâches fastidieuses, comme la recherche de personnes dans de larges bases de données, de manière efficiente, dans un temps réduit et sans monopoliser d'opérateur humain. De nombreux systèmes de reconnaissance faciale automatique existent et sont encore en développement, et sont basés sur des structures et fonctionnements différents. Lesquels répondent au mieux aux besoins opérationnels – gain de temps, faible coût, facilité de prise en main – dans le cadre investigatif ? À savoir :

   a) **Algorithmes *rule-based versus deep learning* ?**
   b) **Systèmes commerciaux *versus* systèmes en source ouverte ?**

3) **Quels sont les apports de l'évaluation probabiliste du score de comparaison dans le cadre judiciaire ? Peut-elle être utilisée dans le cadre de l'investigation comme au tribunal ?**

Nous postulons que les avantages, notamment en termes de fiabilité et de reproductibilité, du calcul de SLR en phase d'expertise forensique peuvent (et devraient) être étendus à la phase d'enquête. Cela renforcerait le poids des résultats de recherche de POI dès le début de l'instruction. Le choix d'une approche de calcul de SLR adéquate est le premier point essentiel, et cela soulève une problématique secondaire :

   c) **Les avantages du SLR sont-ils suffisants pour une utilisation en phase d'enquête ? Est-il possible de proposer une estimation de SLR d'expertise dès l'investigation ?**

Fournir aux enquêteurs des résultats probabilistes dès la phase d'enquête, à titre d'évaluation préliminaire, permettrait de pondérer et combiner les informations de manière plus pertinente afin que les investigateur.trice.s aient une meilleure idée de ce à quoi ils peuvent s'attendre en termes de résultat d'expertise en impliquant tel ou tel POI.

Ces problématiques cristallisent les axes de recherche autour desquels nous avons développé notre méthodologie. Nous détaillons dans la section suivante le matériel expérimental que nous avons choisi pour acquérir et exploiter des images de qualité variable et qui répondent à des contraintes de scénarios forensiques pertinents, et comparer des systèmes automatiques de reconnaissance faciale de différente structure et méthode de distribution.

IV.2. Scénarios forensiques

Il est primordial que ce projet réponde au mieux aux demandes opérationnelles rencontrées par les services de police et les expert.e.s en comparaison de visages. En Suisse par exemple, les cas pour lesquels le plus d'images sont intégrées aux bases de données de la police concernent des fraudes aux documents d'identité, des opérations frauduleuses au retrait automatique de billets, ainsi que l'exploitation d'images enregistrées par des caméras de type CCTV (Dessimoz et Champod, 2015). Selon ces situations, les images fournies diffèrent principalement par leur qualité, l'angle de prise de vue et les distorsions inhérentes au type d'appareil.

La priorité est d'acquérir différents types d'images d'au moins 25 individus masculins, pour permettre d'étudier les cas les plus majoritairement rencontrés. Au final, 34 volontaires ont participé à ce projet (25 hommes et 9 femmes).

L'objectif de la collecte est de constituer un jeu de données représentatif de cas pour lesquels une expertise en comparaison de visages peut être requise (fraudes aux documents d'identité, retrait frauduleux à un distributeur automatique de billets et à la banque). Les volontaires avaient donc pour mission d'être successivement un suspect et l'auteur des faits pour chaque scénario.

La collecte de données s'est déroulée en 3 phases, chacune destinée à un type d'images : les images standardisées et portraits, les images signalétiques (*mugshots*) et les images de vidéosurveillance.

IV.2.1 Images de question

Les images de question sont les images prises sur la scène lors des faits. Ce projet se concentre sur trois types d'images : des images signalétiques, des images enregistrées par des distributeurs automatiques de billets ainsi que par une caméra de type CCTV.

IV.2.1.1 Fraudes aux documents d'identité

Dans les cas de fraudes aux documents d'identité, notamment lors de contrôles douaniers et de demandes de permis de séjour, les documents fournis à des fins de comparaison sont des passeports ou des cartes d'identité. Ce scénario permet une comparaison facilitée par le fait que les documents de question et de référence sont deux documents d'identité avec des photographies de type signalétique. Dans certains cas plus rares, une photographie de la personne d'intérêt peut avoir été prise par les autorités douanières dans des conditions contrôlées (l'angle, la luminosité et la qualité de l'image ainsi que la pose du sujet sont comparables à ceux de la photographie de référence) et fournie comme image de question.

Collecte d'images standardisées et portraits (scénario « ID »)

Chaque participant.e devait fournir au minimum quatre images plus ou moins standardisées, c'est-à-dire des photographies d'identité ainsi que des portraits et/ou des « selfies ». Toutes ces images ne devaient pas nécessairement être récentes, mais en évitant les images prises pendant l'enfance des participants. Les images étaient fournies avec une résolution minimale de 200x200 pixels aux formats JPEG, PNG, TIFF ou PDF (Les images fournies sous un format PDF ont été ensuite extraites au format TIFF).

a. Photos standardisées : carte d'identité, passeport, permis de séjour, permis de conduire, carte d'université, abonnement de transport, etc.

Figure 9 : Exemples de copies scannées de photos standardisées fournies par les participant.e.s.

b. Photos type portraits, soit toutes images frontales, de bonne qualité et de sources variables, sans accessoires dissimulant partiellement le visage (p. ex. lunettes de soleil, voile, masque, cagoule).

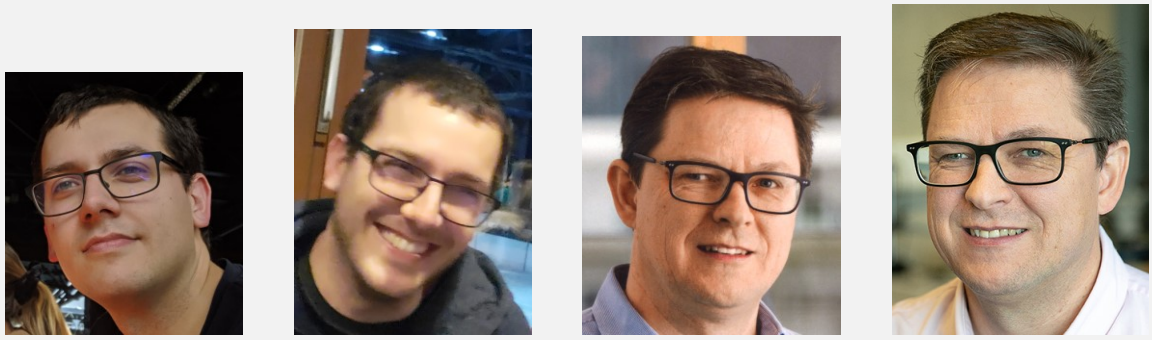


Figure 10 : Exemples de portraits fournis par les participant.e.s : à gauche, deux photographies issues de réseaux sociaux, à droite, deux portraits professionnels.

IV.2.1.2 Caméra-surveillance grand-angle

Avec l'augmentation constante de l'utilisation de caméra de surveillance aussi bien dans le domaine privé (magasins, banque, etc.) que public (transports, voie publique, etc.), ces images font désormais partie des éléments souvent disponibles lors d'une investigation. Dans les cas où un délit frappe une banque ou un distributeur automatique de billets en Suisse, les images sont systématiquement remises à la police par le biais d'un procureur, contrairement aux images

issues de caméras de locaux commerciaux, remises uniquement sur demande de l'enquêteur (Rossy et al. 2013). De plus, dans la majorité des cas, les magasins ne fournissent pas les fichiers bruts contenant les images ou vidéos originales enregistrées mais, comme nous l'avons vu, des captures d'écran ou des images sous un format différent, ce qui pose problème pour l'exploitation de l'image qui peut avoir subi des distorsions lors de l'enregistrement ou une perte d'informations lors de la capture d'écran. Dans le cas des banques, les images brutes sont plus souvent accessibles. Confirmer l'identité d'une personne peut se révéler difficile, principalement à cause de la qualité moyenne voire mauvaise des images enregistrées par ce type d'appareils, principalement lorsque le format de l'image n'est pas le format original. Aussi, ces caméras sont le plus souvent installées en hauteur et enregistrent donc des images en plongée, ce qui peut mener à une déformation des proportions des éléments filmés.

IV.2.1.3 Distributeurs automatiques de billets

Un cas particulier d'images CCTV est celui des images enregistrées par les distributeurs automatiques de billets (ATM). En effet, celles-ci sont enregistrées à très faible distance du sujet avec un objectif hypergone ce qui permet d'avoir un angle de prise de vue assez large malgré la grande proximité de la caméra avec le visage filmé mais qui amène une distorsion de l'image. Cette déformation change toutes les proportions de l'image en élargissant ce qui se trouve au centre par rapport aux bords. Des corrections numériques peuvent être appliquées, mais pourraient rendre plus difficiles les étapes de comparaison et empêcher d'effectuer une comparaison par superposition, à moins d'obtenir des images de comparaison prises sous les mêmes conditions que l'image de question. Les images ATM peuvent être fournies dans des cas tels que le vol pendant lequel une victime se fait dérober sa carte bancaire et/ou billets au cours de son retrait ainsi que dans le cas d'une utilisation frauduleuse d'une carte bancaire après son vol.

Collecte d'images de vidéo surveillance (scénarios « CCTV » et « ATM »)

Les participant.e.s ont pris part à une simulation de retrait d'argent à un ATM de la même manière que dans un établissement bancaire. L'enregistrement d'images de surveillance était effectué à l'aide de deux caméras différentes : une caméra de surveillance AXIS P3225-V MkII placée en hauteur pour une vue générale de la pièce, et une caméra AXIS *pinhole* P1264 MkII spécialement adaptée pour la prise d'images aux ATM (Figure 11). Les vidéos sont enregistrées aux formats MKV, puis cinq images sont extraites dans chaque vidéo pour être utilisées comme traces par la suite.

Figure 11 : Images extraites d'enregistrements par la caméra grand-angle P3225-V (en haut) et par la caméra pour distributeur de billets P1264 (en bas).

IV.2.2 Images de référence

Les images de référence correspondent aux images que l'expert.e mandaté est amené à demander à l'enquêteur pour les besoins de l'expertise lorsque le suspect ou plusieurs images signalétiques sont à disposition. Il peut s'agir de photographies d'identité judiciaire (*mugshots*), capturées par la police dans des conditions contrôlées lors de mises en détention, ou de photographies d'identité.

La BdD judiciaire regroupe 36'392 *mugshots* judiciaires standards des polices cantonales neuchâteloise et vaudoise. Il s'agit de photographies frontales, cadrées autour du visage et buste, prises sur un fond uniforme blanc à une distance focale d'environ 1.5-2m. Les POI sont majoritairement des hommes de type caucasien, puis de type maghrébin, subsaharien et asiatique. Environ 10% de la population est féminine.

Collecte d'images signalétiques (*mugshots*)

Les participant.e.s sont photographié.e.s dans des conditions contrôlées pour acquérir des photographies signalétiques de référence type *mugshots* pris par les forces de l'ordre. Les images étaient prises sur trois jours différents, avec une variation de distance entre le sujet et l'appareil de 1.5m à 2.5m, chaque fois sous cinq angles (face, droite, gauche, haut et bas). Les individus avec des lunettes de vue étaient photographiés avec puis sans leurs lunettes. Ces sessions distinctes permettent de ne pas sous-estimer la variabilité d'apparence de chaque participant en prenant des images successives trop semblables les unes aux autres.

Le but est de reproduire au mieux les conditions réelles de la prise de ces d'images par les forces de police car si ces photographies sont trop différentes de celles déjà présentes dans la BdD judiciaire (aussi bien en termes de luminosité de qualité que du fond utilisé derrière le sujet), le système automatique risque d'exclure trop rapidement les images des bases et de se focaliser uniquement sur les images ajoutées. Dans ce cas, la comparaison ne serait alors plus réellement effectuée contre les N images de la base mais sur le nombre Y d'images ajoutées.

Figure 12 : Photographies d'identité judiciaire (mugshots) prises au cours de trois séances pour 29 POI sans lunettes et 15 POI également avec lunettes.

Les types et nombres de toutes les images collectées et exploitées dans le cadre de ce projet sont résumés dans le Tableau 1, et détaillés en annexes (*cf.* Annexe B).

Tableau 1 : Types et nombres d'images collectées dans ce projet.

Utilisation	Types d'images	Nombre d'images par POI	Nombre total d'images	Particularité	Nombre de POI
Traces	CCTV / ATM	5 images par vidéo (3 vidéos)	480	Sans lunettes	32
			255	Avec lunettes	17
	ID	moyenne=8 (±3)	255	-	31
Références	*mugshots* frontaux	4	116	Sans lunettes	29
			60	Avec lunettes	15
BdD judiciaire	*mugshots* frontaux (Police)	1 à 3	36392	-	29742

IV.3. Matériel technique

IV.3.1 Caméra

Les images CCTV et ATM ont été enregistrées par caméra de surveillance réseau AXIS (respectivement P3225-V MkII et P1264 MkII). Le dispositif AXIS P3225-V MkII est une caméra grand-angle, de résolution HDTV 1080p et dédiée à la vidéosurveillance intérieure dans les commerces et banques, p. ex. La caméra AXIS *pinhole* P1264 MkII a une résolution HDTV 720p et un champ de vision horizontal de 57° et est principalement destinée à une utilisation discrète en guichets et distributeurs. Ces deux dispositifs enregistrent des vidéos au format MPEG-4 (compression H.264). Les caméras Axis sont omniprésentes en Suisse puisqu'elles assurent la surveillance vidéo des transports en commun fédéraux (CFF), mais également de voies publiques, d'industrie et de commerces de détail. Dans le reste du monde, ces dispositifs sont également utilisés dans les aéroports, les universités, les services de police, etc. Utiliser ces caméras permet de capturer des images correspondant directement au type de données que doivent utiliser les services judiciaires.

IV.3.2 Systèmes automatiques

Trois systèmes sont testés dans ce projet : deux générations du système commercial Idemia Morphoface (MFI et MFE) et l'algorithme en source ouverte (*open source*) FaceNet. Ce choix permet de comparer les performances d'algorithmes basés sur des architectures différentes, propriétaires ou publics.

IV.3.2.1 MorphoFace Investigate (v.2.0)

L'algorithme de cette version, de 2014, est basé sur un fonctionnement *rule-based*. Selon les tests effectués par le NIST en 2014, les performances du système MFI le placent au deuxième rang des meilleurs des systèmes commerciaux à l'époque (Grother et Ngan, 2014). Ces tests révèlent que le taux d'erreur au rang 1, c'est-à-dire que le premier candidat de la liste de résultats n'est pas celui correspondant à la trace, est de 9.1% sur une base de données de 1,6 million de photographies signalétiques, et de 7.6% sur une base de 160'000. La taille de la population à comparer a logiquement un effet sur les performances des systèmes automatiques. Dans près de 93% des cas testés par les auteurs, l'image de la personne recherchée fait partie des 50 premiers résultats.

IV.3.2.2 MorphoFace Expert (v.5.0)

Cette version plus récente du système de reconnaissance faciale développé par Idemia est basée sur une architecture de *deep learning*. Cela permet de comparer ces deux fonctionnements sur une méthodologie et des données identiques, pour illustrer l'étendue du progrès apporté par l'utilisation du *deep learning* en reconnaissance faciale. Cette version apparait dans les tests les plus récents du NIST (Grother *et al.*, 2019a ; Grother *et al.*, 2019b), qui la classe parmi les cinq meilleurs systèmes sur les tâches d'identification et d'investigation telles que décrites précédemment. Ce système figure également dans les derniers tests FRVT dédiés à l'impact du port de masques sur les performances des systèmes actuels, en réaction aux contraintes amenées

par la crise sanitaire mondiale. Dans ces tests, les auteurs observent une augmentation des taux d'exclusions incorrectes sur l'ensemble des systèmes, mais le système d'Idemia se classe toujours parmi les plus performants (Ngan *et al.*, 2020).

Bien que ces deux systèmes soient construits sur des architectures très différentes, le calcul de scores fonctionne de la même manière. Les scores de similarité, entre 0 (faible ressemblance) et environ 5000 (forte ressemblance), sont calculés en comparant deux visages puis normalisés en fonction de l'ensemble des scores. De fait, la comparaison de deux images ne fournit pas exactement le même score deux fois de suite si des images sont ajoutées ou enlevées de la base entre temps.

Pour ces deux systèmes, toutes les comparaisons sont effectuées à l'aide de l'outil MBSS (*Multi-Biometric Search Services*) v.6.6.1 qui permet d'effectuer un très grand nombre de comparaisons simultanément, et non par le biais de l'interface utilisateur.

IV.3.2.3 FaceNet

FaceNet est un algorithme en source ouverte développé par Google, construit autour d'une architecture de *deep learning* et entrainé sur 260 millions d'images de 30 millions d'individus (Schroff *et al.*, 2015a). Ce modèle[18] atteint une précision de 99,60% sur LFW et 95,12% sur les données *YouTube Faces*. (Galbally *et al.*, 2019 ; Schroff *et al.*, 2015a). Dans sa forme originale[19], le script permet d'encoder puis comparer de multiples images les unes avec les autres. Il a donc été adapté afin de pouvoir encoder et comparer les images d'un dossier de traces à celui de références puis exporter la liste triée de tous les scores de comparaisons, mais également d'indiquer les images où l'algorithme n'a pu détecter de visage[20]. À l'inverse des systèmes MorphoFace, FaceNet calcule des scores de distance entre 0 et (environ) 3. De fait, plus deux visages se ressemblent, plus proche de 0 est le score. De plus, le score attribué à chaque comparaison est invariable car il n'est pas normalisé en fonction de la base de données.

[18] modèle : 20180402-114759, architecture : Inception ResNet v.1
[19] Disponible sur : https://github.com/davidsandberg/facenet
[20] Le script final utilisé dans ce projet est disponible sur : https://github.com/MlgJcqt/

IV.4. Synthèse

Notre recherche vise à développer des modèles d'interprétation des scores de reconnaissance faciale, à partir de jeux de données représentatives de cas opérationnels et actuels, résumés ci-dessous :

Images faciales	Images de caméra de surveillance (CCTV, ATM), documents officiels d'identité et portraits (32 POI)
	BdD judiciaires (neuchâteloise et vaudoise, CH)
	mugshots (32 POI)
Systèmes automatiques	MorphoFace Expert (Idemia, *deep learning*)
	MorphoFace Investigate (Idemia, *rule based*)
	FaceNet (*open source, deep learning*)
Cadres d'utilisation	Investigation
	Expertise

Chapitre V. Choix méthodologiques

Nous détaillons dans ce chapitre la construction des modèles interprétatifs évalués dans ce projet et discutons nos choix concernant les méthodes de modélisation d'intravariabilité, d'intervariabilité, et de calibration des résultats.

V.1. Modèle interprétatif

Cette section reprend et complète les propositions faites dans l'article Jacquet et Champod (2020) « *Automated face recognition in forensic science : Review and perspectives* » paru dans le journal *Forensic Science International*.

Pour calculer un SLR à partir d'une valeur de score à des fins d'évaluation, il faut en premier lieu répondre à plusieurs questions.

- La qualité de l'image trace permet-elle d'utiliser un système automatique ?
- Quelle base de données utiliser ? Quelles sont les propositions selon l'accusation et la défense ?
- Comment modéliser les distributions d'intravariabilité et intervariabilité en fonction des données disponibles ?

Toutes ces questions sont essentielles pour le développement du modèle de calcul SLR, que nous détaillons dans la Figure 17.

En ce qui concerne la qualité de l'image, plusieurs points doivent être évoqués. Tout d'abord, les systèmes peuvent utiliser un processus d'alignement pour ajuster les images non frontales, mais pour des angles supérieurs à 45°, beaucoup d'informations sont perdues et le taux de résultats incorrects augmente (Dutta, 2015 ; Macarulla Rodriguez *et al.*, 2018). De plus, la plupart des systèmes nécessitent une résolution suffisamment élevée pour distinguer les yeux ainsi que les principales caractéristiques faciales et le visage ne doit pas être obstrué par des lunettes de soleil, un foulard ou d'autres attributs. Comme nous l'avons déjà mentionné, des études ont mis en évidence la corrélation entre la performance de la reconnaissance faciale et la nature des données d'apprentissage utilisées pour développer un modèle (Peng, 2019). Par exemple, les images de traces à basse résolution nécessitent l'utilisation d'un modèle développé avec des données d'entrainement de qualité similairement faible, afin d'assurer des comparaisons plus efficaces. En général, l'auteur suggère que les modèles devraient être construits avec des images qui répondent aux contraintes de la vie réelle des images de traces. Enfin, la distorsion optique due par exemple à la prise de vue très rapprochée, comme dans les images ATM, n'empêche pas le système d'analyser l'image mais peut affecter les résultats de la comparaison en modifiant les proportions des éléments faciaux. La Figure 7 (Section III.2.1 p.29) illustre efficacement les variations morphologiques dues aux changements de distance focale et de distance de prise de vue (Edmond *et al.*, 2009).

Si les données du cas (image trace T et référence S) sont jugées exploitables dans un processus automatique, la ou les traces et la ou les références du suspect sont encodées et comparées par l'algorithme. Ensuite, le score de comparaison est utilisé pour attribuer un SLR en référence à l'intravariabilité et l'intervariabilité appropriées. Ces données permettent de calculer un SLR à un score donné comme dans l'équation (a) (voir la Figure 13 pour un exemple de calcul de SLR).

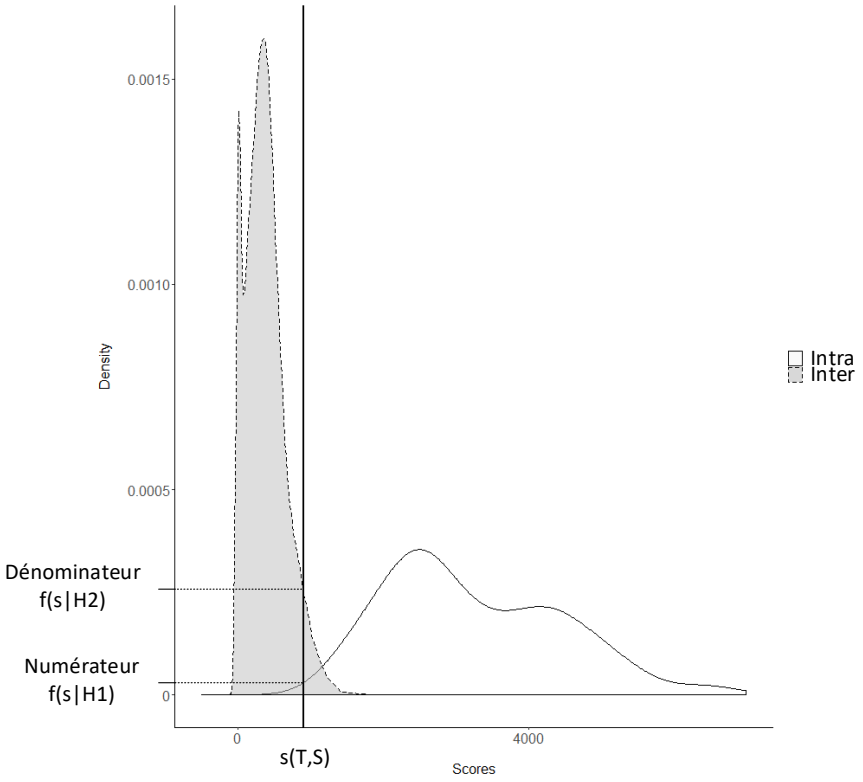


Figure 13 : Méthode de calcul de SLR à partir du score de comparaison s(T,S) reportée sur les distributions de scores d'intravariabilité (blanc) et d'intervariabilité (gris).

Pour générer les distributions des scores d'intravariabilité (variabilité de scores sous H_1) et d'intervariabilité (variabilité de scores sous H_2), l'expert.e doit d'abord déterminer les propositions H_1 et H_2 ainsi que la population pertinente en fonction des données du cas et des informations dont il dispose. La population pertinente est déterminée en fonction de la proposition H_2 selon laquelle l'image de référence et l'image de la requête sont considérées comme représentant deux personnes différentes. Cette population, propre à chaque cas, comprend des personnes qui ne sont pas le suspect mais dont le visage pourrait être celui qui est visible sur la photographie de la question. Dans la plupart des domaines, les témoins seraient ceux qui fournissent des informations potentielles sur l'auteur, comme le genre perçu, la couleur de la peau ou la tranche d'âge, par exemple. Dans les cas de reconnaissance faciale, ce type d'information peut être directement observé sur l'image de question. Mais il faut noter que la population pertinente est basée avant tout sur l'hypothèse de la partie défenderesse. Par conséquent, si H_2 est : "L'image trace ne représente pas la POI mais une autre personne (homme ou femme)", même si les expert.e.s ont supposé à partir des détails de l'image que la personne sur l'image de la requête était une femme, la population pertinente doit contenir des individus de différents genres. Il s'agit de s'assurer que le calcul du SLR prenne en compte les éléments en fonction des propositions établies par les allégations des parties, et non par celles de l'expert.e. Le rôle de l'expert.e est d'adapter la formulation de la proposition en fonction de la méthode choisie pour modéliser les variabilités intra- et intersources avec les données disponibles. Ces méthodes sont détaillées dans la section suivante.

V.1.1 Intravariabilité

La modélisation de l'intravariabilité illustre des variations existantes entre plusieurs photographies d'une même personne. Elle permet l'estimation de la probabilité d'obtenir un score de comparaison entre l'image trace et la référence sous la proposition H_1, selon laquelle la même personne est visible sur les deux images comparées. Les deux principales méthodes utilisées consistent à modéliser l'intravariabilité spécifique au suspect ou de manière générique. Ces approches ont été examinées pour la reconnaissance des visages dans (Ali, 2014)).

V.1.1.1 Spécifique au suspect

Un expert.e peut demander des photographies supplémentaires de la POI afin de modéliser la variabilité intrasource (intravariabilité). Si cela est possible, la personne peut être photographiée directement pour fournir à l'expert.e des *mugshots* de référence (illustrées par des cercles dans les figures Figure 14 et Figure 15). Lors de la comparaison d'images faciales, la période entre l'enregistrement de la trace et les références devient souvent problématique en raison du vieillissement de la personne. Voir Grother *et al.* (2019a), Jain *et al.* (2012) et Kemelmacher-Shlizerman *et al.* (2016) pour des exemples d'études portant sur l'impact du vieillissement sur les performances des systèmes de reconnaissance faciale. Pour la plupart des algorithmes testés, plus l'écart d'âge entre deux images comparées est important, moins le système est efficace. Cependant, l'algorithme FaceNet (Schroff *et al.*, 2015a) fonctionne sur des plages d'âge très variables, lorsqu'il est formé sur des ensembles de données adéquats (Kemelmacher-Shlizerman *et al.*, 2016).

Pour l'intravariabilité spécifique au suspect, la proposition H_1 peut être formulée comme suit : « Le suspect est la personne sur la trace ».

La littérature fait également état d'images de références particulières, appelées « images de contrôle » (Meuwly, 2006). Ces images sont prises dans des conditions comparables à celles de l'image trace pour servir de référence spécifique. Elles sont également appelées « pseudo-traces » dans plusieurs publications (Meuwly, 2006 ; Neumann *et al.*, 2015 ; Ramos *et al.*, 2018) en biométrie en général et plus spécifiquement pour les voix et les empreintes digitales. Disposer à la fois d'images de référence et de contrôle de haute qualité est la situation idéale pour modéliser la variabilité intrasource (intravariabilité) de la POI. Cette méthode consiste à comparer les images de contrôle, d'une part, aux images de référence, d'autre part (Figure 14). L'objectif de la comparaison des images de contrôle avec les images de référence est d'obtenir une distribution de score d'intravariabilité plus représentative et spécifique au cas d'intérêt, en tenant compte de toutes les informations visibles sur la trace. Par exemple, si l'image trace est extraite d'une séquence CCTV, les images de contrôle seront enregistrées avec une caméra de surveillance, idéalement sur la scène où les faits se sont déroulés, ou ailleurs dans des conditions similaires.

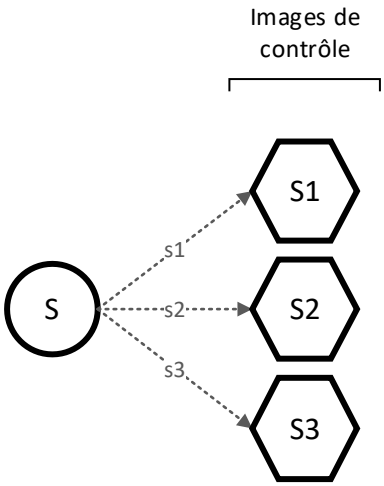


Figure 14 : Méthode de modélisation de l'intravariabilité spécifique à l'aide d'images de référence et de contrôle.

Les images de contrôle constituent un matériel idéal pour la modélisation de l'intravariabilité, mais sont rarement prises dans des cas opérationnels. Dans la plupart des cas, le suspect n'est pas disponible et même lorsqu'il l'est, l'enregistrement de ces images peut être trop coûteux et trop long pour être envisagé. Pour cette raison, ce type d'image ne sera pas exploité dans le présent projet. De fait, lorsque seules des images de référence sont disponibles, les scores d'intravariabilité sont calculés en comparant toutes les références les unes aux autres, comme l'illustre la Figure 15a.

V.1.1.2 Générique

Dans la plupart des cas, l'expert.e ne dispose que d'une seule image de référence, voire d'aucunes. Une approche générique doit alors être utilisée pour pallier le manque de données de référence en utilisant uniquement des photographies de « suspects potentiels » de la population concernée (Botti *et al.*, 2004). Cette méthode exige de calculer les scores d'intravariabilité de tous ces individus pour modéliser une intravariabilité globale (Figure 15b). La proposition H_1 ne peut donc plus être centrée que sur le suspect puisque l'intravariabilité est modélisée avec des images de personnes qui ne sont pas directement liées à l'affaire. La proposition devient alors : « La même personne est visible dans les deux images comparées ».

Dans la littérature est décrit un autre processus pour modéliser une intravariabilité générique, dans le domaine de la reconnaissance du locuteur (Botti *et al.*, 2004). Les données disponibles sont une trace et un seul enregistrement de référence d'un suspect. Les auteurs ont donc créé deux bases de données, l'une contenant les enregistrements de contrôle de dix personnes pris dans des conditions comparables à celles de la trace (cinq enregistrements par personne), et la seconde contenant trois enregistrements de référence pour chaque personne dans des conditions comparables à celles de l'enregistrement de référence du suspect. En transposant ces conditions à un cas de reconnaissance de visage, cela signifierait créer deux bases de données de sources potentielles avec, respectivement, des images correspondant aux conditions de prise de vue de la trace (documents d'identité, séquences de vidéosurveillance, etc.), et des images correspondant

à la description de la photo du suspect (principalement des documents d'identité ou d'autres portraits). Une telle approche reste générique car ni la trace ni la référence du suspect ne sont utilisées par le système. Cependant, elle nécessite soit de disposer déjà d'une grande base de données contenant des images prises dans de multiples conditions différentes, avec plusieurs photos pour chaque individu, soit de trouver de nombreux volontaires pour prendre les photos adéquates pour chaque cas en question. Par conséquent, cette approche ne semble pas facilement applicable à des fins opérationnelles, et ne sera pas exploitée ici.

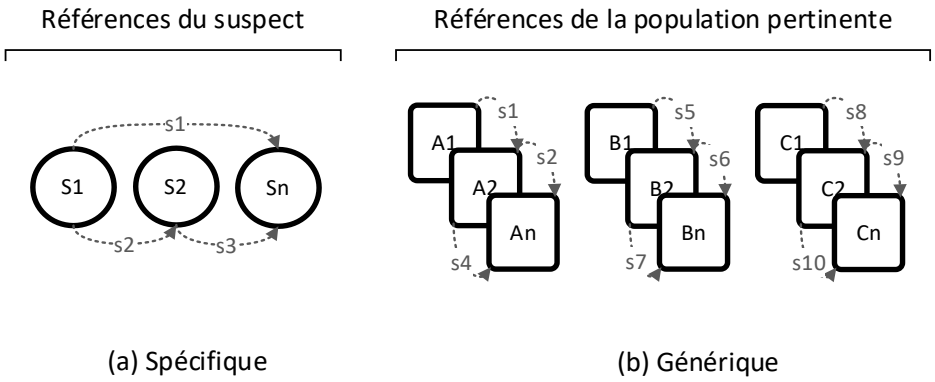


Figure 15 : Méthodes choisies pour la modélisation de l'intravariabilité spécifique (a) et générique (b).

V.1.2 Intervariabilité

L'intervariabilité est modélisée à l'aide des scores de comparaisons sous la proposition alternative H_2, selon laquelle la trace et la référence ne montrent pas la même personne. Cela permet d'estimer le dénominateur du SLR (Équation a). L'objectif est de représenter l'étendue des valeurs de scores lors de la comparaison de photographies d'individus différents. Trois approches de modélisation sont décrites dans la littérature : spécifique à la trace, au suspect, et générique. Seules les deux premières seront exploitées dans le présent projet pour les raisons exposées ci-après.

V.1.2.1 Spécifique à la trace

Cette approche spécifique à la trace décrite dans (Meuwly, 2006) consiste à calculer les scores d'intervariabilité en comparant la trace à chaque image de la base de données, comme l'illustre la Figure 16a. La proposition alternative H_2 indique donc que « l'individu visible sur la trace n'est pas le suspect, mais quelqu'un d'autre de la population pertinente ».

Deux méthodes peuvent être étudiées pour trier la population pertinente en comparant la trace à une base de données non triée. Dans le premier cas (exploité dans le présent projet), la trace est comparée à toutes les images de la base de données et les individus des premiers rangs de la liste de scores composent la population pertinente (15% de la taille de la base de données). L'inconvénient de cette approche est de sous-estimer le SLR en modélisant l'intervariabilité avec les scores aux valeurs les plus proches de ceux d'intravariabilité, selon le système. Dans le second cas, l'expert.e peut trier la population pertinente en sélectionnant des sources potentielles, c'est-

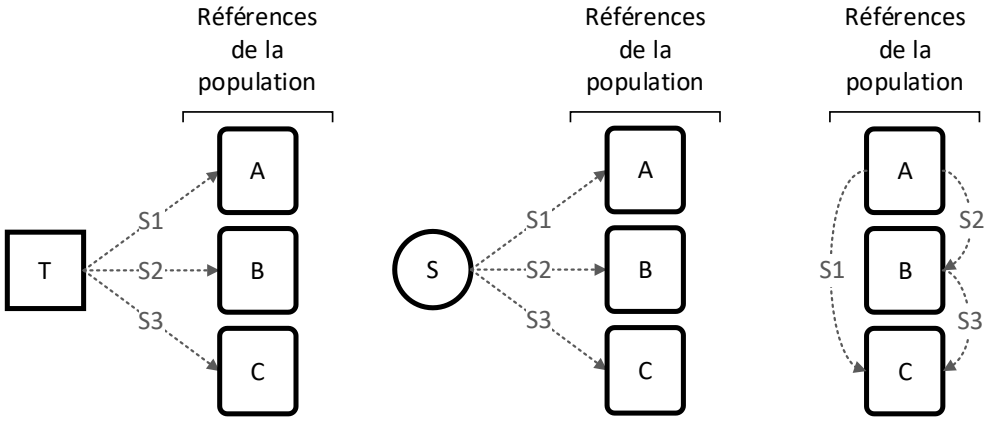
à-dire des individus dont le visage est le plus proche du visage visible sur la trace. Néanmoins, en fonction de la taille et de la structure de la base de données, cette solution peut prendre du temps et dépend fortement de la performance de l'expert.e dans le tri des visages. Cependant, aucune étude n'a encore montré l'impact des variations dans le choix de la population concernée par la reconnaissance automatique des visages, que ce soit à des fins d'enquête ou pour le tribunal.

V.1.2.2 Spécifique au suspect

Cette seconde approche consiste donc à comparer l'image de référence du suspect (appelée « source » dans (Hepler *et al.*, 2012)) avec les individus de la population pertinente (Figure 16b). Dans ce cas, la proposition alternative H_2 est : « Le suspect n'est pas la personne figurant dans l'image de la trace ». L'intervariabilité représente donc la plage de valeurs de scores obtenus lorsque l'on considère que l'image de référence et une image choisie au hasard dans la population représentent des personnes différentes. L'objectif est que le dénominateur du SLR reflète la rareté/fréquence avec laquelle on trouve, dans la population pertinente, un visage similaire à celui du suspect. Cette notion est appelée « *typicity* » par Morrison (2016). Il s'agit d'un critère crucial à inclure dans tout calcul de SLR, par la sélection de la population pertinente correspondant au mieux au cas, comme décrit précédemment, pour l'intervariabilité spécifique à la trace. Dans cette approche, les scores dépendent fortement de la qualité des images du cas. En particulier, la comparaison de traces de type CCTV de faible qualité avec des images frontales de référence standard générerait des scores de similarité plus faibles que si l'on compare deux images frontales de bonne qualité à la fois comme trace et comme référence.

V.1.2.3 Générique

Dans une approche dite générique, toutes les images de la population pertinente sont comparées par paires (Figure 16c). Les scores sont utilisés pour modéliser l'intervariabilité générique, c'est-à-dire non spécifique au cas d'espèce. Ici, l'intervariabilité représente la plage de valeurs de scores en considérant que deux images choisies au hasard dans la population représentent des personnes différentes. Pour cette approche, la proposition qui conditionne le dénominateur du SLR est formulée de manière plus générale : "Deux personnes différentes sont visibles sur les deux images comparées". Comme aucune information ne provenant ni de la trace ni de l'image de référence du suspect n'est prise en compte pour calculer le SLR, cette approche est beaucoup moins instructive que les deux précédentes (Neumann et Margot, 2009). De plus, Les valeurs des scores de similarité seront plus élevées - respectivement plus faibles pour les scores de distance - que lorsque l'on compare une trace CCTV de faible qualité avec des *mugshots* de référence de très bonne qualité. De fait, cette approche ne sera pas exploitée dans le présent projet.



(a) Spécifique à la trace (b) Spécifique au suspect (c) Générique

Figure 16 : Méthodes de modélisations d'intervariabilité spécifique à la trace (a) au suspect (b) et générique (c).

V.1.3 Propositions de travail

Dans le Tableau 2, nous proposons un résumé du conditionnement de chaque formule SLR et des formulations des propositions en fonction des approches de modélisations utilisées dans cette recherche. Pour le dénominateur du SLR, la fonction de probabilité d'obtenir un score s(S,T) est conditionnée par les données utilisées pour la modélisation de l'intervariabilité, respectivement la trace T (approche Trace-spécifique) ou la référence S (approche Suspect-spécifique). En revanche pour l'intravariabilité spécifique, le score s(S,T) au numérateur ne peut être conditionné que sur le suspect. Pour l'intravariabilité générique, le score s(S,T) n'est conditionné ni sur la trace T ni sur le suspect S, car il est calculé en comparant les images des individus de la population pertinente et ne dépend donc que de celle-ci.

Tableau 2 : Formule de SLR et formulation des propositions H_1 et H_2 pour chaque approche de modélisation de l'intravariabilité et de l'intervariabilité.

Approches		Formules SLR	Propositions	
Intravariabilité	Intervariabilité		H1	H2
Spécifique	Suspect-spécifique	$= \frac{f(\mathbf{s(S,T)}\|\mathbf{S},\mathbf{H_1},\mathbf{I})}{f(\mathbf{s(S,T)}\|\mathbf{S},\mathbf{H_2},\mathbf{I})}$	"Le suspect est la personne sur l'image trace"	"Le suspect n'est pas la personne sur l'image trace"
	Trace-spécifique	$= \frac{f(\mathbf{s(S,T)}\|\mathbf{S},\mathbf{H_1},\mathbf{I})}{f(\mathbf{s(S,T)}\|\mathbf{T},\mathbf{H_2},\mathbf{I})}$		"L'individu à la source de la trace n'est pas le suspect, mais une personne de la population d'intérêt"
Générique	Suspect-spécifique	$= \frac{f(\mathbf{s(S,T)}\|\mathbf{H_1},\mathbf{I})}{f(\mathbf{s(S,T)}\|\mathbf{S},\mathbf{H_2},\mathbf{I})}$	"La même personne est visible sur les deux images"	"Le suspect n'est pas la personne sur l'image trace"
	Trace-spécifique	$= \frac{f(\mathbf{s(S,T)}\|\mathbf{H_1},\mathbf{I})}{f(\mathbf{s(S,T)}\|\mathbf{T},\mathbf{H_2},\mathbf{I})}$		"L'individu à la source de la trace n'est pas le suspect, mais une personne de la population d'intérêt"

La formulation des propositions est cruciale car elles conditionnent les résultats que l'expert.e présente au tribunal. Elles doivent représenter au mieux les versions des deux parties opposées dans le procès en utilisant les données disponibles. Par exemple, le manque de matériel de référence conduit à l'utilisation d'une approche générique, utilisant exclusivement des photos d'individus d'une base de données, non impliqués dans l'affaire en question. La conclusion de l'expert.e ne peut alors pas directement inclure le suspect, car aucune de ses photos n'a été utilisée. C'est pourquoi la proposition fait référence à une « même personne » dans les deux images comparées sans mentionner le suspect (Tableau 2). À ce stade, la limitation est que, si une partie demande « La personne sur l'image de la requête peut-elle être le suspect ou non ? », les expert.e.s ne peuvent pas faire de commentaire supplémentaire car leur modèle générique ne leur permet pas d'inclure le suspect dans leur conclusion. Voir par exemple la publication de Champod et Evett (2009) pour une explication approfondie et plus pratique, applicable à tout type de preuve forensique.

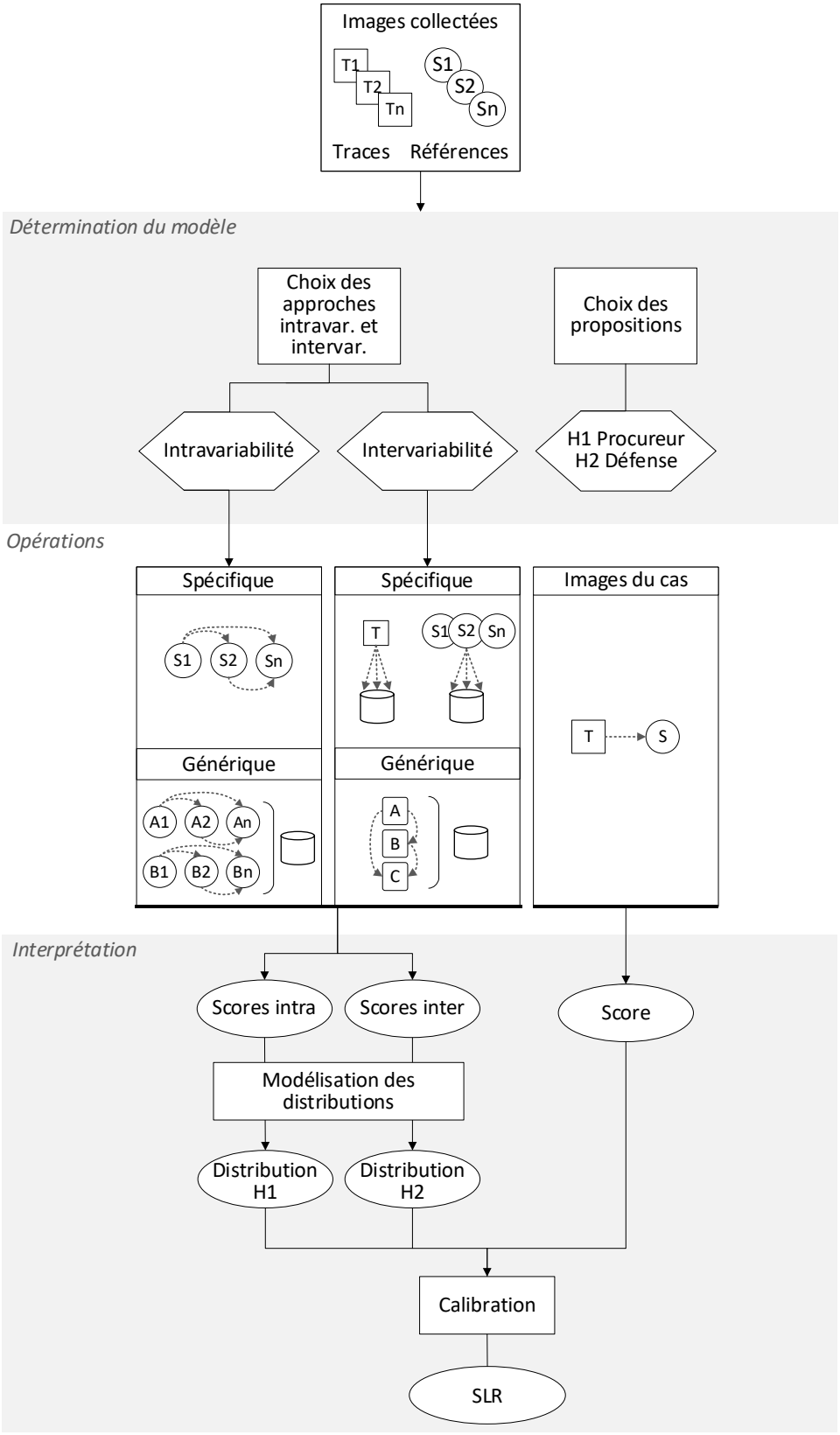


Figure 17 : Processus de développement du modèle évaluatif de calcul de SLR.

Pour conclure, la Figure 17 résume l'ensemble du processus de développement du modèle évaluatif de calcul de SLR. En partant d'un cas forensique hypothétique - si la qualité des données répond aux exigences de systèmes automatiques - l'expert.e définit les spécifications du modèle concernant la modélisation des variabilités intra et intersources, la formulation des propositions

et la détermination de la population pertinente au regard des informations fournies par la trace T et la référence S. Seules les approches de modélisation testées dans le présent projet sont représentées dans le bloc des opérations. Les résultats préliminaires de ce processus sont les scores de comparaisons, qui permettent de modéliser les densités de probabilité en utilisant la méthode la plus adéquate, puis de calculer le SLR suite à la calibration.

V.1.4 Modélisation des fonctions de densité et calibration des LR

Une fois les scores d'intravariabilité et d'intervariabilité générés selon l'approche choisie, les distributions de densités de probabilités sont modélisées de manière à correspondre au mieux aux distributions empiriques des scores.

V.1.4.1 Méthodes choisies

Plusieurs distributions paramétriques et non paramétriques, parmi les plus récurrentes dans la littérature (Ali *et al.*, 2012a ; Egli, 2009 ; Macarulla Rodriguez *et al.*, 2018), ont été testées pour chaque modèle : Lognormal, Normal, Weibull, *Generalized Estimation Value* (GEV), Régression Logistique (Reg Log) et *Kernel Density Estimation* (KDE). Les tests sont effectués sur le logiciel R (v. 4.0.2) par R-studio (v.1.1.383) à l'aide des packages *EnvStats* (Millard, 2013), *fitdistrplus* (Delignette-Muller *et al.*, 2015) et *extRemes* (Gilleland et Katz, 2016). Les méthodes sélectionnées dans cette recherche sont résumées dans le Tableau 3.

Tableau 3 : Choix de méthodes de modélisation des distributions de densités de probabilités d'intravariabilité et d'intervariabilités en fonction des approches et systèmes utilisés.

<table>
<tr><th colspan="2" rowspan="2">Distributions</th><th colspan="3">Modélisation des distributions</th></tr>
<tr><th>Facenet</th><th>MFI</th><th>MFE</th></tr>
<tr><td rowspan="2">Intra-variabilité</td><td>Spécifique</td><td>KDE</td><td rowspan="4">KDE</td><td>KDE</td></tr>
<tr><td>Générique</td><td>Weibull</td><td>GEV</td></tr>
<tr><td rowspan="2">Inter-variabilité</td><td rowspan="2">Trace-spécifique</td><td>(atm, ID) Normal
(cctv) Weibull</td><td rowspan="2">KDE</td></tr>
<tr><td>Suspect-spécifique</td><td>Normal</td></tr>
</table>

Les distributions d'intravariabilité et d'intervariabilité génériques ne dépendant pas de la trace ni du suspect, leur forme ne varie donc pas d'un scénario à l'autre ce qui simplifie la recherche de la distribution adéquate. En revanche, les intravariabilités et les intervariabilités spécifiques sont différentes pour chaque scénario car elles dépendent pour chaque cas de la trace et du matériel de référence utilisé. De fait, les méthodes de modélisation pour ces approches ont été choisies de manière à correspondre au mieux à toutes les distributions de scores de toutes les traces et tous les POI. Ainsi, la KDE est appliquée pour les distributions d'intravariabilité spécifique malgré le faible nombre de scores qui les composent car elles varient fortement d'un POI à l'autre et ne permettent donc pas d'appliquer de méthodes paramétriques. Le risque est d'obtenir des valeurs de numérateurs nulles pour les distributions multimodales. Dans ce cas, pour assurer le calcul d'un SLR, toute valeur égale à zéro a été remplacée par la valeur positive minimale du reste de la distribution.

La figure ci-dessous illustre la diversité de forme des distributions d'intravariabilité spécifique pour tous les POI. Il était évidemment impossible d'appliquer une seule et même méthode pour toutes ces distributions, ce qui explique notre choix de la KDE et ce malgré le faible nombre de scores pour la plupart d'entre eux.

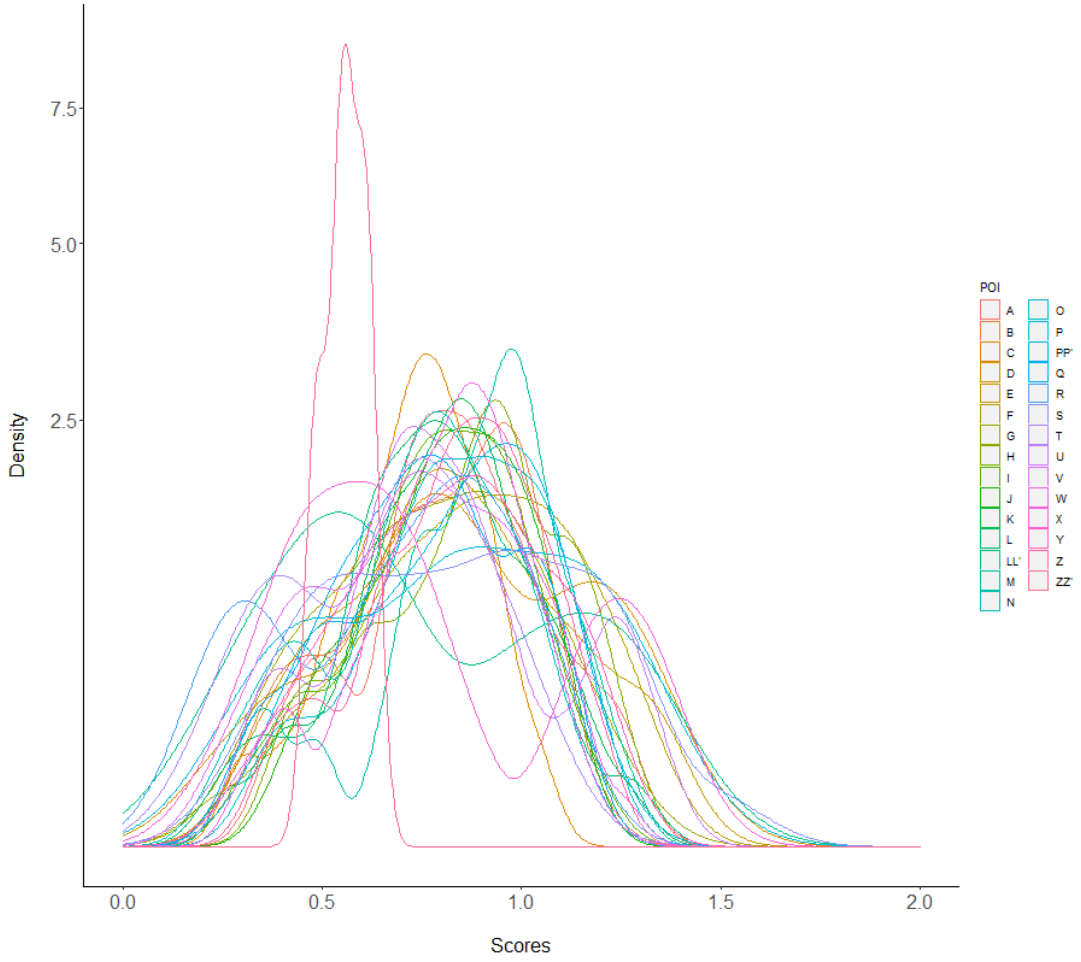


Figure 18 : Densité de distributions des scores d'intravariabilité spécifique pour toutes les POI avec FaceNet.

Avec le système MFE, les scores d'intravariabilité générique suivent une distribution modélisée par GEV *Maximum Likelihood Estimation*. Cette méthode offre une meilleure modélisation aux extrémités des courbes de distributions. Avec FaceNet, l'intravariabilité spécifique et l'intervariabilité trace-spécifique pour les traces CCTV suivent une distribution Weibull, et les scores d'intervariabilité générique et trace-spécifique (pour les scénarios ATM et ID) sont modélisés par une distribution Normale (Figure 19, Figure 20). Pour le MFE, les distributions de scores des deux approches d'intervariabilité nécessitent une modélisation par KDE. Il en est de même pour la totalité des distributions par MFI.

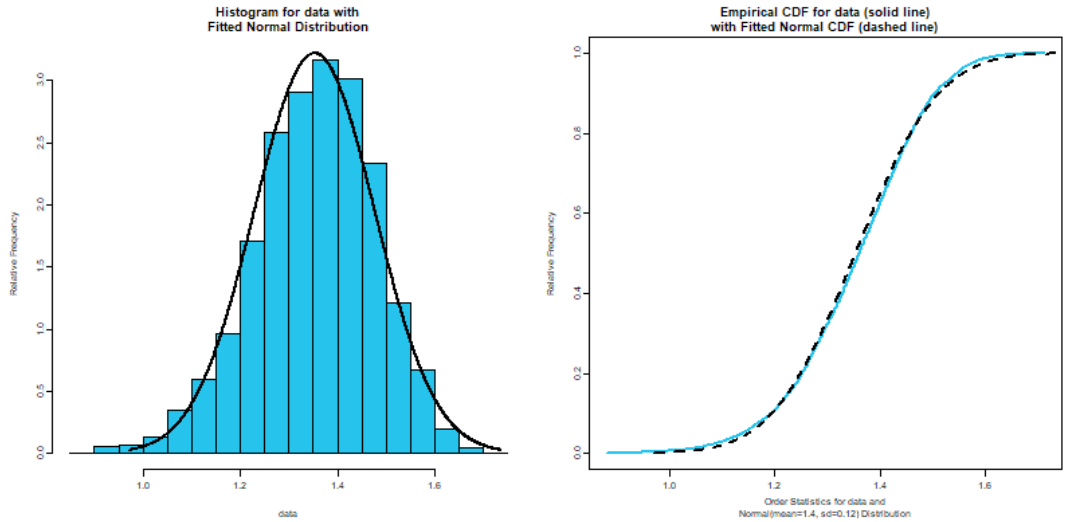


Figure 19 : Modélisation de distribution Normale pour l'intervariabilité générique (FaceNet). En bleu : données empiriques. En noir : données modélisées.

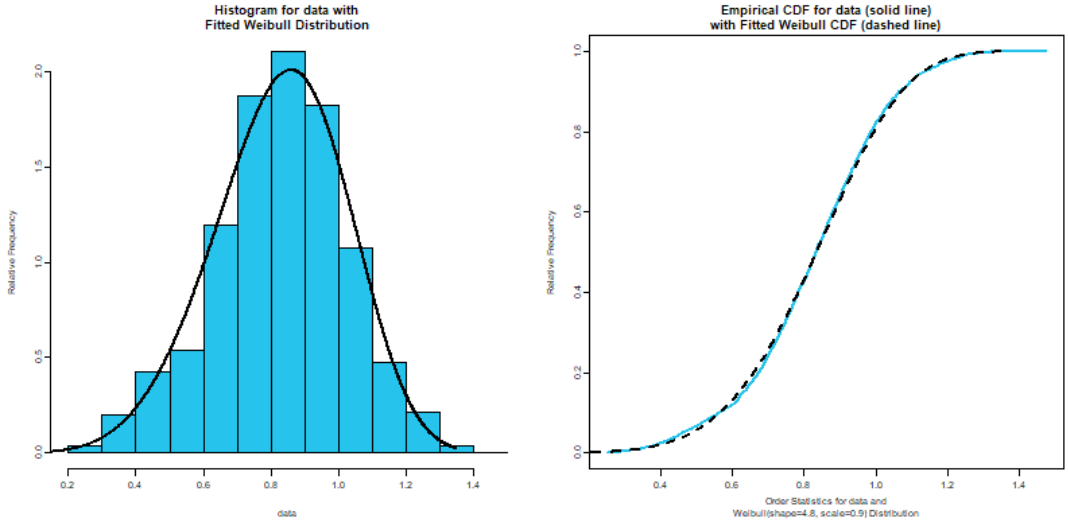


Figure 20 : Modélisation de distribution Normale pour l'intravariabilité générique (FaceNet). En bleu : données empiriques. En noir : données modélisées.

La Figure 21 illustre l'exemple caractéristique du problème rencontré pour la modélisation des distributions d'intervariabilité (générique et spécifiques) générées par le MFE mais également le MFI. Ici, nous montrons la tentative de modélisation de distribution logistique car c'est la méthode paramétrique qui donne la distribution la plus proche des scores empiriques. Au vu de ses mauvais résultats, il a été décidé de modéliser des distributions par KDE.

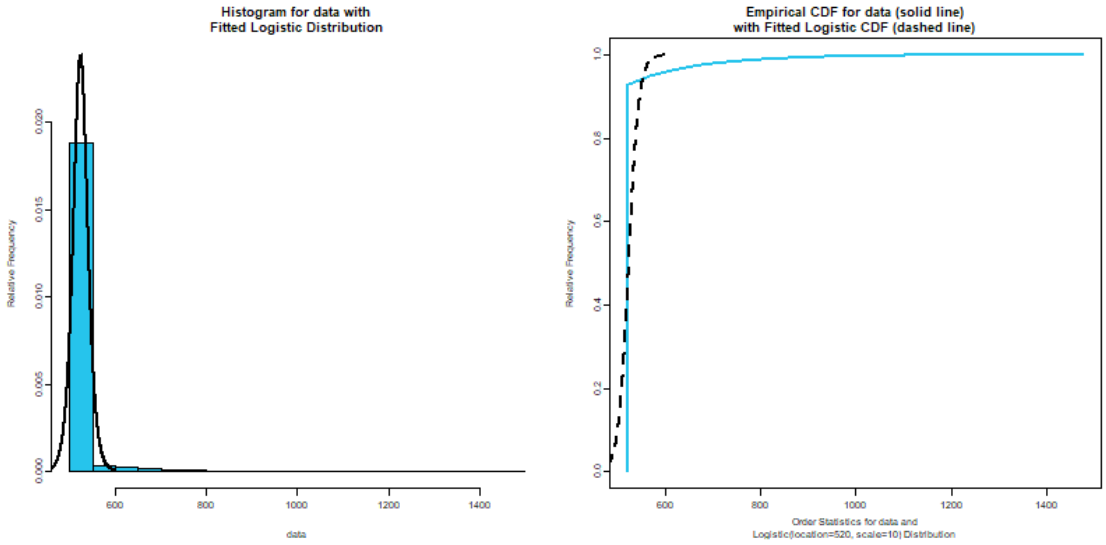


Figure 21 : Modélisation de distribution Logistique pour l'intervariabilité trace-spécifique ATM de la POI A (MFE). En bleu : données empiriques. En noir : données modélisées.

Pour finir, nous détaillons les résultats des tests statistiques associés aux choix de méthodes de modélisation de distribution dans le tableau ci-dessous :

Tableau 4 : Résumé des distributions choisies pour la modélisation paramétrique des différentes intravariabilités et intervariabilités et des résultats des tests statistiques associés.

Système	Distributions		Modélisation des distributions	Paramètres estimés	Taille de l'échantillon	*p-value*
MFE	Intravariabilité	Générique	GEV	location = 4,5e+03 scale = 1,2e+03 shape = 9,7e-02	6432	<0.001
FN	Intravariabilité	Générique	Weibull	mean = 4,8 sd = 0,9	3202	<0.001
	Intervariabilité	Trace-spécifique (ATM)	Normal	mean = 1,343 sd = 0,096	382140	<0.001
		Trace-spécifique (CCTV)	Weibull	mean = 15,8 sd = 1,4	232620	<0.001
		Trace-spécifique (ID)	Normal	mean = 1,32 sd = 0,12	117422	<0.001
		Suspect-spécifique	Normal	mean = 1,36 sd = 0,12	51700	<0.001

V.2. Validation du modèle

V.2.1 Évaluation des performances du modèle

Pour s'assurer que le modèle développé produit des SLR robustes, deux caractéristiques principales doivent être prises en compte : le pouvoir discriminant et l'état de calibration du modèle (Meuwly *et al.*, 2017 ; Robertson *et al.*, 2016). Dans cette section, nous décrivons

successivement ces deux caractéristiques de performance ainsi que les métriques et représentations graphiques utilisées pour les visualiser dans le présent projet.

V.2.1.1 Discrimination

Le pouvoir discriminant d'un modèle illustre sa capacité à discriminer les deux propositions mutuellement exclusives. Cette caractéristique dépend directement de la structure de l'algorithme, et ne peut donc être améliorée qu'en modifiant le code du système (Morrison *et al.*, 2021 ; Robertson *et al.*, 2016). Le pouvoir discriminant peut être mesuré à l'aide de mesures telles que le EER (taux d'erreur égal) et le minCllr (coût minimum du SLR logarithmique) (Meuwly *et al.*, 2017) ainsi que les taux de fausses acceptations et de faux rejets (respectivement, FAR et FRR) (Ramos-Castro, 2007). Dans le cadre de calculs de rapports de vraisemblance, le FAR et le FRR sont désignés respectivement comme le taux de résultats soutenant à tort l'accusation (RMEP) et le taux de résultats soutenant à tort la défense (RMED) (Neumann *et al.*, 2006).

V.2.1.2 Calibration

Contrairement à la discrimination, la calibration peut être intégrée au modèle de calcul de SLR, sans avoir à accéder à l'algorithme du système. En termes généraux, l'état de calibration d'un modèle se réfère à la proximité de la valeur calculée par rapport à la valeur connue. Dans notre cas, cependant, il n'y a pas de valeur connue à viser. Par conséquent, la calibration traduit la mesure dans laquelle le SLR pointe vers la proposition correcte. Les valeurs des SLR peuvent être toutes ajustées en même temps en appliquant le même décalage, soit être modifiées séparément, tant que leur rang dans la liste des SLR résultants ne change pas (Robertson *et al.*, 2016).

La calibration a été décrite en termes généraux dans Lindley *et al.* (1979) et abordé ultérieurement dans le cadre bayésien (voir DeGroot et Fienberg (1983)). La méthode de calibration la plus largement utilisée et étudiée est l'algorithme des « *Pool-Adjacent Violators Algorithm* » (PAVA) (Brümmer, 2010 ; Brümmer et du Preez, 2006 ; Drygajlo *et al.*, 2016 ; Ramos et Gonzalez-Rodriguez, 2013 ; Zadora *et al.*, 2014), majoritairement pour la reconnaissance de locuteur. La seconde méthode largement utilisée dans la littérature et également testée dans cette recherche est la régression logistique (Brümmer, 2010 ; Morrison, 2013 ; Ramos-Castro, 2007).

Il est important de noter que le terme « calibration » est utilisé dans la littérature pour décrire deux processus (pas si distincts). Il fait souvent référence au terme *calibration score-to-SLR*, c'est-à-dire à l'étape de calcul du SLR (que nous appellerons « SLR brut »), comme indiqué dans Ali *et al.* (2012a), mais il se réfère également dans d'autres publications au processus suivant, la calibration du SLR, comme dans Ramos *et al.* (2007), lorsque cela est nécessaire pour des modèles à haut pouvoir discriminant mais mal calibrés. Selon nous, cette deuxième étape est essentielle pour améliorer la performance globale d'un modèle, c'est pourquoi dans la suite de ce projet la calibration est considérée comme partie intégrante du calcul du SLR.

V.2.1.3 ECE

Proposées dans Ramos-Castro (2007), les courbes *Empirical Cross Entropy* (ECE) représentent les coûts associés aux valeurs de SLR générées par un système. Plus la valeur maximale est élevée,

plus le SLR a besoin d'informations pour indiquer la proposition correcte. De fait, un modèle performant présente un taux minimal de pertes en cas d'erreur. Sur une courbe ECE, les valeurs $P(H_1)$ en abscisses représentent les valeurs de probabilité *a priori* sur les propositions en jeu. La métrique Cllr, représentant le coût moyen associé à un modèle, est représentée comme la valeur ECE à l'intersection de l'axe $Log_{10}(P(H_1)) = 0$ et de la courbe du modèle empirique (Brümmer et du Preez, 2006). Le minCllr est le coût associé au modèle calibré et quantifie directement son pouvoir discriminant.

V.2.2 Discussion des choix méthodologiques

V.2.2.1 Modèles et matériel

Parmi les études des différentes approches de modélisation de variabilité intra et intersources (Ali, 2014 ; Botti *et al.*, 2004 ; Hepler *et al.*, 2012 ; Meuwly, 2006 ; Morrison, 2016 ; Neumann et Margot, 2009), la majorité utilisent des données de qualité variable conçues pour la recherche, qui diffèrent des données forensiques collectées dans le cadre opérationnel. Il est donc essentiel d'appliquer les méthodes développées sur les données de recherches, à des scénarios qui reflètent au mieux les contraintes rencontrées dans la pratique notamment la faible qualité mais également faible quantité de données. Par exemple, dans le domaine de la reconnaissance faciale, l'impact d'une approche générique sur les valeurs de SLR par rapport à une approche spécifique a été étudié dans Ali *et al.* (2013). Pour ce faire, les auteurs utilisent les photographies de cinq personnes choisies au hasard dans la base de données du FRGC (*Face Recognition Grand Challenge*) (Phillips *et al.*, 2005). Pour chacun d'entre eux, 36 images sont utilisées, dont la moitié sert de base de données de référence, et l'autre moitié est dégradée artificiellement (réduction de la définition et de la résolution) pour mieux correspondre aux conditions non contrôlées dans lesquelles les images des traces de types CCTV sont prises. Ces photographies dégradées constituent la base de données de contrôle de chaque individu. À partir de ces données, les auteurs modélisent les variabilités intra et intersources par des approches spécifiques et génériques. Les ordres de grandeur des valeurs de SLR sont comparés pour les deux approches en utilisant l'échelle verbale adaptée de (Evett *et al.*, 2000). Dans 59,2% des cas en moyenne, les SLR obtenus par l'approche générique correspondent à ceux de l'approche spécifique. Ces résultats soutiennent l'utilisation d'une telle approche dans le domaine de la reconnaissance faciale. Cependant, pour deux sujets sur cinq, l'approche générique est plus efficace et le pourcentage de SLR correspondant au même niveau verbal pour les deux approches diminue à 28,5%. Afin d'évaluer plus précisément l'impact des deux approches sur les valeurs de SLR, il est donc nécessaire de confirmer ces résultats avec davantage de sujets et de véritables images de contrôle de faible qualité, au lieu de sous-échantillonner les images de haute qualité, comme le suggère également (Peng, 2019). C'est pourquoi dans la présente recherche, nous proposons de comparer les approches spécifiques et génériques à partir d'images de 29 à 32 individus, collectées dans les conditions et avec du matériel correspondant au mieux à ceux rencontrés dans des cas forensiques.

V.2.2.2 Population pertinente

La population pertinente peut être définie par plusieurs caractéristiques apparentes que l'expert.e peut noter lors de la phase d'analyse des traces, comme décrit précédemment. Cette population doit donc inclure tous les individus de la base de données dont le visage peut être celui qui est visible sur la trace, selon l'expert.e. Cette population comprendrait donc tous les individus de la base de données dont le visage pourrait être celui qui est visible sur la trace, selon l'expert.e (p. ex. tous les hommes ou toutes les femmes, tous avec ou sans tatouage apparent). Cependant, nous pensons que, pour la reconnaissance des visages, l'utilisation de caractéristiques observables pour composer la population concernée peut souffrir de multiples restrictions. Premièrement, les marques telles que les tatouages, les grains de beauté, les cicatrices, etc. sont des caractéristiques difficiles à modéliser statistiquement, car l'expert.e devrait trier la base de données manuellement pour trouver les individus présentant des caractéristiques similaires (ce qui prend du temps), ou trier automatiquement les individus en utilisant des étiquettes associées aux caractéristiques visibles. Seule cette dernière solution permet d'appliquer l'approche par trace, mais une telle base de données avec des images étiquetées doit être construite au préalable. Deuxièmement, l'origine ethnique est principalement supposée sur la base de la couleur de la peau et des cheveux. Cette détermination est sujette à des erreurs, dues notamment au mélange génétique des populations, ou au fait que ces caractéristiques peuvent apparaitre différemment selon la qualité de l'image et les conditions de prise de vue. Par conséquent, il convient également d'éviter de présumer de l'origine ethnique d'un individu. Un tri peut néanmoins rester souhaitable dans le cas où, par exemple, l'individu sur la trace semble appartenir à une large population très peu présente dans la BdD. Le tri de la BdD permettrait alors de diminuer le nombre de comparaisons et par conséquent le risque de retrouver la POI recherché loin dans la liste de scores. Enfin, la tranche d'âge pourrait être utilisée pour composer la population concernée, à condition qu'elle soit basée sur l'observation combinée de plusieurs caractéristiques pertinentes. À la lumière de ce qui précède, l'approche par trace ne pourrait pas être appliquée si nous sélectionnons la base de données pertinente sur la base d'une analyse visuelle de la biométrie douce effectuée par l'expert.e.

En outre, dans Dror (2020), l'auteur met en avant l'importance des taux d'erreurs dans l'évaluation de la fiabilité d'une méthode et préconise l'utilisation de base de données appropriées à l'objectif de la méthode. Notamment sous H_2, la base de données doit contenir des cas difficiles (« *look-alike known non-match* »), c'est-à-dire des cas où la ressemblance entre deux personnes différentes complique leur comparaison. Dans de tels cas, il est donc attendu que la courbe des résultats d'intervariabilité se rapproche voire chevauche la courbe d'intravariabilité. Cela permet de ne pas surestimer les performances d'une méthode en la développant sur un jeu de données optimal. Les méthodes proposées dans le présent projet prennent en compte ces observations dans le choix des populations pertinentes. En effet, la courbe d'intervariabilité est modélisée à partir des meilleurs scores de comparaison de sources différentes (15% du nombre total de scores générés). Cette solution permet de considérer les résultats des comparaisons les plus complexes d'après le système biométrique.

V.2.2.3 Calibration

Il est important de noter que cette méthode de calibration présente certaines faiblesses qui peuvent s'avérer très limitantes pour une utilisation forensique. En effet, la méthode PAVA ne génère des valeurs de SLR finies qu'à partir des données où les distributions se chevauchent. Toutes les valeurs de SLR au-delà de cette zone, c'est-à-dire en dessous de la valeur la plus faible sous H_1 et au-dessus de la valeur la plus élevée sous H_2, seront égales à zéro et à l'infini, respectivement. Une solution a été proposée pour résoudre ce problème (Brümmer et du Preez, 2006). En bref, elle consiste à ajouter des SLR fictifs pour tromper l'algorithme et générer des valeurs finies même en dehors de la zone de chevauchement. Selon Brümmer et du Preez (2006), plus les données utilisées pour modéliser les distributions sont nombreuses, plus l'effet des scores factices est faible, mais cela reste un point critique supplémentaire à aborder dans le cadre d'une expertise forensique. Par conséquent, la régression logistique peut être préférée dans certains cas car elle produit un SLR calibré sur un intervalle fini.

V.3. Synthèse

En nous basant sur la littérature actuelle et nos objectifs, nous avons développé une méthodologie permettant d'évaluer les performances de plusieurs modèles probabilistes adaptés aux scénarios décrits dans le chapitre précédent. Nos choix peuvent être résumés ainsi :

Modèles probabilistes	Intravariabilité	Spécifique au suspect
		Générique
	Intervariabilité	Spécifique au suspect
		Spécifique à la trace
	Calibration	PAVA
		Régression logistique
Performances	Métriques	Discrimination (RMEP, RMED)
		Calibration (minCllr)
	Représentations graphique	Distributions
		ECE plots

Chapitre VI. Contribution de la reconnaissance faciale à l'enquête

Dans ce chapitre, nous nous focalisons sur l'utilisation des systèmes automatiques dans le cadre investigatif, tel que résumé dans la Figure 22. L'objectif est d'évaluer les performances de chaque système dans les tâches de recherche de POI dans une large BdD judiciaire, par l'utilisation des scores de comparaison « bruts » fournis par les systèmes dans un premier temps, puis avec le calcul de SLR par le modèle adéquat.

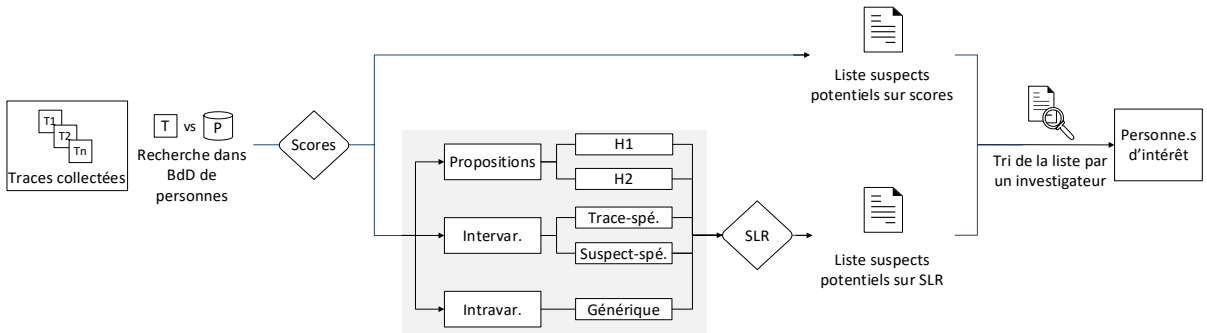


Figure 22 : Processus de développement et application de modèles de calculs de SLR dans le cadre investigatif.

VI.1. Performances de la recherche de POI dans une base de données judiciaire

Pour chaque scénario, le nombre de transactions correspond au nombre de fois où une trace (ATM, ID ou CCTV) est comparée à la base de données (BdD), sachant que la POI visible sur la trace est bien présente dans la BdD. Le nombre de transactions ne traduit pas directement le nombre de traces, car chaque trace est comparée à plusieurs images de référence du même individu connus. Au total, le nombre total de recherches est de 3980 et 1468 pour les traces ATM-CCTV et ID, respectivement. Le nombre de traces, de références et de recherches par scénario sont résumés dans le tableau ci-dessous.

Tableau 5 : Résumé du nombre de données et de recherches de POI pour chaque scénario.

Scénario	Nombre d'individus	Nombre moyen de traces par individu	Nombre de référence par individu	Nombre de recherches de POI	Taille de BdD (judiciaire + références)
ATM-CCTV	32	23 (±7)	4 ou 8	3980	36392 + 168
ID	34	8 (±4)		1468	

À chaque recherche, le système compare la trace avec chaque individu de la BdD puis produit la liste des scores de comparaison, à partir du visage le plus ressemblant (Rang 1) puis par ordre décroissant. Le but est ensuite d'observer à quel rang la POI recherchée apparait dans la liste pour la première fois.

VI.1.1 FaceNet

La Figure 23 résume les performances de l'algorithme FaceNet lors de la recherche de POI dans une base de données de 36'560 images (36'392 photos d'identité judiciaire + 168 *mugshots* des POI), pour les trois scénarios étudiés : ATM, ID et CCTV.

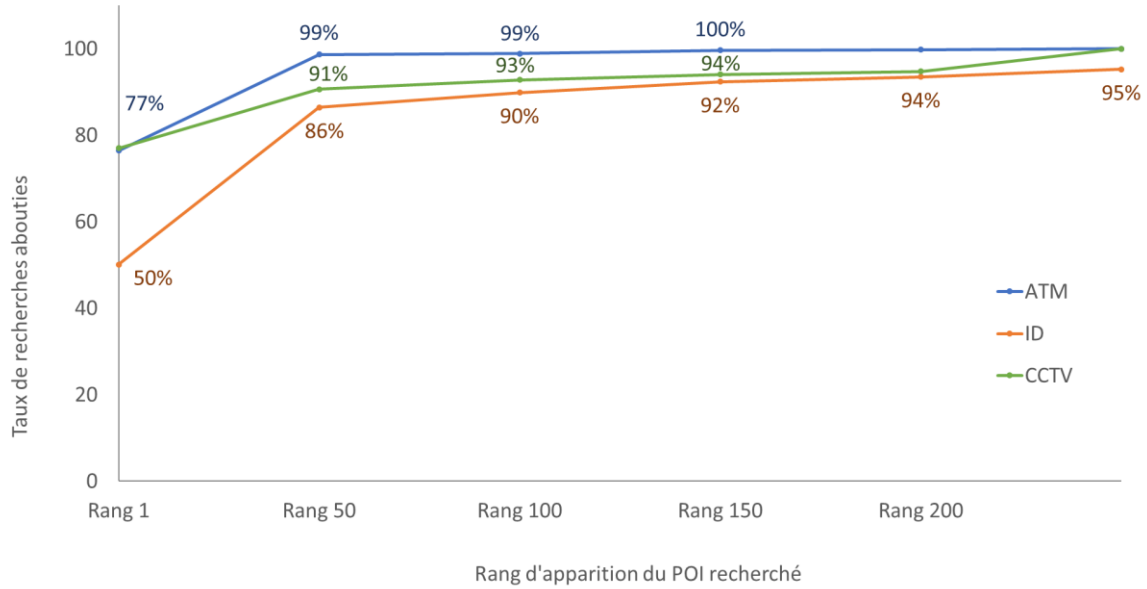
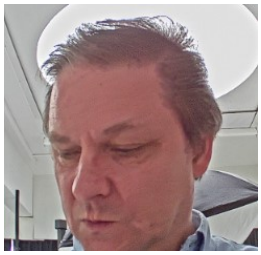


Figure 23 : Performances de FaceNet sur des tâches de recherche de POI dans une base de données sur des traces de type ATM (bleu), CCTV (vert), ID (orange).

VI.1.1.1 ATM

Avec l'algorithme FaceNet, les meilleurs résultats proviennent des recherches à partir de trace de type ATM. Dans 45% des recherches, l'image de référence de l'individu à reconnaitre est le premier résultat de la liste de scores (rang 1), c'est-à-dire, le système a identifié le bon individu. Dans 42% des comparaisons le système place l'individu visé entre les rangs 2 et 50. Dans la pratique, cela implique que, lorsque les investigateur.trice.s recherchent une personne dans leur BdD à partir d'une image issue d'ATM, le système FaceNet trie la totalité d'une BdD de plusieurs milliers d'individus, produit la liste des 50 plus hauts scores et, dans 87% des recherches, il est attendu que cette liste de candidats potentiels contient la POI visible sur la trace (s'il existe dans la base). L'opérateur peut alors vérifier chacun d'entre eux manuellement afin de sélectionner la ou les personnes à considérer dans l'enquête. En augmentant la taille de la liste aux 200 premiers résultats, la POI est retrouvée dans 94% des cas. Alors que 200 images peut sembler être un nombre conséquent de comparaisons visuelles à effectuer, il s'agit en réalité de comparaisons rapides, car sur l'ensemble de ces résultats, seuls quelques-uns peuvent s'avérer difficiles à comparer du fait d'une ressemblance fortuite avec la POI de la trace. En effet, les visages « les plus ressemblants » d'après le système ne sont pas nécessairement les plus ressemblants d'après la perception humaine. Enfin, dans 6% des recherches d'individus, l'identité attendue n'apparait qu'au-delà des 200 premiers résultats.

La Figure 24 permet de rentrer dans les détails de ces performances à partir d'un exemple de recherche d'un POI (appelée POI Z) à partir de la trace ci-dessous.



Trace « ATM »

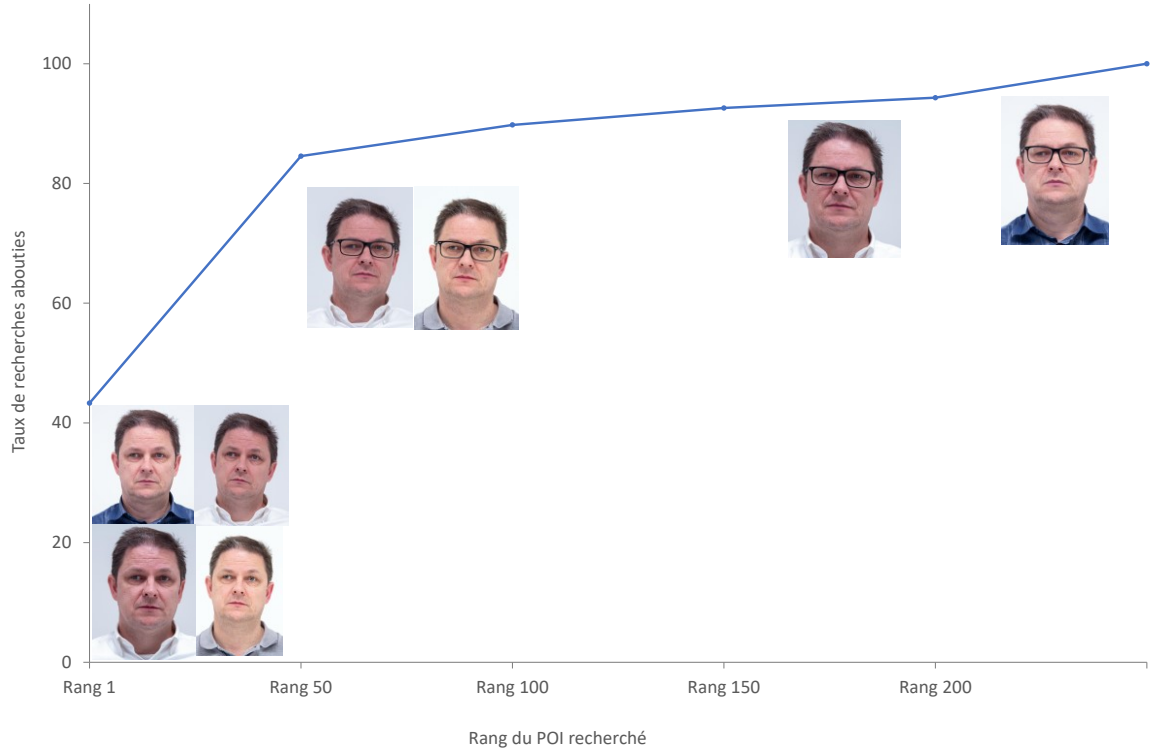


Figure 24 : Visualisation du rang auquel chaque mugshot de la POI Z apparait dans la liste de résultats de comparaison avec la trace ATM ci-dessus.

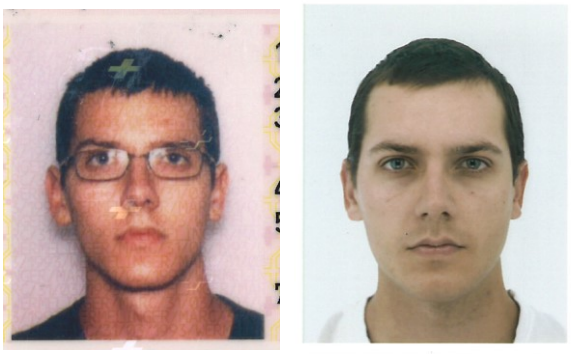
Ces exemples sont intéressants car ils soulèvent plusieurs questions sur les éléments que l'algorithme privilégie pour la comparaison de visages, à partir d'une trace de faible qualité du fait de l'angle de prise de vue.

Les quatre images les mieux placées (à gauche) sont situées entre les rangs 1 et 9 puis les quatre suivantes sont apparues respectivement aux rangs 56, 76, 178 et 246. On remarque immédiatement que les images les mieux placées sont celles où la POI ne porte pas de lunettes, comme c'est le cas sur la trace ATM. Il semble donc que FaceNet inclut non seulement les lunettes dans la comparaison de visages mais leur accorde un poids prédominant. En outre, les *mugshots* sur lesquels la POI porte ses lunettes apparaissent ne présente pas de différence significative pour l'œil humain et apparaissent pourtant à des rangs très variables, dont un au-delà du rang 200. S'il est difficile de définir avec précision tous les critères diminuant les performances de FaceNet lors de la recherche de POI dans de larges BdD, il apparait essentiel de prendre des *mugshots* des individus interpellés avec et sans lunettes dès que cela est possible, pour parer la chute de performances de l'algorithme dans une recherche future du même individu.

VI.1.1.2 ID

Concernant le scénario ID, la POI recherchée apparait au rang 1 dans 42% et entre les rangs 2 et 50 dans 35% des recherches. En tout, les listes des 100 et 200 meilleurs résultats contiennent l'individu recherché dans respectivement 81% et 82% des cas. L'utilisation de l'algorithme FaceNet s'avère donc utile également dans les cas où la trace est un portrait. Cependant, certains types d'images ont mené à des scores plus faibles que ce qui pouvait être attendu de traces de bonne qualité. Dans 15% des cas, la POI recherchée se classe au-delà du 200^{e} rang, et 3% des recherches n'ont pas abouti. Cela signifie que FaceNet n'a pas réussi à détecter de visage dans la trace.

La Figure 25 oppose deux traces d'un même POI. Les recherches à partir de la photo (a), une numérisation de carte d'identité datant de 2011, aboutissent à une classification au-delà du rang 1300. Pour l'image (b), une numérisation de photographie standardisée de 2018 pas encore utilisée sur un document officiel, les références de la POI en question sont toutes classées aux rangs 1 et 2.



(a) (b)

Figure 25 : Traces ID de la POI H apparaissant (a) au-delà du rang 1300 et (b) aux rangs 1 et 2 pour toutes les recherches en BdD.

Ces images peuvent toutes les deux être considérées de bonne qualité par un opérateur humain et pourtant l'algorithme ne réussit à remonter l'individu que sur la seconde. En observant plus attentivement l'image (a), plusieurs éléments peuvent être à l'origine de cette chute de performance : les reflets blancs sur le verre des lunettes peuvent gêner la détection correcte des yeux, les éléments de sécurité (ici, des croix réfléchissantes) spécifiques aux documents officiels d'identité peuvent être détectés et analysés par le système comme des caractéristiques faciales, les variations morphologiques de la POI causées par son vieillissement (sept années séparent les deux clichés) et la différence de prise de vue . En effet en se focalisant sur la visibilité des oreilles de l'individu, et les proportions des différents éléments faciaux les uns par rapport aux autres, l'image (b) semble avoir été prise avec une plus faible distance focale que l'image (a).

Enfin, neuf traces ID (3%) n'ont pas pu être comparées à la BdD car FaceNet n'y a pas détecté de visage au préalable. Les trois images de la Figure 26 donnent un bon aperçu des éléments qui semblent bloquer l'algorithme : obstruction partielle du visage (image (a)), très faible distance

focale, variations de luminosité et expression faciale (image (b)), et la faible résolution et les éléments de sécurité de documents d'identité (image (c)).

(a) (b) (c)

Figure 26 : Traces ID de trois POI différentes sur lesquelles FaceNet ne détecte pas de visage.

En définitive, l'ajout d'éléments de sécurité, la copie de mauvaise qualité, la grande différence d'âge entre la trace et la référence, etc. sont autant de critères à prendre en considération lors de l'utilisation opérationnelle de ce système. Dans un cas où les investigateur.trice.s disposent de documents d'identité récents et de bonne qualité, la recherche dans la base de données donnera le bon résultat avant le rang 200, alors qu'à partir d'un document plus ancien et/ou usé, il sera nécessaire de vérifier une liste plus longue et les risques de non-détection de visage augmentent.

VI.1.1.3 CCTV

Dans les cas de recherche de personne à partir d'enregistrements CCTV, l'algorithme a détecté correctement les visages de toutes les traces, contrairement à ce que l'on pouvait attendre du fait de la faible qualité de ce type d'images. Dans 50% des cas, la POI recherchée figure entre les 1 et 50 de la liste de scores, mais pour 38% d'entre eux la POI se retrouve au-delà du rang 200. Il est attendu que les causes à l'origine de cette chute des performances soient la plus faible résolution du visage sur la trace à cause de la plus grande distance séparant la POI de la caméra, ainsi que les variations de l'angle de prise de vue.

La Figure 27 oppose deux traces d'un même POI. Le visage de la POI sur la trace (a) est photographié avec un angle latéral important (environ 40°) et en plongée. De fait, la moitié du visage est dissimulée et les proportions du haut du crâne sont accentuées par rapport au bas du visage. Néanmoins, l'algorithme classe cinq références de la POI en question aux rangs 1 à 6 et la sixième au rang 160. Sur l'image (b), la POI est plus éloignée de la caméra, ce qui réduit le nombre de pixels de la zone du visage, mais il est de face et la distance atténue également l'effet de prise de vue en plongée. Néanmoins, toutes les recherches à partir de l'image (b) aboutissent à une classification de la POI au-delà du rang 2900. Cela suggère que l'algorithme gère mieux les défauts d'alignement que les faibles résolutions. Pour un usage opérationnel, cela implique que les investigateur.trice.s doivent privilégier les traces sur lesquelles la POI est proche de la caméra, même si le visage est tourné (en deçà de 40°), plutôt que des images avec un meilleur alignement

mais plus lointaine, ce qui diminue la taille (et donc la résolution) du visage. Il serait intéressant de se focaliser sur les solutions en termes de traitement d'image, en particulier pour la gestion des faibles résolutions.

(a) (b)

Figure 27 : Traces CCTV de la POI E apparaissant (a) aux rangs 1 à 6 puis 160 et (b) au-delà du rang 2900 pour toutes les recherches en BdD.

De la même manière que pour les traces ID, la Figure 28 permet de visualiser l'étendue des valeurs de rangs auxquels apparaissent les références de la POI E à partir d'une même trace CCTV. Les quatre images les mieux placées (à gauche) sont situées aux rangs 1 et 8 puis les quatre suivantes sont apparues respectivement aux rangs 46, 86, 137 et 284.

Ce résultat soulève quelques questions, car du fait de l'angle gauche et en plongée important (environ 45°), et de la dissimilation partielle de la zone oculaire par des lunettes de vue, cinq des six références sont classées dans les 150 premiers résultats de la liste de scores. En outre, les deux meilleurs classements, au rang 1, résultent de la comparaison avec les deux références sur lesquelles la POI porte des lunettes identiques à celles portées sur la trace, ce qui laisse à penser que l'algorithme analyse et compare cet élément et lui accorde un poids important. Néanmoins, la POI porte également ces mêmes lunettes sur la trace (b) de la Figure 27, et cela ne semble pas avoir pesé significativement dans la comparaison avec ces deux références.

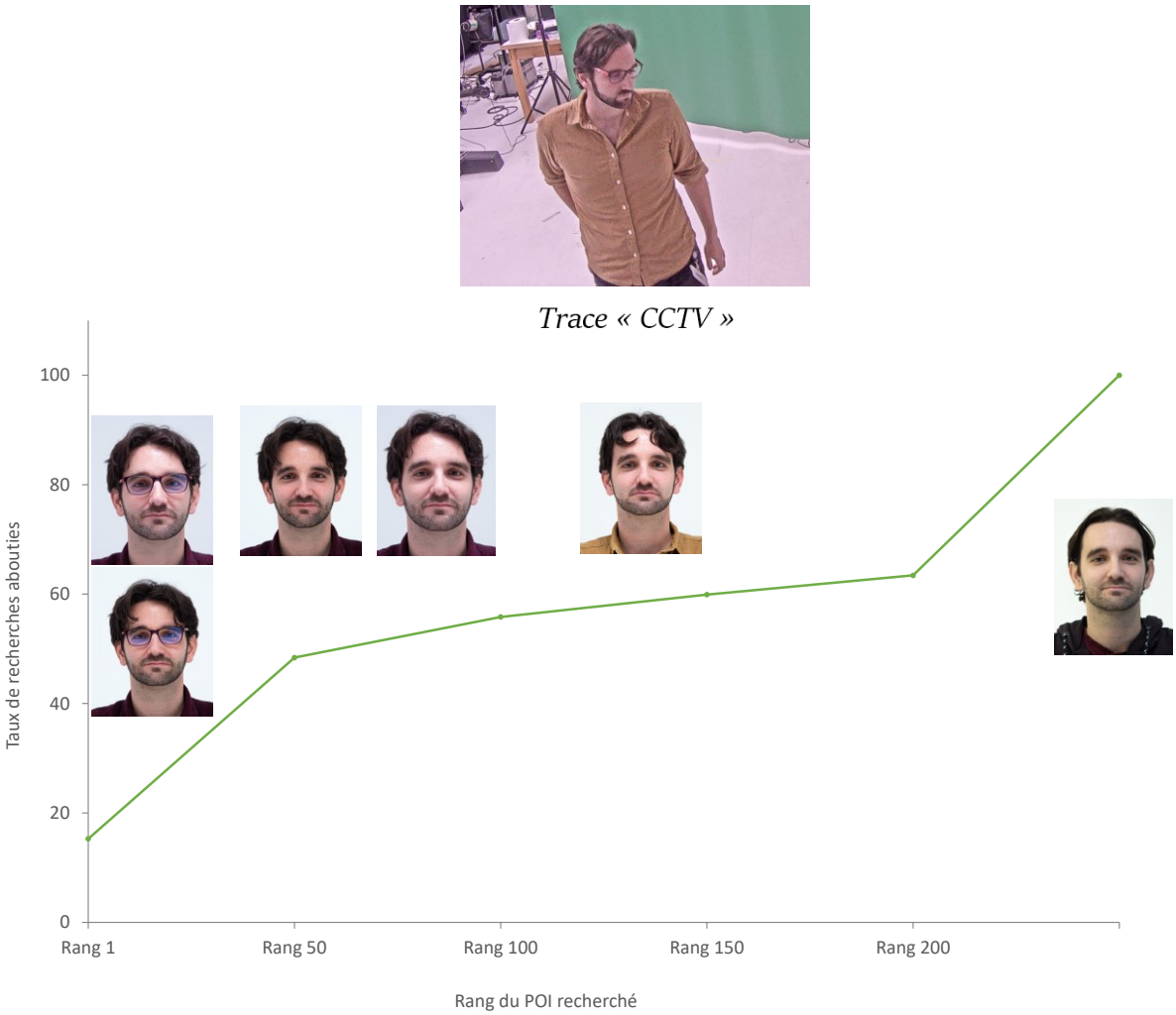


Figure 28 : Visualisation du rang auquel chaque mugshot de la POI E apparait dans la liste de résultats de comparaison avec la trace CCTV ci-dessus.

VI.1.2 MFI

Avant d'analyser ces résultats et à la vue des taux très élevés de recherches non abouties pour le MFI, il faut noter que les listes de résultats générer par les systèmes MorphoFace (aussi bien MFI que MFE) ne renvoient pas l'intégralité des résultats pour toutes les recherches. Dans leur état opérationnel typique, ces systèmes sont développés pour sortir une liste réduite à quelques dizaines de candidats potentiels maximum en fonction du « seuil d'identification » fixé. Pour les besoins du présent projet, les algorithmes MorphoFace ont été au préalable modifiés par Idemia afin de fournir le plus de candidats possible, quels que soient les scores de comparaison, à savoir 4'000 candidats pour chaque recherche. Les candidats au-delà du rang 4'000 ne figurent donc pas dans les listes finales, et les recherches non abouties ne correspondent donc pas uniquement aux images que le système n'aurait pas réussi à analyser (à cause de leur trop faible qualité) mais également aux résultats figurant au-delà du 4'000^{e} rang pour chaque recherche.

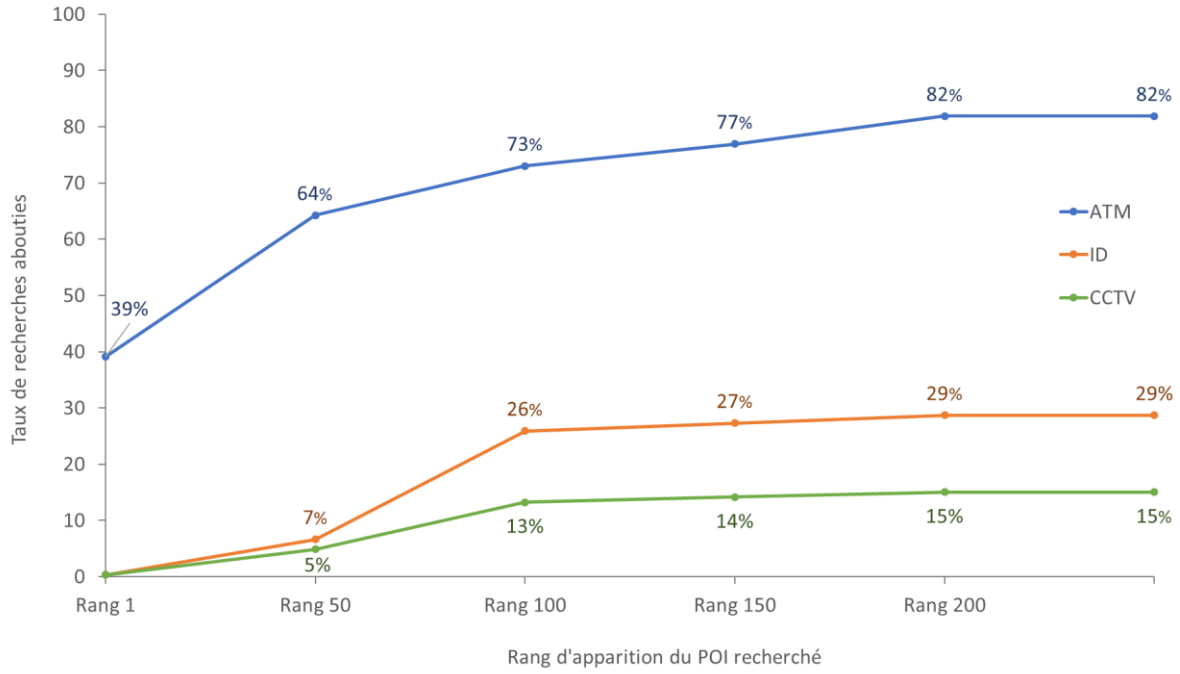


Figure 29 : Performances du système MFI sur des tâches de recherche de POI dans une base de données sur des traces de type ATM (bleu), CCTV (vert), ID (orange).

Les performances du MFI sont très inférieures à celles de FaceNet pour les trois scénarios. Les performances à partir de traces ATM sont néanmoins largement supérieures à celles des scénarios CCTV et ID. Dans 39% des recherches, le suspect recherché se retrouve au premier rang des résultats, et dans 78% des cas il apparait dans les 200 premiers rangs. Les 22% de recherches restants n'ont renvoyé aucun résultat. Comme nous l'avons précisé précédemment, cela peut être dû à deux facteurs : l'algorithme n'a pas su détecter de visage dans l'image trace, ou le résultat se trouvait au-delà du 4'000ème rang, plafond du nombre de résultats retournés par le système à l'opérateur. Ce sont exactement 38 traces ATM, 497 CCTV et 27 ID qui ne figurent pas parmi les résultats de comparaisons. Pour les scénarios CCTV et ID, cela correspondant respectivement à 86% et 73% des recherches. Lorsque le MFI classe la POI recherchée dans la liste de résultats, il le place dans les 50 premiers rangs dans 6 et 5% des cas, respectivement, et au rang 1 dans 0.3% des cas uniquement.

En investiguant plus en détail les traces ATM manquantes, il apparait que leur seul point commun est la faible visibilité des yeux, que ce soit à cause de l'angle de prise de vue, du port de lunettes ou de la direction du regard. Concernant les images CCTV, la proportion de traces absentes de la liste finale de scores représente près de 92% du nombre total de traces enregistrées. Il apparait donc plus pertinent de directement observer les images que le MFI a réussi à comparer, dont 0.1% ont mené à une identification correcte au rang 1, puis 2%, 6% et 1% ont mené à des résultats en deçà des rangs 50, 100 et 200, respectivement.

La Figure 30 présente des comparaisons pour lesquelles l'individu recherché apparait au rang 1. On observe avant tout que pour chaque paire, la trace CCTV a mené à l'identification de l'individu sur des *mugshots* où il porte ou non des lunettes, en accord avec l'image trace. Il se peut donc

que les lunettes fassent partie des éléments analysés et comparés par le MFI, comme observé avec l'algorithme FaceNet, d'autant plus que tous les résultats d'identification correcte entre les rangs 1 et 29 (77 recherches) proviennent de comparaison d'images où la présence/absence de lunettes est constante entre la trace et la référence.

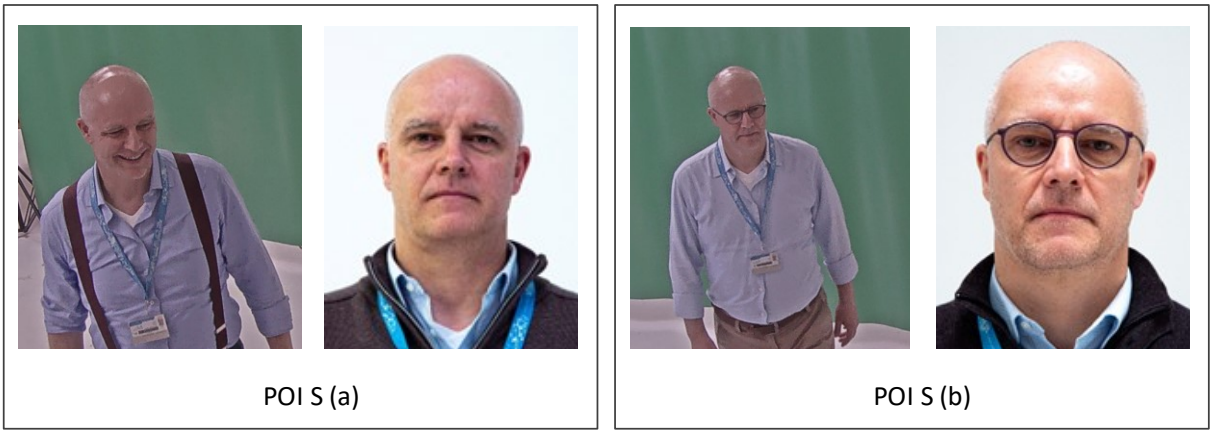


Figure 30 : Traces CCTV de la POI S apparaissant aux rang 1 lors des recherches en BdD.

Enfin, concernant les traces ID absentes des listes de résultats, le MFI semble être freiné par plusieurs variables, résumées par les exemples de la Figure 31. La première image est un portrait de très bonne qualité qui a dû donc être correctement analysé par l'algorithme, mais près de 40 ans séparent cette trace et les références intégrées à la base de données ; la comparaison a donc dû générer un score minimal, se situant au-delà des 4'000 premiers résultats. La seconde trace est également de qualité suffisante pour être considérée comme analysable par le MFI, mais l'expression du visage – différente de l'expression neutre des POI sur les *mugshots* en base, l'angle du visage et la fermeture quasi totale des yeux de l'individu peuvent avoir compliqué la comparaison et engendré un score également trop faible pour figurer dans la liste finale. Enfin, le troisième exemple illustre les dégradations subies par les photographies sur les documents officiels d'identité, causés par l'ajout d'élément de sécurité attestant de l'authenticité des documents. Dans ce cas, il est possible qu'un score de comparaison trop faible ait pu être généré, mais cette trace a pu également être considérée de qualité insuffisante pour comparaison par l'algorithme.

Figure 31 : Traces ID de trois POI différents, non détectées dans les listes de résultats suite à la recherche de POI dans la BdD par le système MFI.

VI.1.3 MFE

Lors de tâches d'identification 1vN sur notre BdD de 57'977 images signalétiques, le système MFE reconnait la personne recherchée au rang 1 dans toutes les comparaisons de traces ATM, et dans 98% des comparaisons de traces ID (Figure 32). Bien que – logiquement - légèrement inférieurs, les résultats à partir de traces CCTV sont également très élevés, puisque dans 85% des recherches le suspect recherché est identifié au rang 1, et dans 14% entre les rangs 2 et 50.

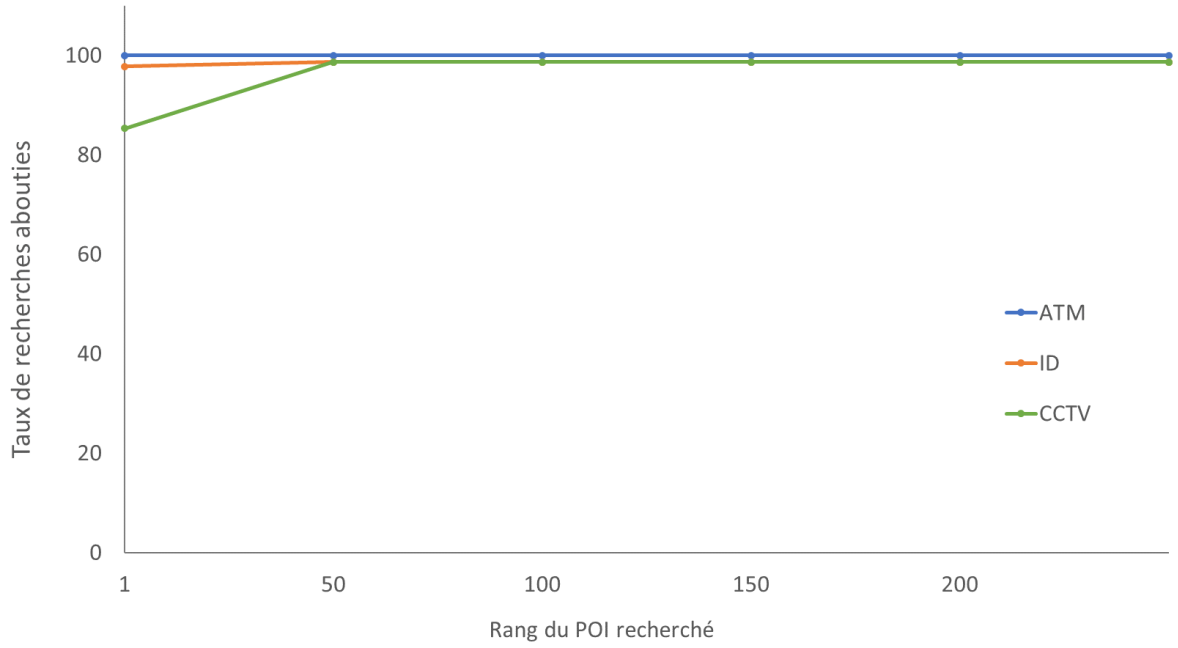


Figure 32 : Performances du système MFE sur des tâches de recherche de POI dans une base de données sur des traces de type ATM (bleu), CCTV (vert), ID (orange).

Pour les scénarios CCTV et ID, 1% des comparaisons n'apparaissent pas dans la liste finale des scores. Plus précisément, cela représente 20 traces CCTV où, comme présenté dans les exemples ci-dessous (Figure 33), les yeux de la POI ne sont pas ou peu visibles à cause de l'angle de prise de vue, et 1 trace ID issue du scan de très mauvaise qualité d'une carte de transport (très faible qualité de détail, manque de texture du visage, flou, etc.).

Figure 33 : Exemple de traces CCTV non détectées dans les listes de scores générés par MFE lors de la recherche de POI en base de données.

À la vue de tels résultats, nous avons jugé judicieux de vérifier les premières dizaines de rangs, pour nous assurer que le MFE ne discriminait pas principalement les *mugshots* des POI de ce projet (pris dans des conditions contrôlées par les auteurs) de ceux fournis par les services de police. Parmi les 200 premiers résultats pour chaque recherche, aucune tendance de la sorte n'est apparue : les *mugshots* de POI du projet n'apparaissent pas en proportions plus importantes par rapport aux références judiciaires.

VI.2. Apport du calcul de SLR dès la phase d'enquête

L'évaluation est couramment associée à la phase d'interprétation des résultats d'expertise à présenter au tribunal (Jackson *et al.*, 2006). Cependant, nous postulons que l'évaluation n'est pas un objectif visé exclusivement par l'expert.e dans ce cadre, mais davantage une méthode qui peut servir dans n'importe quelle phase. Cette notion a été abordée dans plusieurs publications en 2020 (Baechler *et al.*, 2020 ; Jacquet et Champod, 2020 ; Ryser *et al.*, 2020b ; Stoney *et al.*, 2020), ce qui démontre l'intérêt porté à l'adoucissement de la sectorisation des phases de l'instruction judiciaire, dans le but d'améliorer les performances et la communication entre les acteurs des différentes phases. Nous abordons cette notion plus en détail dans le chapitre suivant (*cf.* VII.3.2 p.104).

Dans cette section, nous étudions l'impact potentiel de l'application d'un modèle évaluatif dès la phase d'enquête, en comparant les performances des systèmes aux taches de recherche de POI basées sur les listes de scores à celles obtenues à partir de listes de candidats potentiels directement triées en fonction de SLR.

VI.2.1 Spécificités du SLR investigatif

Le SLR investigatif est un cas spécial du calcul de SLR. Lors de l'enquête, l'investigateur.trice effectue une recherche de POI dans une base de données, et le système lui renvoie une liste de candidats à vérifier manuellement. Ces résultats sont le plus souvent communiqués de manière informelle pour guider les enquêteurs, cependant il serait utile de leur fournir des résultats pondérés dès la phase d'enquête. À cette étape, les seules données disponibles sont la trace du cas de question et les références des individus de la BdD judiciaire. Certains individus peuvent être présents sur plusieurs images de la base, mais dans la majorité des cas elle contient 1 à 2 images de chaque individu. Pour cette évaluation, le calcul du SLR investigatif est donc basé sur la modélisation d'une intravariabilité générique afin de pallier le manque d'images de référence des POI référencées. Cela implique que la POI ne figure pas directement dans la proposition H_1, alors formulée : "La même personne est visible sur les deux images comparées" (Tableau 6). L'intervariabilité quant à elle peut être spécifique à la trace ou à l'image de référence du POI.

Tableau 6 : Formule de SLR et conditionnement des propositions H_1 et H_2 dans le calcul de SLR investigatifs.

Approches		Formules SLR	Propositions	
Intravariabilité	Intervariabilité		H1	H2
Générique	Suspect-spécifique	$=\frac{f(s(S,T)\|H_1,I)}{f(s(S,T)\|S,H_2,I)}$	"La même personne est visible sur les deux images comparées"	"Le suspect n'est pas la personnes sur l'image trace"
	Trace-spécifique	$=\frac{f(s(S,T)\|H_1,I)}{f(s(S,T)\|T,H_2,I)}$		"L'individu à la source de la trace n'est pas le suspect, mais une personne de la population d'intérêt"

VI.2.2 Performances des modèles

Dans cette section, nous analysons et discutons la validation des résultats de calibration des SLR, pour toutes les approches et scénarios testés selon le Tableau 6. Le Tableau 7, Tableau 8 et Tableau 9 résument les taux de SLR soutenant à tort la proposition de l'accusation H_1 (RMEP) et de la défense H_2 (RMED), ainsi que les métriques liées aux coûts associés à chaque modèle (Cllr/minCllr). Dans chaque cas, les impacts respectifs des deux méthodes de calibrations – PAVA et Reg Log – sont mis en évidence par la comparaison avec les SLR bruts (c.-à-d. les SLR reportés directement sur les valeurs de scores générées par les systèmes). En outre, les valeurs de SLR sont également comparées en regard des deux approches d'intervariabilité testées pour identifier la plus performante.

Lors de la phase investigative, le modèle évaluatif doit garantir des taux minimaux de RMED afin de limiter le risque de ne pas détecter la POI visée. À ce stade, les SLR soutenant à tort H_1 sont moins préjudiciables car l'implication incorrecte d'une POI peut être rapidement exclue grâce à d'autres informations d'enquête.

Tous les résultats exposés par la suite détaillent à la fois les valeurs des SLR « bruts » et des SLR calibrés par PAVA et RegLog ainsi que les taux d'erreurs associés à chacun. Les SLR bruts doivent être vus comme des valeurs intermédiaires, et cela permet d'être plus transparent sur l'évolution de nos résultats, et cela montre plus explicitement l'effet de chaque calibration.

VI.2.2.1 FaceNet

Le Tableau 7 résume les performances du calcul de SLR à l'aide des scores générés par l'algorithme FaceNet.

Tableau 7 : Performances du système FaceNet pour les modèles investigatifs de calcul de SLR bruts et calibrés.

Intravar.	Intervar.	Métriques	ATM			CCTV			ID		
			SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA
Générique	Suspect - spécifique	RMEP (%)	4,85 ± 0,01	1,82 ± 0,01	1,93 ± 0,01	4,85 ± 0,01	7,00 ± 0,01	6,59 ± 0,02	4,85 ± 0,01	3,51 ± 0,01	3,02 ± 0,02
		RMED (%)	0,51 ± 0,00	1,85 ± 0,00	1,61 ± 0,01	12,88 ± 0,00	9,56 ± 4,4E-03	9,80 ± 0,01	4,92 ± 0,00	5,74 ± 1,3E-03	6,04 ± 0,02
		(min)Cllr	0,116 ± 2,4E-04	0,074 ± 1,9E-04	0,069 ± 1,7E-04	0,322 ± 2,4E-04	0,305 ± 2,6E-04	0,296 ± 2,5E-04	0,176 ± 2,3E-04	0,171 ± 2,0E-04	0,162 ± 1,8E-04
	Trace - spécifique	RMEP (%)	4,79 ± 0,01	1,62 ± 0,01	1,59 ± 0,01	4,18 ± 0,01	7,86 ± 0,01	7,00 ± 0,03	4,29 ± 0,01	3,09 ± 0,01	2,57 ± 0,01
		RMED (%)	0,39 ± 0,00	1,43 ± 2,6E-03	1,42 ± 0,01	13,35 ± 0,00	8,37 ± 4,8E-03	9,01 ± 0,03	4,65 ± 0,00	5,41 ± 2,5E-03	5,64 ± 0,01
		(min)Cllr	0,112 ± 2,3E-04	0,065 ± 1,8E-04	0,061 ± 1,6E-04	0,326 ± 1,5E-04	0,299 ± 2,2E-04	0,290 ± 2,3E-04	0,160 ± 2,3E-04	0,156 ± 1,9E-04	0,147 ± 1,8E-04

En observant tout d'abord les SLR bruts pour chaque scénario, les valeurs de RMEP sont identiques (4.85% ±0.01), alors que les RMED varient significativement selon le scénario (0.51% pour les traces ATM, 12.88% pour les CCTV et 4.92% pour les ID). W<

L'analyse des résultats sous l'angle investigatif révèle que les meilleurs RMED sont obtenus sur les SLR bruts (sans calibration Reg log ni PAVA), à partir d'images de bonne qualité, à savoir les traces ATM et ID. Pour ces deux scénarios, les deux méthodes de calibrations diminuent les RMEP, mais augmentent les RMED. À partir de traces CCTV, les meilleurs résultats sont obtenus par la calibration Reg Log car elle permet de réduire les RMED de 13% à 9% environ, malgré une augmentation des RMEP (de 4-5% à 7-8%).

Pour comparer les résultats en fonction de l'approche de modélisation d'intervariabilité, il est nécessaire de prendre en compte les distributions des valeurs de SLR calibrés. En effet, l'observation seules des valeurs du Tableau 2 indique que le modèle à l'intervariabilité trace-spécifique est à privilégier pour le scénario ID, et que les deux modèles se valent pour les scénarios ATM et CCTV. Pour ces derniers, la seule différence permettant de discriminer les deux modèles est non décelable à partir des valeurs du Tableau 7 uniquement. Il s'agit de l'impact de la calibration sur les valeurs de SLR, notamment sous H_1. Dans l'exemple du scénario CCTV (Figure 34), les SLR bruts sous H_1 atteignent des valeurs env. $Log_{10}(SLR) = 10$ pour le modèle suspect-spécifique, et plafonnent à env. $Log_{10}(SLR) = 5$ pour le modèle trace-spécifique.

Ce résultat est cohérent en considérant que, dans ce scénario, l'intervariabilité trace-spécifique est modéliser par la comparaison des traces CCTV (de qualité faible à moyenne) au *mugshots* de la BdD, alors que l'approche suspect-spécifique compare la référence du suspect de qualité comparable à celle des images de la BdD. Il est donc logique d'obtenir de meilleures performances avec le modèle suspect-spécifique.

En outre, les valeurs des coûts associés aux modèles calibrés (minCllr) sont systématiquement plus faibles pour le traitement PAVA. Bien que la différence entre les coûts des modèles RegLog et PAVA soient très faibles pour les trois scénarios, cela orienterait vers le choix de la calibration PAVA.

Il est important de noter que le traitement PAVA permet d'obtenir des SLR calibrés sur les mêmes intervalles pour les deux approches (trace et suspect-spécifique), alors que la calibration RegLog conserve les formes des distributions et donc leur fort écart des valeurs maximales de SLR sous H_1.

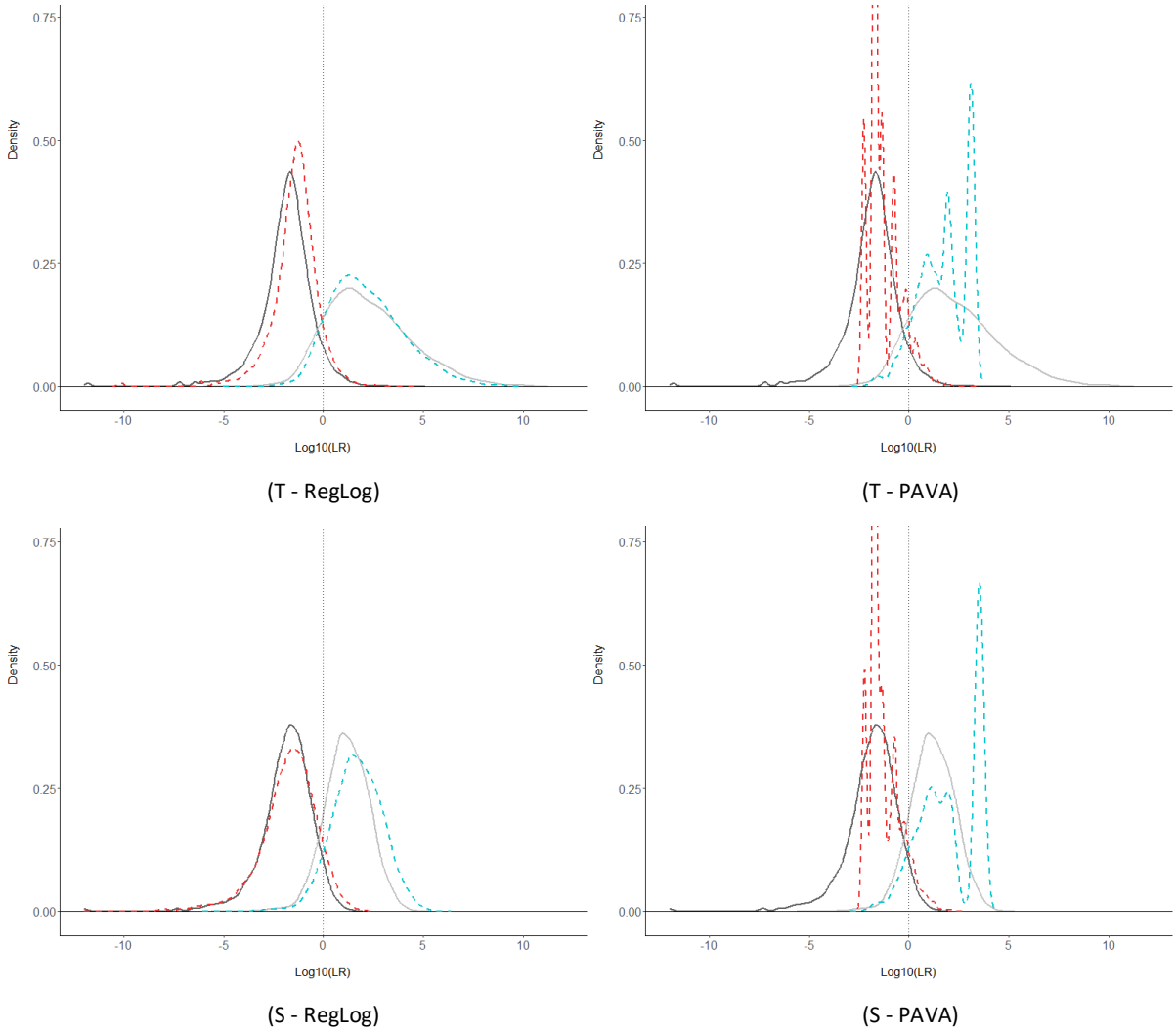


Figure 34 : Comparaison des distributions des SLR bruts (H_1 traits pleins clairs et H_2 traits pleins foncés) et des SLR calibrés (H_1 traits pointillés bleus et H_2 traits pointillés rouges) pour le scénario CCTV en fonction de la calibration (PAVA à droite et régression logistique à gauche) et du modèle d'intervariabilité (trace-spécifique en haut et suspect-spécifique en bas) avec FaceNet.

VI.2.2.2 MFI

Les valeurs minCllr surlignées en bleu dans le Tableau 8 sont celles approchant 1, ce qui traduit la neutralité, et donc le manque de pertinence, du modèle. De fait, l'ensemble des résultats du MFI pour le scénario CCTV montrent que, même calibrés, les SLR ne peuvent pas être utilisés dans le cadre investigatif.

Tableau 8 : Performances du système MFI pour les modèles investigatifs de calcul de SLR bruts et calibrés.

Intravar.	Intervar.	Métriques	ATM			CCTV			ID		
			SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA
Générique	Suspect-spécifique	RMEP (%)	2,64 ± 0,01	13,78 ± 1,3E-02	10,66 ± 3,4E-02	2,68 ± 9,6E-03	24,74 ± 0,03	16,72 ± 9,0E-02	2,66 ± 9,5E-03	16,59 ± 0,01	8,57 ± 0,11
		RMED (%)	30,16 ± 0,00	14,92 ± 6,4E-03	17,22 ± 2,9E-02	91,81 ± 0,00	68,43 ± 0,00	72,74 ± 9,1E-02	35,98 ± 0,00	18,05 ± 0,01	25,03 ± 0,11
		(min)Cllr	1,23 ± 1,6E-04	0,49 ± 1,9E-04	0,46 ± 1,9E-04	5,99 ± 1,6E-04	0,99 ± 4,9E-05	0,94 ± 1,1E-04	1,59 ± 1,5E-04	0,56 ± 1,9E-04	0,52 ± 1,8E-04
	Trace-spécifique	RMEP (%)	3,09 ± 0,01	14,02 ± 1,3E-02	10,78 ± 6,3E-02	2,79 ± 0,01	16,39 ± 0,03	22,75 ± 0,05	3,03 ± 9,9E-03	16,53 ± 0,01	8,78 ± 0,04
		RMED (%)	29,90 ± 0,00	15,53 ± 6,9E-03	18,12 ± 6,4E-02	87,55 ± 0,00	61,20 ± 0,00	55,78 ± 0,05	36,27 ± 0,00	19,84 ± 5,6E-03	25,72 ± 0,03
		(min)Cllr	1,26 ± 1,6E-04	0,50 ± 1,8E-04	0,47 ± 1,7E-04	5,17 ± 1,6E-04	0,95 ± 1,1E-04	0,93 ± 1,2E-04	1,62 ± 1,5E-04	0,58 ± 1,9E-04	0,54 ± 1,7E-04

Pour les scénarios ATM et ID, les SLR bruts sont associés à de faibles RMEP, env. 3%, et de hauts RMED, entre 30 et 36% (pour les deux approches). Les calibrations équilibrent les taux d'erreurs en augmentant significativement les RMEP, jusqu'à environ 14% pour les ATM et 17% pour les ID, et diminuant de moitié les RMED. Pour une utilisation dans le cadre investigatif, en privilégiant donc les RMED les plus faibles, le choix se porterait donc vers un modèle calibré par RegLog.

La Figure 35 montre plus concrètement l'étendue des zones de chevauchement des distributions de SLR RegLog sous H_1 et H_2, pour les scénarios ATM et ID. Les SLR logarithmes (Log_{10}(SLR)) dont la valeur se situe entre env. -1 et +1 sont très peu fiables. Or, cela représente la quasi-totalité des SLR obtenus sous H_2. De fait, seuls les SLR sous H_1 au-delà de Log_{10}(SLR) = 3 ont une fiabilité accrue.

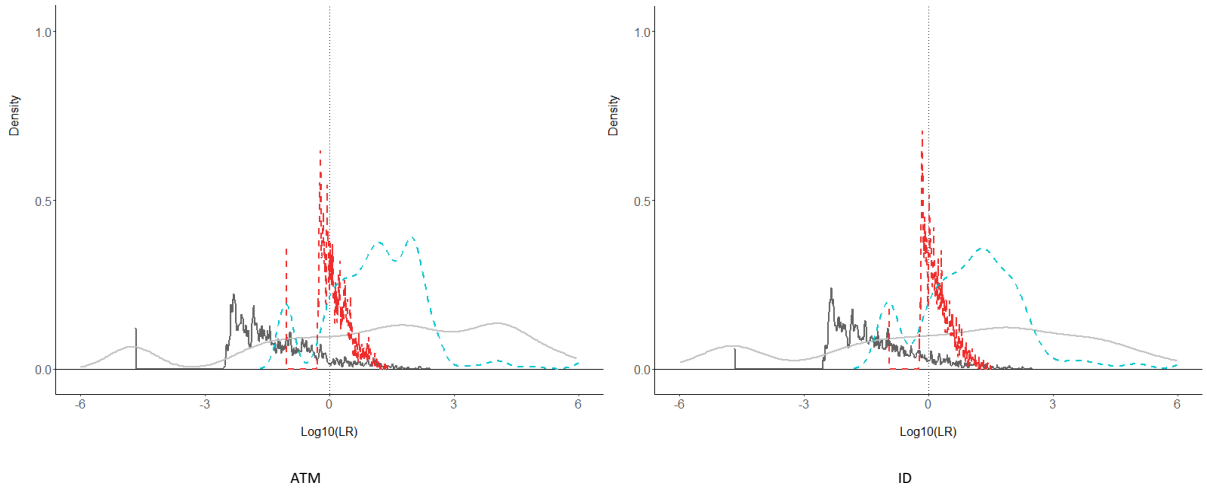


Figure 35 : Comparaison des distributions des SLR bruts (H_1 traits pleins clairs et H_2 traits pleins foncés) et des SLR calibrés par régression logistique (H_1 traits pointillés bleus et H_2 traits pointillés rouges) en fonction du scénario (ATM à gauche et ID à droite).

En regard des faibles performances du système MFI dans le calcul de SLR, il n'a pas été utilisé dans les tests suivants.

VI.2.2.3 MFE

Les résultats obtenus par le système MFE sont considérablement plus fiables que ceux du MFI (Tableau 9), ce qui était attendu.

Tableau 9 : Performances du système MFE pour les modèles investigatifs de calcul de SLR bruts et calibrés.

Intravar.	Intervar.	Métriques	ATM			CCTV			ID		
			SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA
Générique	Suspect-spécifique	RMEP (%)	5,5E-03 ± 4,6E-04	0,04 ± 9,7E-04	0,04 ± 1,2E-03	5,8E-03 ± 4,7E-04	5,81 ± 8,9E-03	3,40 ± 0,01	5,7E-03 ± 4,7E-04	0,49 ± 2,7E-03	0,42 ± 4,7E-03
		RMED (%)	0,23 ± 0,00	0,03 ± 7,2E-04	9,2E-03 ± 6,6E-04	40,69 ± 0,00	13,80 ± 5,3E-03	15,52 ± 5,9E-03	3,33 ± 0,00	0,67 ± 9,8E-04	0,65 ± 3,4E-03
		(min)Cllr	3,3E-03 ± 2,8E-05	1,3E-03 ± 3,2E-05	1,1E-03 ± 2,1E-05	2,82 ± 9,7E-06	0,40 ± 1,2E-04	0,36 ± 1,4E-04	0,13 ± 1,8E-05	0,03 ± 5,3E-05	0,03 ± 5,9E-05
	Trace-spécifique	RMEP (%)	0,00 ± 4,6E-04	0,02 ± 9,1E-04	0,04 ± 1,2E-03	8,5E-03 ± 5,5E-04	2,91 ± 6,1E-03	2,65 ± 0,02	0,00 ± 0,00	0,27 ± 2,0E-03	0,24 ± 2,9E-03
		RMED (%)	0,08 ± 0,00	0,04 ± 7,2E-04	0,00 ± 6,6E-04	23,84 ± 0,00	7,20 ± 2,1E-03	7,34 ± 0,02	2,13 ± 0,00	0,48 ± 1,3E-03	0,47 ± 1,2E-03
		(min)Cllr	0,001 ± 2,2E-05	0,001 ± 3,7E-05	0,001 ± 2,1E-05	1,59 ± 1,9E-05	0,23 ± 1,0E-04	0,22 ± 1,1E-04	0,09 ± 0,00	0,02 ± 3,5E-05	0,02 ± 4,1E-05

En privilégiant les bas taux de RMED, l'approche trace-spécifique est la plus performante pour tous types de traces et ce sur tous les SLR calibrés. Néanmoins, pour les traces ATM et ID, les variations de taux d'erreurs sont infimes pour les deux approches. Le choix du meilleur modèle de calcul de SLR est donc anecdotique pour ces deux scénarios.

La calibration a plus d'impact sur les résultats liés aux CCTV (Figure 36). L'équilibrage des taux d'erreurs augmente les RMEP (d'un ordre de grandeur de 10^{-3} sans calibration à 3-6% pour tous les SLR calibrés) mais divise par trois les valeurs de RMED, qui chutent à 13-15% par l'approche suspect-spécifique et environ 7% pour la trace-spécifique.

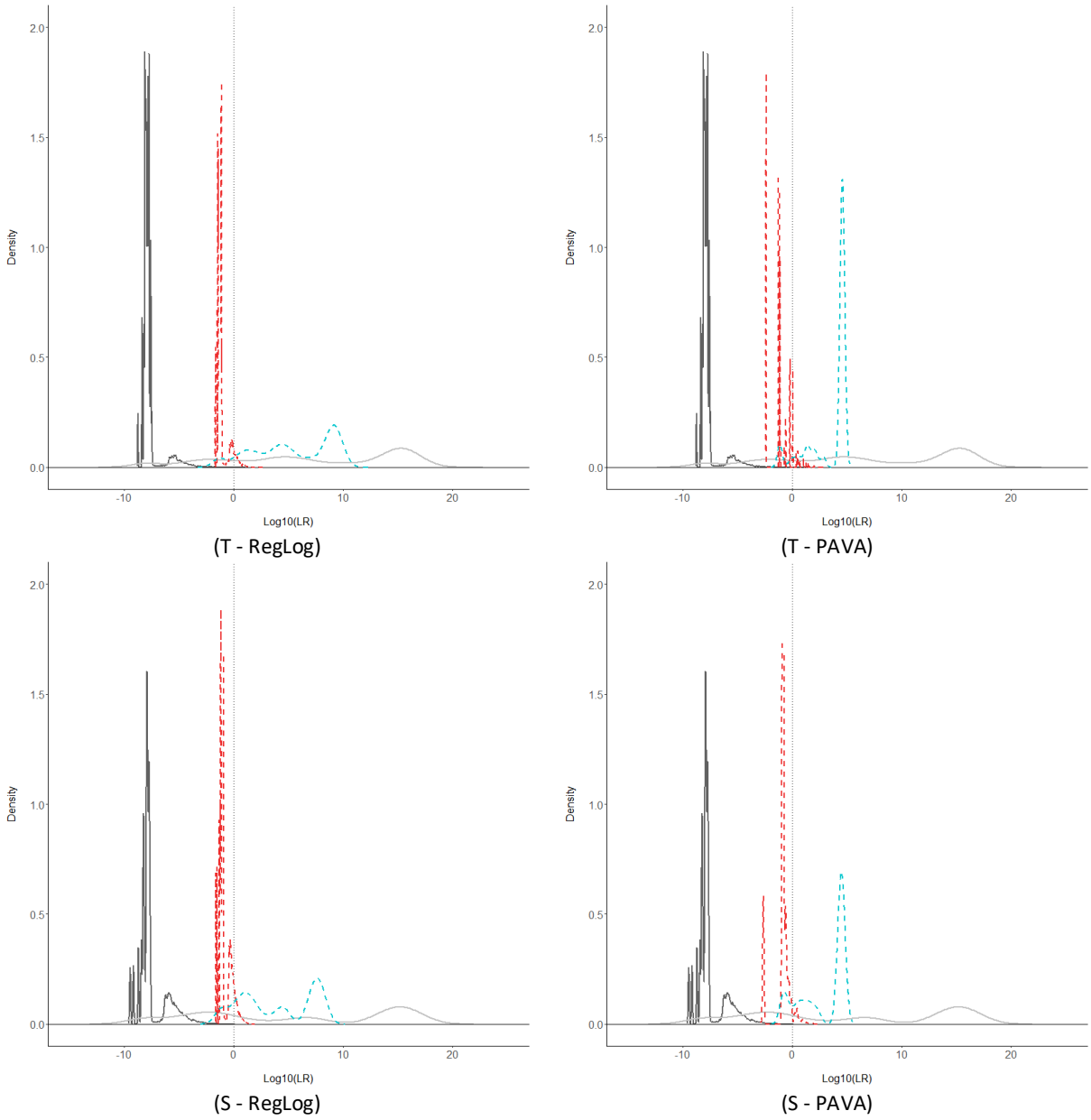


Figure 36 : Comparaison des distributions des SLR bruts (H_1 traits pleins clairs et H_2 traits pleins foncés) et des SLR calibrés (H_1 traits pointillés bleus et H_2 traits pointillés rouges) pour le scénario CCTV en fonction de la calibration (PAVA à droite et régression logistique à gauche) et du modèle d'intervariabilité (trace-spécifique en haut et suspect-spécifique en bas) avec le MFE.

VI.3. Recherche de candidats potentiels en triant par SLR

VI.3.1 FaceNet

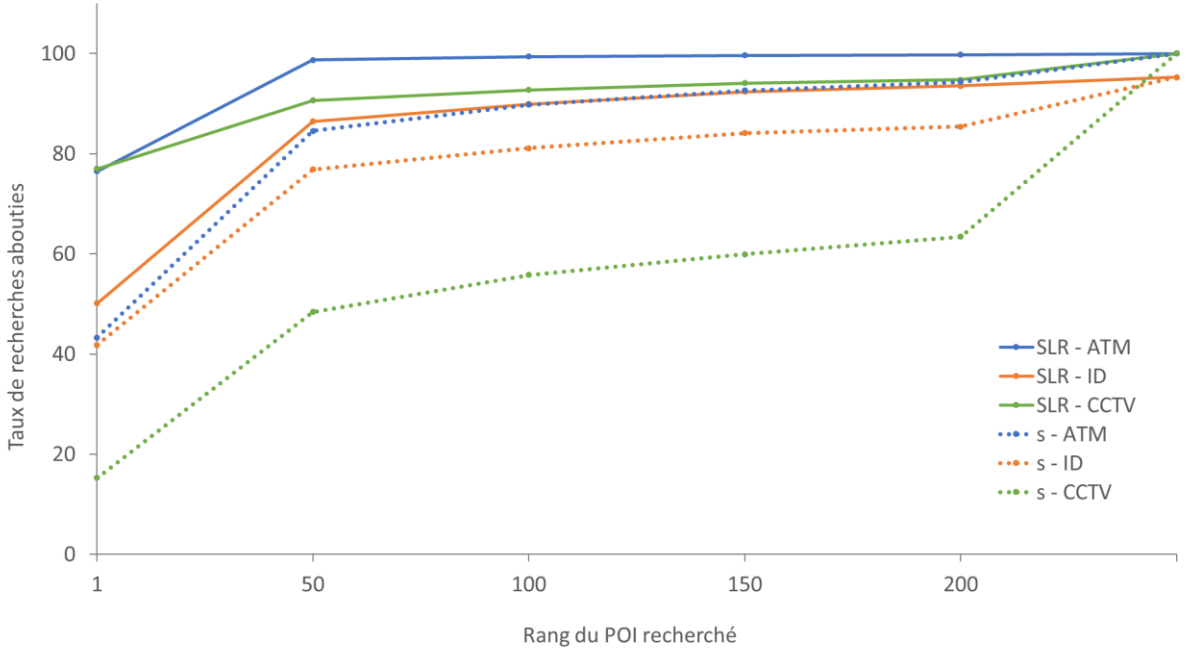


Figure 37 : Comparaison des performances de FaceNet lors de la recherche de POI à partir de scores (traits pleins) et de SLR (pointillés) en fonction du type de traces (ATM en bleu, CCTV en vert et ID en orange).

Le calcul de SLR augmente considérablement les performances de FaceNet dans la recherche de POI, comme le montre la Figure 37. L'augmentation la plus importante concerne le scénario CCTV, qui était de loin le moins performant en utilisant les listes de scores. L'utilisation du SLR pour trier les candidats potentiels permet de voir apparaitre la POI recherchée au premier rang dans 78% des cas, contre 16% avec le score, et dans les 200 premiers rangs dans 95% des cas, contre 63% avec le score.

Les recherches à partir de traces ATM bénéficient également du SLR comme métrique de tri, particulièrement pour les résultats au rang 1, passant de 45% par tri basé sur le score à 77% avec le SLR. Aucun résultat reporté n'excède le rang 100.

Enfin, l'impact du calcul de SLR est moins significatif pour le scénario ID. 50% des POI recherchées ressortent au rang 1 (contre 42% avec le score), puis pas de variation entre les rangs 2 et 200, mais le taux de résultats au-delà du rang 200 diminue environ de moitié (de 15% à 7%).

Cette comparaison met en évidence l'apport important du calcul de SLR dès l'étape de recherche de POI dans l'enquête. Pour les trois scénarios, plus de 80% des recherches classent la POI visée dans les 50 premiers rangs de la liste de candidats potentiels, et au de-là du rang 200 dans uniquement 5% des cas CCTV et 7% des cas ID.

La Figure 38 illustre les performances de FaceNet lors de la recherche de chaque POI présente à travers les trois scénarios en exploitant le SLR. L'axe horizontal pointillé fixe le rang 50 (Ln(rang) = 3.9) pour améliorer la comparaison entre scénarios.

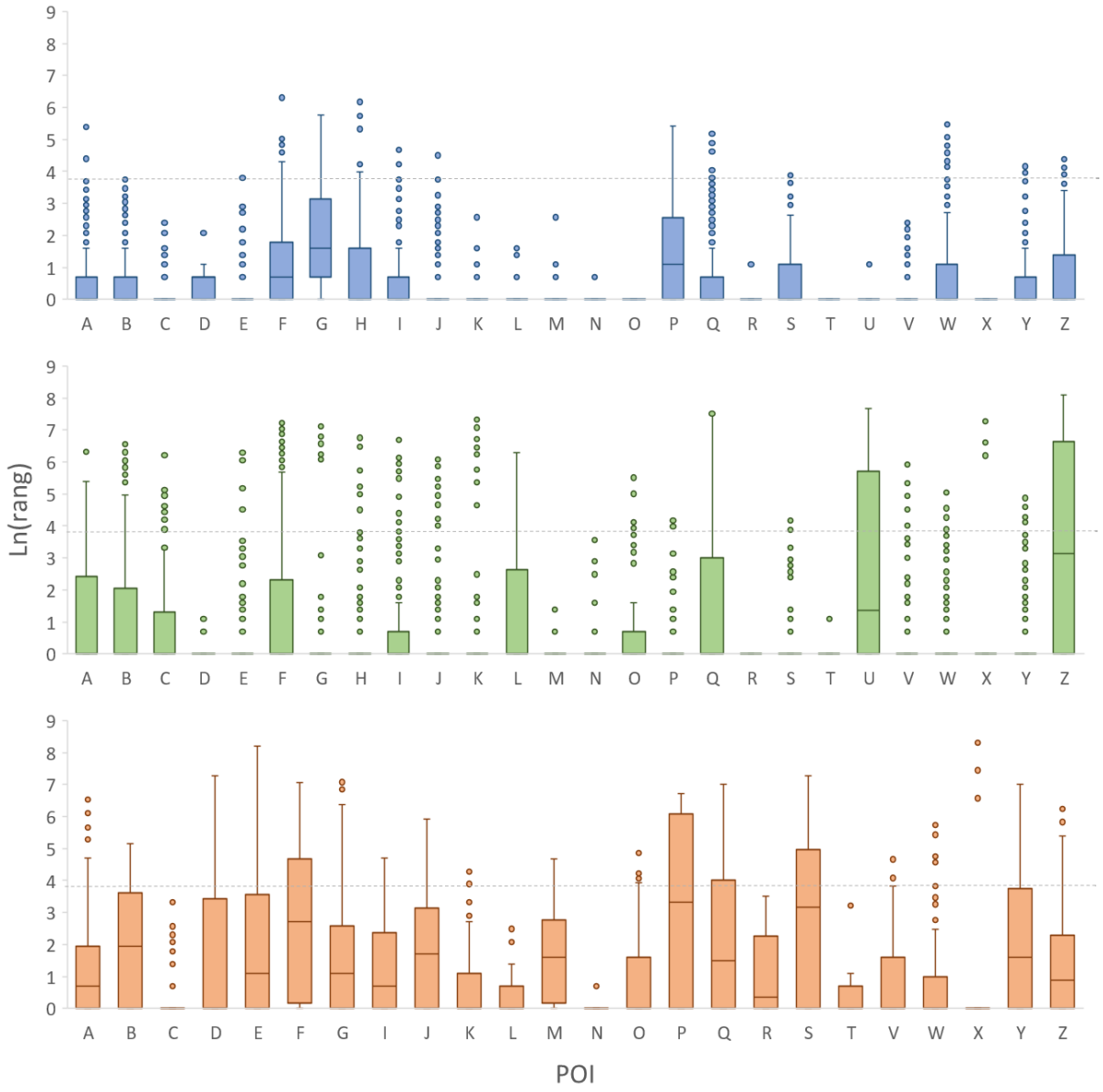


Figure 38 : Variabilités des rangs d'apparition de 26 POI triées par SLR pour les trois scénarios (ATM en bleu, CCTV en vert, ID en orange).

On observe que les variations sont plus étendues pour plusieurs POI à partir de traces ID. Cela est cohérent au vu de la grande variabilité de la qualité des images ID fournies par certains participants (résolution, éléments de sécurité, expression faciale, etc.). Cela démontre également que certaines POI sont plus facilement discriminées des individus de la BdD que d'autres. Par exemple, les POI C, N et X apparaissent majoritairement au rang 1 (Figure 39), ce qui illustre l'augmentation des performances de FaceNet lors de leur reconnaissance.

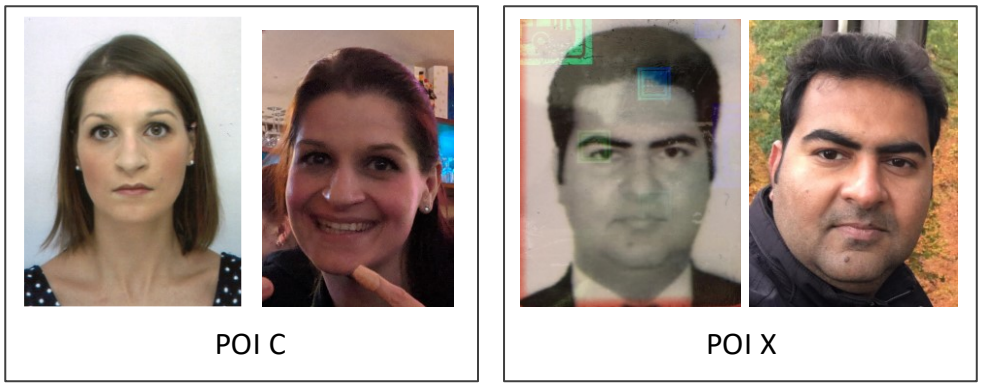


Figure 39 : Individus les plus souvent classés au rang 1 par l'algorithme FaceNet à partir d'images ID.

À l'inverse, certains individus sont retournés à des valeurs très étendues de rang, et dans la majorité des cas plus proches du rang 50, tels que les POI F, P et S (Figure 40).

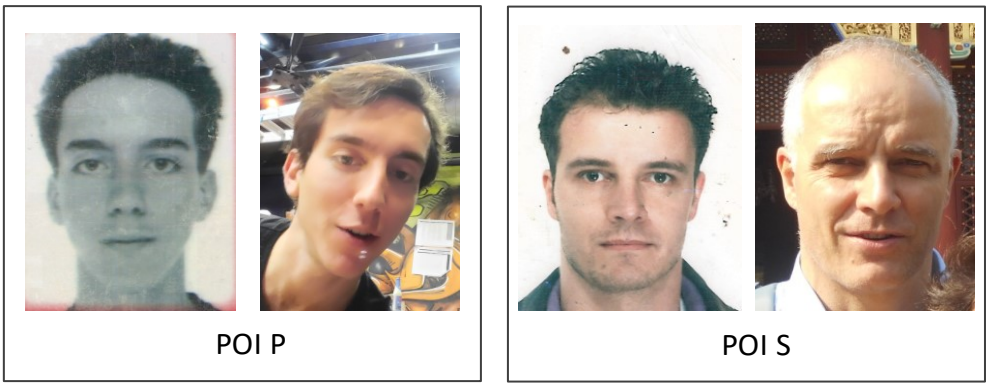


Figure 40 : Individus aux plus grandes variations dans le rang auquel les classe l'algorithme FaceNet à partir d'images ID.

Cependant, ces résultats dépendent directement du nombre d'images collectées, qui varient fortement d'une POI à l'autre. Par exemple, quatre images de la POI C et 8 images de la POI S sont utilisées, alors que les POI W et Z ont fourni chacun 18 et 15 images, respectivement. Afin de mieux étudier les variations de performances de système automatiques en fonction des individus recherchés, il serait nécessaire d'utiliser pour tous le même nombre d'images de référence prises dans des conditions contrôlées comparables.

VI.3.2 MFE

Utiliser le SLR comme métrique de tri est également positif lors de la recherche de POI par le système MFE. Néanmoins, les performances initiales de classement – à partir des scores – étant déjà très importantes, il n'a pas été jugé nécessaire de représenter graphiquement leur amélioration. Il en est de même pour la visualisation des variabilités des valeurs de rang d'apparition des POI. Seules les performances du scénario CCTV bénéficiaient d'une marge de progression, même minime. Celles-ci évoluent de 85% à 98% de classements au rang 1 – respectivement par le score et le SLR.

VI.4. Utilisation de systèmes automatiques dans le cadre civil

Dans cette section, nous nous abordons l'utilisation de système de reconnaissance faciale dans le cadre civil, dans lequel le visage est utilisé pour déverrouiller un appareil ou accorder l'accès à un bâtiment. Dans la majorité de ces applications, l'autorisation ou le rejet de la demande est basé sur la comparaison du score de comparaison à un score seuil, fixé en fonction des taux de fausses acceptations et fausses exclusions souhaités. Nous évaluons ici les performances brutes des systèmes en regard des tâches pour lesquelles ils sont initialement conçus, ainsi que l'impact du calcul de SLR, utilisé à la place du score et du seuil de détection.

VI.4.1 Performances des systèmes

Dans le cadre civil, à l'inverse du cadre investigatif, la priorité est de minimiser les taux de fausses identifications, car il est préférable de demander à l'utilisateur de soumettre une seconde demande d'accès que de risquer d'accorder l'accès à la mauvaise personne. Ces deux cadres ont des exigences opposées, nous allons donc analyser chaque approche différemment en regard de chaque application.

Avec FaceNet (Tableau 7), les meilleurs RMEP sont respectivement sans calibration pour les CCTV (4%), avec un traitement Reg Log pour les ATM (1,5%) et la PAVA pour les ID (2,5%). Ces taux sont comparables pour les deux approches d'intervariabilité, trace et suspect-spécifique.

Avec le système MFI (Tableau 8), afin de garantir le plus faible taux de RMEP, et donc de refus incorrect d'identification, il est nécessaire de garder les SLR bruts, et d'utiliser l'approche suspect-spécifique. Ce modèle mène cependant à 30.16% et 35.98% de RMED à partir de trace ATM et ID respectivement, c'est-à-dire qu'environ une demande d'accès sur trois sera incorrectement refusée.

Enfin, l'application visée lors du développement du système MFE est très claire, car pour les trois types de traces les RMEP bruts restent nulles, et seules les RMED varient en fonction du scénario, pour l'approche trace-spécifique : 0.08% pour les ATM, 23.84% pour les CCTV et 2.13% pour les ID (Tableau 9). Ces résultats sont applicables tels quels dans le cadre civil (objectif original de ce système).

VI.4.2 Apport de la métrique SLR lors de la demande d'accès

Dans la pratique, la détermination du seuil dépend de plusieurs critères tels que la qualité de la base de données, le cadre d'utilisation et des taux de fausses acceptations et fausses exclusions visés.

De manière générale, le seuil « par défaut » des systèmes Idemia est fixé autour de s = 3000, c'est pourquoi nous utilisons cet exemple ici. Il est à noter que ce seuil délimitant les « Hits » (identification du requérant comme étant l'individu autorisé) des « No Hits » (rejet de l'identification) ne sert que pour les systèmes simples. Les utilisateurs disposent plus souvent de deux seuils, ce qui permet d'ajouter une « zone grise » où les scores indiquent un résultat ambigu nécessitant une validation humaine (Comm. pers. Idemia).

L'utilisation d'un seul seuil nous semble obsolète car il augmente inexorablement le taux d'erreur. De plus, la vérification des résultats par l'humain peut prendre beaucoup plus de temps, et n'est pas sans risque d'erreur non plus. C'est pourquoi nous proposons d'utiliser le SLR, et plus particulièrement pour des scores de la « zone grise », à haut potentiel d'erreur.

La Figure 41 représente les distributions de scores de comparaisons de traces et de références d'un même individu par le système MFE pour les trois scénarios - ATM, CCTV et ID. Le trait pointillé note le score seuil utilisé par les deux algorithmes pour délimiter les zones de « Hit » et « No-Hit » (s = 3000).

On observe que, pour les traces ATM, un peu moins de la moitié des comparaisons génèrent un « No-Hit » lors de la comparaison avec le *mugshot* de référence du même individu, traduisant un taux de fausse exclusion d'environ 40%. Pour les traces CCTV, ce taux d'erreur s'élève à près de 80%. Les couleurs permettent de visualiser les conclusions apportées par le calcul de SLR sur les mêmes comparaisons. Les distributions bleues sont les scores générant les SLR soutenant la proposition H_1, alors que les scores rouges mènent à des SLR en faveur de H_2. Les scores gris mènent à des SLR neutres. En reprenant les exemples précédents, cette nouvelle lecture montre que pour les traces ATM, la quasi-totalité des SLR soutient justement la proposition H_1, alors que, pour les traces CCTV, 40% des SLR soutiennent à tort H_2.

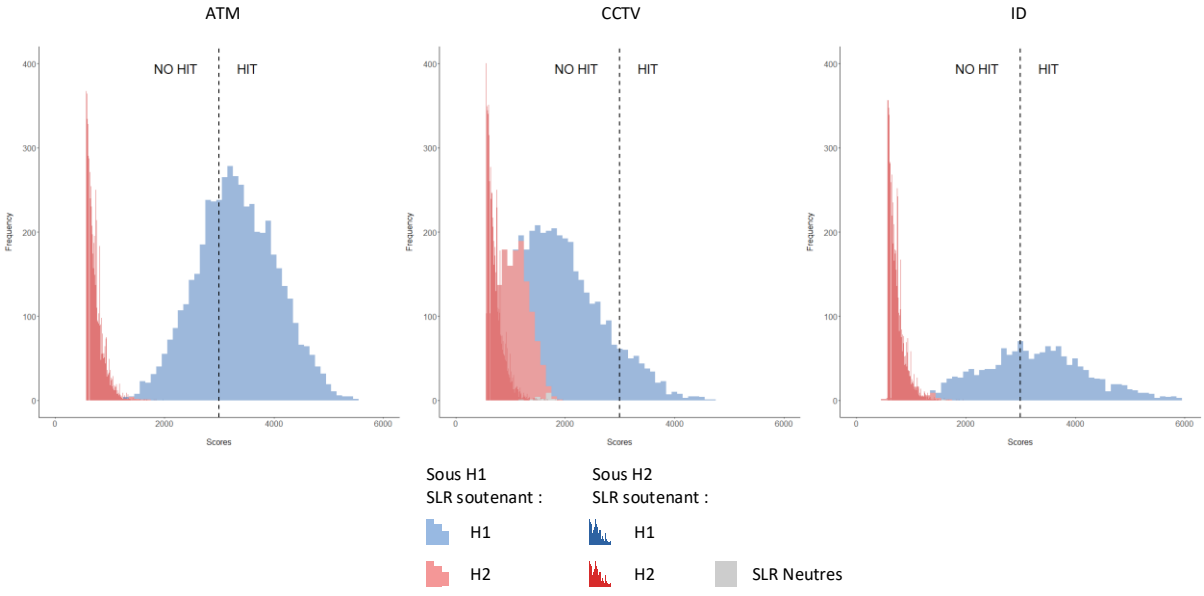


Figure 41 : Comparaison des performances du MFE par l'utilisation d'un score seuil (s = 3000, trait pointillé) et par le calcul de SLR (soutenant H_1 en bleu, soutenant justement H_2 en rouge, et neutre en gris) pour les trois scénarios (ATM à gauche, CCTV au milieu, ID à droite).

D'après la lecture de ces résultats, il apparait que le SLR augmente les performances du système MFE également dans des tâches de vérification d'identité. Il serait intéressant d'approfondir l'étude de l'utilisation du SLR dans le cadre civil pour envisager de l'inclure aux processus de vérification d'identité pour en augmenter les performances sans avoir besoin de modifier le code de l'algorithme d'analyse et comparaison d'images.

VI.5. Synthèse

VI.5.1 Choix des approches de modélisation d'intravariabilité et d'intervariabilité

Dans les cadres investigatif et civil, la contrainte principale est le manque de données de référence du POI. Nous avons donc choisi la modélisation d'intravariabilité générique, à partir de références d'autres individus. Les deux approches de modélisation d'intervariabilité, spécifique à la trace et au suspect respectivement, se révèlent équivalentes dans la majorité des scénarios avec les systèmes MFE et FaceNet. En termes de calibration, les SLR bruts se révèlent logiquement bien calibrés pour le cadre civil, et les SLR bruts et la calibration RegLog peuvent être privilégiés dans le cadre investigatif. Le tableau ci-dessous conclut les choix de modèles - modélisation de variabilité et calibration – faits des résultats exposés dans ce chapitre.

Tableau 10 : Choix de modèles de calcul de SLR pour les cadres investigatif et civil pour les trois systèmes.

<table>
<tr><th></th><th></th><th colspan="2">Investigation</th><th colspan="2">Civil</th></tr>
<tr><th></th><th></th><th>Intervariabilité</th><th>Calibration</th><th>Intervariabilité</th><th>Calibration</th></tr>
<tr><td rowspan="3">FN</td><td>ATM</td><td>Suspect ou Trace-spécifique</td><td>SLR bruts</td><td rowspan="2">Suspect ou Trace-spécifique</td><td>Reg Log</td></tr>
<tr><td>CCTV</td><td>Suspect-spécifique</td><td>Reg Log</td><td>SLR bruts</td></tr>
<tr><td>ID</td><td>Trace-spécifique</td><td>SLR bruts</td><td>Trace-spécifique</td><td>PAVA</td></tr>
<tr><td rowspan="3">MFI</td><td>ATM</td><td>Suspect-spécifique</td><td>Reg Log</td><td>Suspect-spécifique</td><td>SLR bruts</td></tr>
<tr><td>CCTV</td><td colspan="2">Non-adaptée</td><td colspan="2">Non-adaptée</td></tr>
<tr><td>ID</td><td>Suspect-spécifique</td><td>Reg Log</td><td>Suspect-spécifique</td><td>SLR bruts</td></tr>
<tr><td rowspan="3">MFE</td><td>ATM</td><td rowspan="3">Suspect ou Trace-spécifique</td><td rowspan="3">PAVA</td><td rowspan="3">Suspect ou Trace-spécifique</td><td rowspan="3">SLR bruts</td></tr>
<tr><td>CCTV</td></tr>
<tr><td>ID</td></tr>
</table>

VI.5.2 Impact du calcul de SLR dans le cadre investigatif

Lors de la phase d'enquête, les systèmes automatiques permettent de rechercher des POI dans de larges BdD en peu de temps. Nos tests ont mis en évidence des performances très bonnes pour les systèmes FaceNet et MFE sur cette tâche, et médiocres pour le MFI.

Traditionnellement, le système compare la trace à toutes les références en base, puis trie les scores de comparaison pour générer une liste de candidats potentiels, vérifiée par l'opérateur. À travers ce projet, nous avons développé un modèle probabiliste adapté aux contraintes

investigatives et qui permet de trier cette liste de candidats potentiels en se basant sur les valeurs de SLR. Cette méthode augmente considérablement les performances des systèmes sur les tâches de recherche de POI, comme illustré dans la Figure 42 pour le système FaceNet.

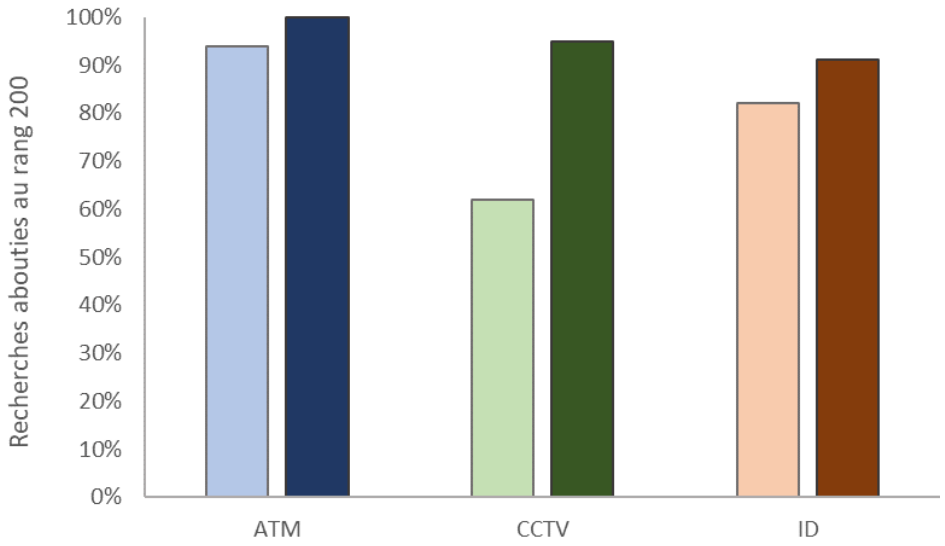


Figure 42 : Amélioration des performances de FaceNet lors de la recherche de POI dans les 200 premiers rangs de la liste de candidats par le tri basé sur les scores (barres claires) et sur les SLR (barres foncées).

Exploiter le SLR améliore les performances de FaceNet pour les trois scénarios. Lorsque le tri est basé sur les valeurs de scores de comparaison, la POI recherchée apparait dans les 200 premiers rangs de la liste dans 94%, 62% et 82% des cas ATM, CCTV et ID, respectivement. En utilisant les valeurs de SLR à la place des scores, ces performances augmentent à 100%, 95% et 91%, respectivement. Il est à noter que pour le scénario ATM, le calcul de SLR permet de réduire la liste de candidats à vérifier aux 100 premiers résultats.

La recherche de POI basée sur le score par le système MFE est particulièrement performante, et laisse peu de marge d'amélioration pour le calcul de SLR. En utilisant le score, la POI recherchée apparait au rang 1 dans 100% des cas ATM et 98% des cas ID. Dans le scénario CCTV, la POI se trouve dans les 50 premiers rangs dans 99% des cas, dont 85% au rang 1. L'impact du SLR n'est visible que pour ce scénario, car il augmente le taux d'apparition de la POI au rang 1 à 98%. Sur l'ensemble des recherches, 1% de CCTV et 1% d'ID n'ont pas permis de retrouver la POI dans la liste totale de 4000 candidats. Cela concerne les CCTV à fort angle de prise de vue ou faible résolution de la zone du visage, et les ID très anciennes ou dégradées par leur impression dans un document d'identité sécurisé.

Les performances du système MFI sont très inférieurs à celles des systèmes FaceNet et MFE. Or, le MFI est basé sur un algorithme *rule-based*, alors que les deux autres sont basés sur un processus d'apprentissage de *deep learning*. Nos résultats indiquent donc que les algorithmes basés sur l'apprentissage sont d'ores et déjà de meilleures options pour une utilisation dans le cadre investigatif, avec ou sans calcul de SLR.

VI.5.3 Conclusion

Pour conclure, les performances du MFE en font un outil de choix pour une implémentation comme aide à l'enquête, à la fois par ses performances et la rapidité de l'algorithme de comparaison. Cependant, le coût d'un tel produit peut être un frein à son acquisition, en particulier par des petits services régionaux/cantonaux de police. C'est pourquoi l'outil FaceNet, disponible en source ouverte, offre une alternative intéressante, d'autant plus en le combinant à une approche exploitant un SLR. Néanmoins, les performances et l'utilisation de FaceNet amènent des contraintes supplémentaires :

- L'opérateur doit réviser des listes d'au moins 200 candidats (au lieu de 50 avec le MFE) pour retrouver la POI recherchée avec des probabilités équivalentes.
- Certains visages ne sont pas détectés sur les images de plus faible qualité (basse résolution, petites images).
- Le processus d'analyse et comparaisons des traces et références est beaucoup plus lent car, en l'état, l'algorithme ne permet pas d'encoder les images puis de stocker uniquement les *templates*, comme c'est le cas pour le MFE. De fait, chaque comparaison nécessite de répéter la phase d'encodage qui prend d'autant plus de temps que l'image est lourde (comme c'est généralement le cas pour les *mugshots* de référence)
- FaceNet ne dispose à ce jour d'aucune interface utilisateur et doit être développé et utilisé uniquement dans un terminal de commandes

Chapitre VII. La reconnaissance faciale comme élément de preuve au tribunal

À travers ce chapitre, nous analysons les performances des systèmes FaceNet, MFI et MFE dans le développement de modèle évaluatif pour la présentation des résultats de reconnaissance faciale automatique comme élément de preuve au tribunal (Figure 43). Dans le continuum judiciaire, cette phase d'expertise suit l'enquête traitée dans le chapitre précédent (Chapitre VI p.65).

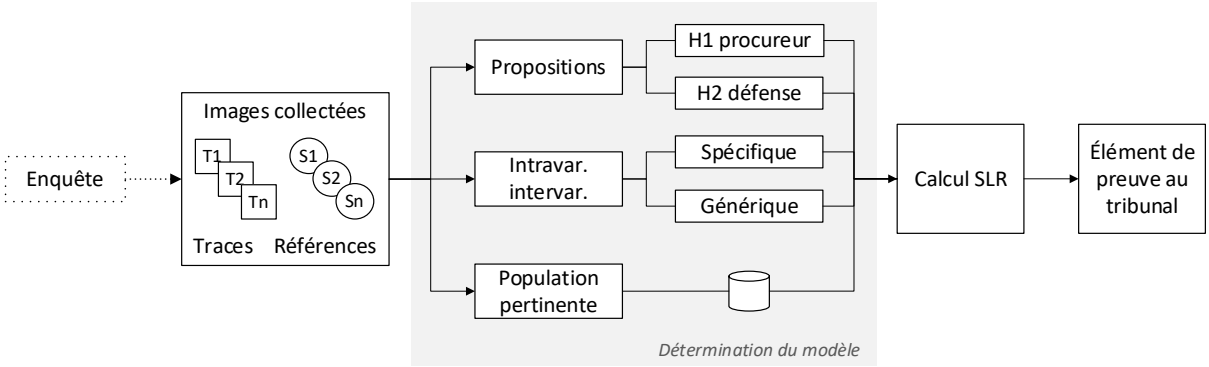


Figure 43 : Processus de développement et application de modèles de calculs de SLR dans le cadre d'expertise pour le tribunal

VII.1. Adaptation des approches de calcul SLR

À cette étape, l'enquête a permis de collecter plusieurs images de référence des POI, maintenant considérés comme suspect. Les traces et références sont transmises pour expertise par l'instruction. L'expert.e doit alors développer un modèle évaluatif adapté aux données du cas de question. Lors de cette étape, l'expert.e dispose le plus souvent de plusieurs images de référence du suspect, aussi bien d'images d'identité judiciaire (*mugshots*) et photographies d'identité. Cependant, dans certains cas, le matériel de référence peut se révéler encore insuffisant pour modéliser une intravariabilité spécifique au suspect. Il peut donc être toujours nécessaire d'exploiter les approches d'intravariabilité générique dans le cadre évaluatif.

Le tableau ci-dessous résume l'ensemble des approches testées ainsi que les propositions et formules de SLR associées.

Tableau 11 : Formule de SLR et formulation des propositions H_1 et H_2 pour chaque approche de modélisation de l'intravariabilité et de l'intervariabilité.

Approches		Formules SLR	Propositions	
Intravariabilité	Intervariabilité		H1	H2
Spécifique	Suspect-spécifique	$= \frac{f(\mathbf{s}(\mathbf{S},\mathbf{T})\|\mathbf{S},\mathbf{H_1},\mathbf{I})}{f(\mathbf{s}(\mathbf{S},\mathbf{T})\|\mathbf{S},\mathbf{H_2},\mathbf{I})}$	"Le suspect est la personne sur l'image trace"	"Le suspect n'est pas la personnes sur l'image trace"
	Trace-spécifique	$= \frac{f(\mathbf{s}(\mathbf{S},\mathbf{T})\|\mathbf{S},\mathbf{H_1},\mathbf{I})}{f(\mathbf{s}(\mathbf{S},\mathbf{T})\|\mathbf{T},\mathbf{H_2},\mathbf{I})}$		"L'individu à la source de la trace n'est pas le suspect, mais une personne de la population d'intérêt"
Générique	Suspect-spécifique	$= \frac{f(\mathbf{s}(\mathbf{S},\mathbf{T})\|\mathbf{H_1},\mathbf{I})}{f(\mathbf{s}(\mathbf{S},\mathbf{T})\|\mathbf{S},\mathbf{H_2},\mathbf{I})}$	"La même personne est visible sur les deux images comparées"	"Le suspect n'est pas la personnes sur l'image trace"
	Trace-spécifique	$= \frac{f(\mathbf{s}(\mathbf{S},\mathbf{T})\|\mathbf{H_1},\mathbf{I})}{f(\mathbf{s}(\mathbf{S},\mathbf{T})\|\mathbf{T},\mathbf{H_2},\mathbf{I})}$		"L'individu à la source de la trace n'est pas le suspect, mais une personne de la population d'intérêt"

VII.2. Performances des modèles évaluatifs

Dans le cadre du tribunal, il est primordial de minimiser les taux RMEP, car les identifications incorrectes à ce stade peuvent amener à condamner à tort un accusé. Le choix de la méthode de calibration est plus ténu que dans le cadre civil (qui minimise également les RMEP), car les valeurs des SLR et les méthodes utilisées doivent être validées et explicables par l'expert.e dans le cadre particulier du tribunal.

Les tableaux 12, 13 et 14 détaillent les pouvoirs discriminants (RMEP et RMED) et calibrations (minCllr) de tous les modèles évaluatifs considérés pour les trois scénarios forensiques, obtenus avec les systèmes FaceNet, MFI et MFE. Les résultats sont discutés selon les besoins et contraintes inhérents au cadre d'application au tribunal.

VII.2.1 FaceNet

Avec FaceNet, les résultats des intervariabilités spécifiques à la trace et au suspect sont similaires lorsque l'intravariabilité est générique, et les SLR les plus fiables sont obtenus par calibration Reg Log pour les traces ATM, par PAVA pour les ID, et sans calibration pour les CCTV (RMEP de 1.5%, 2.5% et 4% respectivement) (Tableau 12).

Tableau 12 : Performances du système FaceNet pour les modèles investigatifs de calcul de SLR bruts et calibrés.

Intravar.	Intervar.	Métriques	ATM			CCTV			ID		
			SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA
Spécifique	Suspect - spécifique	RMEP (%)	3,73 ± 0,01	1,62 ± 0,01	1,70 ± 0,01	3,72 ± 0,01	8,34 ± 0,01	6,25 ± 0,03	3,81 ± 0,01	3,08 ± 0,01	3,78 ± 0,02
		RMED (%)	0,49 ± 0,00	1,51 ± 0,00	1,37 ± 0,01	15,69 ± 0,00	10,06 ± 0,01	11,72 ± 0,03	3,71 ± 0,00	4,91 ± 0,01	3,65 ± 0,02
		(min)Cllr	0,093 ± 2,1E-04	0,067 ± 1,7E-04	0,063 ± 1,5E-04	0,444 ± 2,0E-04	0,354 ± 1,9E-04	0,336 ± 2,1E-04	0,134 ± 2,1E-04	0,133 ± 2,0E-04	0,122 ± 1,8E-04
	Trace - spécifique	RMEP (%)	3,80 ± 0,01	1,82 ± 0,01	2,06 ± 0,01	3,42 ± 0,01	7,91 ± 0,01	5,82 ± 0,02	3,60 ± 0,01	2,82 ± 0,01	3,86 ± 0,02
		RMED (%)	0,68 ± 0,00	1,72 ± 0,00	1,31 ± 0,01	15,49 ± 0,00	10,30 ± 0,01	11,72 ± 0,02	3,37 ± 0,00	4,64 ± 0,01	3,01 ± 0,02
		(min)Cllr	0,102 ± 1,7E-04	0,074 ± 1,5E-04	0,071 ± 1,4E-04	0,443 ± 1,9E-04	0,360 ± 1,8E-04	0,337 ± 2,0E-04	0,126 ± 1,9E-04	0,124 ± 1,9E-04	0,113 ± 1,6E-04
Générique	Suspect - spécifique	RMEP (%)	4,85 ± 0,01	1,82 ± 0,01	1,93 ± 0,01	4,85 ± 0,01	7,00 ± 0,01	6,59 ± 0,02	4,85 ± 0,01	3,51 ± 0,01	3,02 ± 0,02
		RMED (%)	0,51 ± 0,00	1,85 ± 0,00	1,61 ± 0,01	12,88 ± 0,00	9,56 ± 4,4E-03	9,80 ± 0,01	4,92 ± 0,00	5,74 ± 1,3E-03	6,04 ± 0,02
		(min)Cllr	0,116 ± 2,4E-04	0,074 ± 1,9E-04	0,069 ± 1,7E-04	0,322 ± 2,4E-04	0,305 ± 2,6E-04	0,296 ± 2,5E-04	0,176 ± 2,3E-04	0,171 ± 2,0E-04	0,162 ± 1,8E-04
	Trace - spécifique	RMEP (%)	4,79 ± 0,01	1,62 ± 0,01	1,59 ± 0,01	4,18 ± 0,01	7,86 ± 0,01	7,00 ± 0,03	4,29 ± 0,01	3,09 ± 0,01	2,57 ± 0,01
		RMED (%)	0,39 ± 0,00	1,43 ± 2,6E-03	1,42 ± 0,01	13,35 ± 0,00	8,37 ± 4,8E-03	9,01 ± 0,03	4,65 ± 0,00	5,41 ± 2,5E-03	5,64 ± 0,01
		(min)Cllr	0,112 ± 2,3E-04	0,065 ± 1,8E-04	0,061 ± 1,6E-04	0,326 ± 1,5E-04	0,299 ± 2,2E-04	0,290 ± 2,3E-04	0,160 ± 2,3E-04	0,156 ± 1,9E-04	0,147 ± 1,8E-04

Les résultats sont différents pour les approches où l'intravariabilité est spécifique au suspect. Les plus bas taux RMEP sont obtenus par l'approche suspect-spécifique pour le scénario ATM, et trace-spécifique pour les CCTV et ID. Néanmoins, il est nécessaire de prendre en compte également les valeurs de coûts associés aux modèles (minCllr). Pour les traces ATM, les quatre modèles génèrent entre 1.35 et 1.60% de RMEP, ce qui n'apporte pas de variation significative, la valeur minCllr est identique pour tous, et les courbes ECE (Annexe C) n'apportent pas d'information supplémentaire. De fait, les quatre modèles sont utilisables dans la phase d'expertise, pour les traces ATM.

Les résultats du scénario ID présentent également peu de variations entre les différents modèles, calibrés et bruts. Cependant, l'approche trace-spécifique génère les plus bas coûts, particulièrement par calibration PAVA (minCllr = 0.107).

Les valeurs minCllr associées aux calibrations sur le scénario ID montrent un faible impact de la RegLog et de la PAVA car les SLR bruts sont déjà bien calibrés. En outre, la Figure 44 illustre une différence notable dans l'effet de la calibration PAVA. Celle-ci réduit très fortement les valeurs

extrêmes de SLR sous H_1, de $Log_{10}(SLR) = 10$ (courbe pleine grise) à $Log_{10}(SLR) = 5$ (courbe bleue pointillée). Cela n'apporte pas d'information quant aux performances du modèle, et la présentation de telles valeurs extrêmes de SLR au tribunal reste à l'appréciation de l'expert.e. Ce point est abordé plus en détail dans la section VII.3.

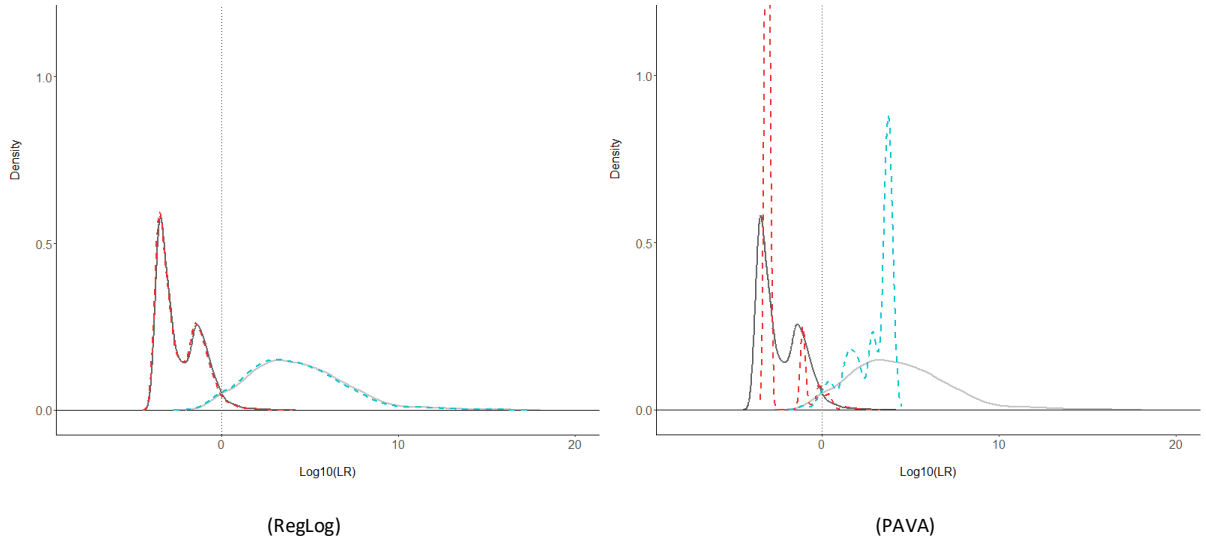


Figure 44 : Comparaison des distributions des SLR bruts (H_1 traits pleins clairs et H_2 traits pleins foncés) et des SLR calibrés (H_1 traits pointillés bleus et H_2 traits pointillés rouges) en fonction de la calibration (PAVA à droite et Régression logistique à gauche) pour le modèle ID intravariabilité spécifique/intervariabilité trace-spécifique.

Pour le scénario CCTV, une courbe ECE donne des informations supplémentaires indispensables. Sur la Figure 45, on observe que la courbe du modèle brut à l'intervariabilité suspect-spécifique est supérieure à celle du modèle neutre pour les valeurs de $\log_{10}(priors) > 1{,}5$. Concrètement, cela veut dire que ce modèle ne doit pas être utilisé lorsque les *priors* en faveur de H_1 sont supérieures à 1.5 ($\log_{10}(Pr(H_1))$). En revanche, la courbe du modèle calibré par PAVA (courbe pointillée) est non seulement plus aplatie sur l'intervalle [-1.5 ; 4], et donc les coûts associés sont moindres, mais elle ne recoupe le modèle neutre qu'au-delà de $Log_{10}(priors) = 2.5$.

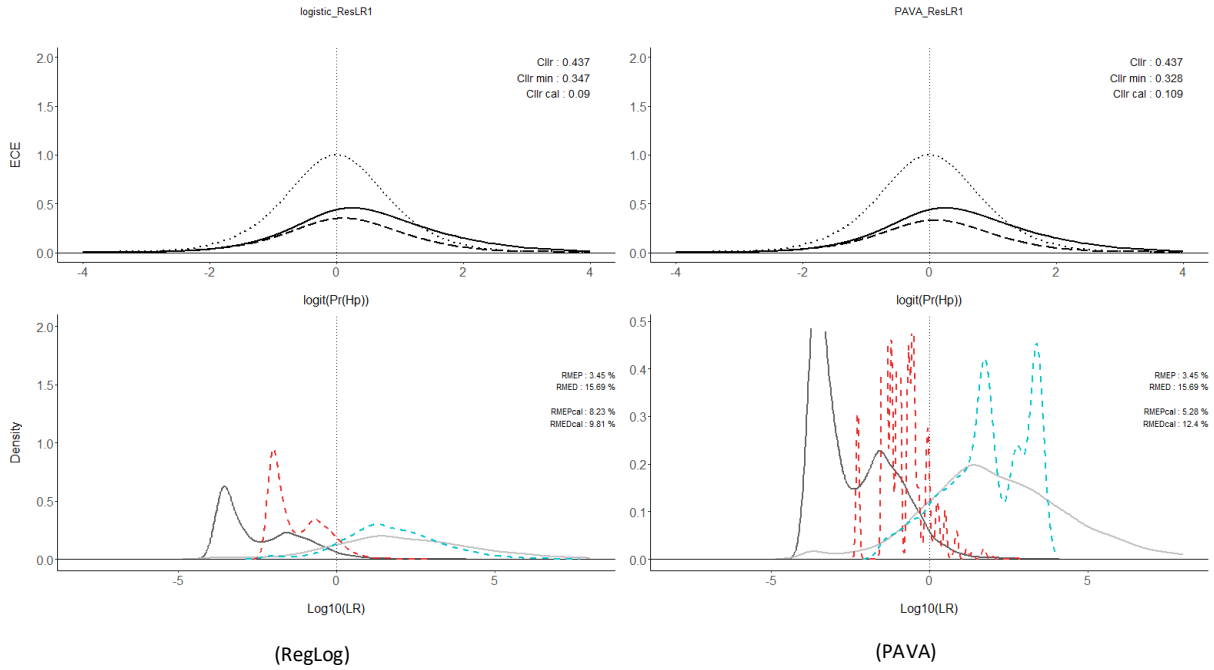


Figure 45 : Impact des calibrations (régression logistique à gauche et PAVA à droite) sur les valeurs de SLR bruts pour le modèle CCTV intravariabilité spécifique/intervariabilité suspect-spécifique (courbe ECE en haut, distributions en bas) avec le système FaceNet.

La calibration PAVA rend le modèle fiable pour des valeurs plus étendues de *priors*, diminue les coûts associés au modèle, et améliore la discrimination des propositions (probabilités à minCllr(PAVA) = 0.328, minCllr(RegLog) = 0.350, Cllr = 0.437). Le modèle intravariabilité spécifique/intervariabilité suspect-spécifique/calibration PAVA est donc finalement préconisé pour le scénario CCTV, au bénéfice de coûts plus faibles.

VII.2.2 MFI

Pour les modèles basés sur une intravariabilité générique, les RMEP des SLR bruts sont très bas, entre 2.46 et 3.29%, mais les hauts taux RMED génèrent des coûts supérieurs à 1 (supérieur à 5 pour les CCTV), les modèles bruts sont donc non pertinents (Tableau 13). De fait, il convient de privilégier le modèle suspect-spécifique/calibration PAVA, associé aux minCllr les plus bas. Les taux d'erreurs sont donc mieux équilibrés que par calibration RegLog. En revanche, tous les modèles calibrés liés aux traces CCTV génèrent des coûts tendant vers 1 (en bleu dans le Tableau 13), ce qui veut dire que ces modèles n'apportent aucune information pertinente. C'est le cas également pour les modèles CCTV basés sur l'intravariabilité spécifique.

Tableau 13 : Performances du système MFI pour les modèles investigatifs de calcul de SLR bruts et calibrés.

Intravar.	Intervar.	Métriques	ATM			CCTV			ID		
			SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA
Spécifique	Suspect-spécifique	RMEP (%)	2,23 ± 8,7E-03	15,25 ± 0,02	12,83 ± 0,02	2,23 ± 9,3E-03	73,21 ± 0,03	19,02 ± 0,02	2,22 ± 9,4E-03	25,88 ± 0,04	8,82 ± 0,08
		RMED (%)	34,03 ± 0,00	13,15 ± 1,8E-03	14,64 ± 3,0E-03	89,90 ± 0,00	12,42 ± 4,1E-03	53,44 ± 0,00	39,49 ± 0,00	18,40 ± 4,2E-03	26,91 ± 0,08
		(min)Cllr	1,67 ± 1,5E-04	0,68 ± 1,2E-04	0,48 ± 2,0E-04	5,70 ± 1,5E-04	0,96 ± 1,0E-04	0,90 ± 1,5E-04	3,32 ± 1,6E-04	0,99 ± 1,7E-04	0,57 ± 2,0E-04
	Trace-spécifique	RMEP (%)	2,19 ± 9,0E-03	17,92 ± 1,7E-02	9,76 ± 3,6E-02	2,06 ± 8,6E-03	70,55 ± 2,8E-02	36,54 ± 0,03	2,40 ± 9,0E-03	23,54 ± 0,05	10,61 ± 0,02
		RMED (%)	34,94 ± 0,00	14,43 ± 9,0E-03	21,00 ± 3,5E-02	93,25 ± 0,00	20,72 ± 8,5E-03	37,11 ± 0,00	39,96 ± 0,00	18,78 ± 5,2E-03	25,78 ± 0,01
		(min)Cllr	1,78 ± 1,3E-04	0,74 ± 1,1E-04	0,50 ± 1,9E-04	5,25 ± 1,2E-04	0,98 ± 6,1E-05	0,93 ± 1,2E-04	3,31 ± 1,4E-04	0,94 ± 1,5E-04	0,57 ± 1,6E-04
Générique	Suspect-spécifique	RMEP (%)	2,64 ± 0,01	13,78 ± 1,3E-02	10,66 ± 3,4E-02	2,68 ± 9,6E-03	24,74 ± 0,03	16,72 ± 9,0E-02	2,66 ± 9,5E-03	16,59 ± 0,01	8,57 ± 0,11
		RMED (%)	30,16 ± 0,00	14,92 ± 6,4E-03	17,22 ± 2,9E-02	91,81 ± 0,00	68,43 ± 0,00	72,74 ± 9,1E-02	35,98 ± 0,00	18,05 ± 0,01	25,03 ± 0,11
		(min)Cllr	1,23 ± 1,6E-04	0,49 ± 1,9E-04	0,46 ± 1,9E-04	5,99 ± 1,6E-04	0,99 ± 4,9E-05	0,94 ± 1,1E-04	1,59 ± 1,5E-04	0,56 ± 1,9E-04	0,52 ± 1,8E-04
	Trace-spécifique	RMEP (%)	3,09 ± 0,01	14,02 ± 1,3E-02	10,78 ± 6,3E-02	2,79 ± 0,01	16,39 ± 0,03	22,75 ± 0,05	3,03 ± 9,9E-03	16,53 ± 0,01	8,78 ± 0,04
		RMED (%)	29,90 ± 0,00	15,53 ± 6,9E-03	18,12 ± 6,4E-02	87,55 ± 0,00	61,20 ± 0,00	55,78 ± 0,05	36,27 ± 0,00	19,84 ± 5,6E-03	25,72 ± 0,03
		(min)Cllr	1,26 ± 1,6E-04	0,50 ± 1,8E-04	0,47 ± 1,7E-04	5,17 ± 1,6E-04	0,95 ± 1,1E-04	0,93 ± 1,2E-04	1,62 ± 1,5E-04	0,58 ± 1,9E-04	0,54 ± 1,7E-04

Le modèle ID à l'intravariabilité spécifique et calibré par RegLog est également neutre. Pour les scénarios ATM et ID, les meilleures performances sont atteintes par le biais de l'approche suspect-spécifique et de la calibration PAVA (minCllr de 0.479 et 0.571, respectivement). Le scénario ATM génère des valeurs RMEP de 13.41% et RMED de 14.61%, et le scénario ID, des RMEP de 9.55% et RMED de 26.14%.

De plus, les courbes ECE et distributions montrent que modèle ATM à l'intervariabilité suspect-spécifique, calibré par Reg Log (Figure 46 gauche), génère des valeurs ECE supérieures à celles du modèle neutre au-dessus de $\log_{10}(H_1) = 1$. Ce modèle est donc inutilisable quand les probabilités *a priori* $\text{Log}_{10}(\text{Pr}(H_1))$ sont supérieures à 1. En opposition, la calibration PAVA (à droite) garantie des valeurs ECE toujours inférieures à ceux du modèle neutre et, de fait, le modèle est utilisable quelques soient les valeurs de *celles-ci*.

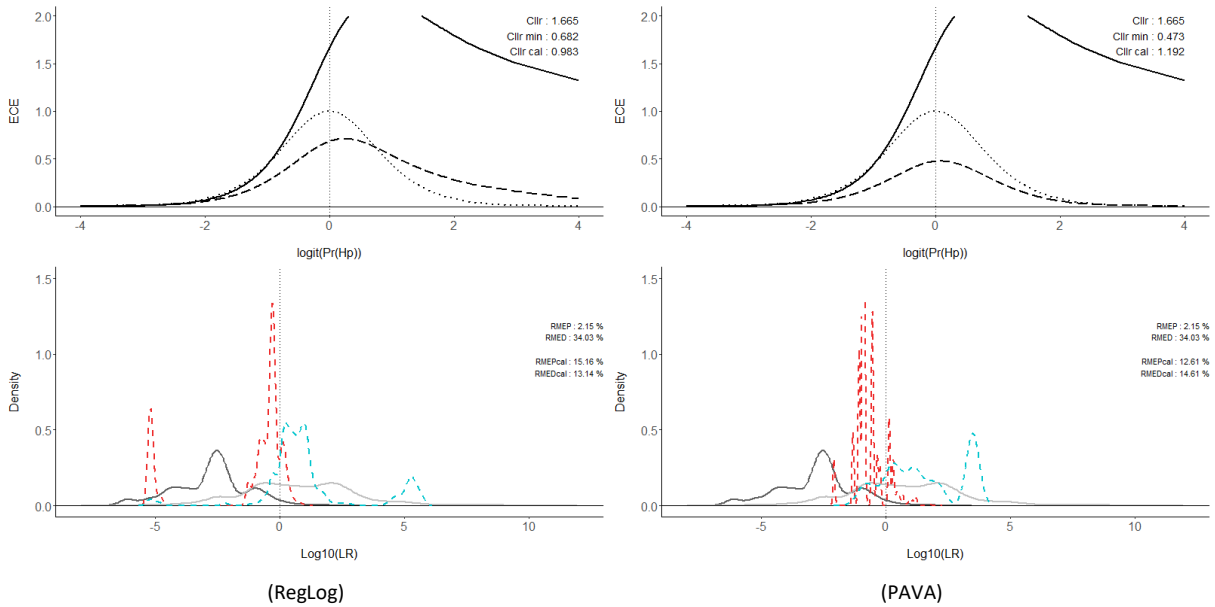


Figure 46 : Impact de la calibration (régression logistique à gauche et PAVA à droite) sur les valeurs de SLR bruts pour le modèle CCTV intravariabilité spécifique/intervariabilité suspect-spécifique (courbe ECE en haut, distributions en bas) avec le système MFI.

Tout comme pour une utilisation dans le cadre investigatif, les performances du système MFI sont trop faibles pour envisager son utilisation dans le système pénal.

VII.2.3 MFE

Les performances du système MFE sont très élevées et comparables pour l'ensemble des modèles testés pour chaque scénario (Tableau 14).

Tableau 14 : Performances du système MFE pour les modèles investigatifs de calcul de SLR bruts et calibrés.

Intravar.	Intervar.	Métriques	ATM			CCTV			ID		
			SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA	SLR bruts	Log Reg	PAVA
Spécifique	Suspect-spécifique	RMEP (%)	0,03 ± 9,9E-04	0,08 ± 1,0E-03	0,11 ± 2,7E-03	0,03 ± 9,9E-04	6,91 ± 0,06	4,72 ± 0,03	0,03 ± 1,0E-03	0,56 ± 2,8E-03	0,85 ± 6,2E-03
		RMED (%)	0,25 ± 0,00	0,19 ± 2,1E-03	0,10 ± 3,0E-03	41,32 ± 0,00	14,34 ± 5,8E-03	15,47 ± 0,03	3,47 ± 0,00	1,10 ± 2,9E-03	0,59 ± 3,4E-03
		(min)Cllr	5,1E-03 ± 1,4E-05	4,0E-03 ± 3,4E-05	3,0E-03 ± 2,3E-05	2,16 ± 2,1E-05	0,44 ± 1,2E-04	0,38 ± 1,6E-04	0,09 ± 2,7E-05	0,04 ± 6,3E-05	0,03 ± 6,5E-05
	Trace-spécifique	RMEP (%)	6,6E-03 ± 5,0E-04	0,07 ± 1,1E-03	0,04 ± 1,8E-03	0,02 ± 9,5E-04	4,40 ± 8,6E-03	2,31 ± 0,01	0,01 ± 6,2E-04	0,38 ± 2,9E-03	0,26 ± 3,1E-03
		RMED (%)	0,16 ± 0,00	0,11 ± 6,6E-04	0,10 ± 1,9E-03	33,46 ± 0,00	10,83 ± 2,2E-04	11,43 ± 9,1E-03	2,00 ± 0,00	0,47 ± 3,6E-04	0,48 ± 1,7E-03
		(min)Cllr	8,0E-03 ± 5,2E-06	3,0E-03 ± 2,5E-05	2,2E-03 ± 2,5E-05	1,67 ± 3,1E-05	0,35 ± 1,1E-04	0,30 ± 1,4E-04	0,06 ± 3,0E-05	0,02 ± 4,7E-05	0,02 ± 4,5E-05
Générique	Suspect-spécifique	RMEP (%)	5,5E-03 ± 4,6E-04	0,04 ± 9,7E-04	0,04 ± 1,2E-03	5,8E-03 ± 4,7E-04	5,81 ± 8,9E-03	3,40 ± 0,01	5,7E-03 ± 4,7E-04	0,49 ± 2,7E-03	0,42 ± 4,7E-03
		RMED (%)	0,23 ± 0,00	0,03 ± 7,2E-04	9,2E-03 ± 6,6E-04	40,69 ± 0,00	13,80 ± 5,3E-03	15,52 ± 5,9E-03	3,33 ± 0,00	0,67 ± 9,8E-04	0,65 ± 3,4E-03
		(min)Cllr	3,3E-03 ± 2,8E-05	1,3E-03 ± 3,2E-05	1,1E-03 ± 2,1E-05	2,82 ± 9,7E-06	0,40 ± 1,2E-04	0,36 ± 1,4E-04	0,13 ± 1,8E-05	0,03 ± 5,3E-05	0,03 ± 5,9E-05
	Trace-spécifique	RMEP (%)	0,00 ± 4,6E-04	0,02 ± 9,1E-04	0,04 ± 1,2E-03	8,5E-03 ± 5,5E-04	2,91 ± 6,1E-03	2,65 ± 0,02	0,00 ± 0,00	0,27 ± 2,0E-03	0,24 ± 2,9E-03
		RMED (%)	0,08 ± 0,00	0,04 ± 7,2E-04	0,00 ± 6,6E-04	23,84 ± 0,00	7,20 ± 2,1E-03	7,34 ± 0,02	2,13 ± 0,00	0,48 ± 1,3E-03	0,47 ± 1,2E-03
		(min)Cllr	0,001 ± 2,2E-05	0,001 ± 3,7E-05	0,001 ± 2,1E-05	1,59 ± 1,9E-05	0,23 ± 1,0E-04	0,22 ± 1,1E-04	0,09 ± 0,00	0,02 ± 3,5E-05	0,02 ± 4,1E-05

Les modèles ATM et ID ont des performances très hautes et semblables quel que soit le modèle, mais l'approche trace-spécifique est sensiblement meilleure pour tous les modèles, excepté pour la combinaison ATM et intravariabilité générique.

En privilégiant les plus bas taux de RMEP, les SLR bruts paraissent plus fiables que les SLR calibrés. Cependant les calibrations diminuent dans tous les cas les coûts associés, sans différences significatives dans valeurs de minCllr entre PAVA et RegLog. Les modèles les plus adaptés à une utilisation destinée au tribunal combinent une intervariabilité trace-spécifique et calibration pour toutes les traces.

Concernant le scénario CCTV, la calibration a un impact plus significatif car les performances brutes sont moins élevées que pour les deux autres scénarios, et les coûts associés sont supérieurs à 1.5. La calibration PAVA augment les RMEP - de 0.04% à 2.18% pour l'approche intra-spécifique et de 0.01% à 2.8% pour l'intra-générique - mais divise par trois les RMED - passant respectivement de 33.46 à 11.50% et de 23.84 à 7.23%. Cet équilibrage des taux d'erreurs accompagne la forte diminution des coûts pour ces deux modèles, à 0.30 et 0.22 respectivement.

Pour conclure, ces résultats démontrent que les taux d'erreurs du MFI sont incompatibles avec son intégration dans le processus judiciaire. En revanche, les modèles évaluatifs basés sur les scores de comparaison générés par les systèmes FaceNet et MFE sont assez robustes pour être utilisés comme élément de preuves au tribunal.

Le nombre suffisant de données de référence de la POI en phase d'expertise permet de baser nos modèles sur une intravariabilité spécifique. Au tribunal, cette approche à l'avantage de permettre à l'expert.e de se prononcer sur des propositions incluant directement le suspect (H_1 : « Le suspect est la personne sur l'image trace »). En termes de performance, le tableau ci-dessous résume les choix de modèles - modélisation de variabilité et calibration – à la lumière des résultats exposés dans ce chapitre.

Tableau 15 : Choix de modèles de calcul de SLR pour la phase d'expertise pour nos trois systèmes.

		Tribunal	
		Intervariabilité	*Calibration*
FN	*ATM*	Trace-spécifique	PAVA ou RegLog
	CCTV	Suspect-spécifique	PAVA
	ID	Trace-spécifique	PAVA ou RegLog
MFI	*ATM*	Suspect-spécifique	PAVA
	CCTV	Non-adaptée	
	ID	Suspect-spécifique	PAVA
MFE	*ATM*	Trace-spécifique	PAVA ou RegLog
	CCTV		
	ID		

VII.3. Discussion

Cette section s'articule autour de trois points de discussion que soulèvent les valeurs de SLR destinées à être présentées au tribunal. Nous discutons successivement les points suivants :

a. La gestion des valeurs extrêmes de SLR et la communication des SLR soutenant très faiblement l'une ou l'autre des propositions
b. L'évolution dans un cas forensique des valeurs SLR pour entre la phase investigation et dans le cadre de l'expertise

VII.3.1 Gestion des valeurs extrêmes de SLR

Dans cette section, nous souhaitons développer les problématiques inhérentes à la communication de SLR au tribunal. Notre discussion s'articule autour de deux sujets : la

justification des valeurs extrêmes de SLR (élevés en soutien à H_1 et faibles en soutien à H_2, respectivement) et la communication des valeurs très faibles de SLR.

VII.3.1.1 Justification des hauts SLR

Parmi nos résultats, certains modèles produisent des SLR bruts ou calibrés par RegLog dont les valeurs atteignent par exemple $Log_{10}(SLR)=10$. Cela implique de devoir présenter au tribunal un SLR d'une valeur de 10^{10} en faveur de H_1, ce qui se traduit au tribunal de la manière suivante :

« La probabilité d'obtenir ce résultat si le suspect n'est pas la personne sur l'image trace est de 1 sur 10 milliards »[21].

De telles valeurs soulèvent deux problématiques, l'une générale et l'autre spécifique à l'exploitation d'images faciale comme trace. D'une part, quel que soit le domaine d'application, les valeurs extrêmes de LR sont difficilement justifiables au tribunal puisqu'elles ne peuvent être démontrables sur des données empiriques de taille équivalente.

En outre, pour aborder plus explicitement cette problématique dans le cas spécifique du visage, nous proposons un parallèle avec l'ADN.

L'ADN fait partie des caractéristiques individuelles. C'est la combinaison de plusieurs éléments (les allèles) dont les fréquences d'apparition – et donc la rareté - varient en sein de la population d'intérêt. La combinaison des valeurs de rareté de chaque élément produit alors une probabilité de coïncidence fortuite souvent infinitésimale, générant souvent des LR de plusieurs milliards.

Une multitude de facteurs affecte l'apparence physique et notamment faciale d'un individu tout au long de sa vie, tels que son environnement (pollution, ensoleillement, humidité, etc.) et son mode de vie (activité physique, tabagisme, alimentation, etc.). Ainsi, même deux jumeaux/jumelles monozygotes peuvent se retrouver à l'âge adulte avec des caractéristiques physiques, et plus particulièrement faciales, très différentes (à l'inverse de leur ADN qui demeure identique). De plus, le visage seul peut être à la source de ressemblances fortuites entre deux inconnu.e.s[22].

Le visage seul devrait donc logiquement avoir un pouvoir discriminant bien plus faible que celui de l'ADN.

Il n'existe pas d'analyse de caractéristiques faciales aussi spécifique que celle des combinaisons d'allèles de régions précises de l'ADN, par exemple. Pour approcher un tel processus, il serait nécessaire de sélectionner l'ensemble des éléments communs à tout visage et d'en étudier les

[21] Cette formulation, choisie par l'auteure, illustre une manière très simplifiée d'expliquer verbalement le SLR au tribunal. Elle n'est pas exhaustive car n'explicite pas la proposition alternative, ce qui risque d'amener plus facilement à transposer le conditionnel. Une formulation exhaustive serait : « Il est 10 milliard de fois plus probable d'observer ce résultat si le suspect est la personne sur l'image plutôt qu'un autre individu inconnu ». Bien que plus précise du point de vue scientifique, cette formulation est plus complexe et risque d'amener les magistrats à demander ou produire eux-mêmes une reformulation simplifiée.

[22] Voir par exemple le projet du photographe François Brunnelle qui immortalise les ressemblances de « sosies », inconnu.e.s non apparenté.e.s (Source : http://www.francoisbrunelle.com/webn/f-projet.html)

fréquences d'apparitions au sein de différentes populations. Le fonctionnement du *deep learning* permet de se rapprocher de ce genre de processus grâce à la phase d'apprentissage. Par exemple, en utilisant une certaine sous population dans le jeu de données d'entrainement, le système apprend à détecter et trier les caractéristiques les plus fréquentes des plus rares.

Dans le cas des résultats présentés dans cette recherche, nous avons remarqué qu'à performances équivalentes avec la calibration RegLog, la PAVA produit des valeurs de SLR moins étendues. Le soutien pour l'une ou l'autre des propositions est amoindri mais peut sembler plus « compatible » avec le pouvoir discriminant supposé du visage humain. Nous concluons à l'heure actuelle que, faute de données empiriques suffisantes dans de nombreux domaines, le choix de la méthode de calibration appartient à l'expert.e qui doit en présenter les résultats.

VII.3.1.2 Communication des SLR amenant un soutien faible

La problématique liée à la communication des SLR faibles est abordée notamment dans le domaine de l'évaluation de traces génétiques dans Samie Foucart (2019).

La Figure 47 met en évidence l'étendue des valeurs de SLR soutenant à tort H^2 (gris clair) et soutenant à tort H_1(gris foncé) s'étendant sur l'intervalle $Log10(SLR) \in [-2\,,2.5]$, c.-à-d. SLR entre 0.01 et 300. Cela correspond aux SLR pour lesquels des erreurs sont observées empiriquement.

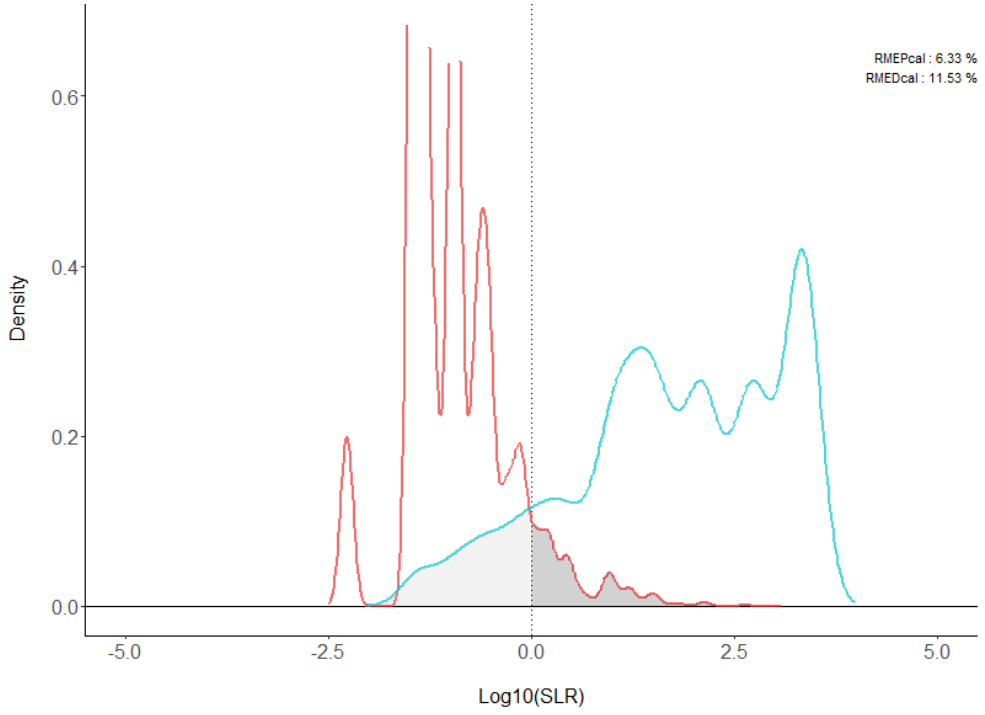


Figure 47 : Intervalle de SLR soutenant à tort H_1 (gris foncé) et soutenant à tort H^2 (gris clair) pour le modèle FaceNet CCTV/intra spécifique/inter suspect-spécifique/PAVA.

Présenter une méthodologie scientifique au tribunal nécessite de pouvoir en présenter les taux d'erreurs. Mais la manière de communiquer – ou non – les résultats d'une expertise et les risques d'erreurs associés est un sujet débattu au sein même de la communauté scientifique.

La problématique que nous souhaitons discuter ici est la suivante : Toutes les valeurs de SLR peuvent-elles (doivent-elles) être présentées au tribunal ?

Dans Samie Foucart (2019), l'auteure propose trois gestions des faibles SLR :

a. Communiquer les valeurs seules des SLR et laisser leur interprétation aux magistrats chargés de la prise de décision
b. Transmettre les valeurs de SLR en précisant les taux d'erreurs
c. Ne pas communiquer de SLR non discriminant, en justifiant qu'ils n'apportent pas d'informations

Prenons pour exemple la comparaison d'une trace, extraite d'une vidéo témoin retrouvée sur un téléphone portable volé, et d'un *mugshot* de référence du suspect ciblé par l'enquête (Figure 6).

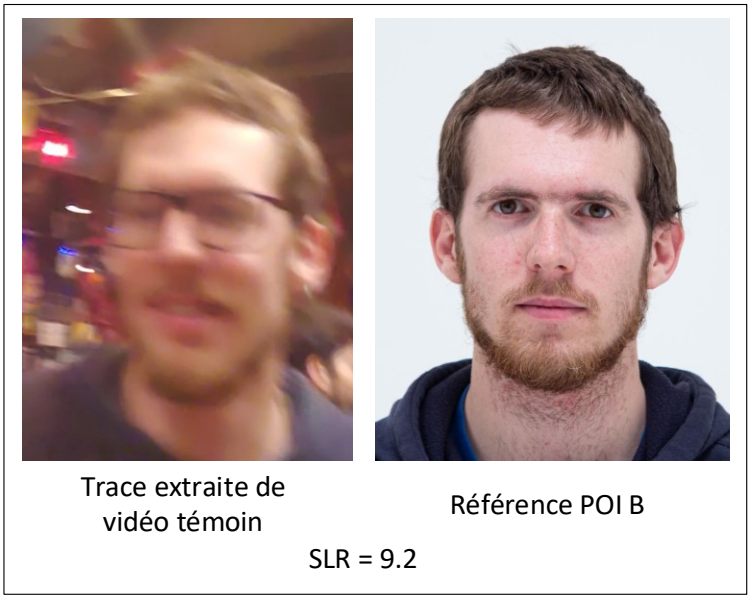


Figure 48 : Exemple de SLR supportant faiblement H_1, issu de la comparaison d'une trace extraite de vidéo témoin de faible qualité (gauche) et du mugshot POI B (gauche) par modèle intra spécifique/inter suspect-spécifique/PAVA.

Le SLR d'expertise pour cette comparaison par le modèle ci-dessus apporte un soutien faible[23] pour la proposition H_1 avec une valeur de 9.2 (Log_{10}(SLR) = 0.96 sur la Figure 5).

Avec ces informations en main, selon les propositions ci-dessus, l'expert.e pourrait donc :

[23] D'après l'échelle verbale établie dans le rapport de l'(ENFSI, 2015)

a. Reporter le poids de ce résultat pour la proposition H_1 en utilisant la valeur numérique et/ou une échelle verbale. Selon l'échelle verbale employée ici (ENFSI, 2015), un LR de 9,2 offre un soutien faible pour la proposition H_1.
b. Présenter le résultat comme pour le point a, en détaillant les taux d'erreurs et en laissant de fait le magistrat interpréter l'ensemble de ces données. Dans ce cas, l'expert.e est un « instrument » scientifique qui présente de manière pragmatique en toute transparence ces résultats sans avis sur leur valeur.
c. S'abstenir de communiquer ce résultat en regard de l'ensemble des informations dont il dispose. Dans ce cas, l'expert.e prend une décision, il n'est donc plus uniquement un instrument car il choisit, en amont du procès, de ne pas inclure ces résultats aux discussions du tribunal.

Notre opinion est que l'expert.e ne devrait pas prendre de décision à la place du magistrat. La seule lui appartenant est d'accepter ou non d'effectuer l'expertise lorsque celle-ci lui est proposée au premier abord. S'il juge exploitable le matériel fourni à ce moment par l'autorité en charge de l'instruction, il ne peut pas effectuer une expertise puis choisir si oui ou non ses résultats doivent être communiqués.

Dans notre exemple (Figure 48), l'expert.e reçoit une trace de faible qualité, surtout causée par le mouvement de l'individu, la luminosité hétérogène et la résolution moyenne. L'expert.e sait qu'un algorithme (ici, FaceNet) peut analyser cette image et détecter un visage à comparer aux références judiciaires. Cependant, d'après les performances des modèles évalués sur des images CCTV dont la qualité se rapproche de celle de la trace de question, il est attendu d'obtenir un SLR de faible poids pour l'une ou l'autre des propositions. En refusant l'expertise de prime abord, et en en expliquant les raisons, l'expert.e n'empiète pas sur le rôle décisionnaire du juriste et propose d'exclure la trace concernée du procès à cause des risques qu'elle soulève. En acceptant l'expertise malgré tout, l'expert.e s'engage à exploiter ladite trace pour fournir des résultats au tribunal afin de peser dans la décision juridique finale, quelle que soit la conclusion de son analyse. In fine, notre choix est de présenter ce résultat et les taux d'erreurs associés (point b) décrits ci-dessus).

Néanmoins, le risque de communiquer un résultat de faible soutien pour l'une ou l'autre des propositions est l'interprétation que les magistrats vont en faire (avocat de la défense ou avocat général selon le résultat), en lui accordant une valeur dépassant la portée du résultat. C'est pourquoi il peut être compréhensible que l'expert.e veuille retenir ces informations, de peur de voir ses conclusions extrapolées voire dénaturées.

À cela peuvent s'ajouter des risques de biais qui accompagnerait la communication d'informations circonstancielles. L'expertise devrait en théorie reposer uniquement sur l'évaluation des risques d'erreurs, et quand l'expert.e ne connaît pas de détails du cas, il est beaucoup plus facile de rester impartial. Mais dans certains cas les informations circonstancielles non essentielles à l'expertise en question sont partagées, et la connaissance d'éléments d'enquête risque de parasiter involontairement la réflexion de l'expert.e. Sera-t-il à l'aise de présenter un SLR de 10 en faveur de l'accusation (donc à charge) s'il sait que la trace est liée à un cas de viol

sur mineurs ? (Et avec un SLR de 0.1 en faveur de la défense (donc à décharge) ?) Beaucoup de paramètres intrinsèques, liés à la personnalité et l'expérience de l'expert.e, risquent d'ajouter des biais à son expertise, si le choix de reporter le chiffre lui est laissé.

De fait, ne serait-il pas pertinent de supprimer au maximum l'humain du processus d'évaluation pour le tribunal ? Pour l'investigation, il est essentiel de garder la partie humaine, mais pour le tribunal, elle pourrait rester exclusivement du côté des magistrats.

Dans (Swofford et Champod, 2021), les auteurs proposent une classification des méthodes d'évaluation selon le partage des rôles entre l'humain et l'algorithme (support, contrôle de qualité, évaluation, etc.). Selon cette échelle du niveau 0 (méthode exclusivement humaine) à 5 (méthode exclusivement automatique), nos modèles évaluatifs pour la phase investigative correspondent au niveau 3, dans lequel l'algorithme est utilisé en première instance puis l'expert.e utilise ses conclusions pour forger son opinion (trier la liste de candidats potentiels pour choisir les POI vers lesquelles orienter l'enquête). Dans ce cas, l'intervention humaine est essentielle mais ajoute un risque lié à l'interprétation subjective de l'opérateur. Une évaluation pour le tribunal correspond au niveau 4, dans lequel le rôle de l'humain est réduit à la supervision du processus, sans impact sur la conclusion fournie par l'algorithme (le SLR). À ce niveau, l'intervention humaine ne fait plus partie des préoccupations lors de la présentation des conclusions de l'évaluation. L'élément le plus limitant est le fait que ces conclusions dépendent exclusivement du fonctionnement boîte-noire de l'algorithme, qui doit être expliqué par l'expert.e au tribunal. Si l'expert.e interprète la conclusion de l'algorithme pour finalement prendre la décision de présenter ou non ces résultats au tribunal, cette méthode n'est plus de niveau 4, et les limitations humaines s'ajoutent au manque de transparence du système.

VII.3.2 Comparaison des valeurs de SLR investigatifs et de SLR d'expertise

Comme abordé précédemment (*cf*. VI.2.1), lors de l'investigation, les résultats de comparaison d'images sont le plus souvent communiqués de manière informelle pour guider les enquêteurs, sous la forme d'une liste de candidats potentiels ou d'une brève comparaison de la trace avec l'image d'un POI, conduisant à une possible "correspondance" ou "non-correspondance".

Nous postulons qu'il serait utile de fournir aux enquêteurs des résultats probabilistes et pondérés dès la phase d'enquête, à titre d'évaluation préliminaire. L'importance de cette problématique est mise en avant par la récente littérature l'abordant à travers des domaines précis (la reconnaissance faciale (Jacquet et Champod, 2020), les traces papillaires (Stoney *et al.*, 2020) et les traces numériques (Ryser *et al.*, 2020b)) ou plus globalement (Baechler *et al.*, 2020).

Pendant la phase d'investigation, seules la trace et les images de références d'un ou plusieurs individus potentiels sont disponibles. Chaque score de comparaison d'une trace et d'une référence permet de calculer un SLR investigatif. En se basant sur nos résultats, il est possible d'associer à ce SLR investigatif une valeur attendue de SLR d'expertise, en fonction du type de trace et de la population pertinente à laquelle appartient le POI. Dans ce cas, nous utilisons le SLR investigatif comme valeur préliminaire en vue du calcul d'un SLR plus spécifique en phase

d'expertise. Lorsqu'utilisé dans ce cadre, nous appelons donc le SLR investigatif « SLR préliminaire » ou « pré-SLR ».

Le calcul de pré-SLR pour soutenir les investigations amène naturellement la question suivante : Est-il possible de communiquer dès l'investigation une « prédiction » quant à la valeur de SLR d'expertise qui peut être attendue, dans le cas où l'individu ciblé par l'enquête deviendrait la personne d'intérêt dans le cadre d'une expertise ?

Dans les graphiques de la Figure 49, chaque point correspond à un cas, c'est-à-dire la combinaison du pré-SLR et du SLR d'expertise d'une comparaison trace-référence pour chaque scénario (ATM en bleu, CCTV en vert et ID et orange). Les courbes claires représentent un échantillon aléatoire (n=10000) de cas testés sous H_2 et les courbes foncées montrent l'intégralité des cas sous H_1.

Nous analysons successivement l'évolution des valeurs de SLR exclusivement avec le système MFE (Figure 49, colonne de gauche) puis avec l'utilisation de FaceNet en phase investigative et du MFE en phase d'expertise (Figure 49, colonne droite). Nous n'exposons pas les résultats exploitant uniquement l'algorithme FaceNet car, comme nous l'avons vu, les performances du MFE en font l'outil à privilégier pour le tribunal.

L'intervalle de chevauchement correspond aux valeurs logarithmiques de SLR obtenues à la fois pour des cas sous H_1 et sous H_2 (Figure 49, zones grises). Obtenir des pré-SLR dans ces intervalles ne permet pas d'apporter d'information pertinente aux investigateur.trice.s. Par exemple, pour le scénario ATM avec FaceNet, un pré-SLR de 1000 peut amener à deux valeurs de SLR d'expertise, l'une soutenant H_1 (SLR d'expertise=1000), l'autre soutenant H_2 (SLR d'expertise =0.01). Dans ce cas, le pré-SLR peut être utilisé pour orienter l'enquête vers la POI comparée puisqu'il soutient H_1, mais il est nécessaire d'informer les investigateur.trice.s que ce résultat préliminaire ne permet pas d'estimer la valeur du SLR d'expertise, qui pourrait soutenir H_1 ou H_2.

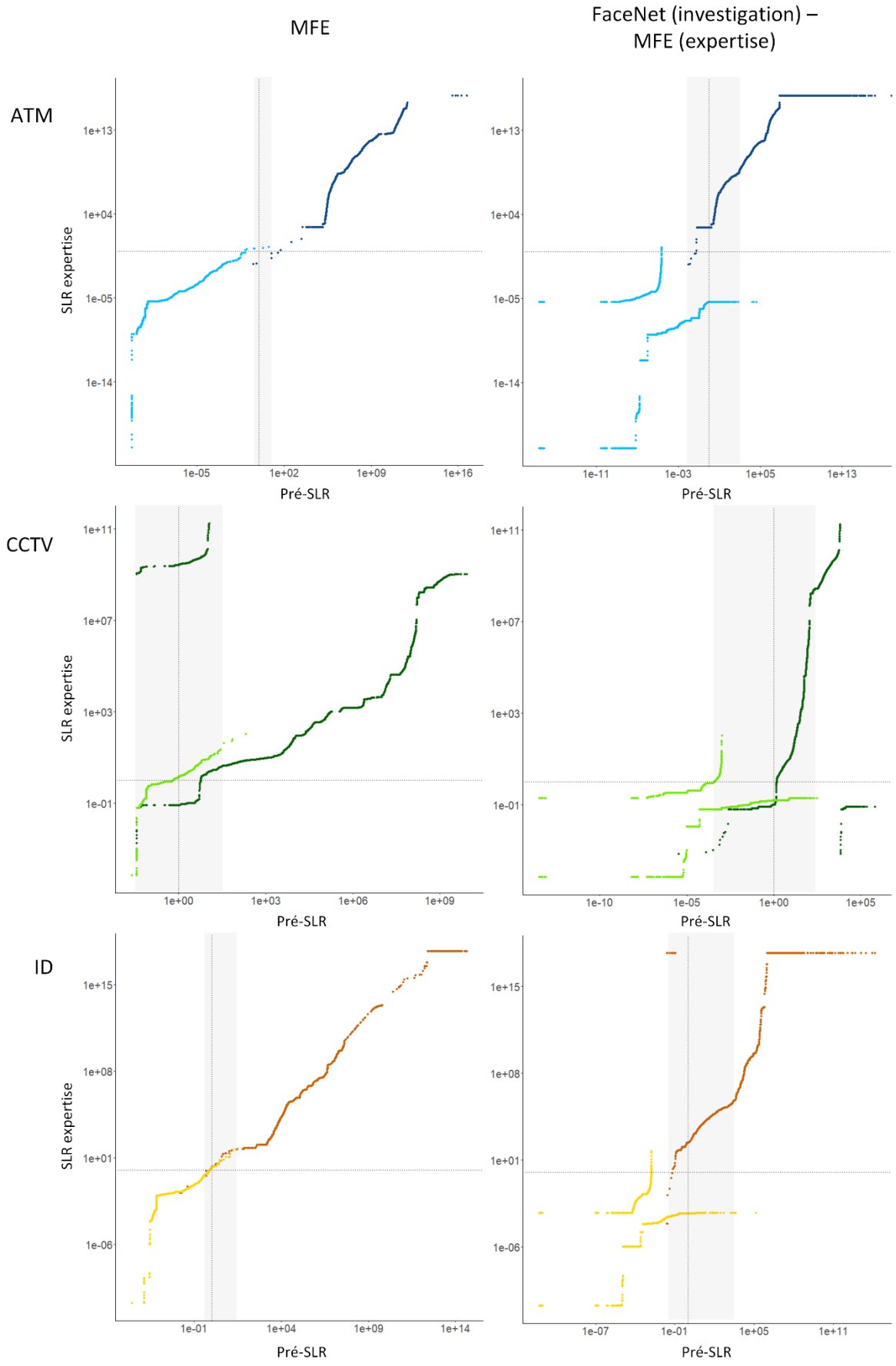


Figure 49 : Représentation des évolutions de valeurs de SLR entre les phases d'investigation (Pré SLR) et d'expertise (SLR d'expertise) pour chaque scénario (ATM en bleu, CCTV en vert et ID et orange), sous H_1 et H_2 (courbes foncées et claires, respectivement).

En utilisant le MFE de bout en bout, les intervalles de chevauchements sont logiquement moins étendus qu'en le combinant aux pré-SLR de FaceNet.

À partir de traces ATM et ID, l'évolution des valeurs de SLR est linéaire pour les cas sous H_1 et H_2. Pour les cas CCTV en revanche, les pré-SLR jusqu'à 10^{+3}, ce qui incluant tous les pré-SLR sous H_2, ne peuvent pas mener à une prédiction sur la valeur de SLR d'expertise. En outre, pour les cas sous H_1 (courbes vert foncé) dans cet intervalle, deux groupes de cas aux pré-SLR entre 10^{-2} et 10^{+3} sont associées à des valeurs de SLR d'expertise entre 10^{-1} et 1 d'une part, et entre 10^{+9} et 10^{+11} d'autre part.

La seconde colonne illustre les cas pour lesquels les services de police utiliseraient FaceNet comme aide à l'enquête avant de confier l'expertise finale à un.e expert.e doté.e du MFE.

Dans ces cas, l'évolution des valeurs des cas ATM et ID sont de nouveau comparables. Dans ces deux scénarios, les valeurs des pré-SLR qui soutiennent correctement H_1 en phase d'expertise et leur évolution sont linéaires. La majorité des cas sous H_2 soutiennent correctement H_2 en phase d'investigation et d'expertise. Cependant, un petit groupe de cas évoluent vers des valeurs soutenant à tort H_1, avec des valeurs de SLR d'expertise entre 1 et 10^{+4}. Dans la majorité des cas CCTV, les valeurs de SLR soutiennent correctement H_2 dans les phases d'investigation puis d'expertise, et les pré-SLR soutenant à tort H_2 mènent à des valeurs de SLR d'expertise soutenant toujours H_2.

VII.4. Synthèse

Les performances du MFE couplé à l'approche « intravariabilité spécifique / intervariabilité trace-spécifique » en font un outil de choix dans le cadre d'expertises à partir de traces CCTV, ATM et de documents d'identité. En revanche, l'algorithme FaceNet ne peut pas être utilisé dans le cadre d'une expertise car aucun des modèles testés n'atteint les exigences de robustesse attendue pour cet objectif lorsque les traces ne sont pas de qualité idéale.

À partir de nos modèles, nous avons mis en évidence l'utilité du calcul du SLR préliminaire en phase d'investigation pour la prédiction de SLR d'expertise. Pour chaque scénario, certains pré-SLR ne permettent pas d'apporter d'information pertinente aux investigateur.trice.s car le SLR d'expertise correspondant ne soutenait pas la même proposition. En dehors de ces zones, les estimations sont d'autant plus robustes lorsque le système MFE est utilisé lors des deux phases, mais il est possible d'estimer un SLR d'expertise, fourni par le MFE, à partir de pré-SLR calculés à partir des scores de FaceNet.

Chapitre VIII. Impact de la qualité des images sur les scores d'intravariabilité et analyse des cas soutenant à tort

Ce chapitre dresse un bilan de la multitude d'éléments impactant les scores de comparaison dans un premier temps, puis les SLR. Dans la première section, nous détaillons plusieurs exemples de comparaisons de références pour lesquelles le score – de similarité ou de distance - étend considérablement l'intravariabilité du POI. Dans un second temps nous nous focalisons sur les comparaisons dont le SLR soutient à tort l'une ou l'autre des propositions. Enfin, nous abordons l'impact du choix de la population pertinente sur le calcul de SLR.

VIII.1. Impact de la qualité des images de référence sur les scores d'intravariabilité

Dans un premier temps, nous nous focalisons sur les variations des valeurs de scores uniquement, lors des comparaisons d'images de référence de POI pour la modélisation de leur intravariabilité spécifique. L'objectif est de mettre en évidence les paramètres qui augmentent les scores de distance (respectivement diminuent les scores de similarité) dans la comparaison d'images d'un même individu, pour distinguer les paramètres pertinents (c'est-à-dire qui offrent une représentation réelle des variations de scores pour l'individu concerné) des paramètres inhérents à la qualité de l'image, ayant pour impact de surestimer ou sous-estimer l'intravariabilité de l'individu.

Nous présentons des exemples pour lesquels les systèmes FaceNet et MFE génèrent des scores imputables à la comparaison de deux individus différents. Cela donne un aperçu des éléments qui compliquent la reconnaissance d'individus par ces algorithmes lors de l'établissement d'intravariabilités sur des images frontales de qualité moyenne à bonne. Au vu des faibles performances du système MFI, nous ne jugeons pas pertinent d'en analyser les résultats ici.

VIII.1.1 FaceNet

Dans Schroff *et al.* (2015b), les auteurs indiquent que pour FaceNet, une valeur de score seuil s= 1.1 permettrait de séparer les scores de comparaison d'images d'un même individu des scores de comparaison d'individus distincts. À partir de ce constat, nous choisissons de présenter dans cette partie des comparaisons d'images d'un même individu pour la modélisation de son intravariabilité dont les scores approchent ou dépassent ce seuil.

Dans la Figure 50, un portrait récent est comparé à une ancienne photographie d'identité pour deux individus. Les deux images de la POI S sont de bonne qualité, sans élément obstruant, ni différence de pilosité, etc. Le résultat semble donc majoritairement lié à la différence d'âge de la POI sur ces deux clichés. Pour la POI Z en revanche, la photo ancienne est en niveaux de gris uniquement, et sur le portrait récent la POI porte des lunettes, ce qui peut avoir également un impact négatif sur les performances de FaceNet. Néanmoins, l'algorithme a classé ces deux exemples avec le même score (±0.01), ce qui ne traduit pas directement les différences que nous relevons à l'œil nu. Par exemple, il est possible que les couleurs de l'image n'entrent pas en considération dans l'encodage de l'image par l'algorithme.

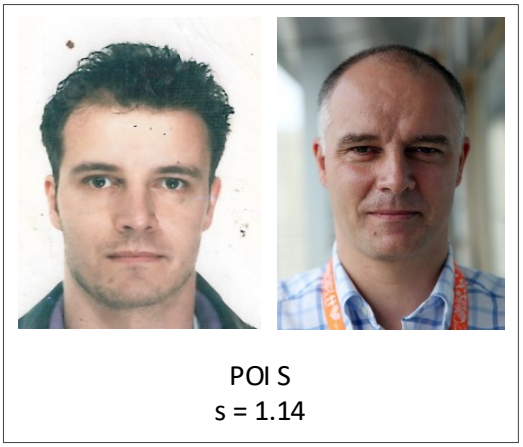


Figure 50 : Scores de distance (s) les plus élevés dans les distributions d'intravariabilités des POI S et Z lors de comparaisons de portraits récents avec des pièces d'identité anciennes par FaceNet.

Dans le premier exemple de la Figure 51 (gauche), la valeur élevée du score (1.09) peut être due à plusieurs facteurs. En premier lieu, le port de la barbe obstrue partiellement le visage et peut donc empêcher l'algorithme de correctement détecter les contours inférieurs du visage. De plus, l'individu ne porte de lunettes que sur l'un des portraits, ce qui peut avoir un impact sur le résultat selon la manière dont elles sont traitées par l'algorithme. Enfin, la luminosité hétérogène sur le second portrait peut avoir également un impact si l'algorithme détecte les limites entre les zones sombres et claires comme des caractéristiques faciales.

Dans la comparaison de droite, la seconde image est un portrait de qualité moyenne du fait de la plus basse résolution et l'ombre couvrant la moitié du visage, et la POI ne porte cette fois ni lunette ni barbe. Le score est logiquement plus élevé que dans la comparaison précédente (1.24). Cela appuie l'hypothèse d'un fort impact du port de la barbe, d'autant plus si elle n'est pas portée sur les deux images.

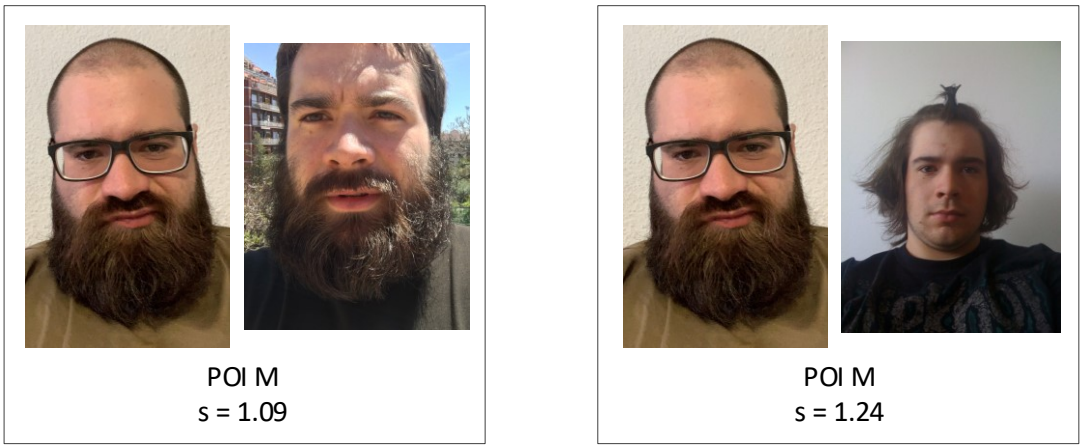


Figure 51 : Scores de distance (s) les plus élevés dans la distribution d'intravariabilité de la POI M par FaceNet lors de comparaisons de portraits avec variations de luminosités et de pilosité faciale.

Les deux comparaisons de la Figure 52 illustrent des exemples de scores élevés à partir d'images frontales de qualité moyenne à faible. Le premier exemple (POI H') utilise une image de

document officiel, de faible qualité, comparée à un portrait de bonne qualité. La faible résolution de la première image « lisse » les caractéristiques faciales et diminue le degré de détails analysables par le système, ce qui peut expliquer la valeur ambiguë du score. Pour la POI E, l'une des images est en niveau de gris, scannée à partir de documents d'identité et de faible résolution. La seconde image est un portrait pris par téléphone portable, à forte pixellisation et luminosité altérée (lumière jaune, surexposition du front et de nez). Le score très élevé associé à cette comparaison (1.23) peut être majoritairement attribué aux défauts de qualité des deux images, malgré le fait qu'il s'agisse de deux portraits frontaux où le visage est entièrement visible.

Figure 52 : Scores de distance (s) les plus élevés dans les distributions d'intravariabilités des POI H' et E par FaceNet lors de comparaisons de portraits et de pièces d'identité de qualité variable.

Les deux exemples suivants illustrent clairement l'impact de la dégradation des photographies d'identité par les éléments de sécurité ajoutés sur les pièces officielles, lors de comparaisons avec des portraits frontaux de bonne qualité (Figure 53). Les deux images scannées directement à partir de documents d'identité sont à gauche pour chaque comparaison. Dans un cas (POI P'), un dessin et une trame de fond sont imprimés par-dessus la photographie, ce qui recouvre partiellement le visage et augmente le grainage de l'image. Pour la POI S, seules des lignes courbes irrégulières recoupent le visage horizontalement, et la photographie (originalement en couleur) a été incluse au document en niveaux de gris et à plus faible résolution. Les scores, respectivement 1.02 et 1.22, traduisent la difficulté de FaceNet à gérer l'ajout d'éléments externes qui recoupent le visage, en particulier pour la seconde comparaison pour laquelle le visage reste parfaitement délimité et les caractéristiques faciales bien visibles pour l'œil humain. Néanmoins, il eut été attendu que FaceNet performe moins bien sur la comparaison de la POI P' du fait de la très faible qualité du document d'identité. Le plus haut score attribué au POI S peut être imputé à la présence d'une barbe sur une seule image, et/ou au fait que l'algorithme gère moins bien certains éléments de sécurité comme les lignes irrégulières.

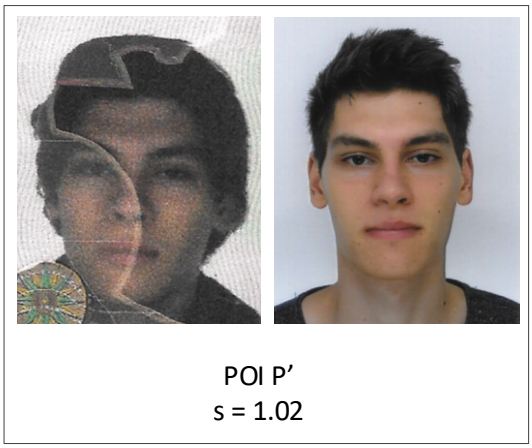


Figure 53 : Scores de distance (s) les plus élevés dans les distributions d'intravariabilités des POI P' et S par FaceNet lors de comparaisons de portraits avec des pièces d'identité altérées par les éléments de sécurité.

La Figure 54 présente deux comparaisons où plusieurs variables peuvent impacter sur le résultat, mais le plus évident étant les variations de luminosité, à la fois entre les images mais également au sein de chaque portrait. Pour la POI J, l'algorithme semble se heurter au changement d'expression du visage (sourire puis neutre) et à la différence de luminosité (prise de vue en journée puis de nuit). Les lunettes sont portées et identiques dans les deux images, mais projettent des ombres sur le visage dans le cliché de nuit, ce qui peut ajouter des éléments encodés à tort par le système. Dans la seconde comparaison (POI S), une photographie est un portrait où la tête est couverte par un casque dont l'ombre couvre également la moitié du visage, et l'autre image est une photographie d'identité très claire mais de bonne qualité. Le casque n'obstrue pas directement le visage, mais l'ombre qu'il projette empêche de bien analyser les caractéristiques faciales, même à l'œil nu, ce qui peut expliquer le haut score de cette comparaison (1.19).

Figure 54 : Scores de distance (s) les plus élevés dans les distributions d'intravariabilités des POI J et S par FaceNet lors de comparaisons de portraits avec variations de luminosités et d'expression faciales.

À travers ces exemples, nous avons pu détecter plusieurs variables problématiques pour l'analyse et la comparaison d'images frontales pourtant généralement considérées de bonne qualité du fait de leur provenance (documents d'identité) ou de leur résolution. Ces observations sont à prendre

en compte en amont de l'expertise, lors de la collecte d'images de référence d'une POI par les investigateur.trice.s, car ces comparaisons entrent ensuite dans la modélisation d'intravariabilité et ont donc un impact direct sur le calcul de SLR. Il est nécessaire de ne pas s'arrêter aux critères de bases (images frontales, visages non obstrués, etc.), mais également prendre en compte les variations de luminosité, la pilosité et les expressions faciales et toutes les dégradations postérieures à la prise de vue, lors de la collecte des références.

VIII.1.2 MFE

Comme expliquer pour le cadre civil (Chapitre VI.4), un seuil est couramment fixé à s = 3500 pour l'attribution de « Hit » à la comparaison d'images d'une même personne. Nous nous sommes alors basés sur cette valeur pour illustrer plusieurs comparaisons dont le faible score est classiquement imputable à la comparaison de visages différents.

La Figure 55 expose trois comparaisons dans lesquelles figurent les mêmes POI que précédemment, parmi les 25 comparaisons dont les scores de similarité sont plus proches des scores de comparaison d'individus différents. Ces exemples résument efficacement les variables sur lesquelles se heurte un système aussi performant que le MFE : le vieillissement de l'individu (POI Z et S) et la dégradation d'images utilisées sur des documents d'identité (POI P'). Il semble que le MFE gère plus efficacement les variabilités dues aux éléments tels que la pilosité faciale, le port de lunettes et la faible définition de l'image.

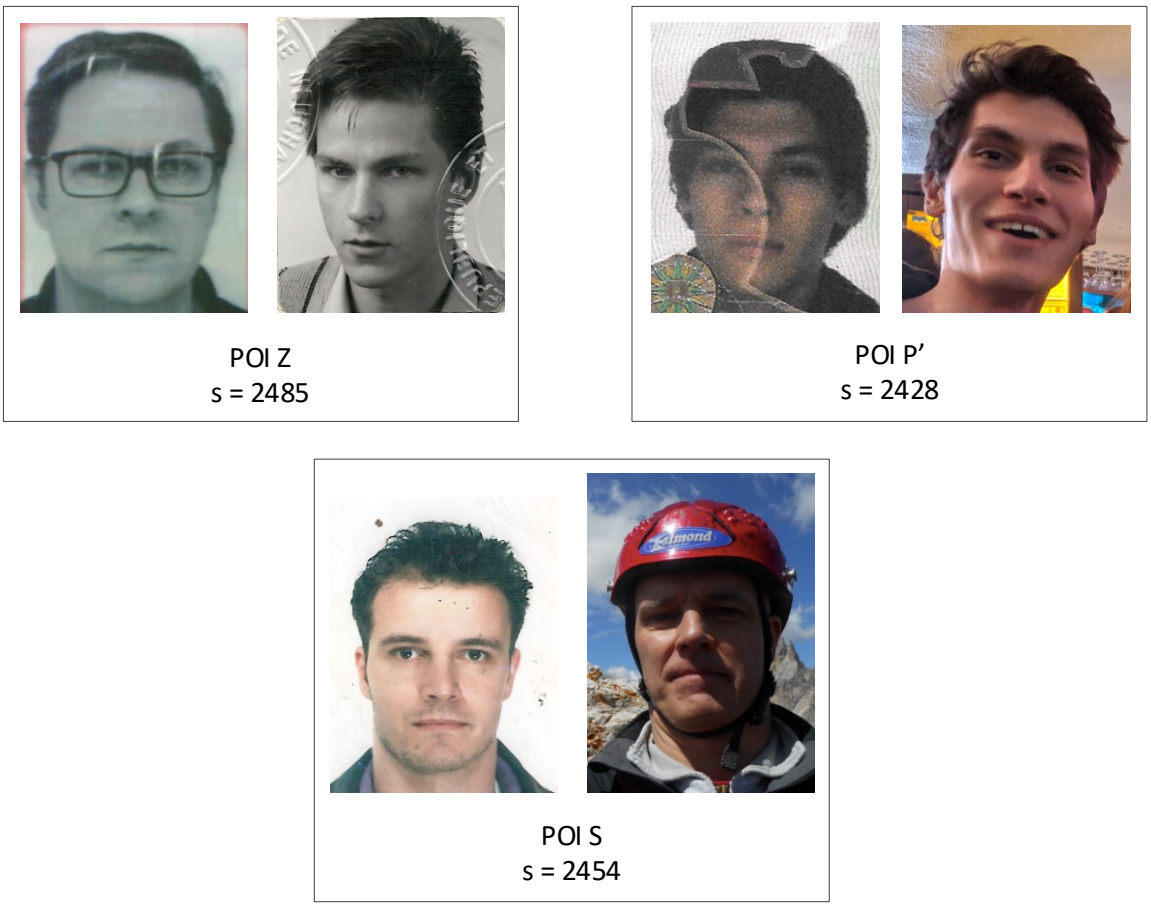


Figure 55 : Scores de similarité (s) les plus faibles dans les distributions d'intravariabilités des POI Z, P' et S par MFE.

VIII.2. Impact de la qualité des traces et références dans les cas orientant à tort

À travers cette section, nous nous focalisons sur les cas dont le SLR d'expertise soutient à tort l'une ou l'autre des propositions. Comme dans la section précédente, nous cherchons à identifier les variables qui peuvent avoir un impact indésirable sur le résultat de la comparaison. En raison de la confidentialité des données judiciaires fournies par les polices cantonales de Vaud et Neuchâtel, les exemples publiés ici montrent exclusivement des traces et *mugshots* des participants photographiés dans le cadre de du présent projet.

VIII.2.1 SLR soutenant à tort H_2

Dans un premier temps, nous analysons les comparaisons entre trace et référence de même source dont le SLR soutient à tort la proposition H_2, avec les systèmes MFE (Figure 56) et FaceNet (Figure 58).

La Figure 56 expose un aperçu de comparaisons traces-*mugshots* des trois scénarios par le système MFE. Les comparaisons de trace CCTV montrent une nouvelle fois que le port de la barbe, la faible résolution et l'angle de prise de vue ont un impact significatif sur le calcul de SLR. La trace ID de la POI S est la seule générant un SLR soutenant à tort H2 à l'aide du MFE, ce qui implique que les SLR assignés aux autres ID anciennes, notamment pour la POI Z, soutenaient correctement H_1. L'utilisation du MFE couplée avec la calibration PAVA augmente donc significativement les performances sur ce type de traces. Enfin, le SLR de comparaison de la trace ATM avec le *mugshot* de la POI E montre que, comme pour les traces CCTV, l'orientation du visage est un facteur limitant car le visage n'est que partiellement visible, mais également que le système analyse et compare les éléments ajoutés tels que les lunettes.

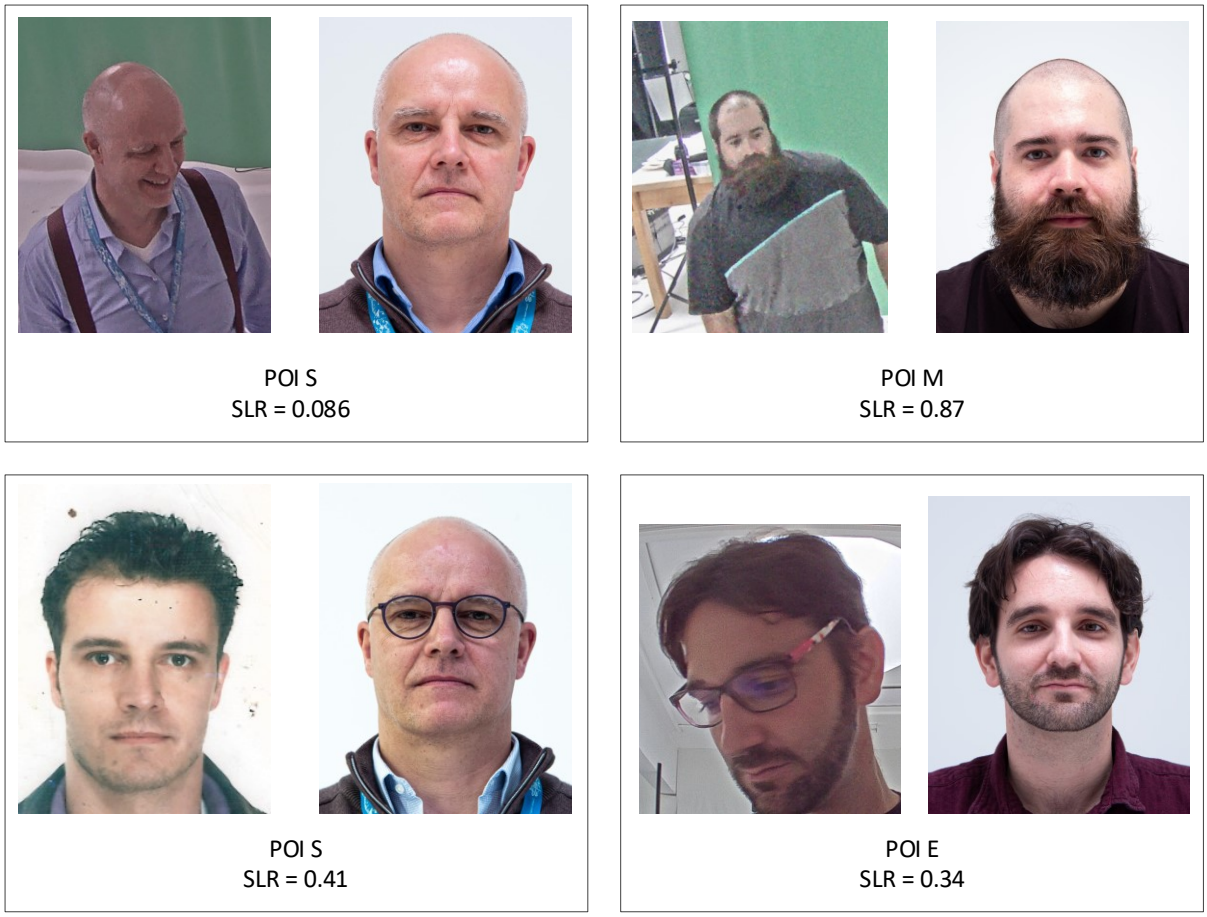


Figure 56 : Comparaisons traces-mugshots parmi les SLR soutenant à tort H_2 générés par le système MFE.

Concernant les résultats de FaceNet, beaucoup de traces CCTV génèrent des SLR soutenant à tort H_2 avec des valeurs de 1 à 10^{-2}. De fait, il nous parait judicieux ici de présenter plusieurs traces CCTV dont toutes les comparaisons aux *mugshots* du même individu ont généré des SLR soutenant à tort H_2 d'ordre de grandeur 10^{-2}. Il est intéressant de noter que les SLR soutenant à tort H_2 viennent majoritairement de plusieurs traces pour seulement quelques individus récurrents (plutôt que d'une ou deux traces pour tous les POI, par exemple). Cela suggère que les performances du système dépendant non seulement de la qualité des images mais également de caractéristiques individuelles. Les trois traces de la Figure 57 montrent la plus grande difficulté de FaceNet à reconnaitre la POI Z, à partir de CCTV dont plusieurs éléments dégradent la qualité : grands angles de prise de vue verticaux et horizontaux, port de lunettes cachant partiellement la zone orbitale, et la dissimulation de la mâchoire inférieure avec la main.

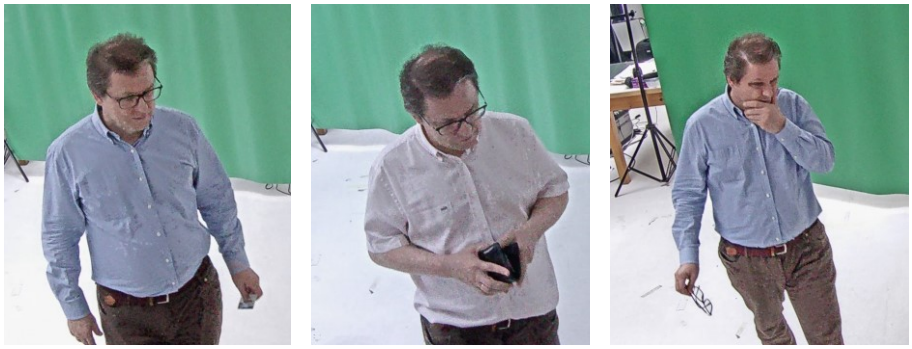


Figure 57 : Traces CCTV de la POI Z menant systématiquement à des SLR soutenant à tort H_2 avec FaceNet.

Les comparaisons ci-dessous montrent des résultats cohérents avec l'ensemble de nos observations, puisqu'elles génèrent des SLR soutenant à tort H_2 à partir d'image ID et ATM de qualité moyenne.

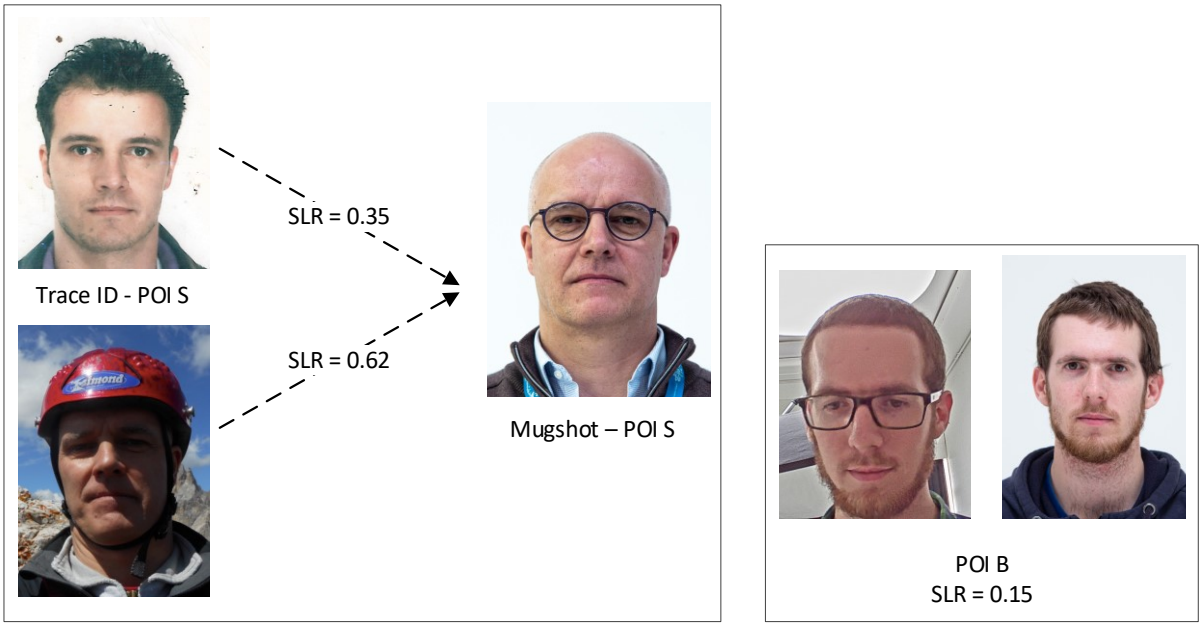
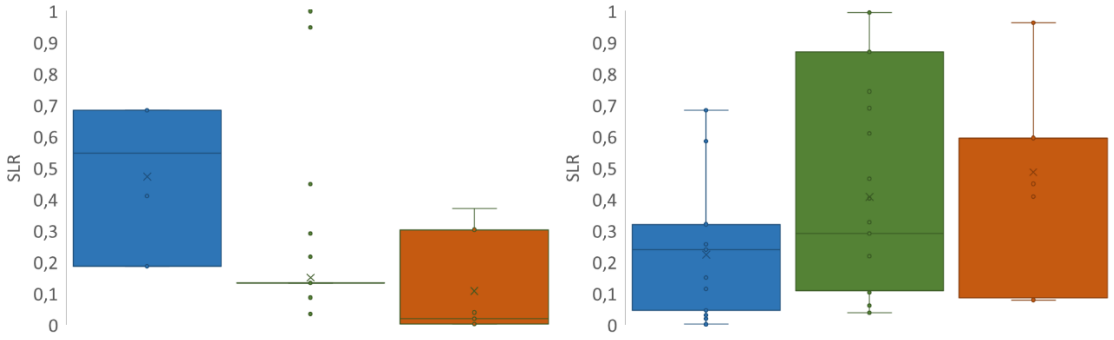


Figure 58 : Comparaison traces-mugshots générant des SLR soutenant à tort H_2 pour des traces ID et ATM avec FaceNet.

La figure ci-dessous résume la situation des SLR soutenant à tort H_2 à partir des modèles choisis précédemment pour chaque scénario, pour le MFE et FaceNet. Les *boxplots* illustrent l'étendue de la majorité des SLR pour chaque scénario (ATM en bleu, CCTV en vert et ID en orange), et le tableau complète ces données avec le pourcentage de SLR concerné et la valeur minimum.



	MFE			Facenet		
	ATM	*CCTV*	*ID*	*ATM*	*CCTV*	*ID*
SLR soutenant à tort H2 (%)	0,1	10,8	0,5	1,3	11,7	3
SLR min	0,186	0,035	0,002	0,002	0,038	0,081

Figure 59 : Bilan des SLR soutenant à tort H_2 des modèles choisis pour une utilisation au tribunal du MFE et de FaceNet

Dans la section suivante, nous étudions les comparaisons pour lesquelles les SLR soutiennent à tort la proposition H_1. Cela nous permet de comparer les types de variables qui orientent à tort les valeurs du SLR sous les deux propositions.

VIII.2.2 SLR soutenant à tort H_1

Lors du développement de la méthodologie pour une utilisation au tribunal, les modèles privilégiés sont ceux qui génèrent les plus faibles taux de RMEP, donc les plus faibles taux de SLR soutenant à tort H_1, inférieurs au taux de SLR soutenant à tort H_2. De fait, peu de comparaisons sons analysables dans cette section, aussi bien pour le système MFE que pour FaceNet.

Dans la Figure 60, le système MFE accorde des SLR faibles en faveur de H_1 lors de la comparaison de traces ATM, ID et CCTV de la POI D, avec deux *mugshots* de la POI L. Les deux traces sont de bonne qualité, et le fait qu'il s'agisse des seuls exemples disponibles pour les participants à ce projet démontre que le MFE ne se heurte ici qu'à la ressemblance fortuite des deux POI sur ces images en particulier.

Figure 60 : Comparaisons traces – mugshots générant des SLR soutenant à tort H_1 par la ressemblance fortuite des POI D et L selon le système MFE.

Avec FaceNet, la majorité des SLR soutenant à tort H_1 pour les scénarios ATM et CCTV proviennent de comparaisons avec des individus des bases de données de police. Il est donc impossible d'en montrer ces exemples afin de préserver la confidentialité des données. La Figure 61 présente donc certains exemples de comparaisons n'impliquant que les volontaires engagés dans ce projet. Les SLR obtenus apportent plus un poids neutre, c'est-à-dire ne soutenant aucune des deux propositions. Comme précédemment, on remarque la présence/absence de lunettes dans les deux images comparées.

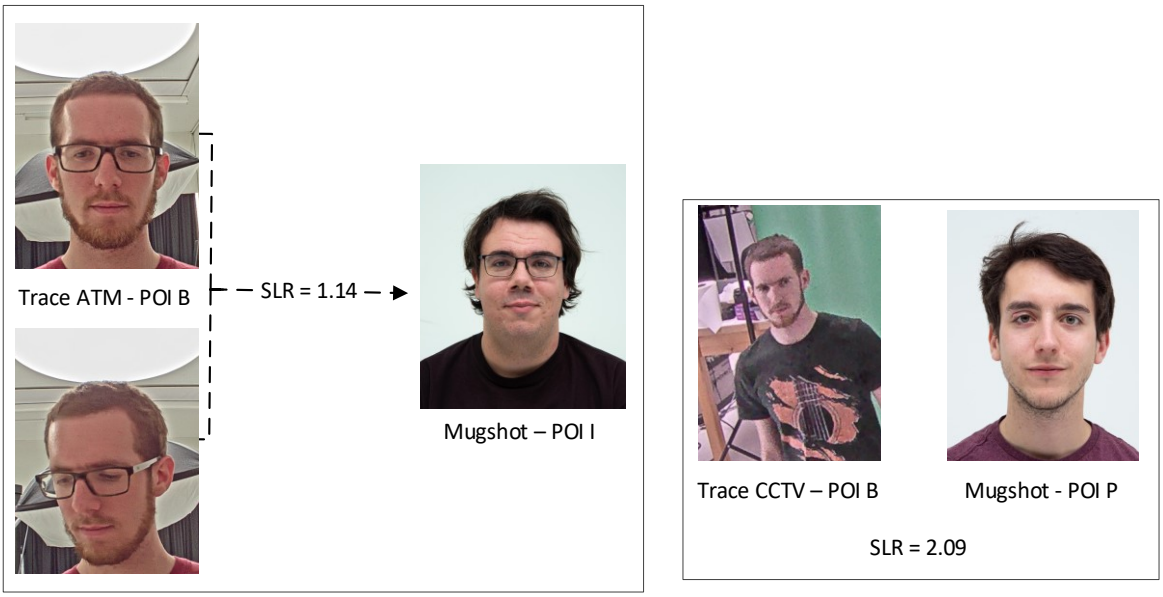


Figure 61 : Comparaisons traces–mugshots générant des SLR soutenant à tort H_1 selon le système FaceNet à partir d'images ATM et CCTV.

Concernant les traces ID (Figure 62), les comparaisons traces-*mugshots* des POI volontaires générant des SLR soutenant à tort H_1 concernent les traces dont la qualité est dégradée par la faible résolution, l'obstruction du visage, le port de lunette et la densité de la pilosité faciale (barbe et cheveux). L'algorithme semble prendre en compte ses éléments, ce qui augmentent les valeurs de SLR lorsque l'individu de référence les présente également.

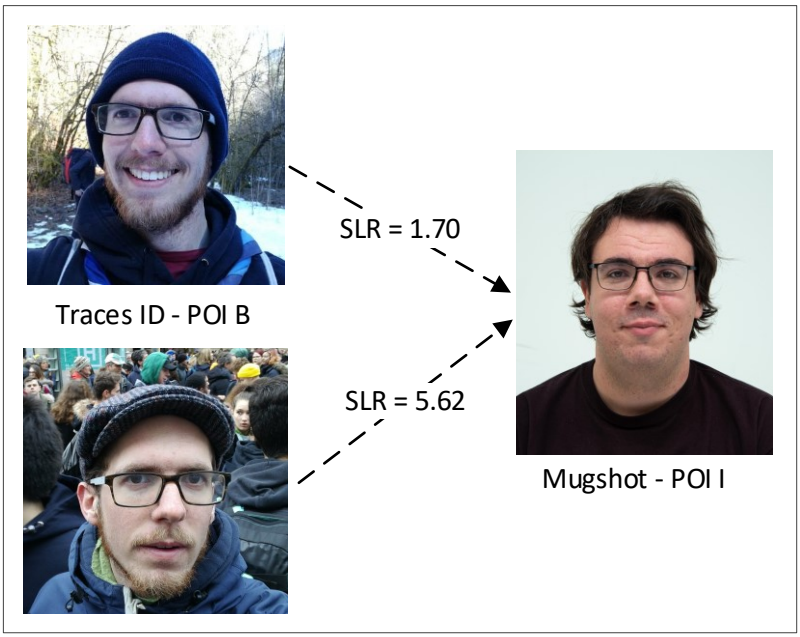


Figure 62 : Comparaisons traces ID–mugshots générant des SLR soutenant à tort H_1 selon le système FaceNet.

Dans la Figure 63, on constate chez FaceNet la même difficulté que le MFE pour différencier les POI D et L à partir de cette même paire d'images. Les comparaisons de gauche génèrent également des SLR soutenant à tort H_1 à cause de la ressemblance fortuite de deux POI. Il est important de remarquer la valeur très élevée du SLR, surtout pour le premier exemple (haut). Dans la pratique, malgré le soutien fort qu'apporte ce résultat à la proposition H_1, il doit évidemment être exclu des discussions autour du cas, puisque les traces ID ont été prises près de 30 ans avant les *mugshots*. Dans cet exemple, les deux POI ne sont pas apparentées, il s'agit donc de ressemblance fortuite.

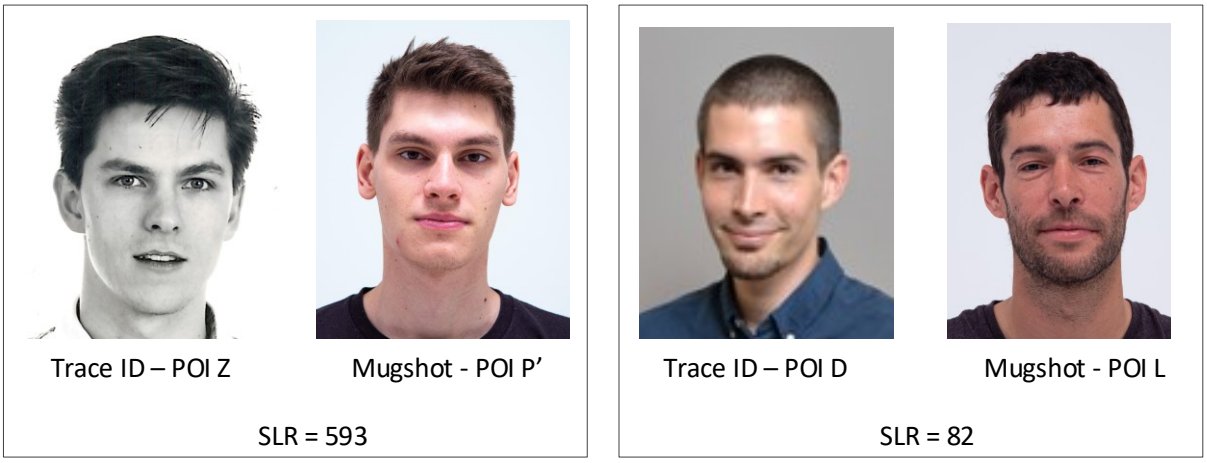
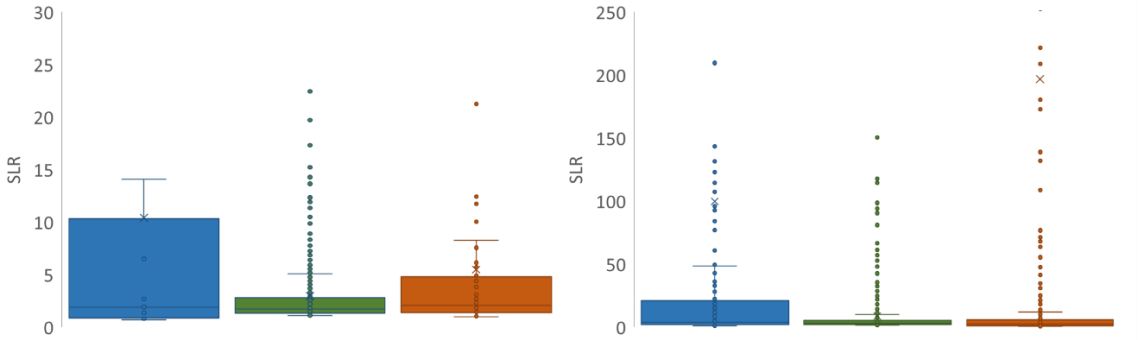


Figure 63 : Comparaisons traces–mugshots générant des SLR soutenant à tort H_1 par la ressemblance fortuite des POI D et L, et des POI Z et P' selon le système FaceNet.

Comme précédemment, la figure ci-dessous résume la situation des SLR soutenant à tort H_1 à partir des modèles choisis pour chaque scénario, avec le MFE et FaceNet. Les *boxplots* illustrent l'étendue de la majorité des SLR et le tableau complète ces données avec le pourcentage de SLR concerné et la valeur maximale.



	MFE			Facenet		
	ATM	*CCTV*	*ID*	*ATM*	*CCTV*	*ID*
SLR soutenant à tort H1 (%)	0,1	4,4	0,3	2,1	6,3	3,9
SLR max	65	74	77	6983	501	63266

Figure 64 : Bilan des SLR soutenant à tort H_1 des modèles choisis pour une utilisation au tribunal du MFE et de FaceNet.

VIII.3.Impact du choix de la population pertinente

Afin de visualiser l'impact du tri de la population pertinente sur les performances des trois systèmes, nous montrons les variations de valeurs de RMEP, RMED et (min)Cllr pour les meilleurs modèles évaluatifs pour l'investigation et le tribunal (Tableau 16). Nous appelons « tri » de la BdD la segmentation de celle–ci en plusieurs sous-populations sur la base de critères de genre perçu, de tranche d'âge et de type ethnique. Ce tri est effectué manuellement par l'analyse rapide de tous les *mugshots* de la BdD. Le terme « population triée » se réfère donc à ces sous-populations.

Dans le tableau ci-dessous, les valeurs sont positives (en vert) lorsque les performances des modèles sont meilleures avec la population non triée, et négatives (en rouge) lorsque ces performances sont meilleures avec la population triée.

Il est attendu que le tri de la population pertinente diminue les performances de tous les modèles, car ce tri consiste à sélectionner des individus plus ressemblants à l'œil nu par rapport au suspect. Concrètement, dans notre exemple, les populations pertinentes étaient composées d'images d'individus hommes ou femmes selon le cas, de type caucasien (exceptée pour la POI X, homme d'origine Pakistanaise). Ce tri diminue considérablement le nombre de POI dans la population pertinente et les intervariabilités. Cependant, les résultats ci-dessous ne montrent pas une diminution très significative des performances après ce tri.

En analysant plus spécifiquement ces résultats, on remarque avant tout que les variations de performances du MFI sont plus fortes que pour les deux autres systèmes, mais totalement hétérogènes. Le tri de la population pertinente diminue très faiblement la calibration de tous les modèles, et les impacts sur les valeurs de risques d'erreurs varient d'un modèle à l'autre.

Pour les systèmes FaceNet et MFE, toutes les différences sont positives, ce qui montre la diminution des performances avec une population triée, spécifique au POI. Néanmoins, ces variations sont peu significatives avec FaceNet, toujours inférieures à 1.5% pour les RMEP et 3% pour les RMED. Avec le MFE, les variations dépendent plus du scénario : moins de 1% de différence pour les RMEP et RMED pour le scénario ATM, entre 0.76 et 6.42% pour les CCTV.

La différence entre le MFI d'une part et le MFE et FaceNet d'autre part peut être imputée à la structure de chaque algorithme. Pour rappel, le MFI est basé sur un fonctionnement *rule-based* alors que le MFE et FaceNet sont construits sur le principe du *deep learning*. Leurs performances sont donc étroitement liées aux données utilisées lors que la phase d'apprentissage, alors que le MFI compare des visages en suivant la méthodologie immuable dictée par son algorithme.

Tableau 16 : Variations des valeurs de RMEP, RMED et (min)Cllr lors du tri de la population pertinente.

Phase	Métriques	FN			MFI			MFE		
		ATM	CCTV	ID	ATM	CCTV	ID	ATM	CCTV	ID
Investigation	RMEP (%)	0,21	0,89	0,65	-6,44	5,77	-2,46	0,64	2,69	0,36
	RMED (%)	1,42	2,24	2,09	14,90	6,73	8,33	0,11	6,42	2,26
	(min)Cllr	0,03	0,06	0,05	0,10	0,04	0,07	0,01	0,13	0,05
Tribunal	RMEP (%)	0,76	0,58	1,50	-9,81	22,97	1,81	0,16	2,92	0,32
	RMED (%)	0,59	1,04	0,81	21,90	-12,65	7,38	0,27	0,76	1,73
	(min)Cllr	0,03	0,04	0,05	0,11	0,01	0,12	0,01	0,05	0,03

Le tri de la population pertinente est une problématique importante dans cadre du tribunal où l'expert.e doit expliquer et justifier l'ensemble de sa méthodologie. Ce n'est pas le cas dans le cadre de l'investigation, où le SLR sert à orienter l'enquête. Nos résultats montrent le faible impact du tri de la population pertinente pour les cas impliquant des individus caucasiens, hommes et femmes. Il sera essentiel d'étudier cette problématique à partir de populations plus variées pour recenser plus précisément les impacts du tri sur chaque population.

VIII.4.Synthèse

À travers cette section, nous avons identifié plusieurs variables qui augmentent le risque de résultats induisant en erreur, tout d'abord pour les scores d'intravariabilité puis pour les SLR.

Les images de référence collectées pour générer les scores d'intravariabilité sont par définition considérées de bonne qualité. Malgré cela, nous avons mis en évidence plusieurs éléments pour lesquels les systèmes FaceNet et MFE génèrent des scores d'intravariabilité imputables à la comparaison de deux individus différents.

- Vieillissement des individus
- Dissimulation partielle du visage, notamment par le port de la barbe
- Ajout d'éléments tels que des lunettes de vue
- Dégradation de la qualité des photographies standards des documents d'identité officiels par l'ajout des éléments de sécurité recouvrant partiellement ou totalement l'image
- Variation des conditions de prise de vue (luminosité, angle de prise de vue, etc.)
- Modification des traits selon les expressions faciales
- Ressemblance fortuite

Nos résultats montrent que tous ces éléments impactent les performances de FaceNet, alors que le MFE n'est limité que par le vieillissement des individus et la dégradation des documents officiels par les éléments de sécurité, ou par la combinaison de plusieurs autres critères. Il est nécessaire de prendre en compte ces critères non seulement lors de la collecte de traces mais également lors de la prise de références. Par exemple, si le suspect porte des lunettes de vue sur la trace, il est utile de prendre des clichés de références des POI impliquées avec et sans lunettes.

Les cas pour lesquels le SLR soutient à tort (le plus souvent H_1) concernent les traces CCTV (le grand angle diminuant la taille de la zone du visage) ainsi que les variations d'angle de prise de vue et d'expression faciale.

Enfin, nos résultats montrent le faible impact du tri de la population pertinente pour les cas impliquant des individus caucasiens, hommes et femmes.

Chapitre IX. Réponses apportées aux problématiques de recherche et perspectives futures

À travers ce projet, nous avons évalué le potentiel d'application de systèmes automatiques de reconnaissance faciale dans les phases d'enquête et dans l'expertise à destination des tribunaux. Pour cela, nous avons testé trois systèmes : Google FaceNet (*open source*, *deep learning*), Idemia Morphoface Investigate (MFI, commercial, *rule-based*) et Idemia Morphoface Expert (MFE, commercial, *deep learning*). Notre méthodologie se base sur l'utilisation d'une méthode d'évaluation probabiliste pour standardiser et valider l'interprétation des scores de comparaison de systèmes de reconnaissance faciale différents.

Ce chapitre conclut l'ensemble de la recherche en résumant avant tout ses principales contributions, puis les réponses apportées aux problématiques initiales et fait état des perspectives futures.

IX.1. Contributions de la recherche

Notre recherche prend sa source sur une revue critique de la littérature qui a mis en lumière les problématiques principales jalonnant l'ensemble du processus judiciaire, depuis la collecte des images sur le terrain, jusqu'à leur exploitation tant à des fins d'investigation et de présentation d'éléments de preuve au tribunal. Pour ce faire, nous développons une méthodologie fondée sur des expérimentations en conditions réelles, destinée à consolider les fondations empiriques nécessaires à une intégration efficace et réaliste de la reconnaissance faciale automatique au sein du système judiciaire.

Quatre principales contributions émergent de cette recherche :

- Au niveau de l'enquête, l'évaluation opérationnelle de différents systèmes automatiques de reconnaissance faciale et la plus-value démontrée de l'utilisation du SLR dans ce contexte constituent une contribution majeure dans notre champ d'études. De même, notre introduction du SLR investigatif comme nouvel outil d'aide à la décision pour les investigateur.trice.s présente des résultats prometteurs en comblant des lacunes qui sont, à notre connaissance, encore peu considérées.
- Au niveau du tribunal, nos résultats constituent des données validées sur lesquelles sont à même de se reposer de futures expertises. Cette recherche répond ainsi à la nécessité de disposer d'études empiriques sur des données opérationnelles afin de permettre l'utilisation de systèmes automatiques comme élément de preuve au tribunal.
- Au niveau de l'ensemble du processus judiciaire, notre méthodologie et les résultats qui en découlent sont un premier pas déterminant vers le décloisonnement des objectifs de l'investigation et de l'évaluation, cette dernière traditionnellement réservée aux expertises destinées au tribunal.
- Au niveau de l'interprétation de résultats forensiques, le développement de notre méthodologie a mené à la clarification de l'usage des différentes approches (modèles probabilistes) de calcul du SLR, et applicable au LR (pas uniquement basé sur un score) plus généralement.

Notre méthodologie, guidée par les problématiques définies en amont et combinée à nos expérimentations en conditions réelles, ouvre de nouvelles et nombreuses perspectives concernant l'utilisation de la reconnaissance faciale dans le système judiciaire. Au-delà de son implication innovante dans ce domaine, cette approche s'inscrit dans un mouvement plus large et susceptible de s'appliquer à l'ensemble des traces d'intérêt forensique.

IX.2. Réponses apportées aux problématiques initiales

Avant tout, nous souhaitons apporter les réponses concrètes et succinctes à la question la plus évidente en conclusion de cette recherche : Les systèmes et modèles testés sont-ils utilisables dès maintenant dans le cadre opérationnel ?

Le MFI est naturellement exclu de cette conclusion car il s'agit d'un système obsolète dont l'intérêt était l'observation des progrès en termes de performances par la comparaison avec les systèmes plus récents MFE et FaceNet.

Le MFE couplé aux modèles choisis est opérationnel et robuste ; il peut donc être utilisé dans l'investigation et dans le cadre d'expertise à partir de traces CCTV, ATM et de documents d'identité. Nos résultats concernent avant tout la population caucasienne et majoritairement masculine, pour laquelle nos modèles sont applicables dans le cadre d'expertise. Pour des populations ethniques différentes, il peut être nécessaire de tester davantage nos modèles. Ce point est développé dans la section suivante (IX.7).

Dans le cadre investigatif, le déploiement du MFE accessible aux forces de l'ordre serait un atout majeur pour la recherche de personnes dans de larges bases de données judiciaires. En Suisse, des tests à grande échelle pourraient être menés en collaboration avec les polices cantonales romandes, qui ont accès au système MFE.

L'algorithme FaceNet ne peut pas être utilisé dans le cadre d'une expertise car tous les modèles testés n'atteignent pas les exigences de fiabilité attendue pour cette application. Il constitue néanmoins un bon outil gratuit et *open source* (de fait, modifiable) pour la recherche de personnes ou la vérification 1:1 à partir d'images faciales de bonne qualité, ce qui exclut les images de types CCTV.

Dans la suite de cette section, nous reprenons les problématiques autour desquelles s'est construite notre recherche pour y répondre succinctement à l'aide des résultats détaillés et discutés précédemment.

4) Comment choisir une approche selon la quantité et la qualité de données à disposition, et quels sont les impacts sur les résultats ?

Nous avons testé deux approches de modélisation d'intravariabilité (spécifique et générique) et d'intervariabilité (trace-spécifique et suspect-spécifique) et de calibration (RegLog et PAVA). Les modèles choisis en fonction du type de trace, du système et du cadre d'utilisation sont résumés dans le Tableau 17.

Tableau 17 : Modèles probabilistes choisis pour chaque phase, scénario et système automatique.

		Tribunal *Intravariabilité spécifique*		Investigation *Intravariabilité générique*	
		Intervariabilité	*Calibration*	*Intervariabilité*	*Calibration*
FN	*ATM*	Trace-spécifique	PAVA	Suspect ou Trace-spécifique	SLR bruts
	CCTV	Suspect-spécifique		Suspect-spécifique	Reg Log
	ID	Trace-spécifique		Trace-spécifique	SLR bruts
MFI	*ATM*	Suspect-spécifique	PAVA	Suspect-spécifique	Reg Log
	CCTV	Non-adaptée		Non-adaptée	
	ID	Suspect-spécifique	PAVA	Suspect-spécifique	Reg Log
MFE	*ATM* *CCTV* *ID*	Trace-spécifique	PAVA	Suspect ou Trace-spécifique	PAVA

La modélisation d'une intravariabilité générique est une solution adéquate pour compenser le manque d'images de référence souvent rencontré en phase investigative. La contrainte principale est la constitution d'une base de données de plusieurs (minimum 3-4) images de référence d'individus appartenant à la même population que le POI.

Dans la littérature, la base de données *Racial Faces in the Wild* (Wang *et al.*, 2019), disponible sur demande auprès des auteurs, regroupe environ 2M d'images de 38'000 célébrités prises dans des conditions non contrôlées (à l'instar de la base LFW (Huang *et al.*, 2008)) d'individus d'origines ethniques différentes (« Caucasien, Indien, Asiatique et Africain »). Il ne s'agit pas d'images frontales de bonne qualité et plusieurs populations n'y sont pas représentées, ce n'est donc pas une solution idéale pour la modélisation d'intravariabilité générique, mais cela permettrait d'entrainer des algorithmes de *deep learning* en *open-source* tels que FaceNet à reconnaitre des individus de populations plus vastes.

5) Le potentiel des systèmes automatiques comme aide à l'investigation

Soumis à des tâches de recherches de POI dans de larges BdD, le MFE a montré des performances optimales sur des traces de qualité bonne à moyenne. Le système compare une trace aux 36'392 *templates* de la base de données en quelques secondes, et l'opérateur peut retrouver la POI recherchée en comparant la trace aux 50 premiers candidats retournés par l'interface utilisateur du système. Le MFE est néanmoins un outil commercial coûteux, qui pourra donc être difficilement implémenté aux niveaux régionaux ou cantonaux afin de servir un grand nombre de services.

Les performances de l'algorithme FaceNet dans cette tâche sont moindres, en particulier pour les traces CCTV de qualité moyenne. D'autant plus que les traces CCTV transmises pour analyses en police peuvent être de qualité très dégradée selon le mode de capture et de diffusion, tel que souligné dans le Chapitre II.2.1 (p.14).

Le fait que l'algorithme soit gratuit et en libre accès offre néanmoins un potentiel plus large d'implémentation, pour la recherche de personnes à partir d'images de bonne qualité telles que des documents d'identité ou portraits. Il n'existe cependant pas d'interface utilisateur, ce qui implique d'être installé et utilisé par des personnes dédiées.

Le tableau ci-dessous résume les principaux « pour et contre » ces deux systèmes pour un déploiement opérationnel dans le cadre investigatif.

Tableau 18 : Avantages et inconvénients relevés pour l'utilisation des systèmes MFE et FN dans le cadre investigatif

	MFE	FaceNet
+	Performances sur des images même de faible qualité : liste de 50 candidats à vérifier Encodage des images puis comparaisons très rapides des *templates* Interface utilisateur	Performances sur des images de bonne qualité : liste de 200 candidats à vérifier Comparaisons relativement rapides selon le poids et le nombre d'images Opensource : algorithme modifiable et apprentissage possible sur données spécifiques Gratuit
−	Algorithme propriétaire non-inspectable Prix	Plus faibles performances sur les images de qualité faible à moyenne (voire visages non détectés) Pas d'interface : connaissances en programmation pour la prise en main et l'utilisation

Préconisations opérationnelles

Par ailleurs, nos résultats nous ont permis de mettre en avant plusieurs pratiques nécessaires à une meilleure utilisation opérationnelle de la reconnaissance faciale dans le cadre de l'enquête et en vue d'une expertise :

- La collecte et la sauvegarde d'images dans leur format d'origine, sans compression ni altération postérieure à l'enregistrement, et la transmission de copies de celles-ci
- La prise de photographies de référence (*mugshots*) de l'individu avec et sans ses lunettes de vue
- L'augmentation de la liste de candidats potentiels à vérifier (par exemple, au-delà du rang 50 avec le MFE ou 200 avec FaceNet) lorsque la trace est de mauvaise qualité, ou en cas de vieillissement possible de l'individu recherché entre la prise du *mugshot* et de la trace

a) Algorithmes *rule-based versus deep learning*

Nos résultats montrent l'étendue des progrès apportés par les algorithmes de *deep learning* (MFE et FaceNet) par rapport au fonctionnement *rule-based* du MFI. Il est à noter que nous n'avons testé qu'un seul système basé sur ce fonctionnement et que son développement date de 2014. Il n'est donc pas question ici d'assurer que le fonctionnement *rule-based* devrait être abandonné, mais que le développement de *deep learning* ouvre une marge de progression que les algorithmes « traditionnels » ont déjà atteinte depuis plusieurs années.

b) Systèmes commerciaux *versus* systèmes *open source*

Le Tableau 18 évoque plusieurs des critères sur lesquels se différencient les systèmes commerciaux et *open source* :

- Les systèmes *open source* sont utilisables gratuitement et sont pour la plupart disponibles sur internet. Les systèmes commerciaux sont payants mais leur prix est très variable, selon les outils disponibles par exemple.
- Les systèmes commerciaux sont basés sur des algorithmes propriétaires boîte-noire, c'est-à-dire que leur structure et leur fonctionnement sont inconnus de l'utilisateur et non modifiables. Au contraire, les algorithmes *open source* sont modifiables au besoin, comme c'est le cas dans le cadre de cette recherche[24].
- Les systèmes commerciaux sont développés pour une utilisation simple par des opérateurs pas nécessairement expert.e.s. Cette ergonomie manque aux systèmes open source dont la prise en main et l'utilisation nécessite des connaissances spécifiques, et le développement ultérieur d'une interface graphique si besoin.
- Les systèmes commerciaux sont souvent plus performants car les entreprises qui les développent (par exemple, Google et Idemia) ont accès à des bases de données propriétaires de millions d'images de populations très diverses, qui assure un apprentissage optimal des algorithmes de *deep learning*.

6) Quels sont les apports de l'évaluation probabiliste du score de comparaison dans le cadre judiciaire ? Peut-elle être utilisée dans le cadre de l'investigation comme au tribunal ?

Dans la totalité de nos tests, le calcul de SLR améliore les performances de recherche de POI dans la BdD judiciaire, c'est-à-dire que, lorsque la liste de candidats est triée par SLR au lieu des scores de comparaison, les POI recherchées sont classées à de meilleurs rangs, pour tous les types de traces. Cela n'apporte pas de grandes améliorations si les services de police bénéficient de

[24] Dans sa forme originale, l'algorithme compare toutes les images d'un dossier et décrit les scores dans le terminal de commandes. Notre version adaptée compare les images de deux dossiers (par exemple traces et références) et génère la liste de toutes les comparaisons avec les scores associés dans un fichier .csv)

systèmes aux performances similaires à celles du MFE mais permet une forte amélioration des performances d'outil comme FaceNet qui, de par l'accessibilité de leur code et leur gratuité, offrent une solution simple et assez performante pour aider dans l'investigation.

c) Les avantages du SLR sont-ils suffisants pour une utilisation en phase d'enquête ? Est-il possible de proposer une estimation de SLR d'expertise dès l'investigation ?

Notre étude a permis d'associer un pré-SLR à une valeur attendue de SLR d'expertise sur des cas simulés sous H_1 et sous H_2. Concrètement, nous avons simulé plusieurs cas dans lesquels l'expert.e recevrait 1) une trace et une liste de suspects potentiels provenant d'une base de données (ou d'un témoin oculaire) pendant la phase d'enquête, et 2) les données pertinentes et les propositions définies pour la phase d'instruction. Nos résultats montrent une évolution linéaire des valeurs de SLR entre les phases investigative et d'expertise lorsque le MFE est utilisé sur des traces de bonne qualité (documents d'identité, portraits et images ATM). En utilisant FaceNet en phase investigative puis le MFE en phase d'expertise, les intervalles de chevauchement (c'est-à-dire les valeurs de SLR obtenues à la fois pour des cas sous H_1 et sous H_2) sont logiquement plus étendus, ce qui augmente le nombre de cas potentiels pour lesquels il n'est pas possible d'apporter une estimation du SLR d'expertise aux investigateur.trice.s dès l'enquête.

Fournir un SLR préliminaire aux investigateur.trice.s ainsi que des indications plus concrètes sur le résultat potentiel que donnerait une expertise s'ils orientaient l'enquête vers telle ou telle POI permet de diriger l'enquête avec des informations plus tangibles. Les investigateur.trice.s ont une meilleure idée de ce à quoi s'attendre mais c'est également l'occasion de mieux communiquer sur le type de données de référence, voire d'autres traces, à rechercher par la suite.

IX.3. Axes de recherches futures

- Déploiement de la méthodologie sur des données opérationnelles

La prochaine étape dans la continuité de ce projet est de déployer les modèles probabilistes développés sur des données opérationnelles, c'est-à-dire à partir de traces collectées dans des cas judiciaires récents et représentant diverses populations ethniques. Pour cela, il est nécessaire d'exploiter des traces et références dont les correspondances ont été avérées par l'enquête (sur la base des images ou d'autres traces telles que l'ADN ou les traces papillaires). Cela permettra d'identifier plus particulièrement les limitations de chaque modèle et les perspectives d'améliorations.

Une première série de tests pour la recherche de personnes sur la base des scores générés par FaceNet et le MFE ont été effectués par l'auteure lors d'un mandat pour la Police cantonale neuchâteloise, courant 2021. A ce jour, ces résultats ne sont pas encore publiés ; nous n'en résumerons donc ici que les principales conclusions.

Pour ce faire, 89 traces CCTV de qualité très dégradée (faible résolution originale et/ou dégradation due à la collecte et à la transmission subséquente) ont été mise à disposition, chacune associée à un *mugshot* de référence d'une POI dont la correspondance a été établi par le biais de l'enquête. Ces *mugshots* ont été recherchés dans une BdD de 3849 références. Au vu de la diversité de populations représentées dans la BdD de références, celle-ci a été segmentée en plusieurs sous-populations sur la base de critères connus (la BdD judiciaire étant construite en connaissances de ces informations) de genre (homme, femme), tranche d'âge (juvénile, adulte, sénior) et type ethnique (africain, maghrébin, caucasien).

Avec et sans tri de la BdD, FaceNet ne classe jamais la POI recherchée au rang 1. Cependant, le tri de populations permet d'augmenter le nombre de POI classées dans les 200 premiers rangs de 3 à 32%, et de 23 à 59% au rang 600. Cela implique que dans près de la moitié des recherches, plus de 600 visages sont jugés plus ressemblant à celui de la trace que celui du POI correspondant, ceci sur une BdD de 3645 visages. Avec le MFE, le tri de populations semble avoir un impact moins extrême, mais les performances sur la BdD totale sont déjà largement supérieures à toutes celles de FaceNet. Avec le tri de population, 82% des POI recherchés sont classés dans les 200 premiers candidats, et 94% dans les 600 premiers.

Pour conclure, l'algorithme en source ouverte FaceNet a montré des performances médiocres à partir des traces CCTV fournies dans le cadre de ce mandat. Cependant, le tri de la BdD en sous-population augmente ces performances, et il serait envisageable d'effectuer des tests plus complets par exemple en entraînant l'algorithme sur des données de qualité comparable à celle des traces disponibles, ainsi qu'en optimisant la collecte préalable des traces pour en améliorer la qualité. Le MFE se montre quant à lui très efficace pour la recherche de personnes et présente un fort potentiel d'utilisation dans un contexte opérationnel.

Dans le cadre de l'expertise, la robustesse des méthodologies scientifiques doit être étayée par des données empiriques. Les performances des modèles testés dans ce projet forment une base solide pour soutenir l'utilisation de systèmes automatique de reconnaissance faciale sur des populations caucasiennes.

Pour les populations non couvertes dans le cadre de ce projet, des tests supplémentaires permettront d'adapter les résultats au cas de question. Les modèles de ce projet ont été tes tés à partir de données très variables, ce qui offre un très bon aperçu des performances globales pour les trois scénarios testés, ATM, CCTV et ID. Il est utile de tester les modèles indépendamment pour chaque population de POI, car il est reconnu que les performances de systèmes automatiques de reconnaissance faciale varient, par exemple, selon les groupes ethniques (Cavazos *et al.*, 2019).

Le Washington post (Jouvenal, 2021) a remis en question de la fiabilité des systèmes automatiques à travers le cas du logiciel TrueAllele pour le domaine des analyses ADN. La réflexion s'articule principalement autour du manque de transparence de ce logiciel boîte-noire, face auquel se dressent les magistrats en renouvelant les demandes d'accès au code du système. L'auteur précise que de tels systèmes sont admissibles dans de nombreux tribunaux, notamment grâce à la disponibilité de données de validation publiées, mais dont l'effet boîte-noire continue de freiner l'acceptation de leurs résultats. Malgré cette admissibilité partiellement reconnue, un nouveau rapport du NIST énonce clairement qu'il n'y a toujours « pas assez de données publiées pour évaluer la fiabilité des méthodes » dans ce domaine (Butler *et al.*, 2021). Cela démontre le besoin encore crucial de validation encore après des années d'utilisations de systèmes comme TrueAllele dans le cadre d'expertises ADN.

Cette discussion commence inévitablement à émerger en reconnaissance faciale, avec un premier cas en Europe, au tribunal de Lyon (France), dont nous avons parlé dans la section II.2.2 (p.18). Le système utilisé était alors qualifié de « robot accusateur » par la défense qui soulevait non seulement le manque de transparence mais également l'absence de validation scientifique publiée. Il est nécessaire de multiplier les études indépendantes, basées sur l'utilisation de données empiriques et opérationnelles, en s'inspirant directement des recommandations existantes pour la standardisation des méthodes automatiques pour l'analyse d'ADN, fournies notamment par le *Forensic Science Regulator*, Royaume-Uni (Forensic Science Regulator, 2020) et le NIST, États-Unis (Butler et al., 2021).

- Recherche de solutions au manque de références d'un suspect

À travers ce projet nous avons présenté une méthodologie qui permet de pallier le manque d'images de référence d'un suspect par l'utilisation d'images d'autres individus. Mais il serait utile d'aborder le problème d'un autre angle, non pas en compensant le manque d'images, mais en en récoltant plusieurs en une seule fois. Lorsqu'un délinquant est répertorié dans la base de données judiciaire, prendre plusieurs *mugshots* frontaux à la suite n'est pas recommandé car cela sous-estime considérable la variabilité de l'individu. Il serait intéressant de prendre plusieurs *mugshots* consécutifs mais avec des angles différents (frontal, gauche, droit, haut et bas) afin de modéliser une intravariabilité spécifique basée sur la variabilité faciale selon l'angle de prise de vue.

En outre, l'utilisation d'images de référence prises avec différents angles permettrait également de diminuer les variations de pose entre traces et références. Par exemple, sur des images de type CCTV, les individus sont enregistrés en vue surélevée et avec un angle horizontal variable. Il serait utile de tester l'impact de l'utilisation de références prises sous un angle comparable à celui de la trace sur les performances des systèmes automatiques.

- Entrainement d'algorithmes *open source*

Le fonctionnement du *deep learning* élargit considérablement les perspectives d'amélioration des performances de la reconnaissance faciale. Il est nécessaire d'étudier plus en détail l'impact de la structure des bases de données d'apprentissage sur les performances algorithmiques, comme récemment abordée par Peng (2019). La base de données *Racial Faces in the Wild* (Wang *et al.*, 2019), décrite précédemment, est un outil intéressant pour l'entrainement d'algorithmes dont les performances chutent sur certaines populations ethniques souvent peu représenter dans les bases de données utilisées.

- Population pertinente

La définition de la population pertinente est une problématique récurrente dans tous les domaines forensiques. En reconnaissance faciale, des études ont montré l'importance de l'impact du biais racial sur les comparaisons automatiques, comme chez l'humain (Cavazos *et al.*, 2019). Il faut étudier plus spécifiquement les origines ethniques souvent représentées dans les BdD judiciaires pour être sûr d'utiliser des modèles adéquats pour chaque population. Nous avons montré le très faible impact du choix des populations pertinentes sur nos résultats, mais, comme précisé précédemment, notre recherche est construite autour d'une population caucasienne majoritairement masculine. Nous ne pouvons pas extrapoler l'impact observé à d'autres type de populations, c'est pourquoi cela devrait faire l'objet d'études dédiées.

Il est établi que les systèmes de reconnaissance faciale sont des outils puissants et leurs performances devraient continuer d'augmenter dans les années à venir. Il est essentiel de se focaliser sur les limitations déjà mises en évidence, tel que le biais racial et l'impact de la qualité des images traces, afin d'accroitre la robustesse des résultats présentée au tribunal car les enjeux qui s'y jouent sont trop importants pour les personnes impliquées. D'autant plus que les polémiques autour de la fiabilité des technologies de reconnaissance faciale dans le système judiciaire affluent (Davies et al., 2018 ; Fussey et Murray, 2019).

Chapitre X. Conclusion

Avec le développement croissant des algorithmes automatiques et de l'intelligence artificielle, la reconnaissance faciale occupe une place toujours plus importante dans les contextes judiciaires et civils. Cependant, les bénéfices en matière de prévention, de renseignement, d'investigation et d'évaluation sont accompagnés d'enjeux et de limitations importants. La reconnaissance faciale doit encore faire l'objet d'études empiriques approfondies pour apporter un cadre scientifique et juridique adéquat dans le cadre investigatif et au tribunal.

Ce constat fixe le fondement de la présente recherche. Nous nous focalisons sur les enjeux liés aux images et à l'utilisation de systèmes automatiques. Pour ce faire, notre objectif a été la validation de modèles de calcul de rapport de vraisemblance (*likelihood ratio*, LR), permettant d'interpréter les scores de comparaison issus de systèmes automatiques de reconnaissance faciale. L'utilisation du LR est préconisée dans de nombreux domaines forensiques, dont ceux des traces digitales, de la reconnaissance vocale, des armes à feu, etc. Plus spécifiquement, le calcul de LR à partir des scores de comparaison générés par des systèmes biométriques – le SLR - a déjà été décrit notamment pour la voix et les traces digitales, avec l'utilisation du système AFIS.

Pour ce faire, nous avons constitué un jeu de données, pour 34 individus, représentatif de cas opérationnels tels que la fraude aux documents d'identité et le retrait frauduleux de billets. Trois types de traces ont donc été récoltées : des images signalétiques et des images de surveillance enregistrées par des distributeurs automatiques de billets ainsi que par une caméra CCTV grand-angle. Les images de référence de type *mugshots* sont prises de manière à correspondre aux images signalétiques collectées et fournies par les polices neuchâteloise et vaudoise dans le cadre du projet. Toutes les comparaisons automatiques sont faites à l'aide de trois systèmes : deux générations du système commercial IDEMIA Morphoface (MFI et MFE) et l'algorithme en source ouverte FaceNet. Ce choix permet de comparer les performances d'algorithmes basés sur des architectures différentes, propriétaires ou publics.

Deux problématiques principales ont été couvertes pour le développement des modèles de calcul de SLR à partir de ces jeux de données : les approches de modélisations d'intravariabilité et d'intervariabilité et la calibration du SLR. La modélisation des variabilités, qui conditionnent respectivement le numérateur et le dénominateur du SLR, doit être adaptée au scénario et aux données disponibles.

Notre premier constat est que le calcul de SLR améliore les performances de recherche de POI dans la BdD judiciaire, c'est-à-dire que, lorsque la liste de candidats est triée par SLR au lieu des scores de comparaison, les POI recherchées sont classées à de meilleurs rangs, pour tous les types de traces.

Dans le cadre de l'enquête, Le SLR permet de fournir aux enquêteurs un résultat pondéré à partir des seules données disponibles à cette étape, c'est-à-dire la trace du cas de question et une référence pour chaque POI. Pour cette évaluation investigative, le calcul du SLR est donc basé sur la modélisation d'une intravariabilité générique afin de pallier le manque d'images de référence. L'intervariabilité quant à elle peut être spécifique à la trace ou à l'image de référence.

Nos résultats démontrent que les taux d'erreurs du MFI sont incompatibles avec son intégration dans le processus judiciaire. Les performances du MFE en font un outil de choix pour une implémentation comme aide à l'enquête, à la fois par ses performances et la rapidité de l'algorithme de comparaison. Cependant, le coût d'un tel produit peut être un frein à son acquisition, en particulier par des petits services régionaux/cantonaux de police. C'est pourquoi l'outil FaceNet, disponible en source ouverte, offre une alternative intéressante. Néanmoins, les performances moindres et la structure très brute de l'algorithme FaceNet amènent des contraintes supplémentaires pour une utilisation par les services de police.

En phase d'expertise, le nombre suffisant de données de référence de la POI permet de calculer un SLR basé sur une intravariabilité spécifique. Cette approche permet à l'expert.e de se prononcer sur des propositions incluant directement l'individu mis en cause (H_1 : « La POI est la personne sur l'image trace »), contrairement à l'approche générique. Nos résultats privilégient des modèles combinant l'intervariabilité trace-spécifique et calibration par PAVA ou régression logistique à partir des scores du MFE. Les performances de FaceNet sur des traces type CCTV sont insuffisantes pour une utilisation dans le cadre du tribunal. Pour les traces de meilleure qualité, les modèles alliant intervariabilité trace-spécifique et calibration (PAVA ou régression logistique) sont les plus performants, mais moins robustes qu'avec le MFE.

Enfin, une problématique importante développée dans notre recherche concerne l'estimation de SLR d'expertise à partir des SLR préliminaires calculés en phase d'enquête. Notre étude a permis d'associer un pré-SLR à une valeur attendue de SLR d'expertise. Nos résultats montrent une évolution linéaire des valeurs de SLR entre les phases d'enquête et d'expertise lorsque le MFE est utilisé sur des traces de bonne qualité (documents d'identité, portraits et images ATM). En revanche, avec FaceNet, les intervalles de chevauchement (c'est-à-dire les valeurs de SLR obtenues à la fois pour des cas sous H_1 et sous H_2) sont trop étendus, ce qui ne permet pas de l'utiliser comme outil pertinent pour le calcul de SLR en phase d'enquête.

Bibliographie

Frye v. United States, 293 F 1013, 1923.

R. v. Turner, 1 All ER 70, 1975.

R. v. Mohan, 2 S.C.R. 9, 1992.

Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579, 1993.

Code de procédure pénale suisse (CPP 312.0), 2007.

702 Federal Rules of Evidence – Testimony by Expert Witnesses, 2020.

Aitken C.G. et Taroni F., *Statistics and the evaluation of evidence for forensic scientists*, 2nd ed. (Statistics in Practice), Chichester: John Wiley & Sons, Ltd., 2004.

Ali T., "Biometric Score Calibration for Forensic Face Recognition", PhD thesis, PhD thesis, University of Twente, The Netherlands, 2014.

Ali T., Spreeuwers L. et Veldhuis R., "A review of calibration methods for biometric systems in forensic applications", *33rd WIC Symposium on Information Theory in the Benelux*, 2012a.

Ali T., Spreeuwers L., Veldhuis R. et Meuwly D., "Effect of calibration data on forensic likelihood ratio from a face recognition system", *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1-8, 2013.

Ali T., Veldhuis R. et Spreeuwers L., "Forensic face recognition: A survey", in *Face Recognition: Methods, Applications and Technology*, Quaglia A. et Epifano C.M. (Eds.) Computer Science, Technology and Applications, Nova Publishers, pp. 9-28, 2012b.

Baechler S. *et al.*, "Breaking the barriers between intelligence, investigation and evaluation: A continuous approach to define the contribution and scope of forensic science", *Forensic Science International,* 309, p. 110213, 2020.

Bertillon A., "De l'identification par les signalements anthropométriques", *Archives de l'anthropologie criminelle et des sciences pénales,* 1, pp. 193-223, 1886.

Bertillon A., *Identification anthropométrique - Instructions signalétiques*, Melun: Imprimerie Administrative, p. 148, 1893.

Bolck A., Ni H. et Lopatka M., "Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison", *Law, Probability and Risk,* 14(3), pp. 243-266, 2015.

Bollé T., Casey E. et Jacquet M., "The role of evaluations in reaching decisions using automated systems supporting forensic analysis", *Forensic Science International: Digital Investigation,* 34, p. 301016, 2020.

Botti F., Alexander A. et Drygajlo A., "An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data", *ODYSSEY04 - The Speaker and Language Recognition Workshop Toledo, Spain*, pp. 63-68, 2004.

Bromberg D.E., Charbonneau É. et Smith A., "Public support for facial recognition via police body-worn cameras: Findings from a list experiment", *Government Information Quarterly,* 37(1), p. 101415, 2020.

Brümmer N., "Measuring, refining and calibrating speaker and language information extracted from speech", PhD Thesis, Department of Electrical and Electronic Engineering, University of Stellenbosch, South Africa, Matieland, 2010.

Brümmer N. et du Preez J., "Application-independent evaluation of speaker detection", *Computer Speech & Language,* 20(2-3), pp. 230-275, 2006.

Burton A.M., Wilson S., Cowan M. et Bruce V., "Face Recognition in Poor-Quality Video: Evidence From Security Surveillance", *Psychological Science,* 10(3), pp. 243-248, 1999.

Butler J.M., Iyer H., Press R., Taylor M.K., Vallone P.M. et Willis S., "(NISTIR 8351-DRAFT) DNA Mixture Interpretation: A NIST Scientific Foundation Review", 2021. Available: https://doi.org/10.6028/NIST.IR.8351-draft [last access: 09.08.2021]

Cagle M. et Ozer N., *Amazon teams up with Government to deploy dangerous new facial recognition technology*, 2018. Available: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazon-teams-government-deploy-dangerous-new [last access: 28.04.2020]

Casey E., "Standardization of forming and expressing preliminary evaluative opinions on digital evidence", *Forensic Science International: Digital Investigation,* 32, p. 200888, 2020.

Castelluccia C. et Le Métayer D., "Analyse des impacts de la reconnaissance faciale Quelques éléments de méthode", Inria Grenoble Rhône-Alpes, Rapport de recherche, 2019. [last access: 02.08.2021]

Cavazos J.G., Phillips P.J., Castillo C.D. et O'Toole A.J., "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?", *ArXiv,* arXiv:1912.07398, 2019.

Champod C. et Evett I.W., "Evidence Interpretation: a Logical Approach", in *Wiley Encyclopedia of Forensic Science*, vol. 2, Moenssens A. et Jamieson A. (Eds.), Chichester, UK: John Wiley & Sons, pp. 968- 976, 2009.

Clarke R.V. et Eck J.E., "Crime Analysis for Problem Solvers In 60 Small Steps", U.S. Department of Justice Office of Community Oriented Policing Services, Center for Problem Oriented Policing, 2005. [last access: 02.08.2021]

CNIL, *Expérimentation de la reconnaissance faciale dans deux lycées : la CNIL précise sa position*, 2019. Available: https://www.cnil.fr/fr/experimentation-de-la-reconnaissance-faciale-dans-deux-lycees-la-cnil-precise-sa-position [last access: 2021.02.22]

Cole S.A. et Dioso-Villa R., "CSI and Its Effects: Media, Juries, and the Burden of Proof", *New England Law Review,* 41, pp. 435-469, 2007.

Davies B., Innes M. et Dawson A., Universities' Police Science Institute, Crime and Security Research Institute, Cardiff University, "An Evaluation of South Wales Police's Use of Automated Facial Recognition", Cardiff, 2018. [last access: 02.08.2021]

DeGroot M.H. et Fienberg S.E., "The Comparison and Evaluation of Forecasters", *The Statistician,* 32, pp. 12-22, 1983.

Delignette-Muller M.-L., Dutang C., Pouillot R., Denis J.-B. et Siberchicot A., "fitdistrplus: An R Package for Fitting Distributions", *Journal of Statistical Software,* 64(4), pp. 1-34, 2015.

Dessimoz D. et Champod C., "A dedicated framework for weak biometrics in forensic science for investigation and intelligence purposes: The case of facial information", *Security Journal,* 29(4), pp. 603-617, 2015.

Dror I.E., "Biases in forensic experts", *Science,* 360(6386), p. 243, 2018.

Drygajlo A., Jessen M., Gfroerer S., Wagner I., Vermeulen J. et Niemi T., *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*, Frankfurt: Verlag für Polizeiwissenshaften, 2016.

Dupont B., Stevens Y., Westermann H. et Joyce M., "Artificial Intelligence in the Context of Crime and Criminal Justice", Canada Research Chair in Cybersecurity, ICCC, Université de Montréal, Canada, Report for the Korean Institute of Criminology, 2018. Available: https://ajcact.openum.ca/en/publications/artificial-intelligence-in-the-context-of-crime-and-criminal-justice/ [last access: 22.02.2021]

Dutta A., "Predicting performance of a face recognition system based on image quality", PhD thesis, University of Twente, Netherlands, 2015.

Edmond G., Biber K., Kemp R. et Porter G., "Law's Looking Glass: Expert.e Identification Evidence Derived from Photographic and Video Images", *Current Issues in Criminal Justice,* 20(3), pp. 337-377, 2009.

Egli N., "Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System", PhD Thesis, Université de Lausanne, Ecole des Sciences Criminelles, Suisse, 2009.

Emami C., Brown D.R. et Smith D.R.G., "Use and acceptance of biometric technologies among victims of identity crime and misuse in Australia", *Trends & issues in crime and criminal justice,*(511), p. 6, 2016.

ENFSI, "ENFSI guideline for evaluative reporting in forensic science - Strengthening the evaluation of forensic results across Europe (STEOFRAE)", 2015. [last access:

ENFSI, European Network of Forensic Science Institutes, "Best practice manual for facial image comparison (ENFSI-BPM-DI-01)", 2018. Available: http://enfsi.eu/documents/best-practice-manuals/ [last access: 02.08.2021]

Evett I.W., Jackson G., Lambert J.A. et McCrossan S., "The Impact of the Principles of Evidence Interpretation on the Structure and Content of Statements", *Science & Justice,* 40(4), pp. 233-239, 2000.

Farrington D.P., Gill M., Waples S.J. et Argomaniz J., "The effects of closed-circuit television on crime: meta-analysis of an English national quasi-experimental multi-site evaluation", *Journal of Experimental Criminology,* 3(1), pp. 21-38, 2007.

FISWG, Facial Identification Scientific Working Group, "Guidelines for Facial Comparison Methods", 2012. Available: https://fiswg.org/documents.html [last access: 02.08.2021]

Forensic Science Regulator, Home Office Forensic Science Regulator, "Expert Report Guidance", Brimingham, issue Issue 3, FSR-G-200, 2019. Available: https://www.gov.uk/government/publications/expert-report-content-issue-3 [last access: 02.08.2021]

Forensic Science Regulator, Home Office Forensic Science Regulator, "Guidance: Validation", Birmingham, 2020. Available: https://www.gov.uk/government/publications/forensic-science-providers-validation [last access: 02.08.2021]

Furl N., Phillips P.J. et O'Toole A.J., "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis", *Cognitive Science,* 26(6), pp. 797-815, 2002.

Fussey P. et Murray D., Human Right Center, University of Essex, United Kingdom, "Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology", The Human Rights, Big Data and Technology Project, 2019. [last access: 22.02.2021]

Galbally J., Ferrara P., Haraksim R., Psyllos A. et Beslay L., "Study on Face Identification Technology for its Implementation in the Schengen Information System", 2019. [last access: 02.08.2021]

Gilleland E. et Katz R., "extRemes 2.0: An Extreme Value Analysis Package in R", *Journal of Statistical Software,* 72(8), pp. 1-39, 2016.

Goodfellow I., Bengio Y. et Courville A., *Deep Learning*, Cambridge: MIT Press, 2016.

Grother P., Ngan M. et Hanaoka K., "Face Recognition Vendor Test (FRVT) Part 2: Identification", National Institute of Standards and Technology - NISTIR 8238, 2019a. [last access: 02.08.2021]

Grother P., Ngan M. et Hanaoka K., "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects", 2019b. [last access: 02.08.2021]

Grother P., Ngan M. et Hanaoka K., "Supplement of the Face Recognition Vendor Test (FRVT) Part 2: Identification", National Institute of Standards and Technology - NISTIR 8238, 2021. [last access: 02.08.2021]

Grother P.J. et Ngan M.L., National Institute for Standards and Technology, "Face Recognition Vendor Test (FRVT) – Performance of Face Identification Algorithms, NIST interagency report 8009", Gaithersburgh, 2014. Available: http://biometrics.nist.gov/cs_links/face/frvt/frvt2013/NIST_8009.pdf [last access: 02.08.2021]

Hane T., "L'intelligence économique au service de la lutte contre le blanchiment de capitaux et le financement du terrorisme", PhD thesis, Droit, Université de Strasbourg, France, 2015.

Hepler A.B., Saunders C.P., Davis L.J. et Buscaglia J., "Score-based likelihood ratios for handwriting evidence", *Forensic Science International,* 219(1-3), pp. 129-140, 2012.

Huang G.B., Ramesh M., Berg T. et Learned-Miller E., "Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments", presented at the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Marseille, France, 2008.

Imwinkelried E.J., "Computer source code: A source of the growing controversy over the reliability of automated forensic techniques", *66 DePaul Law Review,* 66(1), pp. 97-132, 2017.

Institiute B., Biometrics Institiute, "Should we ban facial recognition?", 2021. Available: https://www.biometricsinstitute.org/should-we-ban-facial-recognition/ [last access: 02.08.2021]

Jackson G., Jones S., Booth G., Champod C. et Evett I.W., "The nature of forensic science opinion - a possible framework to guide thinking and practicce in investigation and in court proceedings", *Science & Justice,* 46(1), pp. 33-44, 2006.

Jacquet M. et Champod C., "Automated face recognition in forensic science: Review and perspectives", *Forensic Science International,* 307, p. 110124, 2020.

Jacquet M. et Grossrieder L., "Enjeux et perspectives de la reconnaissance faciale en sciences criminelles", *Criminologie,* 54(1), pp. 135-170, 2021.

Jain A.K., Klare B. et Park U., "Face matching and retrieval in forensics applications", *IEEE Multimedia,* 19(1), pp. 2-10, 2012.

Jendly M., *Prévenir la criminalité : oui... Mais comment ?* (La question), Switzerland: Grolley, 2013.

Jouvenal J., "A secret algorithm is transforming DNA evidence. This defendant could be the first to scrutinize it.", *Washington Post*, Washington DC, 2021.

Kemelmacher-Shlizerman I., Seitz S., Miller D. et Brossard E., "The MegaFace benchmark: 1 million faces for recognition at scale", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4873-4882, 2016.

Kemp R., Towell N. et Pike G., "When Seeing should not be Believing: Photographs, Credit Cards and Fraud", *Applied Cognitive Psychology,* 11, pp. 211-222, 1997.

Lee W.-L., Wilkinson C., Memon A. et Houston K., "Matching unfamiliar faces from poor quality closed-circuit television (CCTV) footage: An evaluation of the effect of training on facial identification ability", *AXIS Online Jounal of Centre for Anatomy and Human Identification,* 1(1), pp. 19-28, 2009.

Lindley D.V., Tversky A. et Brown R.V., "On the Reconciliation of Probability Assessments", *Journal of the Royal Statistical Society. Series A,* 142(2), pp. 146-180, 1979.

Macarulla Rodriguez A., Geradts Z. et Worring M., "Validation of Score-based Likelihood Ratio Estimation for Automated Face Recognition", *20th Irish Machine Vision and Image Processing conference*, Belfast, Northern Ireland, pp. 145-153, 2018.

Margagliotti G. et Bollé T., "Machine learning & forensic science", *Forensic Science International,* 298, pp. 138-139, 2019.

Marzouki M., "Enjeux des techniques de biométrie - Une première approche", *Troisième Conférence internationale des commissaires à la protection des données*, Paris, France, 2001.

Meissner C.A., Brigham J.C. et Butz D.A., "Memory for Own- and Other-Race Faces: a Dual-Process Approach", *Applied Cognitive Psychology,* 19(5), pp. 545-567, 2005.

Meuwly D., "Reconnaissance de locuteurs en sciences forensiques: L'apport d'une approche automatique", PhD Thesis, Université de Lausanne, Ecole des Sciences Criminelles, Suisse, 2001.

Meuwly D., "Forensic Individualisation from Biometric Data", *Science & Justice,* 46(4), pp. 205-213, 2006.

Meuwly D., Ramos D. et Haraksim R., "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation", *Forensic Science International,* 276, pp. 142-153, 2017.

Michalski D. et Snyder G., Defence Science and Technology Group, Australia, "Facial Image Comparison Test 2019 - Results of the ENFSI-DIWG FIC Test", 2019. [last access: 12.02.2020]

Miles F., NYPD uses facial recognition to arrest brazen sex offender accused of attempted rape on subway platform, Fox News, 2020. Available: https://www.foxnews.com/us/nypd-uses-facial-recognition-to-arrest-brazen-sex-offender-accused-of-attempted-rape-on-subway-platform [last access: 02.09.2020]

Millard S.P., *EnvStats: An R Package for Environmental Statistics*: Springer, New York, 2013.

Milliet Q., "Développement d'une méthodologie d'exploitation des images témoins en science forensique", PhD thesis, Ecole des Sciences Criminelles, Université de Lausanne, Lausanne, Suisse, 2017.

Milliet Q., Delemont O. et Margot P., "A forensic science perspective on the role of images in crime investigation and reconstruction", *Science & Justice,* 54(6), pp. 470-80, 2014.

Moreton R. et Morley J., "Investigation into the use of photoanthropometry in facial image comparison", *Forensic Sci Int,* 212(1-3), pp. 231-7, 2011.

Morrison G.S., "Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio", *Australian Journal of Forensic Sciences,* 45(2), pp. 173-197, 2013.

Morrison G.S., "Calculation of forensic likelihood ratios: use of Monte Carlo simulations to compare the output of scrore-based approaches with true likelihood-ratio values", ArXiv, 2016. Available: http://geoff-morrison.net [last access: 02.08.2021]

Morrison G.S. *et al.*, "Consensus on validation of forensic voice comparison", *Science & Justice,* 61(3), pp. 299-309, 2021.

Neumann C. *et al.*, "Computation of likelihood ratios in fingerprint identification for configurations of three minutiae", *Journal of Forensic Sciences,* 51(6), pp. 1255-1266, 2006.

Neumann C., Champod C., Yoo M., Genessay T. et Langenburg G., "Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks", *Forensic Science International,* 248, pp. 154-171, 2015.

Neumann C. et Margot P., "New perspectives in the use of ink evidence in forensic science: Part III: Operational applications and evaluation", *Forensic Science International,* 192(1-3), pp. 29-42, 2009.

Ngan M., Grother P. et Hanaoka K., "Face Recognition Vendor Test (FRVT) - Part 6B: Face recognition accuracy with face masks using post-COVID-19 algorithms", National Institute of Standards and Technology - NISTIR 8231, 2020. [last access: 02.08.2021]

Noyes E. et O'Toole A.J., "Face recognition assessments used in the study of super-recognisers", *ArXiv,* arXiv:1705.04739, 2017.

Noyes E., Phillips P.J. et O'Toole A., "What is a super-recogniser?", in *Face processing: Systems, Disorders, and Cultural Differences*, Bindemann M. et Megreya A.M. (Eds.), New York: Nova Science Publishers Inc, pp. 173-201, 2017.

Pasquale F., *The Black Box Society*: Harvard University Press, 2015.

Peng Y., "Face recognition at a distance: Low-resolution and alignment problems", PhD Thesis, Digital Society Institute, PhD thesis, University of Twente, Enschede, The Netherlands, 2019.

Phillips P.J. *et al.*, "Overview of the Face Recognition Grand Challenge", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 947-954, 2005.

Phillips P.J. et O'Toole A.J., "The great debate: study proves whether people or algorithms are best at facial ID", *Biometric Technology Today,* 2018(9), pp. 5-8, 2018.

Phillips P.J. *et al.*, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms", *Proc Natl Acad Sci U S A,* 115(24), pp. 6171-6176, 2018.

Piquerez G. et Macaluso A., *Procédure pénale suisse - Manuel, 3e Edition*: Schulthess Verlag, 2011.

Rainie L. et Duggan M., Pew Research Center, "Privacy and Information Sharing", Washington DC, 2016, 2016. [last access:

Ramos-Castro D., "Forensic evaluation of the evidence using automatic speaker recognition systems", PhD Thesis, Universidad Autónoma de Madrid, Escuela Politécnica Superior, España, 2007.

Ramos D., Gonzales-Rodriguez J., Zadora G., Zieba-Palus J. et Aitken C., "Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation", *Third International Symposium on Information Assurance and Security*, Manchester, UK, pp. 411-416: IEEE-CS Press, 2007.

Ramos D. et Gonzalez-Rodriguez J., "Reliable support: Measuring calibration of likelihood ratios", *Forensic Science International,* 230(1-3), pp. 156-169, 2013.

Ramos D., Maroñas-Molano J. et Lozano-Diez A., "Bayesian Strategies for Likelihood Ratio Computation in Forensic Voice Comparison with Automatic Systems", *Subsidia 2017: Tools and Resources for Speech Sciences*, Malaga, Spain, 2018.

Ribaux O., *Police scientifique: le renseignement par la trace*, Lausanne: PPUR Presses polytechniques et Universitaires Romandes, 2014.

Ribaux O., Genessay T. et Margot P., "Les processus de veille opérationnelle et science forensique", in *Sphères de surveillance*, Leman-Langlois S. Ed., Montréal: Les Presses de l'Université de Montréal, pp. 137-158, 2011.

Robertson B., Vignaux G.A. et Berger C.E.H., *Interpreting Evidence – Evaluating Forensic Science in the Courtroom*, 2nd ed., Chichester: John Wiley & Sons, Ltd, 2016.

Rossy Q., Ioset S., Dessimoz D. et Ribaux O., "Integrating forensic information in a crime intelligence database", *Forensic Science International,* 230(1-3), pp. 137-146, 2013.

Ryser E., Spichiger H. et Casey E., "Structured decision making in investigations involving digital and multimedia evidence", *Forensic Science International: Digital Investigation,* 34, 2020a.

Ryser E., Spichiger H. et Casey E., "Structured decision making in investigations involving digital and multimedia evidence", *Forensic Science International: Digital Investigation,* 34, p. 301015, 2020b.

Samie Foucart L., "Évaluation des résultats ADN considérant des propositions au niveau de l'activité", PhD thesis, Ecole des Sciences Criminelles, Université de Lausanne, Suisse, 2019.

Schroff F., Kalenichenko D. et Philbin J., "FaceNet: A unified embedding for face recognition and clustering", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 815-823, 2015a.

Schroff F., Kalenichenko D. et Philbin J., "FaceNet: A unified embedding for face recognition and clustering", *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815-823, 2015b.

Smith S.M., Stinson V. et Patry M.W., "Fact or Fiction? The Myth and Reality of the CSI Effect", *Court Review,* 100(47), pp. 4-7, 2011.

Spaun N., "Facial comparisons by subject matter experts: Their role in biometrics and their training", in *Advances in Biometrics*, Tistarelli M. et Nixon M.S. (Eds.), Springer, Berlin, Heidelberg, pp. 161-168, 2009.

Stoney D.A., De Donno M., Champod C., Wertheim P.A. et Stoney P.L., "Occurrence and associative value of non-identifiable fingermarks", *Forensic Science International,* 309, p. 110219, 2020.

Swofford H.J. et Champod C., "Implementation of Algorithms in Pattern & Impression Evidence: A Responsible and Practical Roadmap", *Forensic Science International: Synergy,* 3, p. 100142, 2021.

TELEFI Project, "Summary Report of the project “Towards the European Level Exchange of Facial Images”", Towards the European Level Exchange of Facial Images, 2021. [last access:

Towler A. *et al.*, "Do professional facial image comparison training courses work?", *PLoS One,* 14(2), p. e0211037, 2019.

Towler A., Kemp R.I. et White D., "Unfamiliar Face Matching Systems in Applied Settings", *Face Processing: Systems, Disorders and Cultural Difficulties*, pp. 21-40: Nova Science, 2017.

Towler A., White D. et Kemp R.I., "Evaluating training methods for facial image comparison: the face shape strategy does not work", *Perception,* 43(2-3), pp. 214-8, 2014.

Turing A.M., "On computable numbers, with an application to the entscheidungsproblem", *Proceedings of the London Mathematical Society*, pp. 230-265, 1936.

Unisys, Unisys Australia, "Unisys Security Index Report Australia - Biometrics in Airports", Sydney, Australia, 2014. Available: https://www.unisys.com/unisys-security-index/australia [last access: 02.08.2021]

Wang M., Deng W., Hu J., Tao X. et Huang Y., "Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network", *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

Welsh B.C. et Farrington D.P., "Public Area CCTV and Crime Prevention: An Updated Systematic Review and Meta-Analysis", *Justice Quarterly,* 26(4), pp. 716-745, 2009.

White D., Kemp R.I., Jenkins R. et Burton A.M., "Feedback training for facial image comparison", *Psychon Bull Rev,* 21(1), pp. 100-6, 2014a.

White D., Kemp R.I., Jenkins R., Matheson M. et Burton A.M., "Passport Officers' Errors in Face Matching", *PLoS ONE,* 9(8), 2014b.

Woodward J.D., RAND Corporation, "Super Bowl Surveillance: Facing Up to Biometrics", Santa Monica, CA, 2001. Available: https://www.rand.org/pubs/issue_papers/IP209.html [last access: 22.02.2021]

Zadora G., Martyna A., Ramos D. et Aitken C., *Statistical analysis in forensic science: Evidential value of multivariate physicochemical data*, Chichester: John Wiley & Sons, Ltd, 2014.

Zeinstra C.G., Meuwly D., Ruifrok A.C.C., Veldhuis R.N.J. et Spreeuwers L.J., "Forensic Face Recognition as a Means to Determine Strength of Evidence: A Survey", *Forensic Science Review,* 30, pp. 23-34, 2018.

Zeng D., Veldhuis R.N.J. et Spreeuwers L.J., "A survey of face recognition techniques under occlusion", *ArXiv,* arXiv:2006.11366, 2020.

Annexes

Annexe A. Tables des illustrations et tableaux

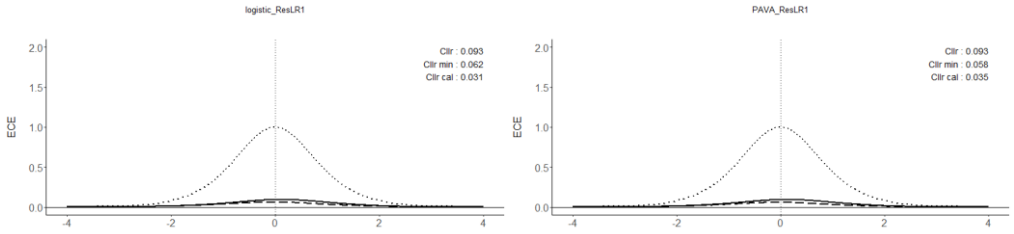
Annexe B. Données : Détails des images récoltées

		Traces			Référence	Nombre de comparaisons		
	POI	Nombre de traces ID (doc identité)	Nombre de traces ID (portraits)	Nombre traces CCTV/ATM	Nombre mugshots	sous H1 (ID) (= Nb traces * nb mugshots)	Sous H1 (cctv/atm) (= Nb traces * nb mugshots)	Intravariabilité spécifique (= Nb mugshots * (nb mugshots-1))
Femmes	A	6	4	30	8	80	240	90
	AA	1	4	15	-	-	-	20
	AA'	5	-	31	-	-	-	20
	C	4	8	20	8	96	160	132
	DD'	1	2	15	-	-	-	6
	EE'	4	1	20	-	-	-	20
	F	6	5	25	8	88	200	110
	G	1	1	15	4	8	60	2
	HH'	4	3	20	-	-	-	42
Hommes	B	3	2	30	8	40	240	20
	D	5	1	15	4	24	60	30
	E	5	1	30	6	36	180	30
	H	4	6	30	8	80	240	90
	I	3	7	30	6	60	180	90
	J	-	10	30	6	60	180	90
	K	5	9	30	6	84	180	182
	LL'	3	4	-	4	28	-	42
	L	4	8	15	4	48	60	132
	M	3	10	15	4	52	60	156
	N	3	4	20	4	28	80	42
	O	4	12	15	4	64	60	240
	PP'	4	3	-	4	28	-	42
	P	2	2	15	4	16	60	12
	Q	4	2	30	8	48	240	30
	R	4	-	15	4	-	60	12
	S	5	3	30	8	64	240	56
	T	2	2	15	4	16	60	12
	U	5	5	15	4	40	60	90
	V	4	1	30	8	40	240	20
	W	5	13	30	8	144	240	306
	X	4	3	15	4	28	60	42
	Y	5	1	30	8	48	240	30
	Z	11	4	30	8	120	240	210
	ZZ'	-	-	15	4	-	60	-
	Total	***129***	***141***	***721***	***168***	***1468***	***3980***	***2448***
	Moyenne	***4***	***5***	***23***	***6***	***54***	***147***	***74***
	Ecart type	***2***	***4***	***7***	***2***	***33***	***82***	***74***

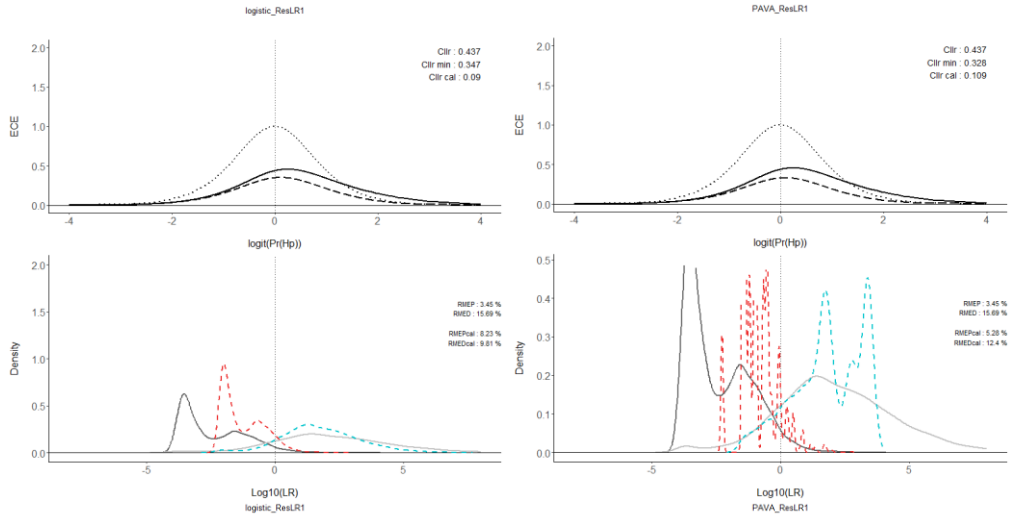
Annexe C. Résultats : Représentations graphiques des performances de tous les modèles choisis pour chaque système et scénario

FaceNet - Intravariabilité Spécifique

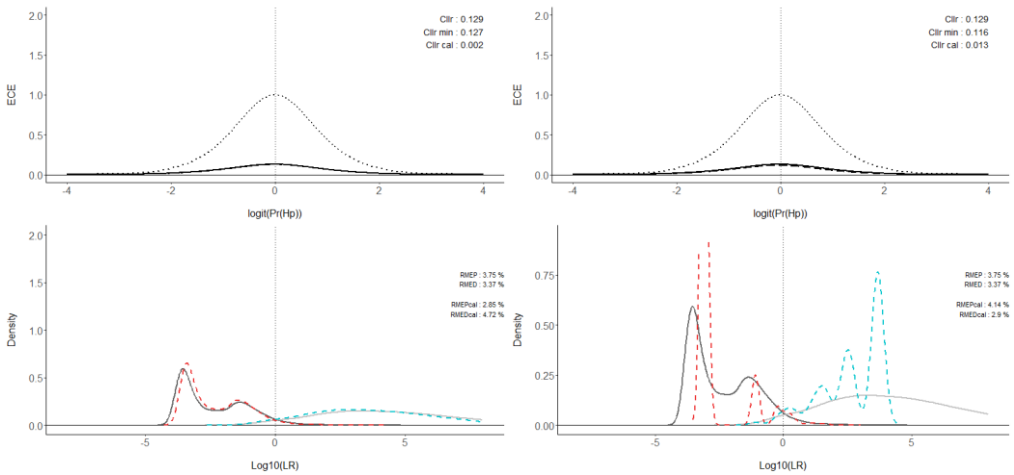
ATM



CCTV

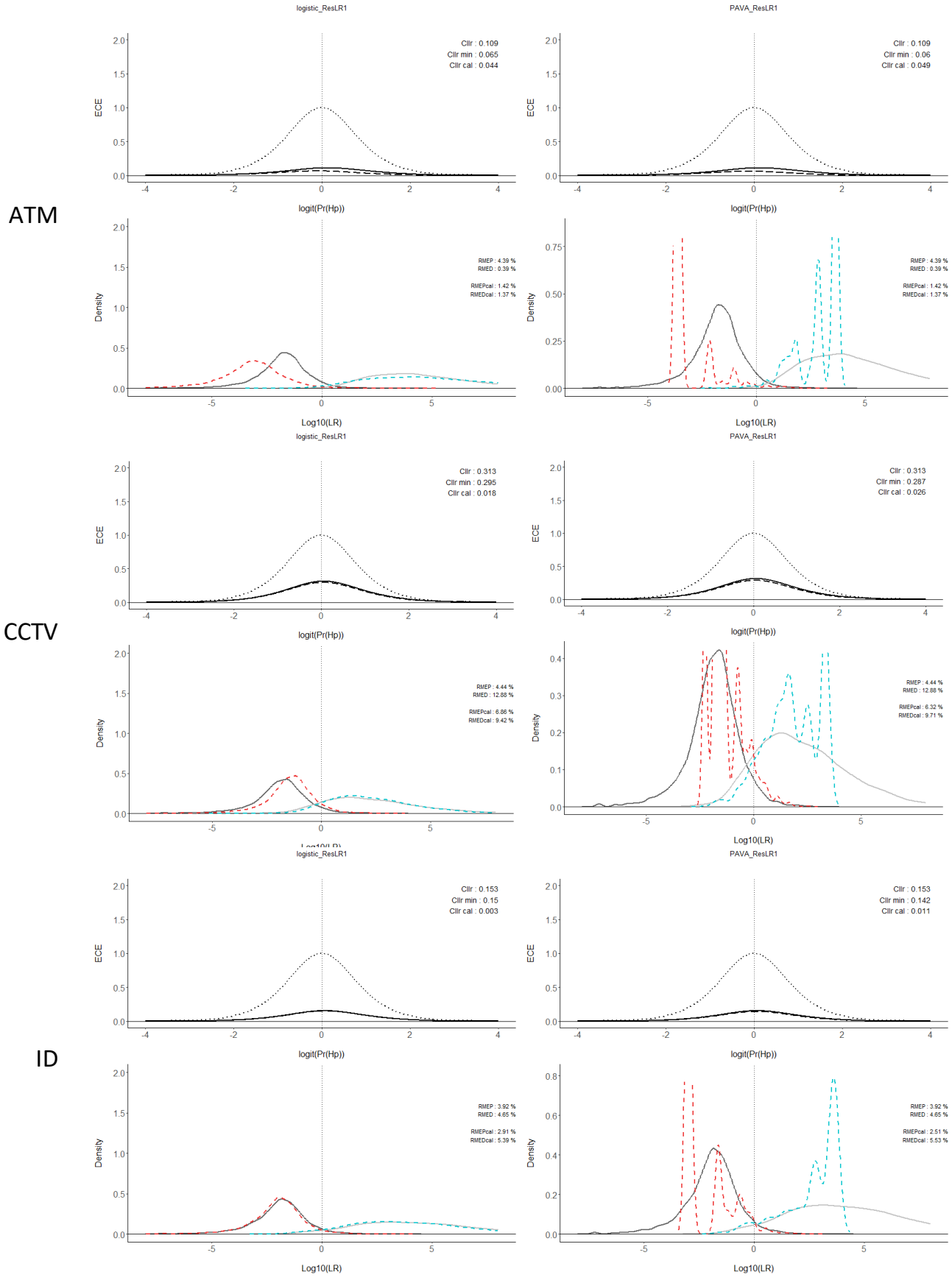


ID

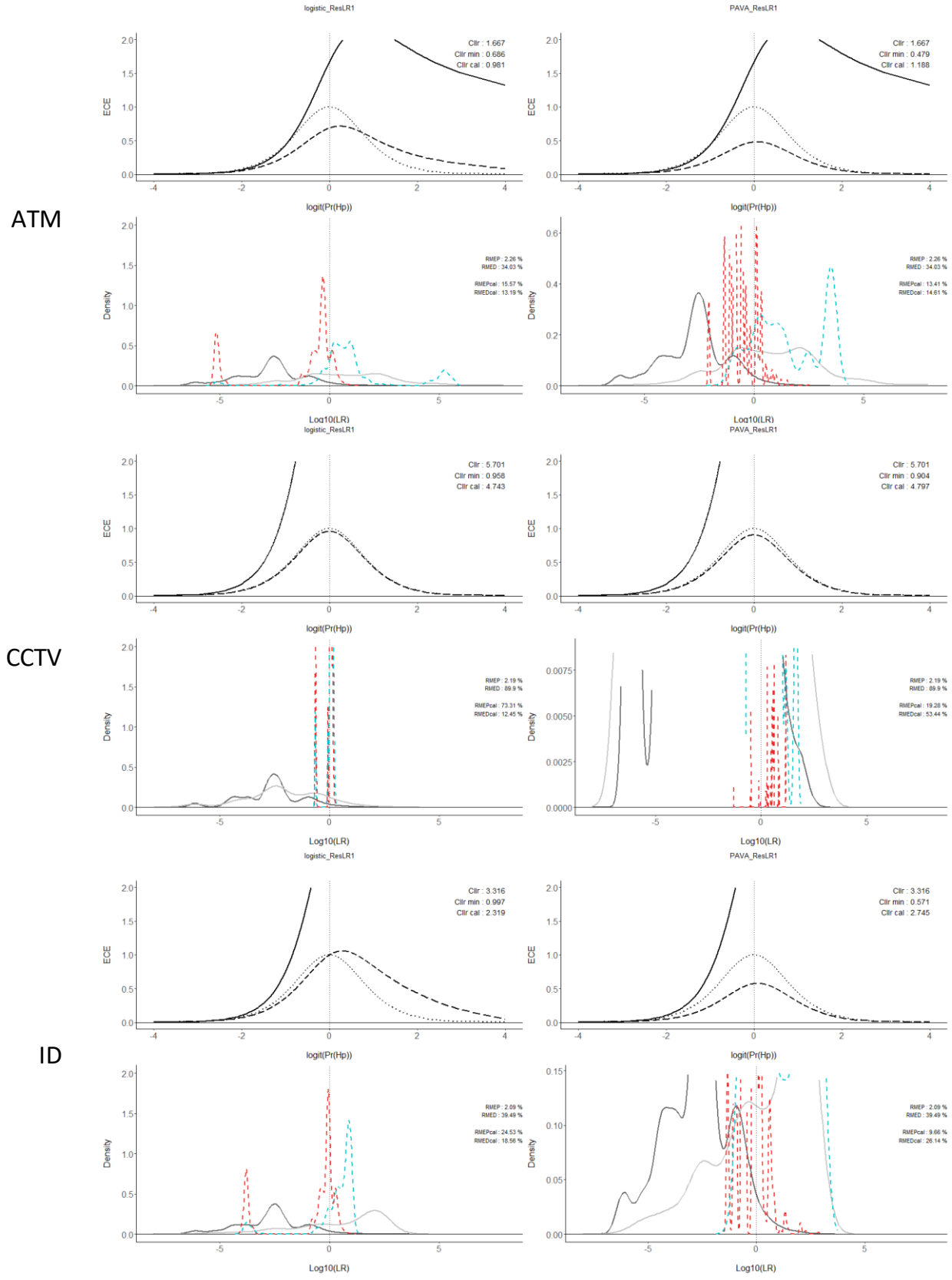


Régression Logistique PAVA

FaceNet - Intravariabilité Générique



MFI - Intravariabilité Spécifique / intervariabilité Suspect-Spécifique



ATM

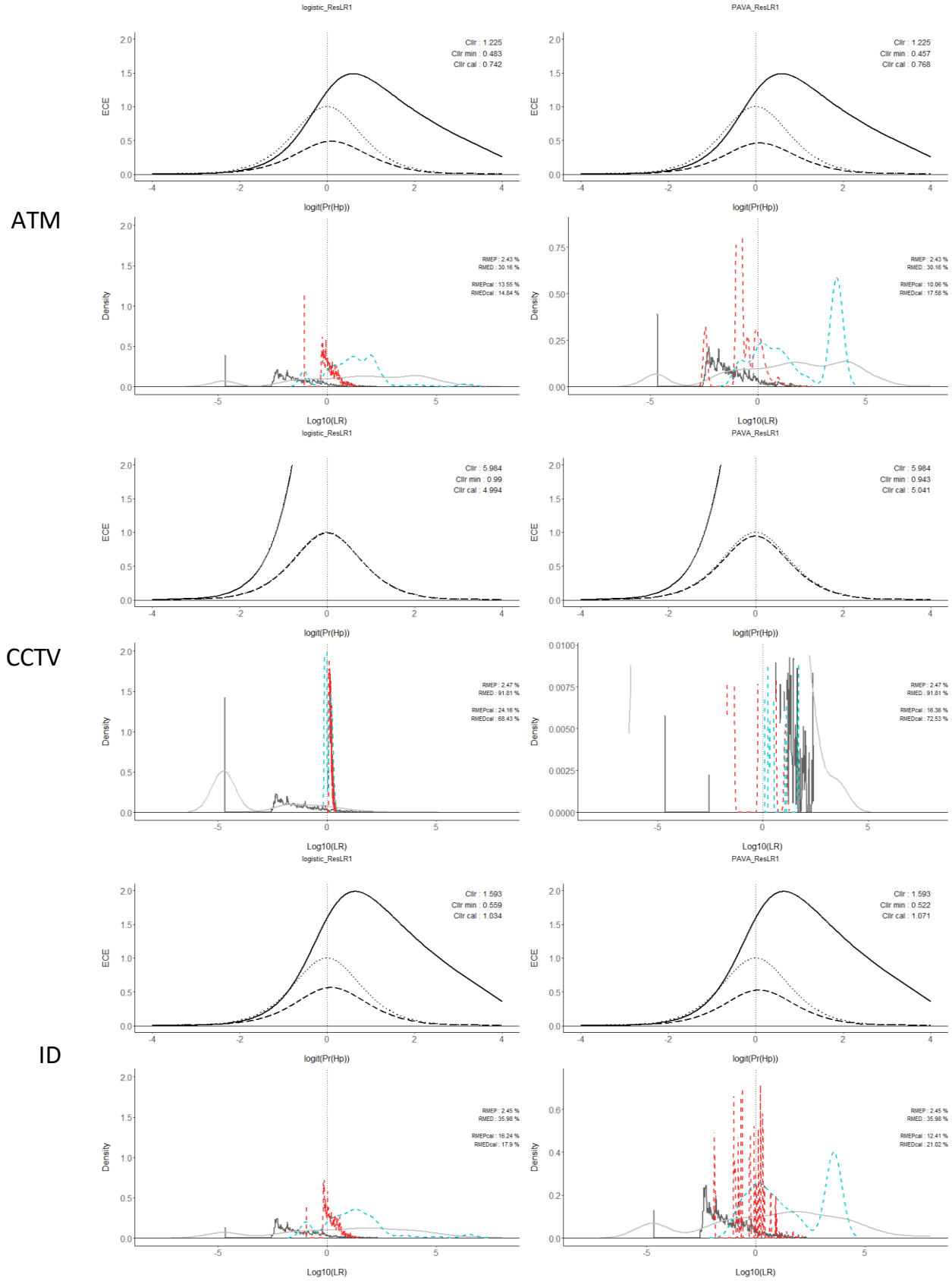
CCTV

ID

Régression Logistique

PAVA

MFI - Intravariabilité Générique / intervariabilité Suspect-Spécifique



ATM

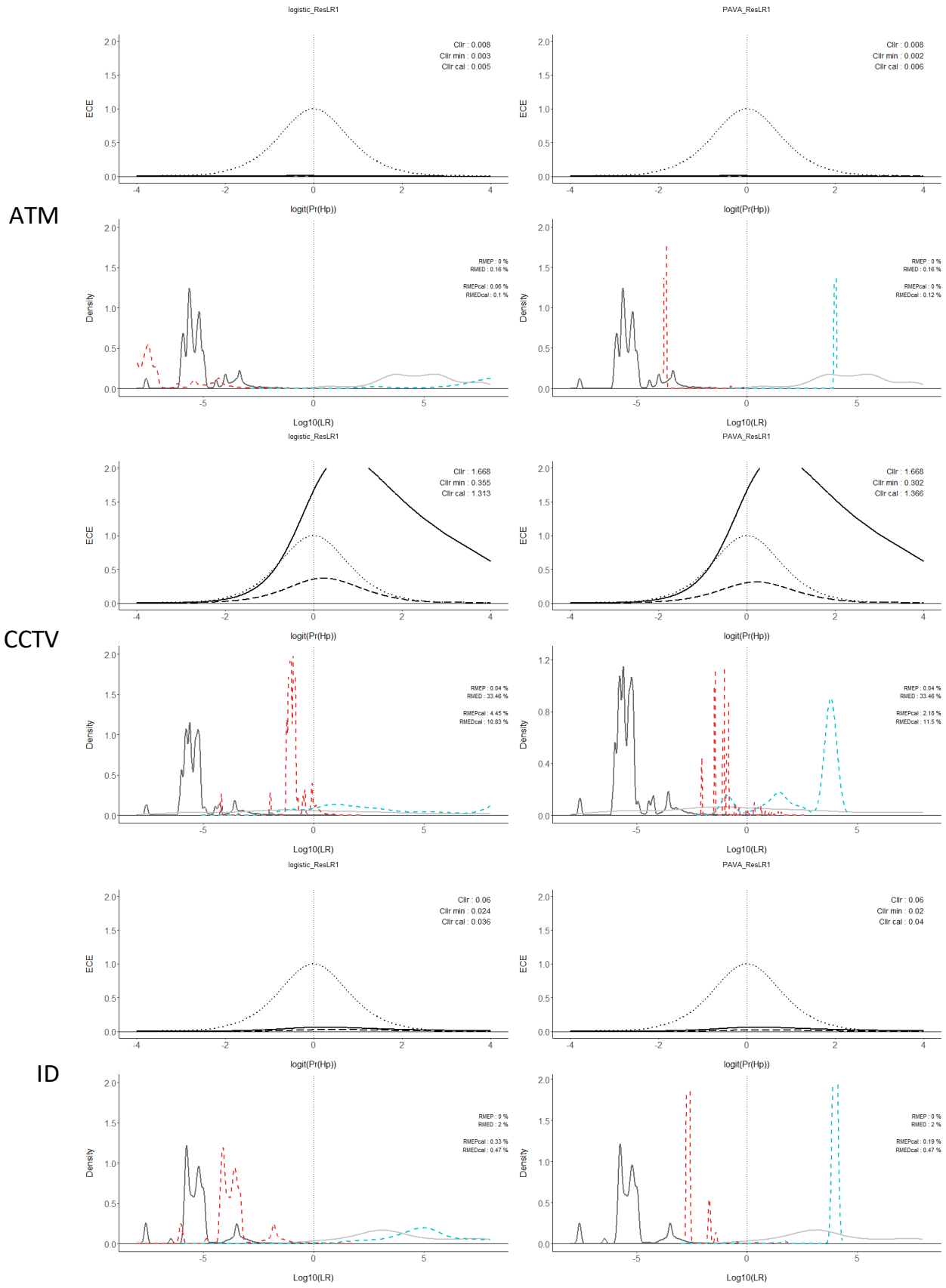
CCTV

ID

Régression Logistique

PAVA

MFE - Intravariabilité Spécifique / intervariabilité Suspect-Spécifique



ATM

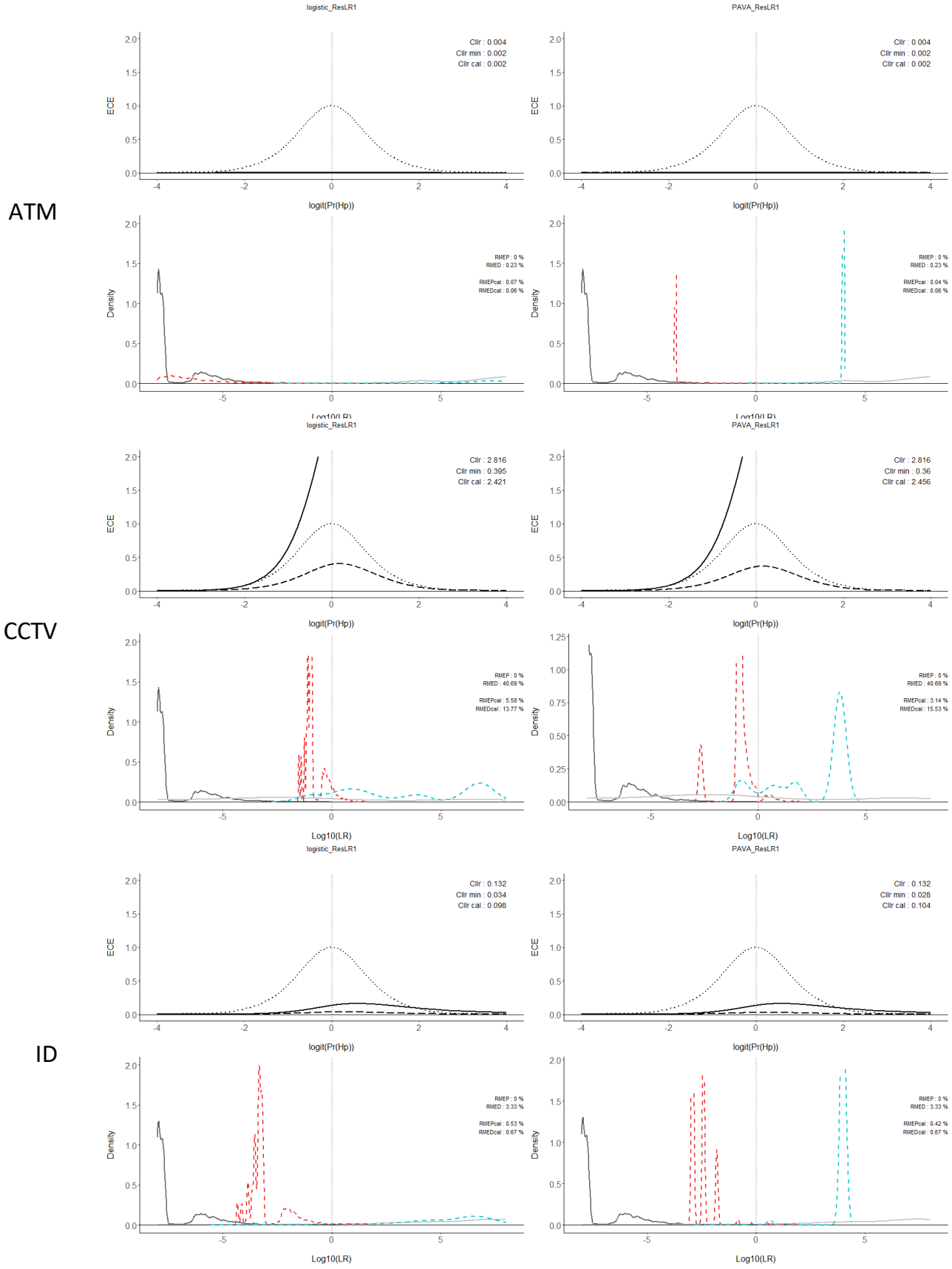
CCTV

ID

Régression Logistique

PAVA

MFE - Intravariabilité Générique / intervariabilité Suspect-Spécifique



Résumé • Avec le développement croissant des algorithmes automatiques et de l'intelligence artificielle, la reconnaissance faciale occupe une place toujours plus importante dans les contextes judiciaires et civils. Cependant, la reconnaissance faciale doit encore faire l'objet d'études empiriques approfondies pour apporter un cadre scientifique et juridique adéquat dans le cadre investigatif et au tribunal. Ce constat fixe le fondement de la présente recherche. Nous nous focalisons sur les enjeux liés aux images et à l'utilisation de systèmes automatiques. Notre objectif est le développement et la validation d'une méthodologie de calcul de rapport de vraisemblance (SLR), permettant d'interpréter les scores de comparaison issus de systèmes automatiques de reconnaissance faciale. Nous étudions deux approches de modélisation de l'intravariabilité, générique et spécifique au suspect, pour adapter la méthodologie à son utilisation dans les cadres de l'investigation et de l'expertise, respectivement. Nous intégrons la calibration au calcul de SLR, en étudiant l'impact de deux méthodes – la régression logistique et la Pool Adjacent Violator Algorithm (PAVA). Nous avons collecté trois types de traces : des portraits (ID), des images de vidéosurveillance enregistrées par des distributeurs automatiques de billets (ATM) et par une caméra grand-angle (CCTV). Des images de référence de type mugshots ont été collectées de manière à correspondre aux images signalétiques d'identité judiciaire mises à disposition par les polices cantonales neuchâteloise et vaudoise dans le cadre de ce projet. Les performances de trois systèmes automatiques de reconnaissance faciale sont comparées : deux générations du système commercial IDEMIA Morphoface (MFI et MFE) et l'algorithme en source ouverte FaceNet.
Nos résultats montrent que le calcul de SLR améliore les performances de recherche de POI dans la BdD judiciaire servant la phase d'enquête. Lorsque la liste de candidats est triée par SLR au lieu des scores de comparaison, les POI recherchées sont classées à de meilleurs rangs. Concernant les différents systèmes testés, les taux d'erreurs du MFI sont incompatibles avec son intégration dans le processus judiciaire. Les performances du MFE en font un outil de choix pour une utilisation dans les cadres de l'enquête et du tribunal, à la fois par ses performances et la rapidité de l'algorithme de comparaison. Les performances de FaceNet sont moindres, principalement sur les traces de plus faible qualité, et se révèlent insuffisantes pour une utilisation dans le cadre du tribunal. Sa disponibilité en source ouverte en fait néanmoins une alternative intéressante pour l'enquête, à partir de traces de bonne qualité. Enfin, une problématique importante développée dans notre recherche concerne l'estimation de SLR d'expertise à partir des SLR préliminaires (pré-SLR) calculés en phase d'enquête. Nous montrons qu'il est possible d'estimer une valeur attendue de SLR d'expertise pour la majorité des valeurs de pré-SLR, particulièrement à partir de scores générés par le MFE lors des deux phases de l'instruction.

Auteure

Maëlig Jacquet
Ecole des Sciences Criminelles, UNIL, Suisse

Comité de Thèse

Prof. Franco Taroni (Président du jury)
Professeur ordinaire, Ecole des Sciences Criminelles, UNIL, Suisse

Dr Simon Baechler (Expert interne)
Responsable de la formation continue, Ecole des Sciences Criminelles, UNIL, Suisse
Chef du domaine traces et analyse criminelle, Police cantonale neuchâteloise, Suisse

Dre Claude Bauzou (Expert externe)
Cheffe de produit, Idemia, France

Prof. Didier Meuwly (Expert externe)
Professeur, Université de Twente, Pays-Bas
Scientifique principal, Netherlands Forensic Institute, Pays-Bas

Prof. Christophe Champod (Directeur de thèse)
Professeur ordinaire, Ecole des Sciences Criminelles, UNIL, Suisse