CONCEPTION DE SYSTÈME DE RECONNAISSANCE DE VISAGES SPATIO-TEMPORELLE SUR VIDÉOS À PARTIR D'UNE SEULE IMAGE DE RÉFÉRENCE

par

Francis CHARETTE MIGNEAULT

RAPPORT DE PROJET PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE EN GÉNIE DE LA PRODUCTION AUTOMATISÉE
M. Ing.

MONTRÉAL, LE 15 DÉCEMBRE 2017

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

Francis Charette Migneault, 2017

Cette licence Creative Commons signifie qu'il est permis de diffuser, d'imprimer ou de sauvegarder sur un autre support une partie ou la totalité de cette œuvre à condition de mentionner l'auteur, que ces utilisations soient faites à des fins non commerciales et que le contenu de l'œuvre n'ait pas été modifié.

PRÉSENTATION DU JURY

CE RAPPORT DE PROJET A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Éric Granger, directeur de projet
Génie de la production automatisée à l'École de technologie supérieure

M. Matthew Toews, évaluateur
Génie de la production automatisée à l'École de technologie supérieure

PRÉSENTÉ AUX CYCLES SUPÉRIEURS LE 15 DÉCEMBRE 2017

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

AVANT-PROPOS

Ce projet s'inscrit parmi un effort de recherche continu visant à l'amélioration de la fiabilité des systèmes de vidéosurveillance, et plus particulièrement, pour le maintien de la sécurité apporté à ceux-ci à l'aide de l'utilisation de modèles et algorithmes de reconnaissance de visages avancés. En effet, l'abondance de sources de données vidéo et de systèmes de sécurité par caméras s'étendant dans les lieux publics, en plus de l'achalandage grandissant, font que des techniques automatisées pour assurer le maintien de la sécurité, pour détecter les individus à danger potentiel, pour analyser ou réagir aux évènements passés capturés par les caméras, ou pour éventuellement prédire et prévenir des situations désastreuses à grande échelle sont maintenant une nécessité. Or, de telles techniques aboutissent toutes à un moment ou un autre à interagir avec les individus, et plus particulièrement leur identité, ce qui renforce le besoin de méthodes qui permettent rapidement et efficacement la reconnaissance des visages.

Malgré que les systèmes proposés en littérature deviennent perpétuellement plus robustes, fiables et performants pour accomplir les tâches de reconnaissance de visages, plusieurs aspects techniques et d'application réels sur le terrain sont souvent omis, ou très simplifiés, ce qui rend l'utilisation formelle et complète de ses systèmes loin d'être encore une solution finale. En outre, beaucoup de propositions considèrent l'aspect de détection de visages sur les images tandis que d'autres adressent le suivi de ceux-ci à travers les séquences vidéo, ou encore les techniques qui permettent de réduire le nombre de fausses détections tout en prenant toujours moins de temps de calcul pour obtenir une réponse en temps réel. D'un autre côté, différents travaux considèrent plus étroitement la problématique de la classification des visages entre les différents individus, afin de les distinguer et les reconnaître lors de nouvelles apparitions sur de nouvelles séquences vidéo ou, encore, proposent des techniques pour également rendre cette tâche plus efficace. Par contre, un nombre relativement limité de publications considèrent le problème dans l'ensemble, c'est-à-dire de produire une reconnaissance de visages d'individus particuliers en considérant que la détection et le suivi de leur visage sur trajectoire peuvent également échouer préalablement. Encore plus, beaucoup de systèmes proposés considèrent l'entraînement de modèles de reconnaissance de visages à partir de larges bases de données qui contiennent plusieurs échantillons d'images représentant

les individus à reconnaître. Par contre, les systèmes traditionnels de surveillance à partir de listes de contrôle (*watchlist screening*) font habituellement davantage référence aux cas où un nombre très limité, voire unique, d'échantillons des individus visés est disponible, comme à l'aide d'une photo disponible dans un passeport ou un visa par exemple. En plus du nombre d'exemples de visages limités s'ajoutent les complexités dues aux conditions de capture comme l'illumination, la pose, l'échelle, les occlusions, et les effets de flou de motion, qui viennent bruiter l'apparence des visages. Une difficulté supplémentaire est donc souvent ajoutée à cause de ces effets, d'autant plus lorsque l'on veut reconnaître des individus à travers des systèmes de caméra de sécurité distribués, alors que les exemples d'entraînement des modèles de reconnaissance de visages proviennent d'un domaine sur image statique qui affiche des conditions d'acquisition complètement différentes. La problématique de la reconnaissance de visages en considérant toutes les contraintes et modalités impliquées dans la mise en application d'un système dans son ensemble, de la détection initiale à la prédiction finale, est ainsi encore loin d'être résolue.

Dans ce travail, plusieurs contributions scientifiques ayant démontré maintes fois les meilleures pratiques et méthodes dans le domaine de la reconnaissance de visages sont alors considérées afin de rapatrier les composantes identifiées comme les plus pertinentes et performantes, de sorte à produire un système surveillance de bout-en-bout à partir de la reconnaissance de visages. Les méthodes sélectionnées le sont également afin de s'assurer de maintenir des performances de reconnaissance extrêmement intéressantes par rapport à l'état-de-l'art. Le projet prend également en considération les besoins applicatifs des techniques employées pour le domaine de la surveillance à partir d'une liste de contrôle, en s'assurant de maintenir les niveaux de performance requis tout en préservant des algorithmes fonctionnant en temps réel, à faible requis de capacités de mémoire et de stockage, et sans nécessiter l'utilisation de plusieurs serveurs à la fine pointe de la technologie pour accomplir le traitement et calcul des résultats.

REMERCIEMENTS

Ce projet de maîtrise, qui, tout en conservant un aspect très applicatif, s'est particulièrement rapproché d'un projet de recherche nécessitant à la fois beaucoup d'heures de lectures de littérature scientifique, d'analyses de données avancées, et d'efforts de programmation constants, a été possible grâce au soutien de nombreux collaborateurs. Je tiens ainsi à prendre ici le moment de souligner l'appui apporté par chacun.

Avant toute mention particulière, je remercie tous les membres du LIVIA[1] qui ont pu avoir participé d'une manière ou d'une autre à ce projet, considérant que plusieurs d'entre nous sont impliqués à travailler sur la même problématique, ainsi que l'atmosphère motivante produite par cette grande équipe multidisciplinaire passionnée par les techniques d'imagerie et d'apprentissage machine.

Éric Granger, directeur du LIVIA, mon directeur de projet de maîtrise, et un immense soutient à travers toutes les initiatives et voies envisagées reliées aux développements de cette étude, je tiens à souligner l'appui et l'attention que tu m'as apportés pour évaluer certaines pistes de résolutions, en envisageant préventivement maintes problématiques, et en fournissant une interprétation précieuse de plusieurs résultats intermédiaires considérant les grandes difficultés, embûches et impasses rencontrées durant une grande partie du projet. Je tiens également à te remercier pour le temps alloué et les explications méthodiques fournisses face à toute la durée du projet, considérant le va-et-vient continuel des membres au laboratoire ayant participé au projet dans l'ensemble, les évènements s'enchaînant les uns après les autres, et les multiples tâches connexes qui pouvaient parfois causer des conflits d'horaires. Cette attention particulière a définitivement été une partie importante et marquante à la réalisation de ce projet. Je tiens aussi à rappeler et te remercier des opportunités offertes, non seulement pour le projet en soit, mais aussi pour la visite de compagnies travaillant directement dans le domaine de la vidéosurveillance et sécurité par reconnaissance de visage, ainsi que l'accès à multiples

[1] LIVIA – Laboratoire d'imagerie, de vision et d'intelligence artificielle, ÉTS – École de technologie supérieure, https://www.etsmtl.ca/Unites-de-recherche/LIVIA/accueil

conférenciers notables lors de séminaires ou conférences, ayant tous apporté un important niveau de concret par rapport aux travaux effectués.

Fania Mokhayeri, une collègue en recherche au LIVIA, merci pour le support offert tant par rapport à certaines explications supplémentaires qui m'ont aidé à mieux comprendre certaines problématiques en lien avec la reconnaissance de visages et la génération de modèles synthétiques, mais également pour les échanges scientifiques stimulants qui ont permis d'apporter de nouvelles perspectives face à certaines situations rencontrées durant le projet. Je te remercie également pour le soutien fourni pour obtenir des données supplémentaires qui ont définitivement aidé à tirer beaucoup plus efficacement des conclusions importantes par rapport aux résultats de certaines techniques employées lors du projet.

Saman Bashbaghi, un autre collègue du LIVIA, merci également pour ton grand soutien qui mon aidés à produire un système très efficace, sans qui les performances seraient définitivement moins impressionnantes. Je te remercie plus particulièrement pour les notables contributions et efforts scientifiques que tu as apportés, qui m'ont permis explorer de nouvelles voies grâces à une architecture initiale d'un système de reconnaissance de visages déjà très performante. Merci, également, pour ton support et tes explications qui ont permis d'envisager et résoudre certaines difficultés rencontrées pendant le projet.

Sam-Nicolai Johnston, merci pour ton appui très significatif apporté lors du développement de l'architecture de suivi de trajectoire. Ton côté très applicatif et tes connaissances de programmation m'ont définitivement aidé à produire extrêmement plus rapidement un code efficace et portable sur plusieurs plateformes, ce qui m'a été très utile pour la poursuite du projet. J'ai fortement apprécié avoir travaillé sur ce projet avec ton aide marquée.

Idrissa Coulibaly, un immense remerciement à toi pour le continuel support que tu m'as apporté tant lors de la rédaction du rapport que pour toutes les autres opportunités où tu pouvais me fournir de très précieux conseils. Ceux-ci m'ont procuré de nouvelles visions des choses qui me seront utiles pour des années à venir. Je souhaite fortement continuer à échanger et à travailler à tes côtés, car il s'agit d'expériences à la fois stimulantes et gratifiantes.

Je tiens à remercier le programme Mitacs Accélération en association avec l'organisation Face4 Systems Inc., avec qui le projet a été initié et financé, ce qui m'a permis de poursuivre plus loin dans le domaine. Je remercie aussi le programme FQRNT (Fonds de recherche du Québec, Nature et technologies) qui a également offert un grand soutien financier pour la réalisation du projet.

Finalement, pour des remerciements plus personnels, je tiens à souligner le support perpétuel que mes proches les plus intimes ont pu m'apporter tout au long de cette épreuve, incluant autant l'ensemble du processus académique que la finalité même associée au rapport final. Je tiens à vous remercier du soutien constant que vous avez fourni, sans quoi, je n'aurais pas atteint le stade où j'en suis. Toute ma gratitude pour votre continuel appui et votre irremplaçable présence.

CONCEPTION DE SYSTÈME DE RECONNAISSANCE DE VISAGES SPATIO-TEMPORELLE SUR VIDÉOS À PARTIR D'UNE SEULE IMAGE DE RÉFÉRENCE

Francis CHARETTE MIGNEAULT

RÉSUMÉ

La reconnaissance de visages (FR, Face Recognition) est une fonction importante des systèmes de vidéosurveillance (FRiVS, FR in Video Surveillance) pour permettre la vérification et l'identification d'individus d'intérêt qui apparaissent dans une scène capturée par un réseau distribué de caméras. Une application recherchée des systèmes de FRiVS est le still-to-video (S2V-FR) où il faut faire correspondre une image de référence d'un individu d'intérêt à ses apparitions sur séquences vidéo. Les systèmes S2V sont sujets à une complexité additionnelle due aux différences entre l'image de référence unique employée pour enregistrer un individu à reconnaître dans la liste de contrôle, par rapport à la scène filmée par les caméras de surveillance, à cause des sévères conditions d'acquisition telles l'illumination et la pose du visage. Dans ce travail, des méthodes d'adaptation de domaine sont employées, pour transférer un visage appris en-ligne par suivi de trajectoire spatio-temporel STRUCK, vers un ensemble de classificateurs SVM exemplaire hautement spécialisés, qui ont été entraînés à l'aide de multiples représentations de visages des individus à reconnaître et de séquences vidéo provenant des caméras. L'adaptation au domaine par le suivi de trajectoire permet aux visages de mieux correspondre à ceux modélisés par les classificateurs, ce qui mène vers une amélioration de la FR. Les modèles sont davantage améliorés avec multiples générations de représentations des visages pour entraîner un nouveau classificateur, Eoℓ-SVM, exploitant leur diversité observable dans l'environnement pour une FR robuste. L'architecture proposée opérant en temps réel est évaluée sur la base de vidéos ChokePoint et comparée à l'état-de-l'art.

Mots-clés : Reconnaissance de visages, suivi de trajectoire, multiples représentations de visage, adaptation de domaine, unique image de référence, contrôle d'accès par vidéosurveillance.

CONCEPTION OF SPATIO-TEMPORAL FACE RECOGNITION SYSTEM ON VIDEO FROM A SINGLE SAMPLE PER PERSON

Francis CHARETTE MIGNEAULT

ABSTRACT

Face recognition (FR) is an increasingly important consideration for video surveillance systems (FRiVS) to allow verification and identification of target individuals of interest appearing on a captured scene over a distributed and continuously operating network of cameras. A particularly sought application of FRiVS systems is still-to-video (S2V), where a reference image of the target individual needs to be matched against his video appearances in the scene. Systems running in S2V configurations are subject to an increased complexity due to the visible disparity between the single sample per person (SSPP) provided for target individuals enrollment to the watchlist compared to harsh acquisition conditions like illumination and pose variations caused by the captured environment. In this work, domain adaptation methods are employed to transfer spatio-temporally grouped faces learnt on-line from a face tracking mechanism, named STRUCK, to extremely specialized ensembles of exemplar-SVM, which are trained using multiple face representations from the reference image sample of every enrolled target individual and non-target video-camera captures. Using the tracker's internal face model adaptation to the individual's trajectory extracted from the surveillance camera environment allows to transfer faces that better match the modelled representations by the face classifiers, which in turn provides better FR robustness against the harsh and variable capture conditions from the scene. Moreover, obtained face models are further improved using multiple face representation generation techniques to train a new classifier, named Eoℓ-SVM, exploiting the diversity of generated face representations for robust FR. The proposed S2V-FR architecture, running in real-time, is evaluated with the ChokePoint video dataset and compared to state-of-the-art systems, showing increased performances in SSPP context using face tracking.

Keywords : Face Recognition, Face Tracking, Multiple Face Representations, Single Sample Per Person, Domain Adaptation, Video Surveillance Watchlist Screening.

TABLE DES MATIÈRES

Page

PRÉSENTATION DU JURY III

AVANT-PROPOS V

REMERCIEMENTS VII

RÉSUMÉ XI

ABSTRACT XIII

TABLE DES MATIÈRES XV

LISTE DES TABLEAUX XIX

LISTE DES FIGURES XXI

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES XXV

LISTE DES SYMBOLES ET UNITÉS DE MESURE XXVII

INTRODUCTION 1

CHAPITRE 1 BASES DE LA RECONNAISSANCE DE VISAGES 7
1.1 Types de représentations du visage 7
1.2 Types de données rencontrés 10
1.2.1 Conditions d'acquisition dans l'environnement 11
1.2.2 Images, vidéos et bases de données 15
1.3 Types de scénarios sur vidéos 17
1.3.1 Video-to-Still (V2S) 18
1.3.2 Video-to-Video (V2V) 19
1.3.3 Still-to-Video (S2V) 19
1.4 Défi de l'image unique de référence (SSPP) 20
1.5 L'information variable à travers les trames vidéo 23
1.6 Architecture de base d'un système de reconnaissance de visages 26

CHAPITRE 2 MODULES D'UN SYSTÈME DE FRiVS ET MÉTHODES CONSIDÉRÉES 29
2.1 Détection de visages 29
2.2 Suivi de trajectoire de visages 33
2.2.1 Suivi de trajectoire en général 33
2.2.2 Approche de combinaison de FD+FT 36
2.3 Extraction de caractéristiques 40

2.3.1 Méthodes par projection 40
2.3.2 Méthodes de vision 41
2.3.3 Méthodes par apprentissage de caractéristiques 44
2.3.4 Combinaison de multiples descripteurs 45
2.4 Classification des visages 47
2.4.1 Approche par CNN 47
2.4.2 Approche par SVM 48
2.4.3 Approche par *e*-SVM 52

CHAPITRE 3 SYSTÈME INITIAL EMPLOYÉ ET OPTIMISATIONS 55
3.1 Système de FRiVS avec ensemble de *e*-SVM et trajectoires de visages 55
3.1.1 Phase d'entraînement 56
3.1.2 Phase d'opération 57
3.2 Prétraitement et extraction de caractéristiques 61
3.3 Ensemble de classificateurs *e*-SVM spécialisés 65
3.4 Étapes de normalisation et de fusion 66
3.4.1 Normalisation des descripteurs 66
3.4.2 Normalisation des scores et fusion par ensemble 67
3.5 Accumulation spatio-temporelle par trajectoire 68
3.6 Optimisations pour le temps réel 69
3.6.1 Modification du code de référence pour le temps réel 70
3.6.2 Modèles binaires et temps d'entraînement 70
3.6.3 Mémorisation et accès rapide des scores pour l'accumulation 71

CHAPITRE 4 SYSTÈME AVEC MULTIPLES REPRÉSENTATIONS DE VISAGE 73
4.1 Problématique de représentation de visages limitée 74
4.2 Système de FRiVS avec l'ajout des visages synthétiques 76
4.2.1 Génération synthétique par transformations géométriques 76
4.2.2 Ajout de visages spécialisés supplémentaires 79
4.2.3 Génération synthétique de pose 81
4.3 Système de FRiVS avec sélection aléatoire de caractéristiques 84
4.4 Système de FRiVS avec adaptation de domaine 86

CHAPITRE 5 ÉVALUATIONS, ANALYSE DE RÉSULTATS ET DISCUSSIONS 89
5.1 Procédures de test 90
5.1.1 Détails de la base de données pour l'évaluation 90
5.1.2 Segmentation de la base de données 94
5.1.3 Variations d'entraînement par méthode 94
5.2 Évaluation de la performance de la reconnaissance de visages 96
5.2.1 Mesures de performance 97
5.2.2 Évaluations par reconnaissance et par identification 101
5.3 Résultats de la reconnaissance de visage 103
5.3.1 Système initial avec suivi de trajectoire 103
5.3.2 Système avec génération synthétique et adaptation de domaine 104
5.3.3 Sommaire des résultats et comparaison à l'état-de-l'art 110

CONCLUSION.....115

RECOMMANDATIONS ET PERSPECTIVES 119

BIBLIOGRAPHIE.....131

LISTE DES TABLEAUX

Page

Tableau 1.1 Quelques bases de données publiques pour la FR [122].....15

Tableau 1.2 Types de données selon le scénario en reconnaissance de visage pour la vidéosurveillance [64].....17

Tableau 2.1 Robustesse de certains descripteurs de caractéristiques face à diverses conditions d'acquisition nuisibles observées en FRiVS [13] 46

Tableau 3.1 Comparaison du nombre de détections manquées (FA), du nombre de fausses détections (FP) et du temps de détection sur toute la base ChokePoint, pour les variantes de FD frontale, de multiples VJ en parallèle et du système de FR utilisé avec suivi de trajectoire 62

Tableau 5.1 Description des combinaisons d'entraînement évaluées par bloc de tests.95

Tableau 5.2 Résultats moyens de 5 réplications de test du système initial de FRiVS avec suivi de trajectoires de visages, pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint.....103

Tableau 5.3 Résultats moyens de 5 réplications de test du système avec les combinaisons ST_{nt}, $ST_{nt}+GT_{nt}$, $GT_{wl}+ST_{nt}$ et $GT_{wl}+ST_{nt}+GT_{nt}$ pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint 104

Tableau 5.4 Résultats moyens de 5 réplications de test du système avec les combinaisons 3DMM à elle seule et $3DMM+GT_{wl}+ST_{nt}+GT_{nt}$ pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint 105

Tableau 5.5 Résultats moyens de 5 réplications de test du système avec les combinaisons $RSM_{(20,128)}$ à elle seule et $RSM_{(20,128)}+GT_{wl}+ST_{nt}+GT_{nt}$ pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint 107

Tableau 5.6 Résultats moyens de 5 réplications de test du système avec la combinaison $DA+GT_{wl}+ST_{nt}+GT_{nt}$ pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint.....109

Tableau 5.7 Résultats moyens de 5 réplications de test du système avec les meilleures combinaisons retenues pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint.....110

Tableau 5.8 Résultats moyens $F\beta$ de 5 réplications de test du système avec les meilleures combinaisons retenues pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint.....110

Tableau 5.9 Comparaison des performances des meilleures variantes du système avec d'autres méthodes en S2V faisant l'état-de-l'art, toutes situations confondues, sur l'intégralité des séquences de ChokePoint 112

Tableau 5.10 Comparaison des performances des meilleures variantes du système avec d'autres méthodes en S2V à partir de SSPP et avec suivi de trajectoire faisant l'état-de-l'art, sur l'intégralité des séquences de ChokePoint, en modes transaction et trajectoire 112

Tableau A.1 Moyennes sur 5 réplications de l'ensemble de tous les résultats, pour toutes les combinaisons d'évaluation considérées du système de FRiVS, pour diverses combinaisons de séquences de ChokePoint, et comparées par performances correspondantes.....121

Tableau A.2 Agrandissement des colonnes de séquences « Caméras frontales » et « Toute la base » du Tableau A.1 pour les évaluations du système de FRiVS avec échelle comparative 122

Tableau A.3 Segmentation de la base ChokePoint pour l'évaluation du système de FRiVS 123

Tableau A.4 Résultats de prédiction de 5 individus d'intérêt enregistrés dans la *watchlist* à l'aide du système de base (*baseline*) présenté au Chapitre 3, et testé sur ChokePoint 129

LISTE DES FIGURES

Page

Figure 1.1 Illustration d'une *fisherface* produite d'un visage d'un individu de la base de données Yale Database [14] 8

Figure 1.2 Extraction de composantes principales d'un visage (*eigenfaces*) [69] 8

Figure 1.3 Extraction de blocs géométriquement paramétrés [87] 9

Figure 1.4 Extraction de caractéristiques locales à partir de *landmarks* [115] détectées par DRMF [7] 9

Figure 1.5 Extraction des « états » ordonnés des sections du visage par HMM [33] ... 9

Figure 1.6 Illustration de la répétabilité des conditions d'acquisition contrôlées observées sur un visage à travers différents individus et images. Individus 00740 et 00750 de la base de données Color FERET [89, 105, 106] 11

Figure 1.7 Illustration de diverses conditions d'acquisition semi-contrôlées observées sur un visage à travers différentes séquences vidéo et caméras. Séquences P1E_S1_C1 et P1E_S3_C1, individu ID0003 de la base de données ChokePoint [139, 140] 12

Figure 1.8 Démonstration d'un portail limitant la trajectoire possible d'un individu en environnement semi-contrôlé. ID0003, séquence P1E_S3_C1 de ChokePoint [139, 140] 13

Figure 1.9 Illustration de diverses conditions d'acquisition difficiles en environnement sans-contraintes 14

Figure 1.10 Comparaison des conditions d'acquisition variables entre l'image *still* d'entraînement (gauche, 2672x4000 px, ratio ~1/15), d'un agrandissement du visage d'intérêt (centre, ratio ~1/3) et d'une trame de caméra de vidéosurveillance (droite, 800x600 px) de l'individu ID0001, séquence P1L_S1_C3 de la base de données ChokePoint [139, 140] 20

Figure 1.11 Architecture fondamentale des systèmes de reconnaissance de visages ... 26

Figure 1.12 Schéma-bloc plus détaillé d'un système de reconnaissance de visages 27

Figure 2.1 Représentations de caractéristiques Haar employées par Viola-Jones [84] 29

Figure 2.2 Classificateurs faibles en cascade de l'algorithme Viola-Jones [130] 30

Figure 2.3 Détections de visages obtenues par Viola-Jones (vert), et de son alternative par caractéristiques LBP [107] (rouge), toutes deux disponibles dans OpenCV [22] 31

Figure 2.4 Cas complexe de visage capté par DRMF [115], malgré l'occlusion prononcée 31

Figure 2.5 Profils de visage complexes pour l'application de la FR, mais trouvés par *HeadHunter* [80] 31

Figure 2.6 Certaines détections de visages très complexes trouvées par CNN en cascade [78], mais parfois inadéquates pour la problématique de la reconnaissance de visages 31

Figure 2.7 Effet de désalignement probable causé par le suivi de trajectoire 35

Figure 2.8 Effet de dérive (*drift*) probable causé par le suivi de trajectoire 35

Figure 2.9 Procédé de suivi de trajectoire en mode *validation-by-tracking* 37

Figure 2.10 Procédé de suivi de trajectoire en mode *tracking-by-detection* 38

Figure 2.11 Schéma-blocs du procédé d'opération de l'algorithme de suivi de trajectoires multiples FAST-DT [34]. 39

Figure 2.12 Encodage des pixels par extraction de caractéristiques LBP [2] 42

Figure 2.13 Histogramme de caractéristiques LBP 42

Figure 2.14 Application du descripteur LBP au domaine de la FR [6] 42

Figure 2.15 Procédé du descripteur HOG 43

Figure 2.16 Caractéristiques apprises par un descripteur à base de CNN (basé sur [102] 45

Figure 2.17 Hyper-plan de séparation SVM dans l'espace de haute dimensionnalité pour séparer des données non linéairement séparables 49

Figure 2.18 Entraînement de SVM en mode *one-against-all* 51

Figure 2.19 Entraînement et test de reconnaissance par SVM spécifique à chaque individu 51

Figure 3.1 Système initial de FRiVS avec E*o*e-SVM et suivi de trajectoire 55

Figure 3.2 Différents niveaux de confidence inférés de deux régions d'intérêt à partir de la mesure *IoU* [114] 58

Figure 3.3 Accumulation de scores spatio-temporels par trajectoire pour un individu recherché, en comparaison avec la reconnaissance des autres personnes .60

Figure 3.4 ROI détectée par Viola-Jones frontal (rouge), ROI obtenue du suivi de trajectoire STRUCK (vert) et zone de recherche locale par Viola-Jones (jaune hachuré).....61

Figure 3.5 Réduction d'arrière-plan pour se concentrer sur les traits importants du visage avec raffinement de 20% de la ROI de la FD Viola-Jones locale ..63

Figure 3.6 Extraction des subdivisions de visage localisé 64

Figure 3.7 Robustesse face aux occlusions localisées.....64

Figure 3.8 Types de normalisations des descripteurs par divers groupement des données 66

Figure 3.9 Déphasage des scores bruts de prédiction des *e*-SVM fortement déséquilibrés vers la classe négative au détriment du positif à représenter 67

Figure 3.10 Représentation de l'accumulation de scores spatio-temporels avec fenêtre glissante par trajectoire-individu.....69

Figure 3.11 Représentation d'une mémoire circulaire (*circular buffer*) pour l'accumulation de scores de reconnaissance de visages 71

Figure 4.1 Représentation du manque de connaissance des Eo*e*-SVM pour définir la portée de la distribution intra-classe positive et classer les nouvelles apparences de test 74

Figure 4.2 Exemples d'échantillons synthétiques (bleus) générés à partir d'opérations géométriques appliquées à l'image de référence unique (rouge).....77

Figure 4.3 Représentation des Eoℓ-SVM mieux adaptés à la distribution des images synthétiques de l'individu à reconnaître (droite) par rapport à l'approche initiale (gauche) 78

Figure 4.4 Représentation des Eo*e*-SVM spécialisés avec l'ajout des autres individus d'intérêts comme contre-exemples négatifs supplémentaires (*targets* annotés avec « ***T*** ») 80

Figure 4.5 Comparaison des distinctions du visage de différents individus d'intérêt selon le type d'environnement 81

Figure 4.6 Comparaison d'un modèle 3D obtenu par génération synthétique avec des visages observés sur trajectoire vidéo dans la scène à tenter de représenter 83

Figure 4.7 Représentation de l'application de RSM sur les vecteurs descripteurs de visage (haut) pour la génération de sous-ensemble de représentations (bas) 85

Figure 4.8 Comparaison de conditions d'acquisition variées par scène captée avec différentes caméras de surveillance, base ChokePoint [139, 140].....86

Figure 5.1 ROI de visage des individus formant la base de test ChokePoint.....90

Figure 5.2 Exemple de séquence qui contient plusieurs individus non indiqués dans les tables de vérité, séquence P1E_S1_C1, individu ID0003, ChokePoint 92

Figure 5.3 Exemple de séquence qui contient plusieurs individus non indiqués dans les tables de vérité, séquence P1E_S2_C3, individu ID0027, ChokePoint 92

Figure 5.4 Exemple de séquence avec détections multiples dues à la réflexion sur la vitre de la porte du portail, séquence P2E_S3_C1.1, individu ID0030, ChokePoint.....92

Figure 5.5 Exemple de réflexion constante d'illumination produisant de fausses détections répétées, séquence P1L_S3_C2, individu ID0003, ChokePoint 92

Figure 5.6 Démonstration du principe de la surface partielle sous la courbe ROC 98

Figure A.1 Courbes ROC et PR en fonction des descripteurs LBP_{59}, HOG_{144} et CNN_{259} 125

Figure A.2 Courbes ROC et PR en fonction des descripteurs LBP_{59} et HOG_{144} appliqués aux ROI de visages minimales et maximales obtenues par Viola-Jones.....126

Figure A.3 Courbes ROC et PR en fonction de visages frontaux (haut) et de visages avec poses (bas), à l'aide de SVM à kernel linéaire entraînés et testés sur la base de visages Color FERET, pour 10 réplications de test avec 100 individus entraînés et testés contre 200 autres.....127

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

3DMM	Méthode de génération de pose 3D par Morphable Models
CNN	Convolutional Neural Network (réseau de neurones convolutif)
CUDA	Bibliothèque « Compute Unified Device Architecture » de Nvidia
DA	Domain Adaptation (adaptation de domaine)
DCP	Directional Corner Points
DPM	Deformable Parts Model
DRMF	Discriminative Response Map Fitting
e-SVM	Exemplar-SVM (SVM entraîné à partir d'un exemplaire)
Eo*e*-SVM	Ensemble of exemplar-SVM (ensemble de *e*-SVM spécifique par individu)
Eoℓ-SVM	Ensemble of linear-kernel SVM (ensemble de SVM à kernel linéaire)
FAST-DT	FAce STructured Detection and Tracking
FC	Face Classification (module de classification de visage)
FD	Face Detection (module de détection de visage)
FE	Feature Extraction (module d'extraction d'un descripteur de caractéristiques)
FR	Face Recognition (ensemble du système de reconnaissance de visage)
FRiVS	FR in Video Surveillance (reconnaissance de visages en vidéosurveillance)
FT	Face Tracking (module de suivi de trajectoire de visage)
HOG	Histogram of Oriented Gradients (histogramme de gradients orientés)
HMM	Hidden Markov Models
LBP	Local Binary Patterns (patrons binaires locaux)
LPQ	Local Phase Quantization (quantification locale de la phase)
OpenCV	Bibliothèque « Open Computer Vision »
OpenMP	Bibliothèque « Open Multi-Processing »

ROC	Receiver Operating Characteristic curve (courbe de performance ROC)
ROI	Region of Interest (région d'intérêt sur une image ou un vidéo)
RSM	Random Subspace Method (méthode de sélection aléatoire de caractéristiques)
S2V	Mode d'utilisation *still-to-video* (images de référence, vidéos de recherche)
S2V-FR	Reconnaissance de visages spécifiquement en mode d'utilisation S2V
SSPP	Single Sample Per Person (référence unique par individu)
STRUCK	Structured Output Tracking with Kernels
SVM	Support Vector Machine (machine à vecteurs de support)
V2S	Mode d'utilisation *video-to-still* (vidéos de référence, images de recherche)
V2V	Mode d'utilisation *video-to-video* (vidéos de référence et pour la recherche)
VJ	Algorithme de détection de visages Viola-Jones

LISTE DES SYMBOLES ET UNITÉS DE MESURE

$\mathbf{a}_j$	vecteur descripteur de test (*probe*)
$\mathbf{a}_t$	vecteur descripteur sur une trajectoire en mode d'opération (*operational phase*)
AUC	Area Under ROC curve (surface sous la courbe ROC)
AUPR	Area Under Precision-Recall curve (surface sous la courbe précision-rappel)
AVG	Average (moyenne)
BEST	Résultat de FR combiné selon la confidence des prédictions par trajectoire
C^+	coût de pénalité de mauvaise classification SVM des échantillons positifs
C^-	coût de pénalité de mauvaise classification SVM des échantillons négatifs
$\mathbf{d}$	vecteur descripteur en mode d'entraînement (*enrollment phase*)
d_i	caractéristique à l'index *i* contenue dans un vecteur descripteur
$d_{i,p}$	caractéristique à l'index *i* d'un vecteur descripteur, pour un *patch p* spécifique
$d_{i,p}^{(norm)}$	valeur normalisée par min-max de la caractéristique $d_{i,p}$
d_{rs}	caractéristique à l'index *rs* contenue dans un sous-descripteur obtenu par RSM
f	indice des trames de vidéos
FA	Fail-to-Acquire (manquement d'acquisition/détection d'un visage présent)
F_β	mesure de performance « F-score » avec pondération des classes β
FN	False Negative (classification de type faux-négatif)
FP	False Positive (classification de type faux-positif)
FPR	False Positive Rate (ratio de faux-positifs, spécificité)
FPS	Frames Per Second (vitesse de défilement des trames vidéo)
γ_s	seuil de reconnaissance associé à un individu donné
GT_{nt}	transformation géométrique appliquée aux *stills* de *non-targets* d'entraînement

GT_{wl}	transformation géométrique appliquée aux *stills* d'individus du *watchlist*
HOG_{588}	variante spécifique du descripteur HOG à 588 *features* employée pour ce projet
N_d	nombre de caractéristiques formant un vecteur descripteur
$N_d^{(rs)}$	nombre de caractéristiques formant un sous-descripteur par RSM
N_D	nombre de descripteurs employés pour l'entraînement des classificateurs
N_f	nombre de cadres vidéo (*frames*) formant l'intervalle de détection
N_{nt}	nombre d'individus de contre-exemples d'entraînement (*non-targets*)
N_{ut}	nombre d'individus inconnus de test en phase d'opération (*unknown test*)
N_{wl}	nombre d'individus enregistrés dans la *watchlist* pour l'évaluation des tests
N_p	nombre de blocs (*patch*) segmentant la ROI de visage d'un individu
N_{rs}	nombre de sélections aléatoires de caractéristiques (*random subspaces*)
N_{sv}	nombre de vecteurs de support modélisé par les SVM pour la classification
p	indice d'une subdivision (*patch*) d'une *ROI* donnée
PPV	Positive Predictive Value (précision des prédictions de FR par le système)
pAUC(*fpr*%)	Partial AUC (surface partielle sous la courbe ROC de zéro jusqu'à *fpr*%)
ROI_d	région d'intérêt d'un visage produite par l'algorithme de détection
ROI_s	région d'intérêt d'un visage correspondant à l'image de référence (*still*)
ROI_t	région d'intérêt d'un visage produite par l'algorithme de suivi de trajectoire
ROI_t^L	région d'intérêt d'un visage validée par détection locale de la trajectoire
ROI_v	région d'intérêt d'un visage sur séquence vidéo (sur un *frame*) d'entraînement
$RSM_{(Nrs,\, N_d)}$	RSM utilisé avec N_{rs} sélections aléatoires de $N_d^{(rs)}$ caractéristiques distinctes
s	indice d'une ROI ou d'un vecteur descripteur provenant d'une image (*still*)

$\mathbf{S}_{s,t,p}$	matrice de scores d'un *patch* de visage d'un individu-trajectoire d'un *frame*
$\mathbf{S}_{s,t}$	matrice de scores par individu-trajectoire à un moment donné (*frame*)
$\mathbf{S}_{s}^{Acc*}$	matrice de scores spatio-temporels accumulés par individu-trajectoire
$\mathbf{S}_{I}^{Acc*}$	matrice de scores d'identification à partir des scores par individu-trajectoire
SE	Standard Error (erreur standard)
ST_{nt}	augmentation de données avec images *still* additionnelles des *non-targets*
$\mathbf{t}_k$	vecteur descripteur de référence pour l'entraînement (*template*)
t	indice de trajectoire de visages, appliqué à une ROI ou un vecteur descripteur
TN	True Negative (classification de type vrai-négatif)
TP	True Positive (classification de type vrai-positif)
TPR	True Positive Rate (ratio de vrais-positifs, sensibilité ou rappel)
v	indice d'une ROI ou d'un vecteur descripteur provenant d'une séquence vidéo
W	dimension de la fenêtre glissante d'accumulation de scores spatio-temporels

INTRODUCTION

Contexte de l'étude

L'étude menant à l'élaboration de ce projet de maîtrise se base sur les besoins en industrie, et des organisations œuvrant dans le domaine de la sécurité, de systèmes technologiques permettant l'authentification d'individus sur séquences vidéo, notablement pour des applications de vidéosurveillance par l'intermédiaire de systèmes distribués de caméras tels que dans les aéroports, aux douanes, ou pour les autres lieux publics en milieu urbain comme les centres d'achat ou les voies routières. En effet, sans de tels systèmes, les opérations habituelles visant au maintien de la sécurité doivent autrement être accomplies par des vidéo-interprètes ou vidéo-analystes qui sont humainement limités en termes de disponibilité et capacité pour analyser le grand nombre de séquences vidéo qui s'accumulent plus rapidement qu'ils ne peuvent les visionner. De plus, plusieurs autres facteurs comme la fatigue et l'interprétation particulière à chaque analyste fait que la reconnaissance de visages sur vidéos est difficilement reproductible par ceux-ci et il devient alors possible d'échouer à détecter d'un individu recherché par inadvertance. Cela est d'autant plus problématique lorsqu'on étant l'ampleur de la tâche d'investigation sur les immenses bases de données de vidéos qui doivent continuellement être inspectées après faits, pour retrouver par exemple, des pistes potentielles de passages d'individus visés dans une scène où un évènement important c'est déroulé, et qui n'avaient encore jamais été identifiées par un interprète.

Ainsi, l'automatisation des systèmes de reconnaissance de visage (FR) est une manière efficace de subvenir simultanément aux requis de l'industrie et à l'ampleur des données, mais cette tâche demeure relativement très complexe en prenant en considération les difficultés engendrées par les variations visibles sur les visages durant l'acquisition des vidéos. En effet, même les méthodologies faisant partie de l'état-de-l'art souffrent encore grandement des effets d'acquisition variés comme la pose du visage, l'illumination sur celui-ci, la résolution et le flou de mouvement causé par les caméras, sans compter les différences observées entre les images d'enregistrement des individus dans la base de référence, et les séquences vidéo observées lors de la mise en fonction du système. Or, l'une des problématiques principales des

systèmes de reconnaissance de visages en vidéosurveillance (FRiVS) est donc de trouver des méthodes permettant d'améliorer la performance de ces systèmes à identifier la présence d'un individu donné dans la scène, sans néanmoins introduire davantage de fausses reconnaissances, et en maintenant ou même en réduisant l'impact en temps de calcul et d'espace mémoire requis pas les algorithmes employés. L'étude est ainsi orientée en fonction de toutes ces prémisses afin d'explorer les voies d'améliorations potentielles.

Objectifs du projet

L'objectif principal est de développer et d'analyser un système de pointe pour la classification de visages dans un contexte de vidéo surveillance. Le système élaboré doit reconnaître le visage d'individus particuliers apparaissant sur des séquences vidéo obtenues par acquisition de caméras de sécurité, en utilisant des modèles de classification entraînés à partir d'une unique image de référence (entraînement en *single sample per person*, SSPP) provenant, par exemple, d'un passeport.

Ainsi, le projet à comme d'ensemble d'effectuer la conception d'un système de S2V-FR (*still-to-video Face Recognition*, voir le Chapitre 1 pour plus de détails) qui s'apparente au maximum des cas d'application réels et des besoins dans l'industrie. Pour ce faire, les objectifs intermédiaires du projet consistaient principalement en des choix technologiques (ex. : type d'algorithmes, choix de bibliothèques de codes, méthodes d'optimisation de détection et de temps d'exécution, etc.) pour le développement d'un système complet de S2V-FR.

De manière plus spécifique, le projet c'est divisé en deux étapes dans globalement, soit (1) exploiter les des méthodes reconnues comme formant l'état-de-l'art afin d'obtenir un système de base, servant à une reconnaissance de visages robuste à partir de trajectoires sur vidéo, et (2) améliorer ce dernier à l'aide de méthodologies variées en explorant à travers les méthodes de génération de visages synthétiques, l'augmentation de données par représentations multiples, ainsi que l'adaptation de domaine.

Afin d'accomplir la première étape du projet, l'architecture du système est fondée sur multiples astuces qui permettent de combiner les notions gagnées par les travaux de Dewan et al. [45, 46, 47, 48], en lien avec l'amélioration de la FR grâce à l'adaptation en-ligne d'un modèle de visage contenu à l'intérieur d'un module de suivi de trajectoire et, des travaux de Bashbaghi et al. [9, 10, 11, 12], qui ont démontré les possibilités de remarquables améliorations de la précision des systèmes S2V-FR, en combinant des ensembles de classificateurs hautement spécialisés entraînés à partir d'une modélisation par multiples représentations provenant d'une image unique de référence. Les études approfondies dans les deux cas ont fait ressortir les meilleures approches pour certaines parties bien spécifiques du système de bout-en-bout, mais il demeurait que celles-ci n'étaient pas encore employées à leur plein potentiel, car il s'agissait d'études distinctes. Ainsi, l'ensemble des travaux précédents sont employés pendant ce projet comme fondement pour construire un système qui regroupe simultanément le suivi de trajectoire par l'adaptation graduelle de celle-ci en fonction des nouvelles apparitions dans la scène, et des multiples représentations de visages pour former des ensembles de classificateurs robustes à travers un large éventail de conditions défavorables subvenant lors des acquittions des visages sur caméras de vidéosurveillance.

Suite à avoir obtenu un système robuste qui combine les deux concepts précédents, la seconde phase du projet se voulait être une exploration de solutions proposées en littérature pour davantage améliorer le système dans son ensemble, considérant que la combinaison des méthodes proposées pour le suivi de trajectoire avec celles pour les ensembles de classificateurs apporte une complexité supplémentaire par rapport à la reconnaissance de visage finale. Autrement dit, il était nécessaire de trouver des manières d'adapter les régions d'intérêt des visages produits par le module de trajectoire adapté en-ligne, vers le domaine différent avec lequel les ensembles de classificateurs hautement spécialisés étaient entraînés, pour faire concorder les visages et ainsi obtenir une prédiction fiable de la FR.

Ainsi, des approches en littérature pour la génération synthétique d'images de visages ont été étudiées afin d'augmenter la robustesse des systèmes proposés, selon les difficultés principales observées par rapport à la pose et l'illumination sur le visage. L'adaptation de domaine et

l'ajout de multiples représentations diverses des visages ont aussi été étudiés, toujours dans l'objectif d'améliorer les performances résultantes du système de SV2-FR pour la vidéosurveillance, résultant de la combinaison d'approches robustes.

Finalement, des considérations selon l'analyse des vidéos en temps réel ainsi qu'en produisant des modèles de visages compacts en mémoire sont maintenues en tant qu'objectif supplémentaire afin de conserver le potentiel applicatif des solutions produites. L'objectif global est ainsi d'obtenir un système de FR de bout-en-bout en temps réel qui permet un niveau plus que satisfaisant de bonne reconnaissance de visages, basé sur les connaissances acquises jusqu'à ce jour des travaux précédents formant l'état-de-l'art.

Justification du projet et méthodologie

Afin d'améliorer les techniques employées pour la sécurité et la surveillance par vidéo-caméra tel que dans les aéroports, la reconnaissance d'individus particuliers selon des listes de contrôle (*watchlist screening*) est un domaine en croissance constante. Par contre, les solutions pour la reconnaissance de visages ne sont pas optimales présentement [41]. Dans plusieurs cas, de simples variations ou déformations des visages captés sur vidéos, entre autres causées par l'intensité lumineuse ou l'angle de capture, suffisent à rendre les algorithmes inefficaces. Le présent projet vise à explorer quelques techniques d'imagerie synthétique afin d'améliorer les performances finales atteintes par le système dans son ensemble, et pour assurer une meilleure concordance entre les visages captés et ceux enregistrés dans la base de référence.

En plus des techniques de génération de multiples représentations du visage, des procédés combinant l'information temporelle contenue dans les trajectoires formées par le passage des individus sur les séquences vidéo sont envisagés, étant donné que l'accumulation dans le temps des reconnaissances offre un potentiel de confidence par rapport aux décisions effectuées sur la présence ou non d'individus recherchés. Étant donné que plusieurs contributions dans la littérature scientifique semblent adresser ces points individuellement, mais que très peu tentent de combiner leurs effets constructifs, ce projet vise finalement à valider ou à infirmer le potentiel de certaines techniques, lorsqu'appliquées dans un système de FRiVS complet.

Pour parvenir aux objectifs considérés par l'étude, diverses solutions de programmation (bibliothèques de codes et algorithmes) ont tout d'abord été explorées afin de déterminer les options les plus prometteuses pour la réalisation du projet. Celles-ci ont été comparées selon des analyses de performance ainsi que plusieurs résultats dans la littérature afin de retenir les options les plus prometteuses en termes de temps d'exécution, de fiabilité des résultats, de facilité d'adaptation du code, etc. Les solutions retenues sont alors employées pour développer un algorithme de reconnaissance de visage de base (*baseline*), afin d'avoir une base de référence pour l'adaptation de solutions plus efficaces à la tâche. Une étude détaillée du taux de succès de reconnaissance des visages selon chaque nouvelle alternative envisagée a servi de base pour comparer de nouvelles solutions d'amélioration explorées. Pour effectuer cette comparaison, des barèmes ont été établis et testés en employant des bases de données publiques de grande taille reconnues et employées en littérature.

Suite à l'élaboration du système *baseline*, des techniques en imagerie ont été étudiées afin de générer des images synthétiques et d'autres représentations possibles du visage qui permettraient potentiellement de bonifier les modèles de visages dans les bases de référence, et d'améliorer la concordance entre le visage recherché et ceux présents dans la scène sur vidéo. Les performances y sont alors comparées avec l'approche initiale sans technique additionnelle d'imagerie pour établir les améliorations atteintes. Étant donné que le but du projet est d'améliorer globalement le système de FRiVS, les performances sont toujours considérées par rapport à l'ensemble du procédé, afin d'y retenir la combinaison la plus optimale.

Retombées et impact industriel

Comme mentionné précédemment, le besoin de techniques de reconnaissance de visages est bien établi dans l'industrie dû au niveau grandissant de sources de données vidéo, et les requis de sécurité dans un milieu de médias de plus en plus interconnecté. Les systèmes de caméras de sécurité s'étendent, et font que des solutions efficaces, mais réalistes, en fonction de leur mode d'utilisation, sont requises avant tout. Ainsi, en considérant les problématiques réelles du milieu visé, c'est-à-dire en s'adressant au mode d'opération en SSPP et en attaquant aux difficultés réelles de variation observées entre les images d'entraînement (*stills*) et les vidéos

capturés, un système de FRiVS basé sur l'approche *still-to-video* risque davantage de produire un véritable impact dans l'industrie. L'application des méthodes proposées pourrait directement être observable à travers les solutions déjà existantes ou en développement dans les entreprises étant donné les considérations réalistes de l'étude.

Principalement, les améliorations qui seront proposées permettront de meilleures performances pour la reconnaissance d'individus particuliers. Cela offrirait comme avantage de rendre les systèmes de sécurité plus robustes, et il serait ainsi possible d'offrir une meilleure surveillance des entrées et sorties des individus marqués sur des listes de contrôle tel qu'utilisées aux douanes.

Présentation des sections

Le contenu de ce rapport est séparé en cinq sections. Tout d'abord, le **Chapitre 1** présentera certains principes fondamentaux reliés aux systèmes de FRiVS employés dans le projet. Les concepts qui viennent impacter la FR y seront discutés de manière à mieux situer les idées qui seront explorées par la suite. Le **Chapitre 2** présentera ensuite plusieurs des méthodes existantes qui permettent de réaliser un système de FR dans l'ensemble. Les techniques qui ont été considérées pour former un système de FRiVS initial utilisé lors de cette étude y seront également à l'aide des justifications associées. Le système initial alors réalisé à partir de ces méthodes sera quant à lui présenté en plus grand détail au **Chapitre 3**. Une élaboration plus fine des choix opérationnels pour chacune des parties du système et des ajustements spécifiques par rapport à l'application des méthodes y sera aussi discutée. Le rapport poursuivra ensuite au **Chapitre 4** avec une présentation des améliorations réalisées et appliquées au système présenté au chapitre précédent. Les détails additionnels seront fournis pour chaque méthode de génération de représentation de visage évaluée durant le projet ainsi que les raisons qui ont incité celles-ci. Le **Chapitre 5** permettra enfin de passer aux résultats et analyses spécifiques des méthodes. Les procédures de test, les données employées, les mesures d'évaluation, et les résultats découlant de celles-ci, seront présentés en grand détail. Le tout sera finalement enchaîné de discussions basées sur les découvertes réalisées et les opportunités futures.

CHAPITRE 1

BASES DE LA RECONNAISSANCE DE VISAGES

Les systèmes de reconnaissance de visages (FR) sont caractérisés par plusieurs composantes tout dépendamment du mode d'application dans lequel ils sont employés, des données utilisées pour l'entraînement des modèles de visages, des données employées durant le mode d'opération, ainsi que des préalables d'utilisation imposés par leur contexte d'application. À ce chapitre, certaines des différentes modalités applicables aux systèmes de FR seront introduites afin de mieux situer le reste du projet, les motivations de certaines approches, et les développements réalisés.

1.1 Types de représentations du visage

Lorsque observés à très haut niveau, les systèmes FR sont habituellement catégorisés en trois grandes familles afin de parvenir à définir une méthode de représentation des visages en mémoire [8, 26, 122, 150], de sorte à pouvoir exploiter ces représentations pour entraîner un système de classification qui pourra à son tour identifier ou vérifier l'identité d'un individu. Ces familles regroupent (1) les approches holistiques (*template matching*), c'est-à-dire dont le visage est globalement représenté par un ensemble de caractéristiques discriminantes d'un visage à l'autre, (2) les approches locales (*feature-based*), qui extraient des caractéristiques particulières à certains points d'intérêt du visage comme les yeux, le nez, la bouche et le pourtour de la mâchoire, et finalement (3) les approches hybrides qui exploitent en symbiose les deux techniques précédentes.

À titre d'exemples pour la première catégorie, soit les approches globales, nous retrouvons particulièrement les techniques PCA, LDA, *eigenfaces*, *fisherfaces*, LBP et HOG (voir le Chapitre 2 pour les détails) qui sont extrêmement rependues en littérature, que pour en nommer que quelques-unes. À noter que cela n'est en effet qu'une liste non exhaustive étant donné non seulement les multiples variations et alternatives existantes à ces techniques mêmes, mais également à l'existence des multiples autres techniques se rapportant à cette catégorie.

Le but des techniques holistiques consiste à extraire les caractéristiques sur l'ensemble du visage qui permettent de les discriminer entre eux, de sorte à ultimement produire un modèle de visage particulier avec lequel un classificateur pourrait être entraîné à les différencier. Par exemple, la Figure 1.1 et la Figure 1.2 présentent des variations de visages représentés par différentes techniques holistiques. Nous pouvons y voir que les techniques permettent de faire ressortir les points saillants du visage, de sorte à permettre d'identifier les caractéristiques discriminantes les plus probables de rendre la classification de différents visages efficace.

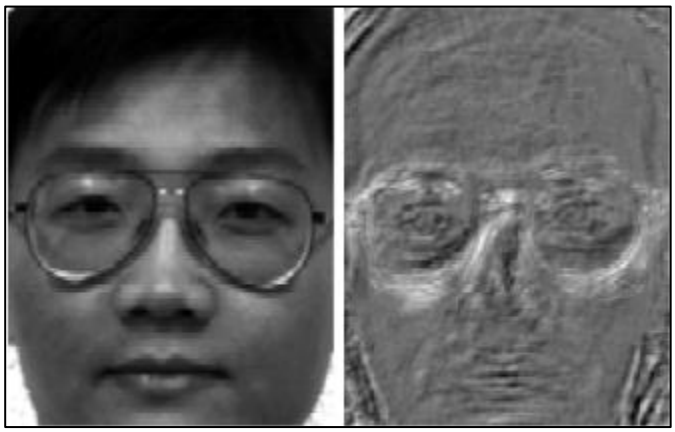

Figure 1.1 Illustration d'une *fisherface* produite d'un visage d'un individu de la base de données Yale Database [14]

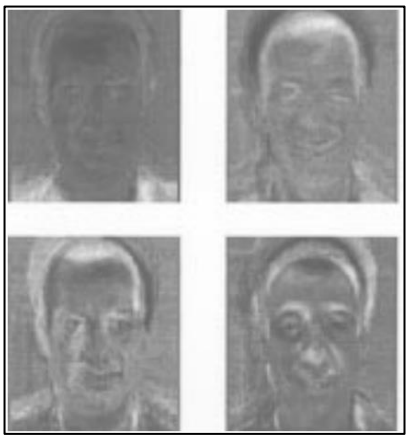

Figure 1.2 Extraction de composantes principales d'un visage (*eigenfaces*) [69]

D'un autre côté, les approches locales abordent la problématique de représentation de visages afin d'en déterminer les traits particuliers du visage qui devraient être plus spécifiquement exploités afin de les distinguer. La difficulté principale reliée à ces techniques consiste à trouver une manière adéquate de représenter les interrelations qui existent entre ces caractéristiques, qu'il s'agisse de positions relatives sur le visage, d'interconnexions géométriques, ou de regroupements par graphes. Parmi les techniques les plus répandues, nous y retrouvons particulièrement les méthodes HMM, DCP, par *landmarks* et les approches purement géométriques. À titre d'exemples, les images représentées de la Figure 1.3 à la Figure 1.5 démontrent des applications appartenant à la famille des techniques locales.

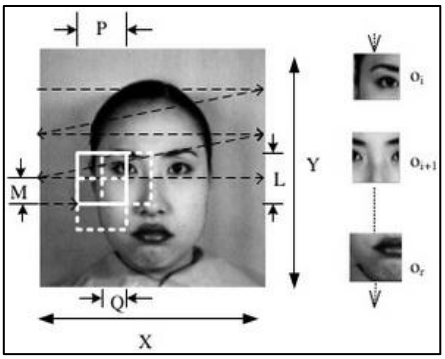

Figure 1.3 Extraction de blocs géométriquement paramétrés [87]

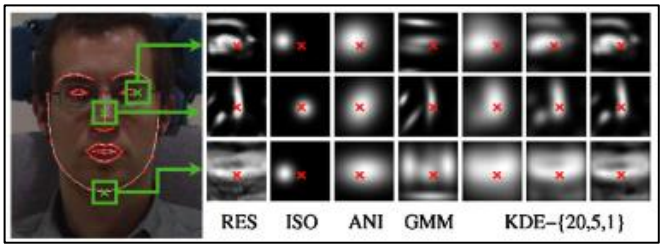

Figure 1.4 Extraction de caractéristiques locales à partir de *landmarks* [115] détectées par DRMF [7]

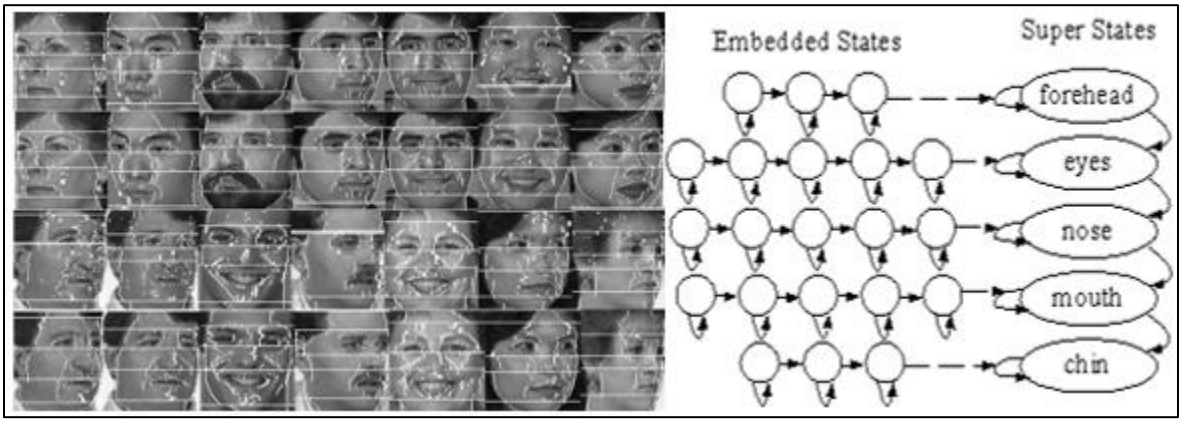

Figure 1.5 Extraction des « états » ordonnés des sections du visage par HMM [33]

Selon les besoins de l'application FR et des difficultés rencontrées lors de sa mise en application, certaines méthodes peuvent être plus adaptées à certains problèmes que d'autres. En outre, les techniques holistiques ont souvent tendance à être plus rapides par rapport au temps de calcul, parce qu'elles ne nécessitent pas d'évaluation des interrelations entre les caractéristiques comme le font les approches locales. Par contre, étant donné qu'elles extraient les caractéristiques à partir de l'ensemble du visage, elles sont également plus sensibles aux petites variations existantes entre la position des pixels d'une image à l'autre ainsi qu'aux positions très variables des points saillants du visage d'un individu à l'autre [146].

À l'inverse, les approches locales s'affranchissent davantage de ces problématiques, étant donné leur nature même qui leur permet une certaine flexibilité sur les positions des caractéristiques discriminantes, ce qui peut les rendre robustes aux désalignements des pixels, mais requiert souvent plus de temps d'évaluation afin de trouver les interactions existantes entre les caractéristiques, ou simplement pour déterminer la position de ces caractéristiques étant donné qu'elles changent sur chaque image. De plus, cette seconde famille à souvent tendance à être nettement plus sensible aux variations causées par la pose du visage sur les images, étant donné la possibilité de variation très brusque de la position des points caractéristiques visibles sur l'image, à un point tel que ces techniques peuvent totalement échouées lorsque l'angle de pose est trop prononcé.

Dans tous les cas, qu'une catégorie de représentation de visage ou l'autre soit employée, plusieurs variations et améliorations des méthodes existent en littérature afin de tenter de réduire l'impact des perturbations principales visibles sur le visage lors de l'acquisition sur images ou vidéos, et qui viennent affecter respectivement chaque technique. Les approches hybrides en sont d'ailleurs un résultat direct. Les principales composantes à prendre en considération lors de la mise en application de ces techniques sont donc de considérer l'environnement dans lequel la méthode choisie sera employée, ainsi que de prévoir les cas problématiques auxquels la méthode devra faire face pour y demeurer fonctionnelle et efficace.

1.2 Types de données rencontrés

Une des propriétés primordiales pour l'application des méthodes de reconnaissance de visage est le contexte dans lequel elle s'effectue, ou plus précisément, si la FR doit s'accomplir sur des images statiques ou des séquences vidéo. En effet, cette distinction a un impact direct tant sur la performance des méthodes, les bases de données disponibles pour leur conception, et peut même venir jouer sur leur applicabilité même. De plus, chaque situation (images ou vidéos) apportera des complexités bien distinctes à considérer pour la conception d'un système de FR complet selon l'environnement de capture rencontré.

1.2.1 Conditions d'acquisition dans l'environnement

Les environnements d'acquisitions d'images et de vidéos sont habituellement catégorisés selon trois types, c'est-à-dire les environnements contrôlés, semi-contrôlés et non-contrôlés (ou sans-contraintes. Selon le type d'environnement où les acquisitions sont effectuées, nous pouvons observer différents facteurs de variation sur les visages avec différents niveaux de changement. Les facteurs de variation principaux observés à travers ces environnements sont notamment l'angle de pose, l'illumination, l'expression du visage, l'échelle, le flou de mouvement et les occlusions [8, 41, 122]. La présence et l'impact de chacun de ces facteurs sont très différents selon le type d'environnement visé par le système de FR.

Dans le premier cas, en environnement contrôlé, il faut penser aux situations correspondant à des sessions photographiques, où les individus capturés sur images sont relativement bien positionnés ou, du moins, sont positionnés dans la scène de manière prévue et répétable, d'où le nom « contrôlé ». Les environnements contrôlés sont la majorité du temps caractérisés par un niveau d'illumination et une répartition des lumières sur le visage de manière relativement constante d'une image à l'autre. De plus, les poses du visage et les expressions affichées sont majoritairement neutres. Dans le cas contraire, les variations de pose et d'expression sont planifiées et répertoriées dans les bases de données où ces images sont collectées. La Figure 1.6 démontre un tel type d'environnement où on peut y voir les caractéristiques décrites.

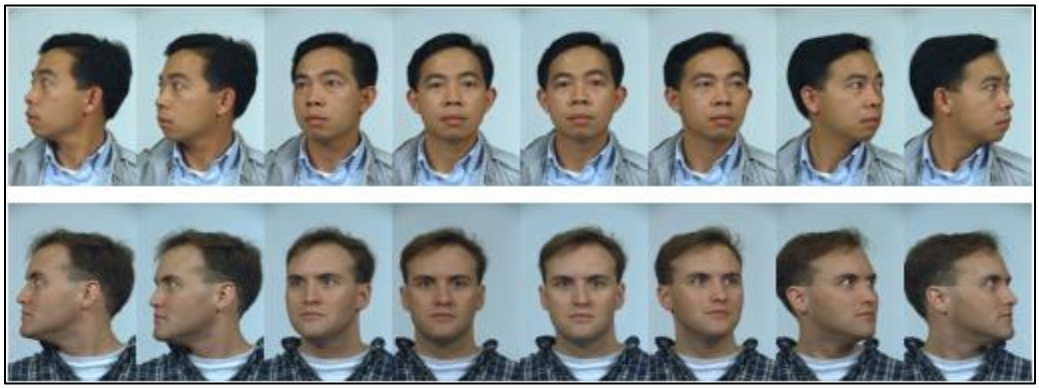

Figure 1.6 Illustration de la répétabilité des conditions d'acquisition contrôlées observées sur un visage à travers différents individus et images. Individus 00740 et 00750 de la base de données Color FERET [89, 105, 106]

Dans le cas des acquisitions en environnement semi-contrôlé, les individus capturés dans la scène ne sont pas nécessairement situés à des positions fixes par rapport à la caméra ni sont imposés de maintenir des poses ou expressions spécifiques du visage au moment requis. Ainsi, les personnes peuvent se déplacer dans la scène d'acquisition et, pour cette raison, la majorité des données collectées de ce type sont sous forme vidéo. La notion ajoutée de mouvement apporte dans ce type d'environnement les effets d'échelle variante et de flou de motion entre les images qui n'étaient pas présentes au cas précédent. Les effets d'illumination et de pose sont notamment plus complexes ici que dans le cas contrôlé, car ils dépendent de la position à chaque instant de l'individu dans la scène, ce qui est plus difficilement prévisible. Par contre, les environnements semi-contrôlés se distinguent des environnements non-contrôlés dans le sens qu'ils forcent les individus à passer à certains points de contrôle spécifique, ce qui vient aider à limiter les variations majeures observables sur les visages dans la scène. Les cas les plus typiques d'environnements semi-contrôlés sont les portails, les corridors ou les points d'inspection dans les aéroports. De bons exemples de ce type d'environnement sont représentés à la Figure 1.7, où on peut y voir le visage d'un même individu qui affiche plusieurs poses, illuminations, niveaux d'échelle et flous d'image selon le cas, mais toujours en conservant un visage relativement visible et orienté de sorte que la caméra puisse le capturer.

Figure 1.7 Illustration de diverses conditions d'acquisition semi-contrôlées observées sur un visage à travers différentes séquences vidéo et caméras. Séquences P1E_S1_C1 et P1E_S3_C1, individu ID0003 de la base de données ChokePoint [139, 140]

La Figure 1.8 suivante permet de mettre en contexte le type de portail décrit précédemment pour un environnement semi-contrôlé, où on peut y voir que l'angle de pose, l'illumination et la dimension du visage varient grandement d'un moment à l'autre, mais dont la trajectoire de l'individu est relativement imposée et prévisible par la contrainte même du portail.

Figure 1.8 Démonstration d'un portail limitant la trajectoire possible d'un individu en environnement semi-contrôlé. ID0003, séquence P1E_S3_C1 de ChokePoint [139, 140]

Finalement, les environnements non-contrôlés, aussi appelés sans-contraintes (*unconstrained environment*), sont caractérisés par l'ensemble des problématiques d'acquisition précédentes en cas semi-contrôlé, mais les individus y sont libres de se déplacer comme bon leur semble. Étant donné qu'il n'y a aucune limitation par rapport la direction, le sens et la manière générale de se déplacer, la position des visages est pratiquement imprévisible. Dans ce genre d'environnement, non seulement nous ne pouvons pas prévoir la position de l'individu incluant la pose et l'illumination sur son visage, mais nous ne pouvons non plus être certain que son visage sera même capturé par la caméra. En effet, si la tête de la personne est tournée à l'opposé du capteur, où qu'une autre personne ou un véhicule passe en avant-plan de l'individu recherché, son visage ne sera simplement jamais capturé. Évidemment, ce type d'environnement est le plus complexe étant donné qu'il y a d'une partie l'ajout d'occlusions possibles, mais également le fait que la scène ne sera jamais la même d'une caméra à l'autre. Ainsi, tout environnement ne respectant pas les conditions des deux cas précédents est habituellement rapatrié dans cette catégorie. Quelques exemples de cet environnement type constituent, entre autres, les caméras situées dans les parcs, les centres d'achat ou presque n'importe quel autre lieu public où une foule peut s'y former.

Afin de mieux comprendre la difficulté engendrée par ce dernier genre d'environnement lors de l'acquisition de données, la Figure 1.9 est offerte à titre d'exemple. Celle-ci démontre de manière non exhaustive les diverses conditions variables d'acquisition en milieu sans-contraintes qui peuvent être obtenues à partir de multiples caméras à différents points de vue, angle d'acquisition, moments de la journée, conditions météorologiques et résolution d'image. Nous pouvons aussi y voir que les scènes sont aussi généralement plus achalandées que par rapport aux autres types d'environnement, ce qui complexifie davantage le problème de FR relié à ces situations. En considérant toutes les variables impliquées dans ce dernier type d'environnement, les algorithmes de FR doivent absolument être robustes à un ensemble de paramètres difficiles à prévoir. Ainsi, ces derniers cas démontrent clairement pourquoi le domaine de la FRiVS demeure irrésolu à ce jour.

Figure 1.9 Illustration de diverses conditions d'acquisition difficiles en environnement sans-contraintes[2]

[2] Tiré de diverses vidéos publiques venant simuler des séquences de vidéosurveillance, site en ligne ShutterStock, https://www.shutterstock.com/video, recherche par mots-clés connexes à « security, camera, footage, crowd ».

1.2.2 Images, vidéos et bases de données

Dans le cas d'images statiques employées pour la FR, il arrive souvent que les images produites à travers les bases de données publiques correspondent à des images de haute qualité. La raison pour cela est que plusieurs travaux en littérature se concentrent particulièrement sur la reconnaissance d'expressions sur le visage, ce qui requiert habituellement un niveau plus élevé de qualité et une bonne résolution. D'un autre côté, certaines bases de données contiennent beaucoup de variation de profils variés du visage (à plusieurs angles), afin d'étudier les effets de la pose sur les modèles de visages appris. Peu importe le domaine d'étude visé, beaucoup des bases de données à images statiques sont acquises lors de sessions photographiques en rafale avec des centaines, voire des milliers, de sujets. Cela a pour effet de produire des regroupements d'images plus régulières d'un individu à l'autre, s'appartenant à un environnement contrôlé, car elles sont acquises selon des conditions de capture semblables où les individus viennent à tour de rôle prendre leurs quelques acquisitions d'image devant une caméra statique et un arrière-plan bien illuminé. Les travaux dans [122] présentent un bon effort pour lister certaines des bases de données disponibles, tel qu'illustré au Tableau 1.1, mais la liste demeure toujours incomplète et toujours grandissante avec la croissance fulgurante de la disponibilité de données ces dernières années.

Tableau 1.1 Quelques bases de données publiques pour la FR [122]

Research pointers	
Face recognition homepage	www.cs.rug.nl/~peterkr/FACE/frhp.html
Face detection homepage	home.t-online.de/home/Robert.Frischholz/face.htm
Facial analysis homepage	mambo.ucsc.edu/psl/fanl.html
Facial animiation homepage	mambo.ucsc.edu/psl/fan.html
Face databases	
FERET database	http://www.itl.nist.gov/iad/humanid/feret/
XM2TVS database	http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/
UT Dallas database	http://www.utdallas.edu/dept/bbs/FACULTY_PAGES/otoole/database.htm
Notre Dame database	http://www.nd.edu/~cvrl/HID-data.html
MIT face databases	ftp://whitechapel.media.mit.edu/pub/images/
Shimon Edelman's face database	ftp://ftp.wisdom.weizmann.ac.il/pub/FaceBase/
CMU face detection database	www.ius.cs.cmu.edu/IUS/dylan_usr0/har/faces/test/
CMU PIE database	www.ri.cmu.edu/projects/project_418.html
Stirling face database	pics.psych.stir.ac.uk
M2VTS multimodal database	www.tele.ucl.ac.be/M2VTS/
Yale face database	cvc.yale.edu/projects/yalefaces/yalefaces.html
Yale face database B	cvc.yale.edu/projects/yalefacesB/yalefacesB.html
Harvard face database	hrl.harvard.edu/pub/faces
Weizmann face database	www.wisdom.weizmann.ac.il/~yael/
UMIST face database	images.ee.umist.ac.uk/danny/database.html
Purdue University face database	rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html
Olivetti face database	www.cam-orl.co.uk/facedatabase.html
Oulu physics-based face database	www.ee.oulu.fi/research/imag/color/pbfd.html

Il existe évidemment des exceptions aux bases de données d'images statiques à haute résolution, comme la base de données LFW [63] qui offre des images dans un environnement non-contrôlé, afin d'étudier spécifiquement ce type de conditions, mais il demeure que les exemples fournis sont disponibles sous forme d'images qui ne contiennent pas nécessairement toutes les difficultés retrouvées sur vidéos.

D'un autre côté, certaines bases de données s'adressent très spécifiquement au problème de la FRiVS, ce qui permet d'obtenir des données publiques sous forme de séquences vidéo dans des environnements majoritairement semi-contrôlés. En effet, malgré que plusieurs bases de données, comme ChokePoint [139, 140], COX-S2V [64, 65], SCface [58], et plusieurs autres, ont été conçues particulièrement pour la vidéosurveillance, et encore plus, pour le cas de SSPP (*voir la section 1.4, Défi de l'image unique de référence*), celles-ci demeurent limités à des cas semi-contrôlés plutôt que non-contrôlés. Cela est d'ailleurs compréhensible, étant donné la difficulté associée à l'étiquetage d'individus bien spécifiques à travers plusieurs acquisitions vidéo en milieu sans-contraintes, de sorte à permettre d'évaluer les algorithmes de FR.

Il existe bien évidemment des bases de données avec séquences vidéo à travers des environnements totalement sans-contraintes. Pour cela, il suffit de penser aux multiples défis de suivis de trajectoires comme VOT (*Visual Object Tracking Challenge*) [1, 73, 74, 75, 76] et MOT (*Multiple Object Tracking Challenge*) [82] qui offrent une abondance de vidéos pour tester de nouveaux algorithmes. Par contre, comme leurs noms le suggèrent, ceux-ci ont pour objectif le suivi d'objets de toute sorte, et ne possèdent donc pas nécessairement d'étiquettes utiles pour l'évaluation de la reconnaissance d'individus et de visages spécifiquement. Parmi les bases de données en milieu sans-contraintes qui contiennent des étiquettes de visages, on retrouve YouTube Face DB [136, 137]. Par contre, cette dernière source manque souvent de références pour comparer les séquences observées aux individus recherchés, ou ne possède simplement aucune référence. Certains cas peuvent y être inférés étant donné que les personnes observées sont des célébrités, mais le manque d'étiquettes formelles rend la base inadéquate particulièrement au cas de la FRiVS, mais davantage pour la détection de visages en général.

Pour toutes les raisons mentionnées aux paragraphes précédents, la majorité des travaux en littérature, faisant partie de l'état-de-l'art pour la FRiVS, se tournent vers les bases de données en environnement semi-contrôlé, car il est nécessaire d'avoir des références étiquetées pour identifier et vérifier la reconnaissance des individus observés sur vidéos. Ceci est le cas malgré que la vidéosurveillance soit en effet appliquée autant aux environnements semi- que non-contrôlés sur le terrain. Ainsi, beaucoup algorithmes se basent actuellement sur l'hypothèse que la majorité des points de contrôle où une FR robuste est le plus nécessaire sont, en effet, là où l'on tente de contrôler autant que possible les entrées et sorties des individus, comme aux douanes et aux points de passage des frontières dans les aéroports. D'ailleurs, cette hypothèse peut être partiellement justifiée lorsque l'on considère que, lorsqu'un système de vidéo-caméras est mis en service pour la sécurité, celui-ci est toujours finement ajusté par rapport au point de vue et de la plage focale d'acquisition des caméras. Cela est accompli de sorte à capturer les endroits plus simples à analyser, comme les corridors, les portails, les files d'attente ou tout autre emplacement le plus propice à fournir aux caméras un maximum d'apparitions frontales et claires du visage, afin de s'assurer de ne pas le manquer.

1.3 Types de scénarios sur vidéos

Tel que présenté aux sections précédentes, les données de visages disponibles sont présentes autant sous forme d'images statiques que de séquences vidéo. En ce qui concerne les systèmes de FRiVS, une distinction supplémentaire doit être effectuée par rapport aux données employées, c'est-à-dire le type de données lorsque la requête (*query set*) est effectuée, en plus du type visé (*target set*) pour la reconnaissance. Le Tableau 1.2 présente les cas à considérer.

Tableau 1.2 Types de données selon le scénario en reconnaissance de visage pour la vidéosurveillance [64]

Target set \ Query set	Still images	Video sequences
Still images	/	Video-to-Still (V2S)
Video sequences	Still-to-Video (S2V)	Video-to-Video (V2V)

En effet, alors que les séquences vidéo sont presque toujours employées à cause des systèmes de caméra de sécurité distribués, il demeure possible d'exploiter les images statiques simultanément. L'ensemble de requêtes (*query set*) consiste en toutes les images ou vidéos, selon le cas, qui sont utilisées pour spécifier au système de FR, quels individus reconnaître, vérifier la présence, ou identifier à travers une base de données de sécurité précédemment acquise et sauvegardée, ou à travers des caméras qui filment en temps réel. L'ensemble visé (*target set*) est quant à lui cette dernière source où les individus en requête seront recherchés et comparés aux visages trouvés afin d'évaluer la reconnaissance de visage. Autrement dit, les données employées à la requête sont celles exploitées pour l'entraînement de systèmes et l'enregistrement des individus d'intérêt (*enrollment phase*) dans une liste de contrôle (*watchlist*), alors que les données visées sont celles utilisées pour tester de nouveaux visages (*probes*) contre ces systèmes entraînés, lors de leur mise en opération (*operational phase*). Les sous-sections qui suivent présenteront chacun des cas énoncés au Tableau 1.2.

1.3.1 Video-to-Still (V2S)

Le cas *video-to-still* représente la situation où un individu capté sur séquence vidéo doit être retrouvé à travers des bases de données de photos d'identification, par exemple, à partir de l'image retrouvée sur un permis de conduire ou contenue sur des documents officiels de tout genre. Étant donné que la source de requête est une séquence vidéo, il est habituellement possible d'obtenir plusieurs variations du visage d'un même individu afin de le reconnaître sur les images comparées. La qualité des séquences vidéo dépendent de leur source, et peuvent donc autant être de bonne ou moins bonne qualité et résolution, selon l'ancienneté des systèmes de caméra, ou si par exemple, si une personne possède une évidence d'une scène de crime filmée avec un téléphone mobile récent. Les images comparées sont majoritairement à haute résolution puisqu'elles se retrouvent le plus souvent dans des archives officielles, comme des données sur les criminels ou les sources d'agences gouvernementales. Un cas typique d'utilisation de ce mode est ainsi de rechercher un fugitif capturé sur vidéo ayant fui une scène de crime, et dont l'identité doit être déterminée ou retrouvée à partir des bases d'images existantes.

1.3.2 Video-to-Video (V2V)

Le mode d'opération *video-to-video* utilise une ou plusieurs séquences vidéo comme référence de recherche, tout comme le cas précédent, mais la recherche est effectuée à travers d'autres vidéos plutôt que des images. Encore une fois, la qualité des vidéos peut dépendre de leur source, et peut également ne pas concorder entre la requête et les données visées. Un défi majeur de ce mode est donc de faire correspondre les données sur le point de vue de la résolution, le taux de trame et les autres facteurs de perturbations (flou de motion, échelle, illumination et pose) causés par des caméras disposées différemment pour mieux les comparer. Un cas typique d'utilisation est de reconstruire une trajectoire complexe d'un individu à travers différentes scènes captées par multiples caméras d'un système de sécurité distribué, afin de suivre sa trace malgré les occlusions. Par exemple, si une caméra capte un individu alors qu'il transite le long d'une rue, et que celui-ci tourne le coin menant vers une autre rue causant la caméra initiale de le perdre de vue, avec ce mode, on pourrait tenter d'établir la correspondance avec une autre caméra située sur l'autre coin de rue pour continuer le suivi de trajectoire lorsque l'individu revient dans le champ de vision.

1.3.3 Still-to-Video (S2V)

Le dernier mode *still-to-video,* fonctionne quant à lui à l'inverse du premier cas présenté. Ainsi, les images statiques sont employées afin d'indiquer au système de FR quels individus reconnaître à travers des séquences vidéo. Les modalités par rapport à la qualité et la résolution des données sont en tout point identiques au premier mode. Par contre, ce dernier cas d'opération agit habituellement de manière plus proactive que le premier mode qui était réactif. En effet, alors que le mode V2S tente souvent de retrouver et identifier un individu après-fait, la nature d'utilisation des données en S2V fait qu'il est habituellement appliqué pour prévenir des situations en vérifiant le nouveau passage d'un individu dans une scène. Par exemple, l'une des utilisations les plus traditionnelles de ce mode consiste à inscrire dans une liste de contrôle des individus marqués comme potentiellement dangereux ou recherchés, et de les repérer rapidement sur les systèmes de sécurité par caméras, alors qu'ils tentent de s'infiltrer ou de s'échapper de là où nous désirons les en empêcher.

1.4 Défi de l'image unique de référence (SSPP)

L'une des situations très particulière et réaliste de l'application de la FRiVS est le cas du *single sample per person* (SSPP). Ce cas correspond au mode d'utilisation *still-to-video*, présenté plus tôt, spécifiquement là où une seule image de référence est disponible afin d'enregistrer chacun des individus d'intérêt dans la base de référence employée pour effectuer un contrôle de passage (*watchlist screening*). Ainsi, contrairement au cas générique de S2V où plusieurs images d'un même individu pourraient être employées pour entraîner le système de FR à le reconnaître, cette façon de procéder suppose que l'on ne possède qu'une seule référence, souvent de type « *mug-shot* » représentant une image de haute qualité, et acquise dans un environnement extrêmement bien contrôlé. Cette application spécifique apporte plusieurs difficultés au problème de FR déjà relativement complexe.

Une première difficulté rencontrée en SSPP est tout d'abord que l'apparence particulière du visage d'un individu sur l'image d'enregistrement ne concorde en aucun point aux observations effectuées sur séquences vidéo, tel que présenté à la Figure 1.10. En effet, nous pouvons observer avec cette figure que l'image de référence est d'extrêmement haute qualité et possède des détails qui ne pourront jamais être approchés sur la séquence vidéo provenant de la caméra de surveillance, en plus des variations de pose et d'illumination différentes.

Figure 1.10 Comparaison des conditions d'acquisition variables entre l'image *still* d'entraînement (gauche, 2672x4000 px, ratio ~1/15), d'un agrandissement du visage d'intérêt (centre, ratio ~1/3) et d'une trame de caméra de vidéosurveillance (droite, 800x600 px) de l'individu ID0001, séquence P1L_S1_C3 de la base de données ChokePoint [139, 140]

Étant donné que la résolution très fine des images de références des individus d'intérêt (*target*) permet de capter des détails du visage qui ne sont ni même visibles sur les acquisitions vidéo, pratiquement toujours de nettement plus faible résolution, il est nécessaire, lors de la conception d'un système de FR en SSPP, de considérer le fait que les visages observés ne posséderont pas exactement les mêmes signatures. Cela est le cas, peu importe le type de descripteur employé pour représenter les visages en mémoire, puisque ceux-ci sont fondamentalement différents. De plus, certaines poses et formes d'illuminations complètement distinctes et nouvelles seront nécessairement observées, étant donné que la seule référence est frontale et à expression neutre, alors que les caméras produisent très rarement de telles apparences du visage. Cela est produit par le fait qu'elles sont habituellement situées en hauteur et de sorte à ne pas obstruer les entrées, ce qui crée sans surprise des angles de vue de profil sur les visages. Puisque les individus peuvent aussi afficher n'importe quelle expression désirée durant la capture vidéo, cela fait que les apparences de visages y seront également modifiées par rapport à la référence neutre frontale. Ensuite, il faut considérer l'ajout de flou de mouvement qui ne peut aucunement être observé sur l'image unique de référence, et qui est difficilement prévisible en se basant également sur cette unique référence (*i.e.* : un défilement de plusieurs poses de visages de référence pourrait être utile pour accomplir ceci si elles étaient disponibles, mais ce n'est à la base pas le cas en SSPP).

Suite aux problèmes précédents s'ajoutent les difficultés associées au déséquilibre entre les données d'entraînement ainsi que les limitations que celles-ci amènent face aux algorithmes de classification applicables [122]. En effet, étant donné que seulement un exemple de l'individu *target* n'est disponible pour l'entraînement, tout classificateur de visages basé sur celui-ci comme classe positive, contre l'ensemble des autres individus existants *non-target* comme classe négative, sera extrêmement affecté par le manque de données du côté de la classe positive. Plusieurs des méthodes de classification existantes sont d'ailleurs tellement affectées par un tel déséquilibre, quelles ne produiront que des réponses négatives lorsque testées avec des visages sur vidéos, même lorsque ce visage de l'individu voulu se rapproche relativement beaucoup de l'exemple initial sur l'image *still*. L'importance trop marquée de la classe négative viendrait simplement induire le classificateur en erreur dans tous les cas.

Encore plus, la limitation d'un unique échantillon positif fait que certains classificateurs ne sont littéralement pas applicables à la FR en SSPP, ou très difficilement sans modification majeure, car leur méthode d'entraînement même nécessite beaucoup d'exemples pour capturer toute la complexité des variations observables sur le visage pour permettre de bien classifier un individu donné malgré les multiples facteurs d'acquisition imprévisibles. Il ne suffit qu'à penser aux méthodes de *deep learning* traditionnelles qui requièrent autant d'exemples que possible pour ajuster leurs pondérations internes. En SSPP, des approches différentes ou des subtilités judicieuses doivent être appliquées, tel que le *data augmentation*, pour combattre les désavantages apportés par le manque de données d'entraînement de chaque individu d'intérêt. Certaines de ces techniques sont employées pour ce projet, et y seront donc discutées plus loin.

Pour finir, les difficultés ardues rencontrées dans le contexte du SSPP peuvent être d'autant plus mises en emphase lorsque l'on compare cette application spécifique du S2V avec les deux autres modes explorés à la section précédente. Effectivement, contrairement aux S2V, le mode V2S permet un accès à au minimum une séquence vidéo complète ou partielle, et, même dans le cas où seulement une très courte séquence de quelques secondes n'est disponible, étant donné la vitesse d'acquisition habituellement d'au moins 30FPS des caméras de sécurité, cela permet d'avoir un bon ensemble de représentations du visage caractérisant l'individu à reconnaître. L'ajout de représentations supplémentaires avec variations de pose, d'illumination et autres facteurs d'acquisition vient donc grandement aider à l'entraînement de classificateurs dans un tel cas, car les variations intra-classes peuvent être évaluées et considérées durant l'entraînement à l'aide de ces multiples exemples de visages. Alors qu'autant le S2V que le V2S sont affectés par l'effet de transfert de domaine marqué entre les d'images statiques et les vidéos, ce qui rend la correspondance des visages entre les deux environnements très faibles, le mode V2V souffre notablement moins de cet effet, car les sources sont semblables. Évidemment, les domaines *video-to-video* ne sont pas parfaitement correspondants, car les scènes visualisées peuvent différer, mais il n'en est pas moindre que la variation de résolution y sera caractéristiquement moins prononcée. Basé sur les points précédents, le contexte du S2V en SSPP considère donc sans doute l'une des combinaisons de difficultés et d'applications en FRiVS les plus difficiles, sinon la plus complexe de toutes, à résoudre.

1.5 L'information variable à travers les trames vidéo

Une notion importante à considérer lorsque des séquences vidéo sont exploitées en FR est le fait que l'information contenue à travers le groupe d'images séquentiellement logiques et continues évolue spatio-temporellement. Ainsi, une possibilité de combinaison de l'information, ou à l'inverse, un affranchissement de cette modification dans le temps et l'espace, est nécessairement impliquée en FRiVS. Selon les travaux présentés dans [49], il existe trois approches distinctes permettant de caractériser l'exploitation de l'information contenue à travers une séquence de *frames* vidéo, soit :

1. Évaluer la qualité visuelle des multiples images formant la séquence vidéo afin d'en extraire les quelques instants particuliers (*key-frames*) sélectionnés là où les visages possèdent les meilleures représentations pour produire le maximum de bonnes reconnaissances possible seulement à partir de ces cas optimaux.
2. Utiliser l'information redondante qui se reflète à travers un ensemble d'images consécutives, comme le mouvement transitoire de la trajectoire de l'individu, le flux de mouvement général, la transition d'illumination ou la variation de pose graduelle, plutôt que de manière brusque lorsqu'observées à des temps plus éloignés, afin de produire une décision qui se base sur une confidence progressive dans le temps.
3. Déterminer les caractéristiques descriptives peu ou complètement invariantes qui ressortent et se démarquent à travers les images de la séquence observée, afin d'y devenir robuste et de limiter l'impact de variation entre chacune des images.

Dans le premier cas, la méthode dépend grandement de la capacité à déterminer quelles sont les meilleures images disponibles dans la séquence. Cela est un premier désavantage majeur de cette approche parce qu'il peut s'avérer extrêmement difficile de déterminer un indice de qualité qui puisse trouver de manière efficace quelles images sont en effet les meilleures. D'autant plus, il est généralement très complexe d'établir quelles sont les meilleures représentations auxquelles le classificateur de visages employé s'attend, plutôt que seulement ce qui « paraît » comme les meilleures images visuellement. Cette difficulté est encore plus

rehaussée avec, par exemple, les systèmes en apprentissage profond de plus en plus exploités, et dont la compréhension à bas niveau demeure encore loin d'être comprise dans les fins détails. Un second désavantage de cette approche est qu'elle dépend de la présence même d'une telle « image idéale » à travers la séquence vidéo, mais il peut facilement s'avérer (et c'est souvent le cas sur des séquences en vidéosurveillance) qu'aucune des images dans la séquence ne soit de bien bonne qualité. Pour la FRiVS, il arrive en effet souvent que le nombre de bonnes détections de visages ne soit limité qu'à une infime quantité d'instants particuliers et, avec aussi peu d'échantillons, on ne peut se permettre d'être trop sélectif.

La seconde technique exploite la duplication d'information répétée dans le temps et l'espace afin de développer graduellement une confidence à propos de la reconnaissance de visage sur vidéo. L'approche en soit est sensée étant donné que, plus la réponse d'un algorithme est forte, plus l'on considère habituellement que la fiabilité de sa reconnaissance associée est élevée. Ainsi, dans les cas où il pourrait y avoir des variations spontanées de faibles reconnaissances de visage d'un individu visé, ou à l'inverse de fortes reconnaissances erronées correspondant à d'autres personnes que celle recherchée, l'utilisation d'une confidence s'étalant sur plus d'un échantillon rend nécessairement la FR moins sensible aux instants spécifiquement problématiques et imprévisibles en vidéosurveillance. Donc, une telle approche peut sembler efficace autant lorsque les cas problématiques sont dus à des erreurs de détection, du désalignement du visage, des changements de pose ou d'illumination soudains, ou toute autre forme de perturbation, car la redondance d'information vient agir de manière naturelle comme un filtre passe-bas pour adoucir ces changements brusques. Par contre, il est important de constater également que cet effet d'adoucissement peut devenir problématique et contre-productif en faussant les résultats si la réponse du classificateur de visage employé est trop peu discriminante ou est même complètement erronée. Ainsi, cette technique possède un avantage marqué qui ne la rend pas dépendante de certains instants spécifiques sur la vidéo, comme la méthode précédente, mais il ne demeure pas moins que la réponse de la reconnaissance de visage doit rester à la base tout de même robuste et assez forte pour produire une réponse suffisamment discriminante entre les individus.

Le dernier ensemble de méthodes considéré tente, contrairement aux deux autres, de limiter les effets nuisibles visibles à travers les trames vidéo en FR grâce à l'acquisition d'informations pertinentes par une sélection de descripteurs de visages spatio-temporellement discriminants par rapport aux facteurs problématiques traditionnellement observés avec les caméras de surveillance. Une telle approche nécessite donc de produire des algorithmes permettant d'extraire les caractéristiques importantes qui représentent correctement et distinctement le visage de divers individus, sans pour autant changer lorsqu'affectée des variations de pose, d'illumination, de flou ou autres conditions de capture. Malgré que la conception d'un tel algorithme puisse paraître fondamentalement complexe, plusieurs des méthodes existantes pour l'extraction de caractéristiques ont été réalisées spécifiquement à cause de ces besoins marqués. Auparavant, bonne majorité de descripteurs comme LBP, HOG, LPQ, SIFT, etc., que pour en nommer certains, ainsi que leurs multiples variations, ont été appliqués au domaine de la FR afin de réduire les effets de l'illumination, du flou de motion en vidéo ou autres particularités quelconques qui produisent des variations indésirables sur les visages malgré qu'appartenant aux mêmes individus [5, 6, 13, 15, 53]. Ainsi, ces techniques correspondent davantage à des approches conceptualisées à partir de bases en imagerie ou statistiques, parfois interpellées méthodes *handcrafted*, et elles visent des problématiques bien particulières. Plus récemment, des approches d'extraction de caractéristiques générées automatiquement à l'aide d'apprentissage par CNN ont également été exploitées pour trouver de tels descripteurs invariables aux situations néfastes causées par les vidéos [101, 102, 117]. Une autre façon de procédé qui est également d'employée afin de devenir robuste aux variations sur vidéos est d'utiliser des systèmes adaptatifs aux nouvelles apparitions de visages, tel qu'élaboré dans [43, 45, 98, 143], de sorte qu'après un certain temps d'opération, les modèles ou descripteurs utilisés deviennent invariants aux nouvelles données vidéo en ayant adéquatement appris à modéliser et comprendre ces effets variables.

Ainsi, nous pouvons voir qu'il existe définitivement un grand nombre d'approches pour pallier aux conditions rencontrées en FRiVS, et particulièrement sur les séquences vidéo associées, mais il est unanimement nécessaire avec celles-ci de se préoccuper des facteurs d'acquisition et de variations d'apparence du visage qui demeureront toujours des embûches présentes.

1.6 Architecture de base d'un système de reconnaissance de visages

L'ensemble des sections précédentes démontrent clairement à quel point il y possible d'obtenir des systèmes de FR extrêmement variés, malgré que ceux-ci accomplissent tous un objectif de nature relativement similaire. Il en est de même pour les architectures permettant de représenter le séquencement des étapes principales employées en reconnaissance de visages. Dit autrement, malgré que les algorithmes utilisés aux plus bas niveaux de chaque application de FR puissent grandement varier d'un cas à l'autre, en considérant non seulement tous les types de catégories présentées jusqu'à présent, mais également les modes spécifiques d'opérations envisageables, l'architecture à plus haut niveau est en tout point similaire d'un cas à l'autre. La mise en place un système de FR complet requiert trois étapes majeures, soit (1) la détection de visage (*face detection*, FD) à partir des images ou séquences vidéo, (2) l'extraction de caractéristiques (*feature extraction*, FE) permettant de décrire et encoder l'apparence d'un visage sous forme numérique ou binaire, et finalement, (3) la classification de visages (*face classification*, FC) à proprement parlé pour accomplir la FR qui permet d'identifier ou vérifier la présence d'un individu dans scène en entrée [8, 122, 150]. La Figure 1.11 représente cette architecture sous sa forme la plus traditionnelle et simplifiée.

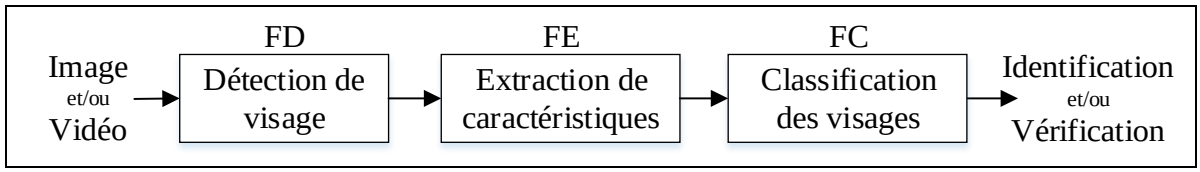

Figure 1.11 Architecture fondamentale des systèmes de reconnaissance de visages

Il arrive que nous retrouvions également une légère variante du schéma précédent dans la littérature, afin de mettre l'emphase sur un module de suivi de trajectoire particulièrement utile pour considérer l'information spatio-temporelle ajoutée aux vidéos, lorsqu'utilisés en entrée, tel que discuté à une section précédente. Cette forme alternative, malgré que très similaire, indique habituellement que le système contient une technique légèrement plus avancée d'association de l'information à travers les trames du vidéo qu'une simple détection à chaque image. L'indication FD est donc souvent remplacée à la Figure 1.11 par (FD+FT) signifiant

que le système emploie les deux méthodes simultanément, ou uniquement FT lorsqu'il fait simplement usage du suivi de trajectoire du visage (*face tracking*, FT).

Les systèmes de FR fonctionnement au minimum en deux étapes, soit la phase d'entraînement (*enrollment phase*) et la phase d'opération (*operational phase*). La première étape consiste à préparer des modèles de visage permettant de discriminer les individus à reconnaître entre eux lors de la seconde étape. Ainsi, la phase initiale effectue d'abord une détection de visage afin d'obtenir des régions d'intérêt (ROI) qui précisent l'emplacement du visage recherché à travers l'intégralité d'une image ou trame de vidéo. Évidemment, lorsqu'il s'agit de représentation globale du visage, la ROI englobe l'ensemble du visage (ou presque selon le type de FD), alors que les approches locales obtiennent les sections du visage spécifiquement à leur application. Ensuite, les descripteurs $\mathbf{t}_k$ représentant les visages détectés y sont extraits, afin d'entraîner des classificateurs de visages performants pour discriminer entre les visages d'individus à reconnaître. Les modèles caractérisant les individus d'intérêt obtenus de cet entraînement sont alors inscrits dans une base de référence intitulée *gallery*. Celle-ci deviendra la base à laquelle les nouvelles représentations de visages encore non identifiés $\mathbf{a}_j$ seront comparés et testés lors de la phase d'opération. Ainsi, lorsque l'on passe à la deuxième phase, les nouveaux visages captés en entrée subissent le même procédé de détection de visage et d'extraction de caractéristiques, de sorte à les comparer selon la même forme que sauvegardés préalablement à la première phase. Finalement, tel qu'illustré à la Figure 1.12, la classification des nouveaux visages contre les modèles de la *gallery* produit des scores de classification qui peuvent être seuillés, afin de prendre une décision finale sur la reconnaissance, la vérification ou l'identification d'un individu particulier en fonction du visage observé dans la scène.

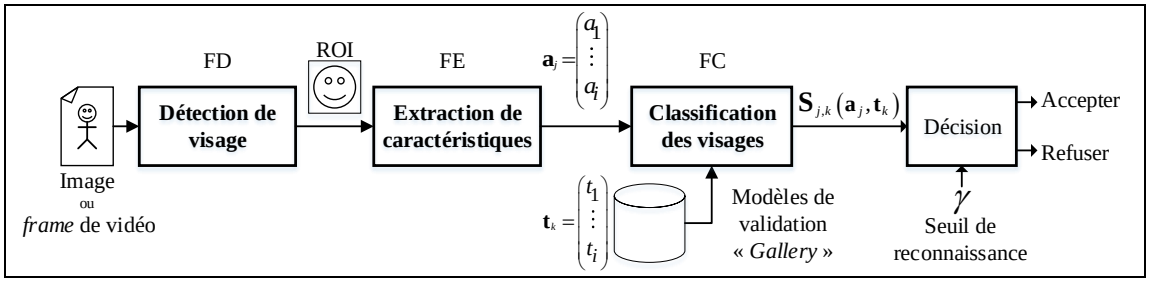

Figure 1.12 Schéma-bloc plus détaillé d'un système de reconnaissance de visages

CHAPITRE 2

MODULES D'UN SYSTÈME DE FRiVS ET MÉTHODES CONSIDÉRÉES

Ce chapitre présentera une vue d'ensemble sur les multiples modules qui composent un système de FRiVS spécifiquement pour le cas de S2V-FR entraîné par SSPP. Certains algorithmes qui ont pu avoir été considérés lors de la réalisation du projet, ainsi que quelques justifications des choix, seront discutés lorsqu'opportuns. Il faut noter que toutes les méthodes ne sont pas employées dans le résultat final, mais seulement les techniques jugées comme plus adéquates selon la problématique visée et les défis rencontrés. Ainsi, cette section donnera un aperçu de méthodes existantes, mais ne sera en aucun point une liste exhaustive de toutes les possibilités associées à chacune des catégories explorées. Les lecteurs seront donc invités à se référer aux références bibliographiques ou autres publications en littérature afin d'obtenir plus de détails sur les technologies et algorithmes existants qui pourraient les intéresser davantage.

2.1 Détection de visages

Il existe un très grand nombre de méthodes permettant d'accomplir la première étape d'un système de FR, c'est-à-dire la détection de la position de régions d'intérêt des visages sur les images ou vidéos. La technique la plus répandue est sans doute l'algorithme proposé par les auteurs Viola-Jones [129, 130, 131], portant le même nom. La détection VJ consiste à appliquer plusieurs faibles classificateurs en cascade (d'où le nom alternatif *cascade classifier* de l'algorithme) qui ont été entraînés à progressivement valider avec de plus en plus de rigueur la présence de caractéristiques Haar bien précises et qui représentent la réponse typique d'un visage. Des exemples de ce type de caractéristiques Haar sont présentés à la Figure 2.1.

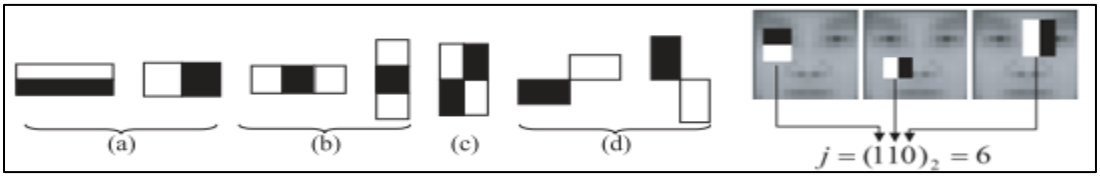

Figure 2.1 Représentations de caractéristiques Haar employées par Viola-Jones [84]

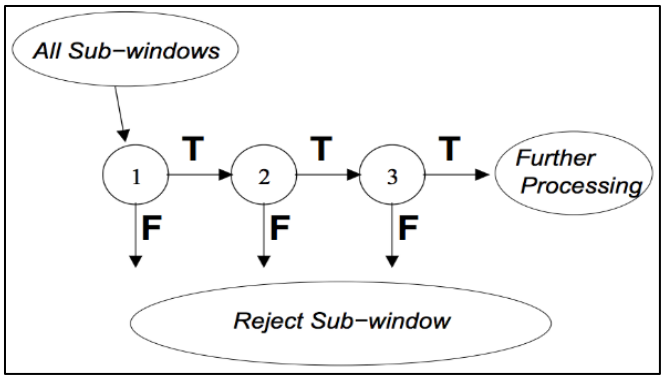

Figure 2.2 Classificateurs faibles en cascade de l'algorithme Viola-Jones [130]

Le concept derrière l'architecture en cascade du modèle VJ, tel qu'illustré à la Figure 2.2, permet d'atteindre des temps de détection se « *rapprochant* » du temps réel (en considérant que les images sont de plus haute résolution à ce jour qu'à l'époque de la publication originale), puisqu'il est possible de rapidement éliminer les cas de non-visages assurés dès les quelques premiers stages de l'architecture. Ainsi, seulement les objets plus complexes et se rapprochant grandement de la signature d'un visage réussissent à atteindre les derniers stages en cascade, ce qui fait que très peu d'évaluations approfondies sont requises pour valider la présence ou non d'un visage sur une image. Les cas nécessitants plus d'attention et qui passent à travers tous les stages sont quant à eux considérés comme des visages. L'approche permet donc de combiner des réponses rapides à un niveau de robustesse impressionnant. En rejetant hâtivement, aucun temps précieux de traitement n'est perdu sur l'évaluation de cas évidents de non-visages, ce qui permet à la technique d'accorder plus de temps aux cas difficiles. Cela a également pour avantage de permettre des détections multi-échelles sur les images sans impact trop prononcé sur le temps de réponse, renforçant alors les détections par contre-validations, ainsi que permettant de retrouver les ROI de visage en fonction de diverses distances de l'individu par rapport aux caméras.

Étant donné que la méthode VJ existe depuis de nombreuses années, l'algorithme profite de plusieurs sources de code optimisées sur GPU ou en parallèle, comme offert avec la bibliothèque OpenCV [22], ainsi que des versions alternatives utilisant, par exemple, des

caractéristiques LBP [107] pour encore plus de rapidité et de précision ou, encore, qui supportent les visages de profil [21, 22], contrairement aux modèles initiaux qui était uniquement pour le quasi-frontal.

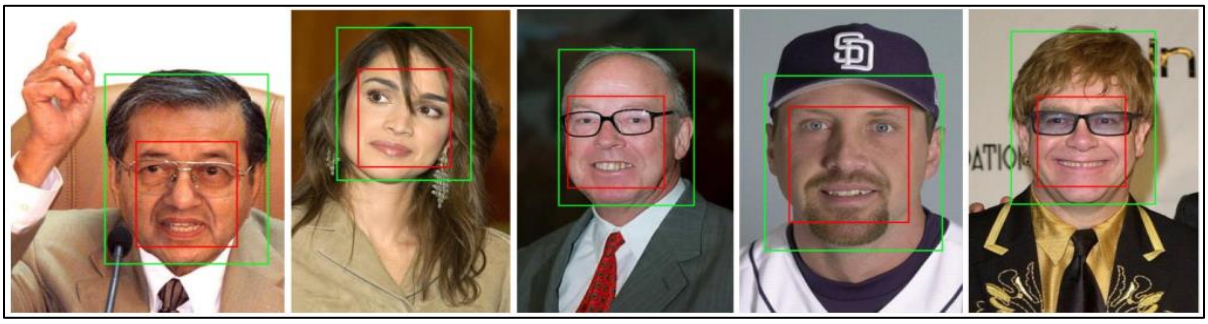

Figure 2.3 Détections de visages obtenues par Viola-Jones (vert), et de son alternative par caractéristiques LBP [107] (rouge), toutes deux disponibles dans OpenCV [22]

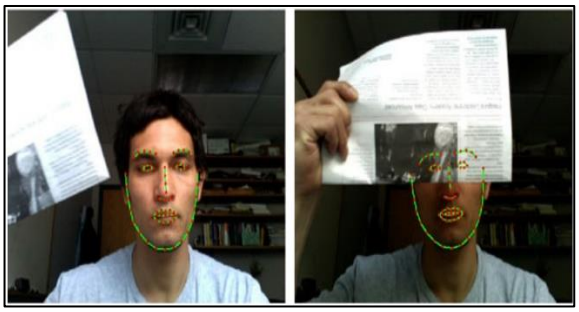

Figure 2.4 Cas complexe de visage capté par DRMF [115], malgré l'occlusion prononcée

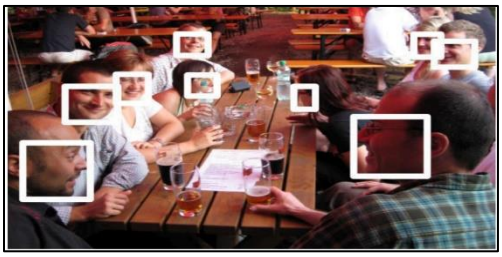

Figure 2.5 Profils de visage complexes pour l'application de la FR, mais trouvés par *HeadHunter* [80]

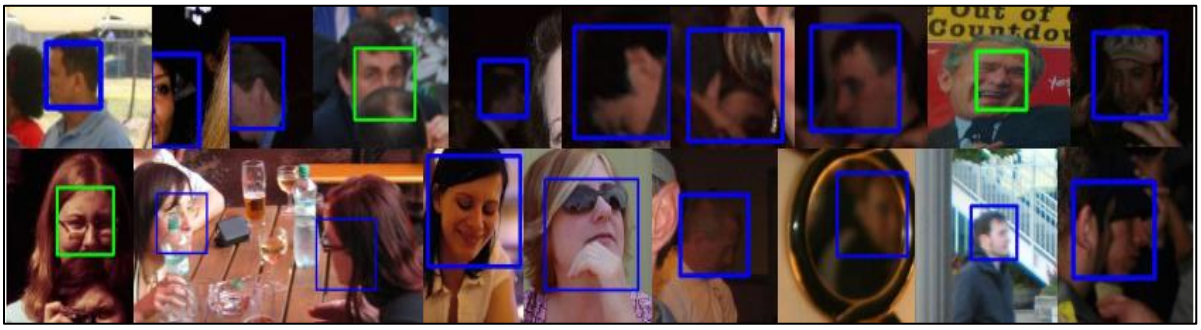

Figure 2.6 Certaines détections de visages très complexes trouvées par CNN en cascade [78], mais parfois inadéquates pour la problématique de la reconnaissance de visages

D’autres techniques existent pour la FD, comme la méthode basée sur les DPM (aussi appelée *HeadHunter*) [80], les DRMF (*landmark* du visage) [7, 115], certains CNN qui se basent sur des approches en cascade comme le fait VJ [78], ainsi que des CNN *end-to-end* qui tentent d’effectuer tant la FD, la FR et l’alignement de pose simultanément [32]. Il y a aussi les multiples variations populaires de CNN entraînés originalement pour la détection d’objets génériques (R-CNN, Fast R-CNN, Faster R-CNN [55, 56, 111] ou YOLO et YOLOv2 [109, 110], parmi tant d’autres), mais pouvant être adaptés pour la détection de visage spécifiquement [68] par *transfer learning*.

Dans chacun des cas précédents, les techniques sont toujours plus rapides et efficaces en temps de calcul les unes que les autres, et ce, à intervalles de publication toujours de plus en plus rapprochés dans le temps. À chaque nouvelle méthode, nous pouvons généralement observer que le nombre de détections de « tête » est bien meilleur qu’avant, mais cela inclut également l’arrière de la tête ou les profils de visage extrêmement prononcés, ce qui n’est pas toujours souhaité pour la FR. Ainsi, une difficulté supplémentaire est souvent ajoutée par ces algorithmes qui surpassent VJ, lorsqu’ils captent davantage les visages présents sous toutes ces formes, car cela apporte toujours des représentations plus difficiles à caractériser.

Donc, surtout pour le cas du SSPP visé par cette étude, où une seule image de référence frontale, de haute qualité selon des conditions d’acquisition contrôlées, et quasi-parfaite est assurée, la performance accrue des méthodes récentes de FD peuvent en effet devenir contre-productive à la tâche de FRiVS, puisque les observations et positions de visages captés s’éloignent de plus en plus du cas de l’environnement semi-contrôlé (voir la progression des visages de la Figure 2.3 à la Figure 2.6 de plus en plus difficiles à caractériser par rapport à une image frontale de référence). Le choix d’algorithmes de FD performants doit donc parfois être considéré avec un certain recul, car ceux-ci peuvent venir produire un effet de lame à doubles tranchants, où les difficultés associées aux problèmes d’acquisitions sur les visages deviennent trop complexes pour simplement concevoir l’application d’un FR de manière efficace.

2.2 Suivi de trajectoire de visages

2.2.1 Suivi de trajectoire en général

Le suivi de trajectoire de visages en FR consiste, de manière générale, à initialiser un ensemble formant une piste (*track*) qui regroupera et mémorisera les régions d'intérêt de visages produits par la trajectoire de passage (*trajectory*) d'un individu à travers plusieurs *frames* d'un vidéo, et ce, grâce à un module de *face tracking* (FT) spécialement conçu pour cette tâche [48]. Ici, nous faisons référence à *track* pour indiquer le conteneur même de ROI extraites et associées de sorte à indiquer lesquelles appartiennent à un même passage. La trajectoire de passage est quant à elle le déplacement physiquement effectué par l'individu dans l'environnement capté par la caméra de surveillance. Cette distinction pour cette section du rapport est soulignée étant donné l'ambiguïté et l'interchangeabilité souvent effectuée pour les termes *track* et *trajectory*.

Le FT diffère de la FD par le fait qu'il associe l'information contenue dans le temps et l'espace à travers les mouvements observés dans la scène, afin d'y déterminer la position du visage, alors que la FD applique son algorithme de manière isolée à chaque *frame*, n'utilisant que les informations visuelles observables à cet instant donné du vidéo. Ainsi, les approches par suivi de trajectoire peuvent inférer un certain contexte par rapport au mouvement effectué par le passage de l'individu, ce qui permet de prédire avec assez grande précision la position probable du visage sur les *frames* à venir. Il est à noter que le FT n'utilise pas une reconnaissance de visage à proprement parler pour associer ceux-ci dans l'ensemble commun d'apparences sur la piste (*track*), mais se base plutôt soit sur l'un ou plusieurs des points suivants :

1. l'apprentissage ou l'adaptation d'un modèle interne de visage caractérisant l'apparence générale d'un « objet » à suivre pour déterminer la nouvelle position;
2. l'observation des régions potentielles situées autour de la position actuelle, par déphasage de la ROI courante dans l'espace de la scène vidéo;
3. l'inférence d'une nouvelle position basée sur le vecteur de mouvement de l'objet suivi entre plusieurs *frames,* pour extrapoler la position future.

Ainsi, la tâche d'un module de FT consiste à retrouver une ROI de visage, qui lui a été assignée à l'initialisation, sur le prochain *frame* vidéo grâce à l'information spatio-temporelle qu'il a accumulée depuis le début (ou aussi loin qu'il considère fiable) jusqu'à sa position actuelle, employant ainsi les informations contenues aux *frames* passés. Étant donné les performances de plus en plus accrues des méthodes de détection de visages employées à elles seules, tel que présenté à la section précédente, certains pourraient se demander en quoi l'application du FT nous devient utile. D'autres pourraient considérer que cela apporte une complexité supplémentaire au système. Malgré que ces points puissent être véridiques pour certains cas, l'application de suivi de trajectoire de visages a démontré en littérature [49] que son addition à un système de bout-en-bout pour la FRiVS vient grandement améliorer les performances de ce dernier pour la reconnaissance produite des visages d'individus d'intérêt, étant donné l'ajout en robustesse apporté face aux conditions d'acquisition inévitables et difficiles à caractériser, comme l'illumination, la pose et le flou de motion affectant l'apparence du visage. Ainsi, étant donné que la FD ne requiert pas de déterminer les associations complexes pour regrouper les visages spatio-temporellement, celle-ci est très sensible aux conditions de capture, ce qui fait qu'il est fort probable que la FD échoue de manière sporadique pour des *frames* particuliers affichant des combinaisons non répétables de bruits nuisibles d'acquisition par les caméras.

Afin d'accomplir le FT, une quantité tout aussi vaste d'algorithmes que pour la FD sont proposés en littérature, sinon plus. Certains des mieux connus, mais relativement dépassés dans le temps, en efficacité et en robustesse, sont CAMshift (Continuously Adaptive Meanshift) [23], Compressive Tracking [145] et TLD (Tracking-Learning-Detection) [70]. Des approches plus récentes basées sur des CNN sont également disponibles, comme GOTURN [61]. D'autres, comme AAMT [45, 47], se basent sur l'adaptation de modèles d'apparence interne du visage suivi afin de modéliser et de graduellement son apparence visuelle à partir des nouvelles apparitions observées dans la scène (*on-line learning*) pour mieux continuer à le suivre. En fait, il existe tellement de nouveaux algorithmes pour la tâche générale de suivi de trajectoire (d'un objet quelconque, et non particulièrement d'un visage), et pouvant s'appliquer à la FT, que des défis récurrent et à part entière, comme VOT (*Visual Object Tracking Challenge*) [1], existent pour comparer les performances de méthodes proposées.

Une considération qui semble ressortir pour les algorithmes de suivi de trajectoire est le fait d'effectuer ou non une mise à jour graduelle (*on-line*) du modèle en mode d'opération à partir des nouvelles observations capturées. Certains cas indiquent que l'adaptation du modèle interne de l'objet suivi en continu vient dégrader ceux-ci, car le modèle devient alors fortement dépendent des données observées et, puisqu'elles sont parfois de mauvaises représentations de l'objet dues à la qualité bruitée de l'image et des conditions d'acquisition variantes dans la scène, cela fait que le modèle ne maintient pas une robustesse contre celles-ci [145]. D'autres applications démontrent par contre que cette approche est à l'inverse plus efficace [45, 60, 70]. Or, il est difficile de déterminer laquelle est la meilleure approche, mais chose certaine est que, les algorithmes de suivi de trajectoire sont habituellement tous affectés par les problématiques potentielles (1) de désalignement de l'objet suivi et, (2) de la dérive (*drift*) de trajectoire. Autrement dit, la transition fluide du mouvement de la position de la ROI obtenue d'un *frame* à l'autre peut être légèrement saccadée, ce qui crée un décalage par rapport au visage pour le premier cas et, lorsque le suivi de trajectoire d'un objet commence à bifurquer, celui-ci s'empire assez facilement sans possibilité de retour pour le deuxième cas. Les deux situations problématiques sont illustrées à la Figure 2.7 et la Figure 2.8 respectivement.

Figure 2.7 Effet de désalignement probable causé par le suivi de trajectoire

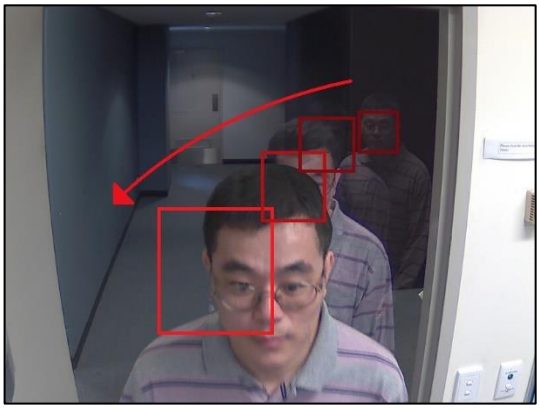

Figure 2.8 Effet de dérive (*drift*) probable causé par le suivi de trajectoire

Les problèmes de suivi de trajectoire font que, lorsqu'un FT est employé, il faut naturellement s'attendre à ce que la FD appliquée sur les vidéos correspondants produise des ROI légèrement divergentes (voir la Figure 3.4 pour un exemple concret).

Une notion également importante à considérer lorsque l'on veut suivre des visages dans une scène de vidéosurveillance est qu'il est tout à fait probable, et même plus souvent le cas qu'autrement, que plusieurs individus y circulent simultanément. Il devient alors important d'évaluer comment le FT gère les croisements de visages dans la scène, lorsqu'un individu passe devant un autre du point de vue de la caméra par exemple. Effectivement, lorsque différents objets sont suivis, il arrive que ceux-ci persistent à être correctement suivis malgré les croisements, car ils sont très distinctifs visuellement. Par contre, pour des visages fondamentalement très similaires en apparence, ne pas considérer les cas d'occlusions mutuelles de ceux-ci résultera très certainement en des pertes de trajectoires ou en des confusions entre les pistes de visages, regroupant alors des visages qui n'ont rien à voir ensemble. Les systèmes de FRiVS sont donc très affectés par la capacité des FT à gérer correctement ces cas particuliers, afin de continuer le suivi de trajectoire des bons individus respectifs, peu importe les interférences de trajectoires et passages d'individus d'intérêt.

Basés sur le principe ci-haut, d'autres types de FT, appelés *multi-object trackers*, sont apparus afin de remédier aux cas de confusion des trajectoires en appliquant diverses techniques telles que l'association d'information mutuelle, l'ajout de confidences associées au suivi des trajectoires, ou toute autre méthode déterminée comme la plus efficace. Ce domaine est tout aussi vaste que par le suivi d'un objet unique, d'où l'apparition de défis annuels également pour ces cas particuliers, comme MOT (*Multiple Object Tracking Benchmark*) [82].

2.2.2 Approche de combinaison de FD+FT

Étant donné que le suivi de trajectoire du visage apporte habituellement des gains en précision, pour déterminer sa ROI, il est considérable d'employer le FT dans un système de FRiVS. Par contre, les *trackers* nécessitent que l'on leur spécifie où débute la trajectoire afin d'y suivre l'objet présent. Ainsi, il devient nécessaire de judicieusement combiner la détection de visages d'un algorithme quelconque, pour obtenir la position initiale, et ainsi la transférer au module de suivi de trajectoire pour qu'il accomplisse le suivi du visage.

Il est possible de considérer deux façons principales pour combiner FD+FT dans le contexte spécifique de la FR sur vidéos [40]. La première approche se veut à titre de *validation-by-tracking* (terme proposé par l'auteur), alors que la deuxième, davantage connut dans le contexte général de suivi de trajectoire d'objets, le mode de *tracking-by-detection* [24, 35, 77, 96].

En *validation-by-tracking*, les visages détectés et représentés par une ROI_d sont simplement obtenues de l'algorithme de FD qui analyse chacune des images formant la séquence vidéo. Les détections sont envoyées à un module de FT qui fonctionne en parallèle pour venir évaluer les correspondances entre les nouvelles détections et celles déjà suivies. Ainsi, le module de FT vient valider, en fonction des *tracks* actuelles qu'il gère, à laquelle la nouvelle détection appartient le plus probablement, et lui associe une référence vers cette piste de visages pour référence ultérieure. Après avoir validé l'appartenance de la nouvelle ROI_d à l'une des *tracks*, le FT n'est plus exploité à en tant que tel. La FR est appliquée directement sur la ROI_d détectée, et le score produit peut alors être accumulé et fusionné grâce à l'information de la *track* qui avait été validée. L'architecture de ce mode d'opération est représentée par la Figure 2.9, et est similairement employé dans [45] à titre d'exemple.

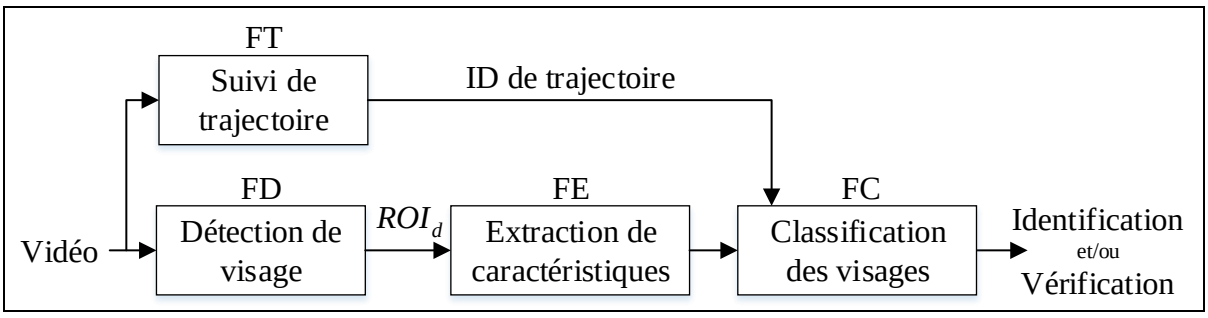

Figure 2.9 Procédé de suivi de trajectoire en mode *validation-by-tracking*

En mode *tracking-by-detection*, la ROI_d, représentant la position détectée d'un visage par le FD sur un *frame* donné de la séquence vidéo, est passée directement au module de FT afin de l'initialiser. Lorsque la *track* est lancée, c'est maintenant le FD qui n'est plus employé, car c'est les ROI_t provenant directement de la trajectoire qui se met à jour à chaque *frame* en suivant le visage qui fournit les images pour la FR.

Donc, le FT maintient les ensembles de ROI correspondantes par trajectoire initialisée pour former chaque *track* individuellement. Le mode *tracking-by-detection* est illustré à l'aide de la Figure 2.10.

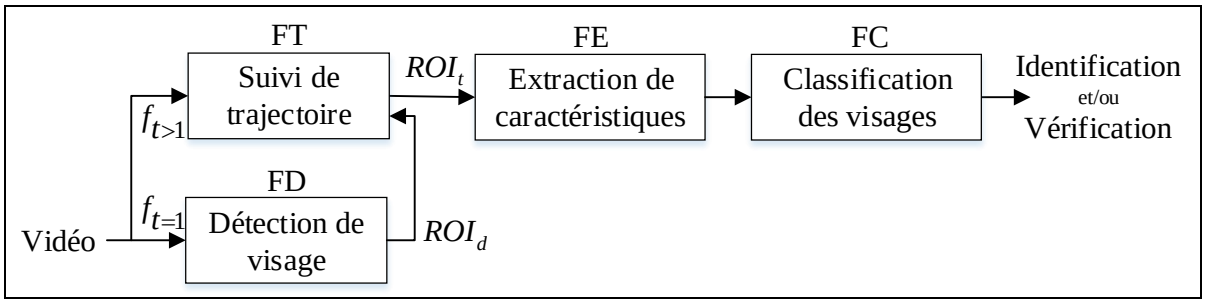

Figure 2.10 Procédé de suivi de trajectoire en mode *tracking-by-detection*

Chacun des deux modes possède leurs avantages et inconvénients. Dans le cas de *validation-by-tracking*, il peut être parfois difficile de définir correctement la relation qui existe entre les ROI_d, et l'ID de trajectoire correspondante à appliquer, car le modèle interne du FT qui évalue les ROI_d en parallèle pourrait bifurquer à une position qui ne correspond aucunement à la nouvelle détection. Si les positions diffèrent trop, il peut même être impossible de définir la correspondance. Aussi, puisque la détection génère les visages pour la FR, si le détecteur échoue pour un moment donné, la trajectoire ne peut simplement pas être associée. Par contre, étant donné que les ROI_d sont employées directement, il n'y a théoriquement aucune différence entre celles produites en phase d'entraînement qu'en phase d'opération du système. Dans le cas du *tracking-by-detection*, si la détection échoue, des ROI sont toujours produites, donc la FR peut tout de même être accomplie. Aussi, étant donné que le FT produit directement les visages pour la FR, l'identification des trajectoires correspondantes est triviale, en plus d'être beaucoup plus rapide que le FD appliqué à chaque *frame*. Par contre, n'étant basé que sur le FT, si le modèle interne venait à bifurquer, la trajectoire de visages de l'individu pourrait totalement être perdue en mode *tracking-by-detection*. Encore pire, si la détection initiale échoue, aucun visage ne sera produit pour toute la *track* et l'individu risque d'être complètement manqué. De plus, les ROI_t employées en phase d'opération pour la FR diffèreront des ROI_d employées pour l'entraînement en S2V (voir un exemple concret présenté à la Figure 3.4), pouvant alors affecter les prédictions de FR.

Étant donné que *tracking-by-detection* fonctionne par une seule détection pour l'initialisation, et que le FT se charge ensuite à part entière de suivre le visage, ce type d'approche semble moins souvent employée en FRiVS. En effet, le problème de l'unique détection rend l'approche moins adaptée aux cas où les individus puissent arriver à n'importe quel moment pendant une séquence vidéo capturée en continu. Aussi, la perte de trajectoire étant une réalité constante avec les FT, se fier uniquement à celui-ci risque de nous faire manquer des individus d'intérêt sans même possibilité de déterminer qu'il est passé (erreur FA pire que FN ou FP).

Or, afin de remédier au problème ci-haut, des adaptations particulières de *tracking-by-detection* sont possibles, tel que FAST-DT [34], parmi plusieurs autres. Celle-ci est basée sur des FD à intervalles réguliers, afin de réévaluer continuellement l'initialisation ou la suppression de trajectoires. Lorsqu'un nouveau visage est détecté, le FT est initialisé adéquatement avec la nouvelle détection. L'architecture proposée par les auteurs, et montrée à la Figure 2.11, emploie également des techniques d'association de données afin de gérer les cas de multi-trajectoires, pour éviter les confusions lorsque plusieurs individus se croisent dans la scène. L'algorithme effectuant le FT est appelé STRUCK [60], retenu pour ces performances marquées à travers le *challenge* VOT mentionné plus tôt. Celui-ci est basé sur des SVM qui s'adaptent *on-line* selon les nouvelles représentations de visages observées. En plus de cela, le code de ce FD+FT, qui a été optimisé pour l'opération en temps réel, est disponible en ligne. Ainsi, considérant tous les avantages apportés par la méthode et les requis du projet, celui-ci a été retenu pour le système de FRiVS.

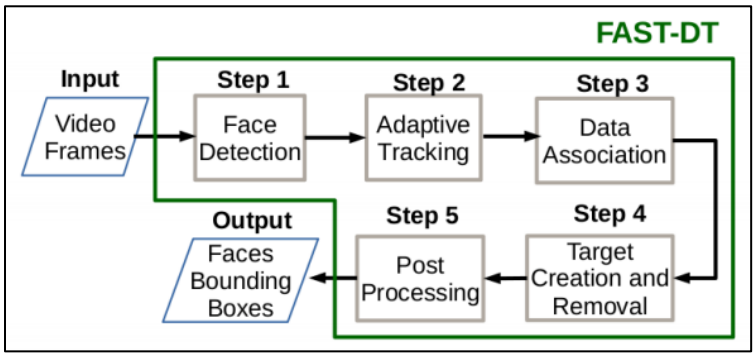

Figure 2.11 Schéma-blocs du procédé d'opération de l'algorithme de suivi de trajectoires multiples FAST-DT [34].

2.3 Extraction de caractéristiques

Afin d'accomplir la reconnaissance de visages, il est nécessaire de représenter ceux-ci dans un hyper-espace augmenté de caractéristiques, dont des classificateurs entraînés à partir d'exemples appropriés pourraient classifier adéquatement. Ainsi, les techniques d'extraction de caractéristiques (FE) discriminantes pour représenter les visages sont requises pour un système de FRiVS. De manière générale, trois sous-classes de descripteurs de la catégorie *template-matching* sont définies en littérature [67, 69], soit (1) les méthodes par projection, (2) les méthodes de vision et, (3) les méthodes par apprentissage de caractéristiques. Celles-ci seront présentées aux sous-sections qui suivent.

2.3.1 Méthodes par projection

Les approches par projection en FR regroupent autant les techniques génériques permettant la réduction de dimensionnalité des vecteurs descripteurs caractérisant les visages, tels que PCA ou LDA, que les méthodes très spécifiques pour la reconnaissance de visages comme les représentations par *eigenfaces* ou *fisherfaces*. Ces techniques ont maintes fois été appliquées au domaine de la reconnaissance de visages [14, 69, 72, 88, 118, 124, 125, 135]. Quelques exemples avaient été fournis au dernier chapitre avec les Figure 1.1 et Figure 1.2. Le principe général des techniques se base sur le fait que plusieurs des caractéristiques existantes dans l'hyper-espace servant à caractériser les visages sont extrêmement redondantes. Ainsi, étant donné la corrélation élevée existant à travers certaines combinaisons de *features*, il devient possible de les projeter selon un système d'axes différent, à plus petite dimensionnalité, et où chaque dimension (ou nouvelle *feature* générée) est bien plus discriminante et non corrélée aux autres données qu'initialement.

Ainsi, les pixels formant l'image de la personne à reconnaître peuvent être similairement projetés afin de trouver des combinaisons pondérées des valeurs de pixels pour former des sous-représentations de celui-ci. En ne sélectionnant que les quelques premières représentations les plus décorrélées, nous obtenons des visages discriminants pouvant être classifiés. Ces méthodes sont particulièrement efficaces pour les cas où la dimensionnalité des

vecteurs générés est trop grande et que l'on veut la réduire. En effet, les grands vecteurs utilisés pour la classification causent habituellement le problème de *curse of dimensionality* [16, 40, 51, 67], soit lorsque les caractéristiques ne deviennent plus suffisamment discriminantes pour identifier des individus particuliers. Cela est causé par le fait que l'information redondante qui est distribuée à travers trop de sous éléments provoque de très faibles variations à travers l'ensemble des valeurs. Les faibles variations de caractéristiques deviennent alors à peine percevables à travers le bruit naturellement généré par les variations de conditions d'acquisition sur un même visage. Cette catégorie de descripteurs est donc principalement basée sur l'analyse des données et les manières de les représenter autrement, ou les compresser, sans toutefois perdre la majorité des informations pertinente à la classification de visages.

2.3.2 Méthodes de vision

Les approches de FE basées sur les méthodes de vision sont celles qui sont fondées sur les variations de gradients des pixels formant l'image du visage, de techniques morphologiques afin d'appliquer des masques sur les pixels, ou des analyses en fonction des couleurs de l'image. Cette catégorie regroupe ainsi les approches traditionnelles de *feature engineering*, qui sont obtenues en se basant sur des propriétés sélectionnées selon des objectifs précis. Étant donné la variété de la catégorie, il existe une abondance de méthodes, mais les quelques exemples suffiront à démontrer pourquoi elles sont considérées *handcrafted*.

Une première approche de ce type qui est souvent ressortie en littérature pour la FR est le descripteur LBP [2, 53, 148, 149, 151, 152]. Celui-ci est conceptualisé dans l'objectif de caractériser des *patterns* de textures très spécifiques, à partir des tons de gris dans l'image. La technique se base sur un seuillage des pixels autour d'une valeur centrale dans une fenêtre localisée, afin d'encoder binairement si les pixels voisins sont supérieurs ou inférieurs à celle-ci, d'où le nom *Local Binary Patterns*. Ce procédé d'encodage est représenté à la Figure 2.12. Les occurrences de combinaisons binaires observées sur l'image sont ensuite accumulées dans un histogramme, tel que montré à la Figure 2.13, après filtrage à un maximum de 59 *features* pour supprimer les patrons irréalistes selon les observations accomplies dans le mode réel.

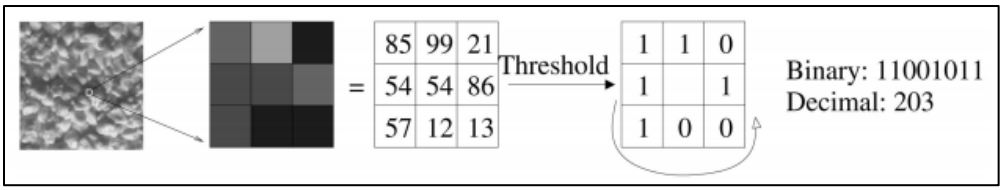

Figure 2.12 Encodage des pixels par extraction de caractéristiques LBP [2]

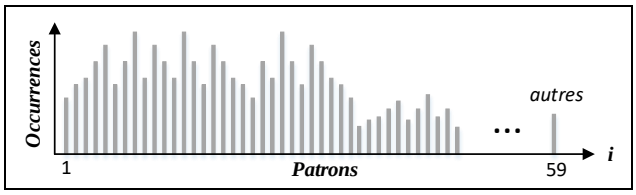

Figure 2.13 Histogramme de caractéristiques LBP

Le descripteur LBP a eu un franc succès dans multiples domaines, ce qui a mené à plusieurs variantes comme LBP-riu2 [94, 151] et LBP-HF [3]. Les nouvelles versions de LBP visent toutes à devenir davantage invariantes aux conditions d'acquisition (surtout d'illumination) afin de produire des descripteurs discriminants. Ceci permet donc d'obtenir des vecteurs très représentatifs de visages, tel que démontré par la Figure 2.14. De plus, étant représenté en valeurs binaires et avec peu de *features*, LBP est très rapide à appliquer pour des applications qui doivent fonctionner en temps réel comme le système de FRiVS visé par le projet.

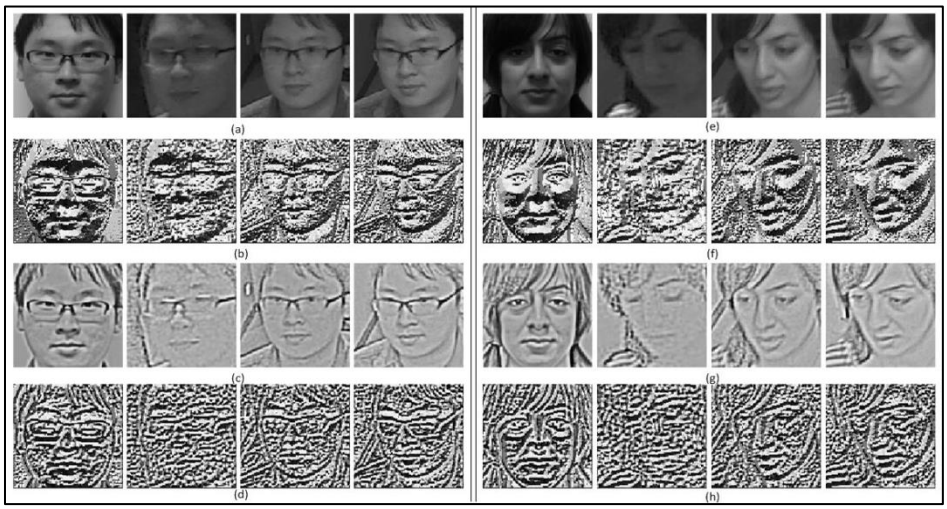

Figure 2.14 Application du descripteur LBP au domaine de la FR [6]

Un autre descripteur marquant faisant partie des méthodes de vision pour l'extraction de caractéristiques est l'algorithme HOG (*Histogram of Oriented Gradient*) [36]. L'extraction de caractéristiques par HOG est aussi grandement employée dans le contexte de la reconnaissance de visages [5, 15, 44]. La méthode a été développée initialement dans le but de détecter les gens (leur corps complet) sur des séquences vidéo, ce qui peut expliquer pourquoi elle est également très efficace en FR. Comme son nom l'indique, celui-ci est basé sur les gradients définis par les pixels de l'image, qui sont judicieusement combinés selon leur orientation correspondante afin de construire un histogramme de caractéristiques pour représenter le visage. Les gradients sont sélectionnés et groupés par orientations communes successivement selon des sous-divisions de l'image en fenêtres, blocs, et cellules, avant d'être concaténés sous forme de vecteur de caractéristiques, tel que démontré à la Figure 2.15.

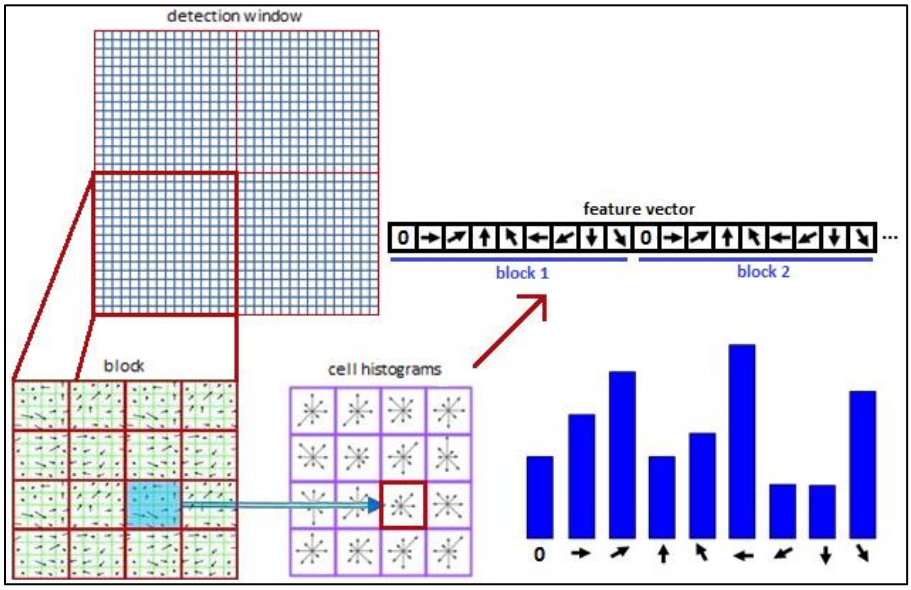

Figure 2.15 Procédé du descripteur HOG

Finalement pour les approches de vision, un dernier descripteur qui avait été considéré pour le projet est LPQ (*Local Phase Quantization*). Cette approche de FE se démarque des autres particulièrement pour sa robustesse par rapport au problème de flou de motion introduit sur les séquences vidéo, tel que démontré dans [95, 108]. Effectivement, les deux autres descripteurs présentés ne sont pas efficaces contre ce problème lorsqu'il est particulièrement prononcé. La

méthode se base sur une transformée de Fourier pour convertir la phase fréquentielle de la réponse contenue dans l'image, qui elle est robuste au flou, sous forme de caractéristiques locales par régions voisines [4]. Avec la représentation par sections voisines sur l'image, il devient possible d'extraire les caractéristiques de certains pixels, afin de former le vecteur descripteur final. Le descripteur LPQ a aussi été appliqué en FR [57, 66], validant la possibilité d'application de celui-ci pour un système de FRiVS fonctionnant particulièrement avec des séquences vidéo sujettes à du flou sur les visages des individus en motion dans la scène.

2.3.3 Méthodes par apprentissage de caractéristiques

La troisième et dernière classe de descripteurs regroupe les approches d'apprentissage, c'est-à-dire lorsqu'un algorithme est employé pour automatiquement apprendre lesquelles des caractéristiques de l'image, ainsi que leurs interrelations sous-jacentes, sont les plus pertinentes pour obtenir des vecteurs les plus discriminants possible en fonction des données fournies pour l'apprentissage. Cette catégorie (*feature learning*) est donc devenue particulièrement intéressante et a été appliquée de plus en plus depuis quelques années à cause de la disponibilité de données massives ainsi que d'architectures de réseaux pour l'apprentissage profond (*deep neural networks*).

Évidemment, les architectures sont très variées en littérature, et sont entraînées sur une multitude de bases de données différentes, mais les vecteurs résultants réussissent tout de même à relativement bien caractériser les visages, peu importe leur application future et domaine spécifique, étant donné la quantité massive d'exemples utilisés pour l'entraînement, qui permet d'apprendre en grande profondeur tous les fins détails discriminants du visage. Afin d'employer une approche de FE par apprentissage de caractéristiques à travers un CNN, la méthode utilisée est habituellement d'enlever les quelques couches situées à la fin du réseau entraîné, représentant les prédictions d'individus, pour accéder à ses représentations internes des individus comme caractéristiques de visage [103], tel qu'illustré à la Figure 2.16. Ainsi, en donnant un nouveau visage en entrée à un tel réseau, la couche de *features* permet d'obtenir sa représentation encodée, pour ensuite entraîner un autre classificateur avec celle-ci.

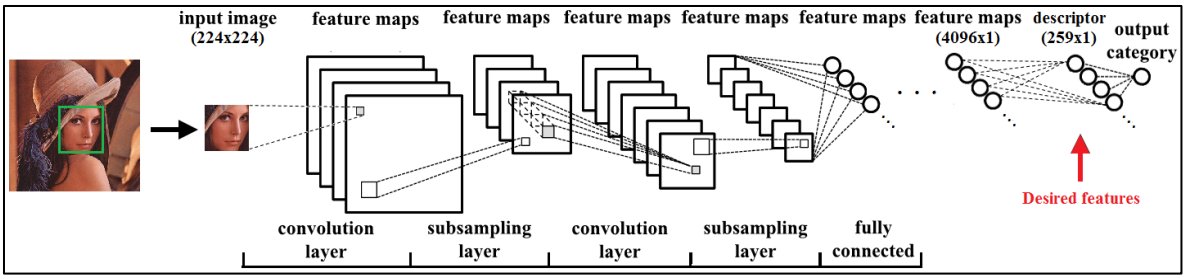

Figure 2.16 Caractéristiques apprises par un descripteur à base de CNN (basé sur [102]

Quelques exemples de systèmes qui ont été entraînés spécifiquement pour la FR par l'utilisation de CNN et d'auto-encodeurs à architecture profonde sont présentés dans les travaux [101, 102, 117]. Avec la génération de tels modèles très performants pour apprendre à reconnaître des individus bien particuliers, il devient possible d'exploiter leurs connaissances acquises afin d'accomplir une FE, en retirant les dernières couches du réseau.

2.3.4 Combinaison de multiples descripteurs

Cette dernière sous-section présente que brièvement un concept possible pour la réalisation de système de FRiVS robustes, soit l'utilisation de multiples représentations de visages à travers différents descripteurs. En effet, tel que présenté dans [13], il est possible d'employer une combinaison de techniques pour représenter le visage selon différentes formes, ce qui permet habituellement d'entraîner des ensembles de classificateurs (un par représentation) plus robustes face aux faiblesses de chaque descripteur en les compensant avec les forces complémentaires des autres. Aussi selon [6], il est démontré que les descripteurs sont souvent spécialisés à demeurer invariants à des problèmes d'acquisition sur caméras bien spécifiques, mais il ne peuvent pas tous les résoudre. Pour cela, il est souvent nécessaire d'appliquer en parallèle divers descripteurs discriminants de manière à profiter de leurs avantages particuliers et complémentaires contre les facteurs d'acquisition nuisibles lors d'applications pour la FR en S2V [71]. Le Tableau 2.1 présente certains descripteurs populaires et leurs forces respectives contre les conditions d'acquisition nuisibles à la reconnaissance d'individus sur caméras de vidéosurveillance. Ceux-ci devraient donc être considérés pour la conception de tout système de FRiVS selon les situations observées comme plus problématiques.

Tableau 2.1 Robustesse de certains descripteurs de caractéristiques face à diverses conditions d'acquisition nuisibles observées en FRiVS [13]

Nuisance factors	**Feature extraction techniques**
Illumination	LDA, PCA, **LBP**, Gabor, LPP, **Haar**, SIFT, SURF, **HOG**
Pose	**Haar**, **HOG**
Scale	SIFT, SURF, Gabor, **HOG**
Motion Blur	**LPQ**
Occlusion	**Haar**, SURF (partial occlusion)

Étant donné que le système visé de FRiVS introduit beaucoup de complexité, à cause du suivi de trajectoire de visage ainsi que de la génération de représentations multiples du visage (présentés aux Chapitre 3 et Chapitre 4 respectivement), il a été décidé de se limiter à un seul type de descripteur pour encoder le visage lors de la conception initiale et l'étude de faisabilité des méthodes. De plus, il était espéré que l'utilisation de génération synthétique apporterait déjà des variations très distinctes du visage à représenter. Or, il serait possible qu'employer plusieurs descripteurs ne fasse qu'alourdir inutilement le système et le rendre plus lent par les opérations supplémentaires induites par plusieurs descripteurs employés en parallèle.

Ainsi, afin de sélectionner le descripteur à retenir en particulier parmi toutes les combinaisons qui s'offraient à nous pour l'application du projet, il a été choisi de se fier aux recommandations dans [10]. Les auteurs semblaient y indiquer que HOG était particulièrement plus pertinent que la majorité des autres descripteurs, par rapport à leurs autres travaux dans [13], qui employaient en combinaison les descripteurs en gras du Tableau 2.1. Avec ce même tableau, nous pouvons aussi voir HOG est simultanément robuste à une multitude de conditions d'acquisition nuisible présentes sur les caméras employées en FRiVS. Ainsi, celui-ci était le candidat le plus prometteur dans l'ensemble, considérant également que plusieurs sources de codes très optimisées étaient disponibles afin d'implémenter celui-ci. De plus, étant une méthode de représentation du visage à caractère global, tel qu'introduit au Chapitre 1 (section 1.1), ce descripteur permet de profiter d'un maximum de rapidité de calcul pour assurer une solution en temps réel.

2.4 Classification des visages

Dans un contexte général de FR sur vidéos, il est possible d'utiliser un grand nombre de classificateurs afin de reconnaître les individus entre eux. Plusieurs réseaux à apprentissage profond pré-entraînés peuvent être employés afin d'obtenir des performances remarquables sans pour autant nécessiter un long processus d'apprentissage ou de conception de système spécialisé à une situation particulière. Quelques exemples notables pour la FR sont DeepFace [121], FaceNet [116] et TBE-CNN [49] et HaarNet [100].

Par contre, dans un contexte de SSPP, plusieurs des méthodes de classification deviennent invalides à cause de l'échantillon unique ou du déséquilibre des classes trop prononcé. Or, pour les systèmes de FRiVS spécifiquement, et en mode d'application S2V, la quantité d'algorithmes qui persistent à produire de hautes performances est grandement réduite. La majorité des solutions se basent alors sur des variantes bien spécifiques de CNN, ou des classificateurs à formule d'exemplaires. Certains des modèles les plus récents considérés pour le projet seront donc présentés ici.

2.4.1 Approche par CNN

Parmi les approches existantes qui supportent le SSPP, et à base de CNN simultanément, nous retrouvons CCM-CNN [99] et CFR-CNN [101]. Ces deux architectures proposent une façon compacte (relativement à d'autres CNN), et à leur propre manière, d'accomplir la reconnaissance d'individus d'intérêt à partir d'une image de référence. Étant donné la parution très récente de ces modèles, ceux-ci n'ont malheureusement pas pu être considérés pour l'implémentation du système de FRiVS de ce projet. À la meilleure connaissance de l'auteur, il n'existe pas à ce jour d'autre méthode produisant de similaires performances dans un contexte de SSPP par réseau profond. Ainsi, ces approches pourront uniquement être employées à titre comparatif pour cette étude. Deux considérations demeurent tout de même à ce qui a trait à l'applicabilité de ces méthodes dans notre contexte, soit le temps requis pour l'enregistrement de nouveaux individus dans le système, et l'aisance pour y parvenir.

En effet, malgré que plusieurs méthodes de *deep learning* ont montré à mainte reprise, à travers un divers ensemble de domaines distincts, qu'ils procurent de nettement meilleures performances de classification par rapport aux anciennes méthodes qui faisait l'état-de-l'art, leur utilisation de manière incrémentale demeure relativement limitée, surtout pour le domaine de la FRiVS. Dans une telle application, il est à considérer que la majorité des individus à enregistrer au système y seront ajoutés incrémentalement, c'est-à-dire que les nouvelles personnes d'intérêt seront insérées à la *watchlist* selon les besoins de reconnaissance et les situations désirées pour accomplir la surveillance des lieux visés, alors que le système est déjà en service. Or, les systèmes à base de CNN pour la FRiVS en SSPP, déjà très limités en quantité, ne montrent pas de manière évidente d'accomplir cet ajout incrémental. Un réentraînement du modèle complet semble alors être nécessaire à chaque fois, ce qui n'est aucunement pratique ni efficace en termes de temps pour une solution dynamique dans le temps. En fonction de ces observations, les CNN, tout aussi performants qu'ils puissent être, ne seront donc pas employés pour ce projet.

2.4.2 Approche par SVM

Les machines à vecteurs de support existent depuis déjà de bien nombreuses années et leur utilisation c'est répandu à travers pratiquement tous les domaines imaginables de classification et d'apprentissage machine, étant donné leur simplicité comparativement à d'autres modèles et leur capacité à accomplir une classification robuste à l'aide d'un nombre très réduit de représentations en mémoire. En effet, les SVM n'ont besoin que des quelques vecteurs de support employés pour séparer les classes et de quelques hyper-paramètres pour être totalement définis et prêt à appliquer, produisant des modèles légers à entraîner. L'utilisation répandue des SVM a fait qu'un très grand nombre de variations ont également été développées afin de répondre à une multitude de situations particulières. En outre, nous retrouvons des SVM pour la classification multi-classes (C-SVC), une variation semblable avec paramètre de régularisation normalisé (ν-SVC), pour les cas à classe unique (OC-SVM), pour la régression plutôt que la classification binaire (ε-SVR et ν-SVR), et ainsi de suite [62]. Diverses variantes emploient aussi différents *kernels* pour l'augmentation dans l'hyper-espace.

Indépendamment du type de SVM employé, l'idée générale est sensiblement la même. Les SVM déterminent la meilleure marge de séparation possible qui permet de distinguer les classes d'entraînement à l'aide du nombre minimal de vecteurs de support pour définir des hyper-plans de séparation. Pour atteindre une marge de séparation maximale entre les classes, les tentent de minimiser une erreur associée à la mauvaise classification des vecteurs donnés en entraînement. La formulation générique de cette fonction de minimisation est donnée par l'équation (2.1), où la partie de gauche maximise les poids de l'hyper-plan permettant une marge visant la meilleure généralisation possible et la partie de droite pénalise les poids de l'hyper-plan qui cause des erreurs de classification avec la fonction de (*hinge loss*) entre les vecteurs d'entraînement $\mathbf{x}_i$ et leur classe associée y_i. L'hyper-paramètre C est quant à lui un terme de compromis entre une grande marge et une petite erreur de classification. Les variations de SVM viennent donc jouer sur ces diverses sections pour ajuster la méthode d'optimisation pour divers objectifs de classification [27, 62].

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i}\max\left(0,1-y_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right) \tag{2.1}$$

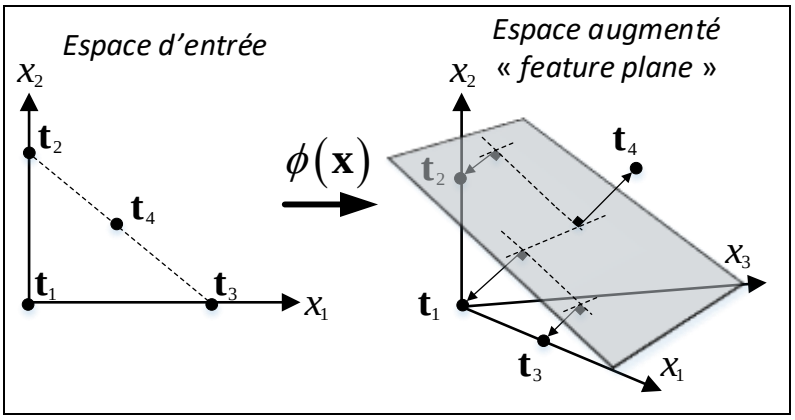

Figure 2.17 Hyper-plan de séparation SVM dans l'espace de haute dimensionnalité pour séparer des données non linéairement séparables

Ainsi, étant donné qu'ils définissent la meilleure marge de séparation possible, selon les contraintes qui leur sont imposées, les SVM assurent une robustesse dans le sens que la

séparation des données sont techniquement optimales par rapport aux configurations et à condition de la possibilité même de classifier les classes. Afin de parvenir à séparer les données non linéairement séparables, les SVM emploient une représentation des données d'entraînement dans un hyper-espace de caractéristiques augmenté, comme illustré de manière simplifiée à la Figure 2.17 précédente, où des vecteurs n'étaient initialement pas séparables linéairement en 2D, mais le deviennent en passant à une représentation alternative en 3D. La fonction de transformation $\phi(\mathbf{x})$ définit ainsi comment cette augmentation est accomplie, en fonction du *kernel* choisi (polynomial, sigmoïde, radial, linéaire ou autre).

Dans un contexte de SSPP, avec plusieurs données négatives et une seule positive, la formulation de la fonction de minimisation d'erreur peut être arrangée selon la forme de l'équation (2.2), afin de permettre des pondérations de mauvaise classification différentes pour chaque classe, de sorte à compenser le déséquilibre des classes [42, 127, 142, 147]. Il est donc important de respecter la condition $C^+ \gg C^-$ en SSPP afin de favoriser la bonne classification de l'individu d'intérêt, même si cela engendre quelques erreurs pour les contre-exemples.

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \|\mathbf{w}\|^2 + C^+ h\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_E + b\right) + C^- \sum_{\mathbf{x}\in\mathcal{N}_E} h\left(-\mathbf{w}^{\mathrm{T}}\mathbf{x} - b\right) \\ &\text{sujet à} \quad h(x) = \max\left(0, 1-x\right) \qquad (\textit{hinge loss}) \end{aligned} \tag{2.2}$$

avec C^+ et C^- définis comme pondérations de régularisation de l'erreur associée à une mauvaise classification des représentations des descripteurs de classes de l'individu d'intérêt et des contre-exemples respectivement, $\mathbf{w}$ la matrice de poids ajustable pour définir l'hyper-plan de séparation, b le terme de biais, $\mathcal{N}_E$ l'ensemble de descripteurs de visages formant la base de contre-exemples, et $\mathbf{x}_E$ la représentation du visage de l'unique image de référence de l'individu d'intérêt.

Le processus d'entraînement des classificateurs SVM selon l'équation ci-haut peut donc être répété pour chaque nouvel individu à entraîner dans le système de FRiVS, étant donné que seulement la référence de cette personne est employée pour un classificateur donné, contre une abondance d'individus différents quelconques et inconnus au système, tel des acteurs employés pour circuler dans la scène des caméras de sécurité pour sa conception et son installation.

Ainsi, la mention précédente peut être visualisée selon le schéma de la Figure 2.18, où chaque nouvel individu d'intérêt ($k = 1$) dont le visage est représenté par un vecteur descripteur $\mathbf{t}_k$, qui provient de l'image de référence, est modélisé par un SVM distinct entraîné en mode *one-against-all*, et dont les « autres » sont les apparences sur vidéos des contre-exemples. La marge de séparation maximale y est donc définie par les vecteurs de support les plus contraignants, soit les cas de visages qui s'apparentent le plus à notre individu à distinguer et reconnaître parmi tous les autres pouvant être présenté au système. L'entraînement de SVM spécifique à chaque nouvel individu à inscrire à la base de modèles du système de FRiVS est représenté à la Figure 2.19, où l'on voit que les vecteurs de support dépendent des cas rencontrés et déterminés comme les plus complexes à classer pour définir la meilleure marge de séparation.

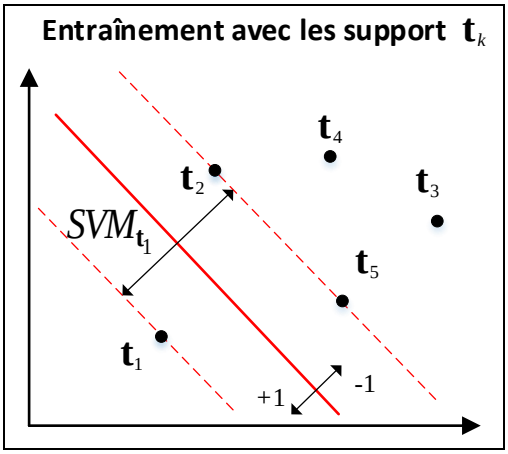

Figure 2.18 Entraînement de SVM en mode *one-against-all*

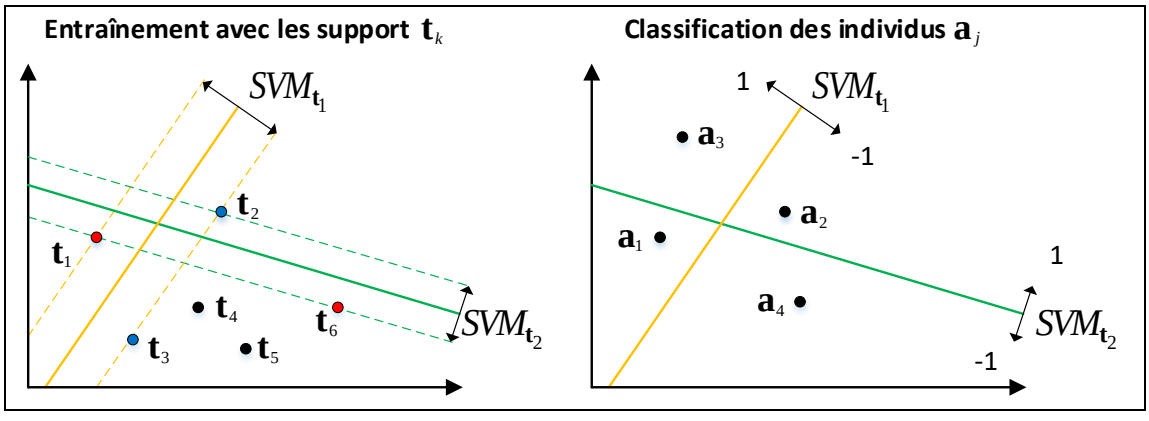

Figure 2.19 Entraînement et test de reconnaissance par SVM spécifique à chaque individu

Nous pouvons aussi voir avec la Figure 2.19 que les prédictions seraient accomplies respectivement à l'hyper-plan de décision de chaque SVM entraîné indépendamment, selon une nouvelle image de visage de test représenté par un vecteur $\mathbf{a}_j$ dans l'hyper-espace. Ceci permet alors d'avoir des SVM plus spécialisés étant donné qu'ils se concentrent sur les complexités associées à leur sous-problème de classification respectif, contrairement à un système entraîné avec un seul modèle qui tenterait de résoudre tous les cas problématiques simultanément. La solution obtenue par des SVM par individu permet aussi d'entre extensible en ajoutant de nouveaux individus au système, simplement à ajoutant un nouveau SVM qui ne vient pas modifier ceux déjà existants d'aucune manière. Ceux-ci sont complètement distincts. Ainsi, la base de modèles du système de FRiVS augmente linéairement en fonction du nombre d'individus à entraîner selon cette approche, mais les modèles étant extrêmement compacts, cela n'est pas particulièrement majeur comme problématique pour une quantité raisonnable d'individus enregistrés dans la base de référence.

Une approche par SVM telle que présentée jusqu'à présent avait été testée lors des phases initiales de conception du système de FRiVS, à partir de la base de données Color FERET [89, 105, 106]. Certains résultats correspondants sont situés à l'ANNEXE IV pour référence. De ce test fort simple, il était déjà possible de voir qu'une approche par SVM traditionnels ne serait pas suffisante étant donné (1) le trop grand déséquilibre des classes, (2) le manque de représentations positives dû au contexte de SSPP, et (3) le fait que ce test présentait des conditions d'acquisition beaucoup moins sévères que pour le l'application en mode S2V.

2.4.3 Approche par *e*-SVM

Considérant les limitations de performance qui avait été observées par rapport à l'utilisation des SVM traditionnels, une approche par exemplaire (*e*-SVM), tel qu'introduit dans [79] a été retenue pour la réalisation des classificateurs du système de FRiVS. L'utilisation de ces classificateurs particuliers est fondée sur les travaux dans [9, 10, 11, 12] qui démontrent que ceux-ci sont extrêmement efficaces pour l'application de FR visée. En effet, les *e*-SVM sont spécifiquement entraînés à partir d'une recherche approfondie des plus difficiles cas négatifs

à classer (*hard negative mining*) parmi des milliers de contre-exemples disponibles, par rapport à une unique référence positive. Il s'agit donc d'une application idéale pour la situation de SSPP visée. Afin d'accomplir la maximisation de la marge de séparation, la formulation de l'équation (2.3) est employée par les *e*-SVM telle que proposée par les auteurs de la solution originale [79]. Or, cette formulation est équivalente à la précédente à l'équation (2.2), mais l'entraînement est effectué pour un ensemble de *e*-SVM plutôt qu'un unique SVM, selon le nombre de cas négatifs complexes rencontrés.

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C^{+}\sum_{y_i=+1}\xi_i + C^{-}\sum_{y_i=-1}\xi_i \\ \text{sujet à} \quad & y_i\left(\mathbf{w}^{\mathrm{T}}\phi\left(\mathbf{x}_i\right)+b\right) \geq 1-\xi_i, \;\; \xi_i \geq 0 \end{aligned} \tag{2.3}$$

avec C^{+} et C^{-} définis comme pondérations de régularisation de l'erreur associée à une mauvaise classification des représentations des descripteurs de classes de l'individu d'intérêt et des contre-exemples respectivement, $\mathbf{w}$ la matrice de poids ajustable pour définir l'hyper-plan de séparation, b le terme de biais, et $\mathbf{x}_i$ sont les représentations du visage de l'unique image de référence de l'individu d'intérêt ainsi que contre-exemples, associés à une classe y_i et évalués par la fonction de transformation linéaire $\phi(\mathbf{x})$ dans l'hyper-espace des représentations de visages.

Pour l'application de l'équation (2.3), la condition de pondération des classes selon $C^{+} \gg C^{-}$ est également préservée pour contrefaire le déséquilibre marqué entre les vidéos négatifs et l'image de référence de l'individu d'intérêt. Celle-ci permet plus spécifiquement de repousser la frontière de décision des *e*-SVM très prêt des contre-exemples, afin de favoriser une réponse bien plus forte de la part des observations correspondant à de vrais-positifs [79].

Une autre distinction entre les SVM traditionnels et l'exemplaire est par rapport à la fonction de score à sa sortie, basé sur une sigmoïde telle que présentée à l'équation (2.4).

$$f\left(\mathbf{x}\middle|\mathbf{w}_E, \alpha_E, \beta_E\right) = \frac{1}{1+e^{-\alpha_E\left(\mathbf{w}_E^{\mathrm{T}}\mathbf{x}-\beta_E\right)}} \tag{2.4}$$

La sigmoïde appliquée à la sortie du *e*-SVM effectue une calibration du score à produire pour le descripteur **x** représentant un visage de test, en fonction des paramètres de régression α_E, β_E appris lors de l'entraînement du visage exemplaire E, et des poids de l'hyper-plan $\mathbf{w}_E$ permettant d'accomplir la décision de reconnaissance par rapport à l'exemplaire et les milliers de contre-exemples.

Ainsi, les *e*-SVM emploie une sortie probabiliste qui permet d'évaluer une certaine forme de proximité entre le vecteur de test en fonction du déséquilibre majeur des classes, qui lui, est caractérisé à travers l'hyper-plan $\mathbf{w}_E$ qui a été repoussé vers les négatifs autant que possible selon les pondérations C^+ et C^- imposées pour la résolution. Ce genre de sortie permet de définir une certaine forme de similarité entre le vecteur testé et l'image de référence exemplaire de l'individu d'intérêt, ce qui est moins radical que la sortie binaire de classes $+1/-1$ du SVM à *kernel* linéaire traditionnel.

CHAPITRE 3

SYSTÈME INITIAL EMPLOYÉ ET OPTIMISATIONS

Ce chapitre introduira le système de base qui a été réalisé et employé pour l'accomplissement de la tâche de FR durant ce projet, ainsi que pour les évaluations qui suivront aux chapitres suivants. Les sections qui enchaîneront vont premièrement fournir un aperçu sur l'ensemble du système accompagné d'une explication globale de son fonctionnement, avant de passer à travers chaque portion plus en détail pour y justifier les choix de conception. Les particularités d'implémentation de chaque module y seront également discutées au besoin.

3.1 Système de FRiVS avec ensemble de *e*-SVM et trajectoires de visages

La Figure 3.1 présente l'ensemble du procédé initial qui a été implémenté de sorte à inclure les algorithmes retenus du Chapitre 2 comme les plus pertinents pour accomplir la reconnaissance de visage globalement, c'est-à-dire en considérant l'efficacité du système dans son ensemble.

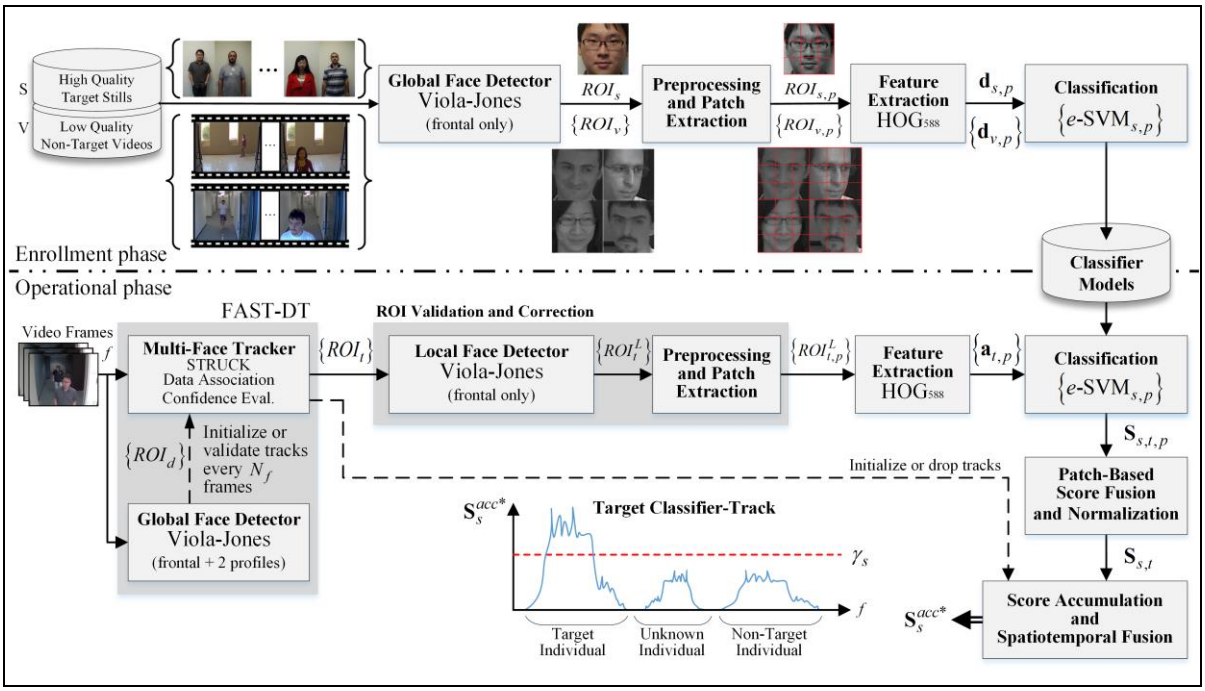

Figure 3.1 Système initial de FRiVS avec E*o*e-SVM et suivi de trajectoire

Parmi les méthodes retenues ici, certaines emploient quelques ajustements ou combinaisons particulières par rapport aux descriptions fournies au dernier chapitre, afin de réaliser un système qui est dans l'ensemble plus performant et rapide d'exécution. Le système proposé est séparé en deux sections, soit la première étape située dans le haut de la Figure 3.1, qui consiste à l'entraînement des classificateurs (*enrollment phase*), et la seconde étape, dans le bas, étant dédiée à la reconnaissance des visages en mode d'opération (*operational phase*) avec les classificateurs entraînés préalablement.

3.1.1 Phase d'entraînement

Lors de la phase d'entraînement, une unique image de référence S annotée et de haute qualité d'un individu à enregistrer dans le système de reconnaissance de visage (*target individual*) est employée, ainsi qu'une grande base de trames de vidéo V de nettement plus faible résolution, provenant de caméra de surveillance, et affichant plusieurs autres individus inconnus (*unknown non-targets*). Tant l'image de référence que les trames vidéo passent à travers un même détecteur de visages frontaux de Viola-Jones afin de retirer uniquement les régions de visages suffisamment favorables a de bonnes représentations des individus observés pour l'entraînement et la reconnaissance. En employant une détection frontale uniquement, seulement les régions localisées des visages adéquats persistent, et donc, les autres poses, illuminations ou qualités jugées insuffisantes par ce détecteur, qui causeraient potentiellement de mauvaises représentations faute d'un manque de caractéristiques discriminantes visibles, sont immédiatement retirées. Cela vient également limiter le bruit dans la base d'apprentissage en permettant d'avoir des visages de *non-targets* sur vidéo valides et suffisamment frontaux pour pouvoir adéquatement les comparer aux visages neutres des images statiques.

Ensuite, l'ensemble des régions d'intérêt collectées $\{ROI_s, \{ROI_v\}\}$, et catégorisées en tant que positive (l'image unique) ou négatives (toutes les autres sur vidéos), subissent le même processus de conversion en ton de gris, d'égalisation d'histogramme pour ajuster le contraste et normaliser les tons de gris des pixels, et sont redimensionnées a 48x48 pixels. La région d'intérêt unique de l'individu à enregistrer ROI_s, ainsi que les multiples contre-exemples ROI_v,

subissent par la suite une segmentation en sous-régions *p* afin d'obtenir $N_p = 9$ *patches* égales et sans chevauchement, tous de 16x16 pixels, indiqués par $\left\{ROI_{s,p}, \{ROI_{v,p}\}\right\}$. Ces imagettes de visages passent finalement par une extraction de caractéristiques HOG ($N_d = 588$ *features*) pour obtenir un ensemble de vecteurs descripteurs discriminants $\left\{\mathbf{d}_{s,p}, \{\mathbf{d}_{v,p}\}\right\}$ qui représentent chacun une petite partie de chaque visage. Ces derniers, avec leurs classes correspondantes $+1/-1$, sont employés pour entraîner des *e*-SVM, soit un pour chaque *patch* p et individu s correspondant, pour ainsi former un nouvel ensemble $\{e\text{-SVM}_{s,p}\}$ respectif à chaque individu à enregistrer dans le système. Pour des fins de simplification de la nomenclature, les références futures à ces ensembles de *e*-SVM par individu exemplaire du *watchlist* seront indiquées par Eo*e*-SVM. Le processus de la phase d'entraînement se termine lorsque tous les modèles sont sauvegardés dans la *gallery* (*classifier models* à la Figure 3.1), soit lorsque l'entraînement répété pour les s individus à enregistrer dans le système est complété. Les vecteurs $\{\mathbf{d}_{v,p}\}$ sur vidéos (les contre-exemples) sont toutefois uniquement générés à la première itération afin de les réutiliser à chaque nouvel enregistrement d'un individu d'intérêt dans le système.

3.1.2 Phase d'opération

Lorsque les classificateurs sont entraînés pour représenter les individus à partir de la base de référence (*classifier models*), il devient possible d'employer le système en mode d'opération. Lors de cette configuration, un flux vidéo caméra de surveillance qui filme en continu est présenté séquentiellement au système une image f (*frame*) à la fois. Les trames passent premièrement par un module FAST-DT effectuant l'évaluation en parallèle de la détection de visage et le suivi de trajectoire. À tous les $N_f = 10$ *frames*, une triple détection Viola-Jones, soit une détection de visage frontal et deux de profil (gauche et droite), sont enforcées afin d'évaluer la présence de nouvelles régions d'intérêt de visages $\{ROI_d\}$ dans la scène. Les régions détectées sont alors envoyées au sous-module de suivi de trajectoire, basé sur l'algorithme STRUCK, afin de (1) initialiser de nouvelles trajectoires dans le cas d'un nouveau visage apparaissant dans la scène, (2) valider la position actuelle d'une trajectoire existante, selon l'association des données évaluées et la confidence associée à la détection par rapport à

chaque individu déjà suivi (*individual-specific trajectories*), et (3) supprimer les trajectoires rendues invalides lorsque l'individu a quitté la scène. L'indice *IoU*, présenté à l'équation (3.1), est quant à lui employé entre chaque détection ROI_d et la position actuelle de chaque trajectoire spécifique ROI_t afin d'inférer la confidence de celles-ci, tel qu'illustré avec la Figure 3.2.

$$IoU = \frac{area(ROI_d) \cap area(ROI_t)}{area(ROI_d) \cup area(ROI_t)} \tag{3.1}$$

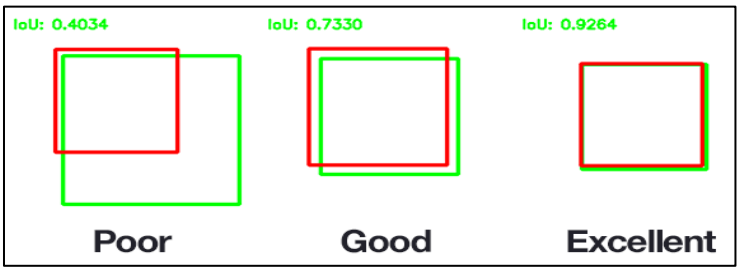

Figure 3.2 Différents niveaux de confidence inférés de deux régions d'intérêt à partir de la mesure *IoU* [114]

La contre validation effectuée entre les détections et les trajectoires permet ainsi d'assurer des positions de visages valides avec haute certitude, puisque trois sources évaluations distinctes sont fusionnées, soit (1) la détection en cascade frontale, (2) les deux détections en cascade de profil (à noter qu'il s'agit fondamentalement de différents modèles que le frontal, ce qui produit différentes réponses), et (3) un modèle adaptatif interne de visage à base de SVM qui s'améliore grâce aux nouvelles apparences pour assurer un meilleur suivi de visage le long de la trajectoire. En plus de ces dernières validations, l'algorithme *Hungarian* est employé afin d'accomplir une association de données entre les multiples ROI trouvées. Cet algorithme détermine la combinaison associative optimale entre les détections ROI_d et les trajectoires ROI_t en fonction de la minimisation des coûts établis par *IoU*. Cela permet donc non seulement de regrouper les régions de visages communes, mais également, de dissocier les ROI appartenant à différentes trajectoires [34]. Il devient ainsi possible d'avoir multiples suivis de trajectoires en parallèle et pouvant se croiser dans la scène (une personne obstruant une autre en passant devant). Le système de FR est alors capable d'effectuer de manière efficace de multi-reconnaissances de visages en simultané à différentes positions dans la scène.

La combinaison des multiples détections en cascade est également accomplie de sorte à repérer les nouveaux visages le plus tôt possible, même si ceux-ci se présentent sous une orientation (pose) très peu favorable à la reconnaissance de visage. Cela permet d'assurer un suivi de trajectoire dès que la personne est captée, plutôt que de risquer d'attendre pour une meilleure apparence frontale, et alors risquer de manquer la détection si le cas ne se présente jamais. Le but est donc de réduire les cas de *fail-to-acquire* (FA) autant que possible. Cette détection hâtive a également pour avantage de fournir davantage d'apparences au modèle de visages interne pour le FT, ce qui construit progressivement sa robustesse contre de futures apparitions sous illuminations et poses variées.

Entre chaque intervalle de N_f images avec détections, le suivi de trajectoire prend en charge la tâche de mise à jour de la position du visage dans la scène, selon sa trajectoire correspondante à chaque nouvelle trame *f*, afin de continuellement produire des régions d'intérêt $\{ROI_t\}$. Les détections VJ ne sont pas effectuées à chaque image, car celles-ci sont notablement coûteuses en temps de calcul, ce pour quoi elles sont étalées dans le temps à travers N_f images. Le sous-module de FT s'assure également de gérer l'initialisation ou la destruction d'accumulateurs de trajectoires mémorisées (voir ligne hachurée à la Figure 3.1), qui eux assureront une association spatio-temporelle ultérieure des scores de reconnaissance de visage.

Par la suite, les régions d'intérêt $\{ROI_t\}$ produites subissent une étape de validation et d'ajustement à l'aide d'une détection frontale Viola-Jones localisée, qui vient produire les régions $\{ROI_t^L\}$. Cet ajustement local permet d'effectuer une normalisation à haut niveau de la position des ROI obtenues par le FT, qui peuvent autrement différer du FD employé à l'entraînement. Plus de détails sont fournis à ce propos à la section 3.2.

Les régions localisées sont par la suite employées avec le même processus de prétraitement, de segmentation en *patches* et d'extraction de descripteurs HOG qu'à la phase d'entraînement du système pour obtenir des vecteurs descripteurs $\{\mathbf{a}_{t,p}\}$ comparables. Ces vecteurs sont donc présentés aux Eo*e*-SVM spécifiques à chaque individu afin de produire une matrice de scores de classification $\mathbf{S}_{s,t,p}$ contenant les valeurs numériques de similarité évaluées entre chaque

patch p de visage de la trajectoire *t* observée sur la trame actuelle *f* par rapport au modèle préalablement entraîné d'un individu *s*. Un processus de fusion des scores par sous-région (*patch-based fusion*) ainsi qu'une normalisation des résultats est accomplie afin d'obtenir une matrice de scores $\mathbf{S}_{s,t}$ situés dans l'intervalle $[0,1]$ et correspondants à chaque combinaison individu-trajectoire. Puis, les scores obtenus à chacun des instants discrets *f* sont accumulés spatio-temporellement selon une fenêtre glissante des $W = 20$ dernières images de la vidéo, et pour chaque trajectoire indépendamment, produisant une matrice de score $\mathbf{S}_s^{acc^*}$ associant la similitude évaluée entre chaque visage suivi et chaque individu enregistré dans le système. Finalement, l'accumulation de scores, comparée à un seuil de reconnaissance γ_s spécifique à chaque individu, permet d'identifier si celui-ci est repéré dans la scène. Ce dernier point était illustré à la Figure 3.1, mais est davantage mis en emphase à la Figure 3.3 suivante, où l'on peut voir que l'accumulation de scores d'individus d'intérêt à reconnaître dépasse le seuil de reconnaissance γ_s lors de leur apparence sur vidéo-caméra, alors que l'accumulation est plus faible pour les autres individus, tant ceux inconnus du système que ceux dans la *gallery*.

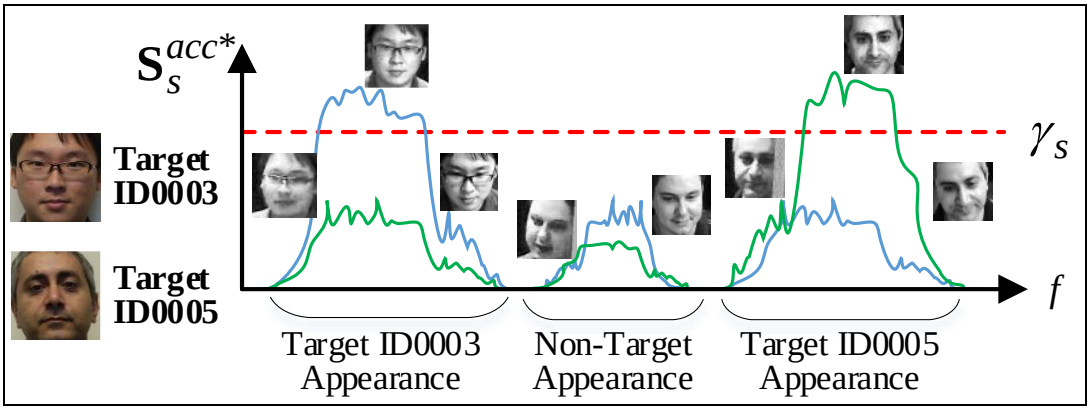

Figure 3.3 Accumulation de scores spatio-temporels par trajectoire pour un individu recherché, en comparaison avec la reconnaissance des autres personnes

Les sections qui suivront apporteront les quelques détails et explications supplémentaires par rapport aux modules présentés à la Figure 3.1, afin de justifier et approfondir sur certains choix d'implémentation effectués.

3.2 Prétraitement et extraction de caractéristiques

Afin d'améliorer autant que possible les chances de bonnes reconnaissances de visages, quelques étapes de prétraitement sont appliquées sur les images et les vidéos d'entrée au système. Tout d'abord, il y a une détection locale VJ qui vient raffiner la ROI_t provenant du module de suivi de trajectoire en une ROI_t^L localisée. Le but de cette opération de FD supplémentaire est double : (1) retenir seulement les ROI suffisamment frontales de sorte qu'elles s'apparentent à celles employées lors de l'entraînement du système, et (2) corriger la dérive de position introduite par le FT pour reproduire la ROI qui serait autrement obtenue si l'algorithme VJ était plutôt appliqué sur l'intégralité de la scène à chaque *frame* du vidéo. En effet, puisqu'un *tracker* entraîné à suivre le visage avec un modèle adaptatif est employé pour générer les ROI de visages pendant l'opération, celui-ci va nécessairement générer des régions différentes du modèle en cascade VJ. Cela est causé par sa capacité à s'adapter qui lui permet de modifier sa région d'analyse à l'interne pour réaliser au mieux possible sa tâche spécifique, soit le suivi de visage, et non la détection générale de visage comme VJ. Ainsi, si le module de FT venait à décider que la région proposée par la FD initiale n'est pas une représentation adéquate pour suivre le visage efficacement à travers le vidéo, la ROI peut y être ajustée pour lui simplifier la tâche de recherche du visage sur le *frame* suivant, et c'est ce que l'on observe en pratique. La Figure 3.4 témoigne de cette variation notable. Ainsi, une recherche localisée sur une région étendue (hachurée en jaune) d'approximativement 25% à 50% par rapport à la ROI produite par le *tracker* est employée pour rechercher et corriger la position du visage.

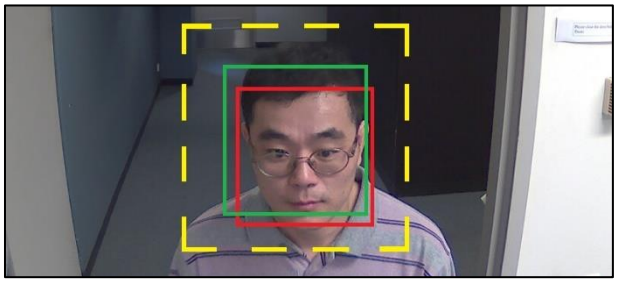

Figure 3.4 ROI détectée par Viola-Jones frontal (rouge),
ROI obtenue du suivi de trajectoire STRUCK (vert) et
zone de recherche locale par Viola-Jones (jaune hachuré)

Étant donné qu'une détection locale doit être utilisée pour corriger par après la dérive potentielle de la région du FT, il serait possible d'argumenter le fait même d'utiliser le suivi de trajectoire. Il existe en fait plusieurs raisons pour lesquelles il est nettement plus important et justifié de conserver celui-ci dans le système, soit :

1. L'ajout de multiples contre-validations robustes qui viennent supprimer rapidement les détections erronées de non-visages dans la scène;
2. Le grand gain en rapidité pour le temps réel que le FT procure par rapport à la FD brute;
3. Le faible impact en complexité, temps et calcul requis par l'application de VJ sur une zone de recherche très localisée;
4. La forte dépendance existante entre l'accumulation de scores et le besoin de regrouper les visages correspondants à travers une trajectoire commune.

Effectivement, puisque la recherche s'effectue localement, pratiquement aucune translation de la fenêtre de recherche des classificateurs en cascade n'est nécessaire, et il en est de même pour le nombre d'échelles à visiter. À l'inverse, la détection globale requiert une recherche exhaustive à travers toute l'image et pour une très vaste quantité de mises à l'échelle, ce qui est trop long pour une utilisation à chaque *frame* en temps réel, encore moins lorsque les trois cascades sont employées en parallèle. Les points ci-haut ont été validés à l'aide d'une petite analyse dont les résultats figurent au Tableau 3.1, et qui supporte la méthode employée.

Tableau 3.1 Comparaison du nombre de détections manquées (FA), du nombre de fausses détections (FP) et du temps de détection[3] sur toute la base ChokePoint, pour les variantes de FD frontale, de multiples VJ en parallèle et du système de FR utilisé avec suivi de trajectoire

	FD(+FT)	**FA**	**FP**	**s/frame (±SE)**	**FPS (±Δfps)**
VJ-1CC (frontal)	63300 (87.5%)	9029 (12.5%)	1448	0.2538 ± 0.14	3.94 ± 1.40
VJ-3CC (frontal+profils)	64817 (89.6%)	7512 (10.4%)	5117	0.2748 ± 0.09	3.63 ± 0.89
VJ-3CC (n_f=10) + STRUCK	67369 (93.9%)	4363 (6.1%)	5763	0.0598 ± 0.01	16.72 ± 2.39

[3] Évalué par une machine équipée de carte Intel® Core™ i7-5500U (2.40Ghz) 16GB RAM, et d'un GPU NVIDIA GeForce GTX 950M, avec algorithmes optimisés par les bibliothèques OpenCV, OpenMP et CUDA.

Il est à noter qu'étant donné que 3 détections sont employées, il est normal d'avoir plus de FP qu'avec une seule. Aussi, puisque l'initialisation des trajectoires dépend des multiples FD, le système ne peut malheureusement pas s'affranchir totalement des FP que celles-ci génèrent. Par contre, le système réussit tout de même à bien reconnaître les visages, car les ROI détectées sont maintenant combinées par trajectoires, plutôt qu'être évaluées individuellement (par transaction) avec seulement VJ. Plusieurs FP sont aussi dues aux problèmes de la base de données, tel que discuté à la section 5.1.1, ce qui fait que ces quantités sont en fait surévaluées.

Suite à l'obtention de la région locale corrigée, où une conversion en niveaux de gris avait été accomplie pour appliquer VJ, un raffinement supplémentaire venant enlever 20% du pourtour de la région est réalisé, ce qui réduit grandement la présence d'arrière-plan dans la ROI. Ceci est basé sur l'amélioration proposée dans [107] (aussi illustré à la Figure 2.3, section 2.1), et démontré ci-bas à la Figure 3.5, de sorte à utiliser un maximum de caractéristiques des vecteurs descripteurs pour représenter davantage les traits saillants du visage, plutôt que la scène de l'arrière-plan fortement bruitée et de faible intérêt.

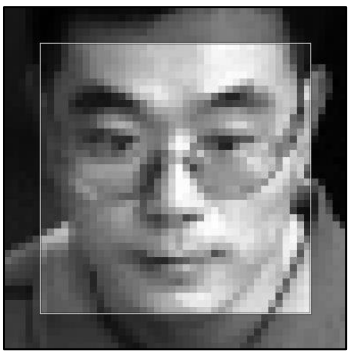

Figure 3.5 Réduction d'arrière-plan pour se concentrer sur les traits importants du visage avec raffinement de 20% de la ROI de la FD Viola-Jones locale

Le reste des étapes de prétraitement consistent en une égalisation d'histogramme qui est effectuée sur les images en niveaux de gris afin de rehausser le contraste visible sur les visages, et un redimensionnement à une taille constante de 48x48 pixels. L'égalisation d'histogramme permet de faire ressortir davantage les nuances qui seront ultérieurement captées par la

caractérisation par gradients du descripteur HOG, tout en réduisant l'impact des variations d'illumination nuisibles à travers les différentes images de la vidéo [9]. Cette même référence est aussi employée comme base pour le choix de certains paramètres.

Tel que proposé par les travaux [9, 10, 11], les descripteurs HOG extraient à partir de 9 subdivisions du visage de 16x16 pixels semble être la meilleure façon de procéder pour avoir une représentation à la fois rapide et efficace des visages, tout en apportant une très forte robustesse aux variations nuisibles de pose et d'illumination. La méthode HOG semble être la plus efficace selon les résultats qui y sont présentés, et il s'agit également de la conclusion tirée avec une rapide évaluation empirique effectuée (voir ANNEXE III). Ce descripteur a donc été retenu pour le système réalisé. L'utilisation de *patches*, telles qu'illustrées à la Figure 3.6, est aussi supportée à travers les auteurs dans [2] et [6], où l'on note que la méthode de subdivision est particulièrement adaptée pour combattre les cas d'occlusion partielle sur les visages en environnement sans-contraintes grâce à la classification locale produite par l'approche. On peut également voir un exemple de telle occlusion produite par un portail à la Figure 3.7, où l'avantage des *patches* montre que seulement les descripteurs de droite seront affectés, plutôt que l'intégralité du descripteur qui serait autrement obtenu par l'extraction globale sur l'ensemble du visage. La méthode est d'autant plus renforcée par la précision accrue apportée avec l'exploitation d'un ensemble de classificateurs qui apprendront chacun à mieux se spécialiser aux problématiques rencontrées sur leur section de visage particulière [9].

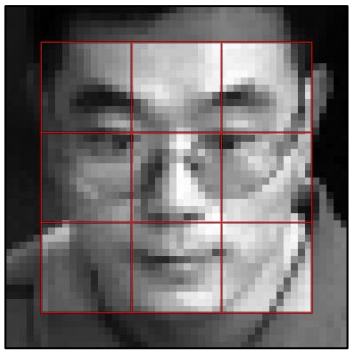

Figure 3.6 Extraction des subdivisions de visage localisé

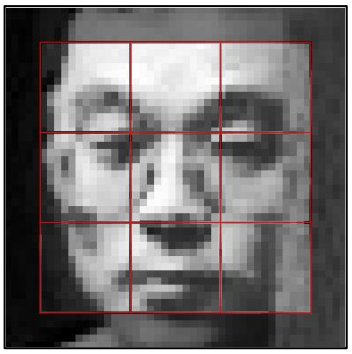

Figure 3.7 Robustesse face aux occlusions localisées

Pour finir, les paramètres employés pour extraire les vecteurs HOG des *patches* de visage sont basés à la fois sur l'étude dans [44], qui évalue multiples combinaisons pour déterminer les meilleurs cas, sur quelques validations expérimentales, et selon les travaux de Bashbaghi et al. (2014-2017) [9, 10, 11, 12, 13]. Le descripteur final retenu ($N_d = 588$ *features*), est produit à partir de cellules 2x2, de blocs 2x2, d'un déphasage 2x2, et de 3 bandes d'histogramme.

3.3 Ensemble de classificateurs *e*-SVM spécialisés

L'approche par Eo*e*-SVM est employée initialement dans le projet selon les caractéristiques intéressantes décrites au Chapitre 2. Les *e*-SVM utilisés pour chaque subdivision d'images distincte sont entraînés à partir d'une très grande quantité d'échantillons négatifs provenant d'individus inconnus apparaissant sur les séquences vidéo où nos *targets* doivent être reconnus pour former le modèle par contre-exemples (*universal background model*). L'abondance de vecteurs négatifs provenant de vidéos permet d'accomplir une recherche approfondie des cas les plus complexes (*hard negative mining*) pour entraîner des *e*-SVM robustes, en définissant la meilleure marge de décision par rapport à l'image de référence de chaque individu visé. L'approche par ensemble vient quant à elle assurer une solution qui se généralise bien aux nouvelles observations de visages dans la scène en réduisant les chances d'*over-fitting* [11].

Étant donné que la classe positive ne contenant qu'un seul exemple est nettement désavantagée lors de l'entraînement par rapport aux milliers de contre-exemples, et qu'il s'agit de celle dont on préfère accorder plus d'importance pour réussir la FR, un coût de mauvaise classification très élevée $C^+ = 1$ est appliquée à la classe positive, contrairement à $C^- = 0.01$ pour la classe négative. Ces hyper-paramètres d'entraînement sont tirés des travaux sur les *e*-SVM discutés dans [12] ainsi que les expérimentations effectuées dans [147] qui, d'une partie, proposent directement les valeurs de ces pondérations, mais qui démontrent également que leur utilisation augmente optimalement la performance atteignable par les SVM. Les valeurs spécifiques ne sont pas nécessairement importantes, tant qu'une pondération suffisante est appliquée pour contrefaire le déséquilibre marqué observé entre les deux classes. Les *e*-SVM sont finalement entraînés avec la sortie probabiliste, pour obtenir des similarités plutôt que les classes binaires.

3.4 Étapes de normalisation et de fusion

Afin d'effectuer une bonne classification des données avec les Eo*e*-SVM, les étapes de normalisation des descripteurs et des scores obtenus sont primordiales. Une attention particulière a donc été apportée à la normalisation tant des *features* servant à l'entraînement des classificateurs que les valeurs obtenues lors de la fusion des scores de prédiction.

3.4.1 Normalisation des descripteurs

Pour normaliser les descripteurs de caractéristiques, plusieurs options s'offrent à nous. Il est possible d'appliquer une normalisation globale sur l'ensemble des caractéristiques, de l'effectuer par vecteur descripteur respectif, ou de l'appliquer par *features* correspondantes, tel qu'illustré avec les regroupements en rouge de la Figure 3.8. En plus de ces trois méthodes, il est possible de considérer une normalisation globale pour l'ensemble des *patches* du visage ou une normalisation respective à chacune.

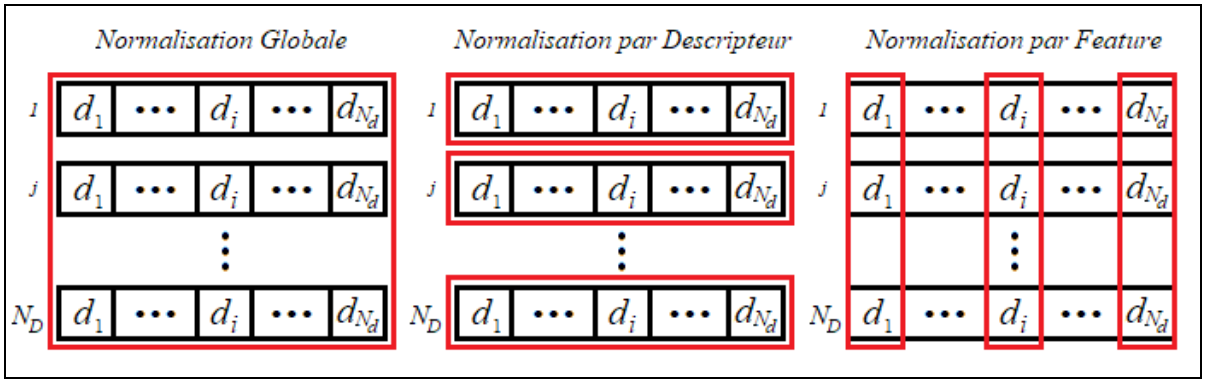

Figure 3.8 Types de normalisations des descripteurs par divers groupement des données

Basé sur des évaluations empiriques, il a été déterminé que la normalisation par *patch* ainsi que par *feature* spécifique était la meilleure des combinaisons. Cette observation semble ainsi venir confirmer que chaque caractéristique correspondante à travers les vecteurs descripteurs reflète des propriétés bien particulières et distinctes sur le visage, ce qui fait qu'elles doivent être séparément normalisées par rapport à leurs propres plages dynamiques. Ceci signifie que

les vecteurs pourront adéquatement être comparés entre différents visages afin d'évaluer leur similarité, et ainsi déterminer le niveau de certitude de la FR. La normalisation par *patch* étant meilleure indique également que les caractéristiques sont bel et bien spécialisées pour chaque partie du visage distinctivement, faisant qu'une caractéristique d_i d'un *patch* ne correspond aucunement à celle d'un autre *patch*. Cette combinaison logique retenue a donc été appliquée à travers le projet. La fonction *min-max* (3.2) est employée pour normaliser, car la plage [0,1] qu'elle procure est garantie, à l'opposé d'une distribution uniforme avec *z-score*, par exemple.

$$d_{i,p}^{(norm)} = \frac{d_{i,p} - \min_{j,p}(d_{i,p})}{\max_{j,p}(d_{i,p}) - \min_{j,p}(d_i)} \quad \forall j \in [0, N_D], \forall p \in [0, N_p] \qquad (3.2)$$

3.4.2 Normalisation des scores et fusion par ensemble

En ce qui a trait aux scores de prédiction des classificateurs, une normalisation est nécessaire étant donné le déséquilibre des classes d'entraînement. En effet, malgré que des pondérations très fortes aient été appliquées pour favoriser les vecteurs de représentations de visages positifs (voir section 3.3), la surabondance d'exemples négatifs amène les *e*-SVM à produire des scores déphasés négativement, tel qu'illustré à la Figure 3.9. Cela n'est toutefois pas problématique pour discriminer les individus d'intérêt du reste parce que leurs apparences sur vidéo ont habituellement tendance à produire des scores « *moins négatifs* » que dans le cas d'apparitions d'individus inconnus ou autres que celui représenté par le *e*-SVM correspondant.

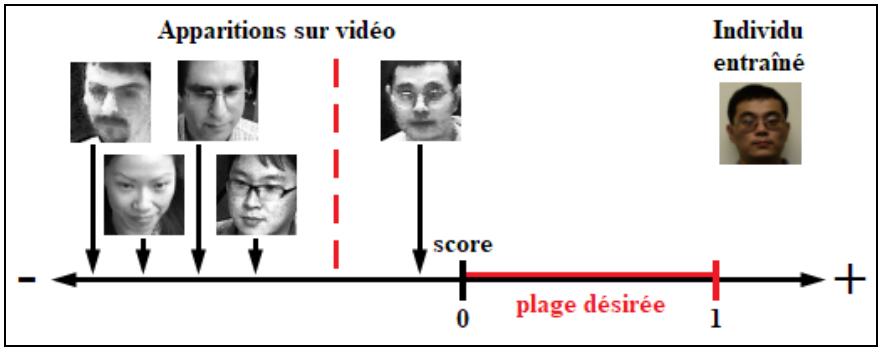

Figure 3.9 Déphasage des scores bruts de prédiction des *e*-SVM fortement déséquilibrés vers la classe négative au détriment du positif à représenter

Les scores obtenus pour chaque e-SVM$_{s,p}$ doivent donc impérativement être normalisés par rapport à leur plage dynamique respective, pour les mêmes raisons que décrites lors de la normalisation des descripteurs de caractéristiques. Après normalisation des $N_p = 9$ *patches*, ainsi que pour tous les individus d'intérêt *s*, nous pouvons passer à la fusion de la matrice de scores $\mathbf{S}_{s,t}$ selon l'équation (3.3). Ceci indiquera la prédiction de FR des individus dans la *watchlist* par rapport à un instant précis (*frame* particulier) le long d'une trajectoire *t* donnée.

$$\mathbf{S}_{s,t} = \frac{1}{N_p} \sum_{p=1}^{N_p} \mathbf{S}_{s,t,p} \tag{3.3}$$

3.5 Accumulation spatio-temporelle par trajectoire

L'accumulation de plusieurs scores correspondants à des apparitions spatio-temporellement reliées par une même trajectoire de visages permet de graduellement assurer plus de confiance aux prédictions effectuées, car elles viennent se valider entre elles. Dans le cas de scores irréguliers dus aux conditions d'acquisition, l'accumulation vient filtrer la réponse bruitée. Cette confiance est notablement importante dans le cas d'illuminations inattendues, de poses trop sévères du visage pour bien le reconnaître, ou lors de flou de motion trop prononcé pour extraire des caractéristiques suffisamment discriminantes servant à distinguer les individus dans la scène. Ainsi, l'accumulation de scores par trajectoires assure une FRiVS beaucoup plus robuste par rapport à ces changements inconsistants sur les visages, offrant une efficacité accrue du système dans l'ensemble [38, 39].

Étant donné que le système fonctionne de manière continue, une fenêtre glissante de $W = 20$ *frames* est utilisée afin de conserver et réutiliser les plus récentes prédictions pour chaque combinaison de trajectoire-individu. L'accumulation de scores s'effectue ainsi selon une somme des derniers scores retenus selon l'équation (3.4), et tel qu'illustré par la Figure 3.10.

$$\mathbf{S}_s^{acc*} = \sum_{w=0}^{W-1} \mathbf{S}_{s,t(W-w)} \quad \in [0, W] \tag{3.4}$$

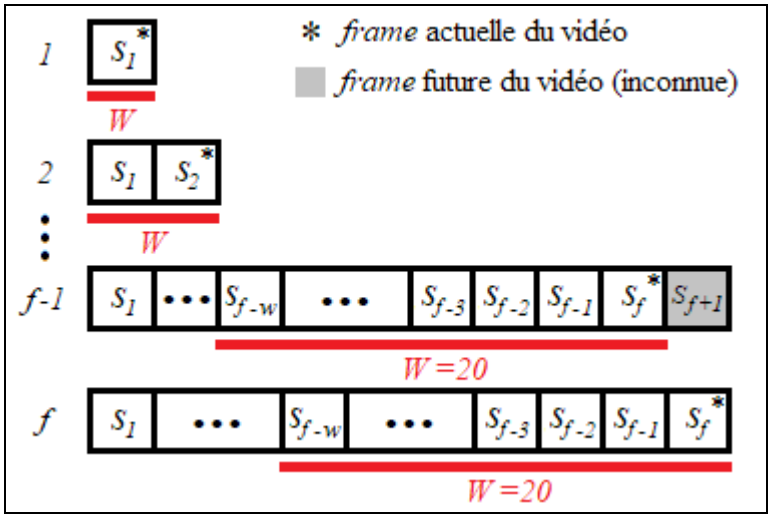

Figure 3.10 Représentation de l'accumulation de scores spatio-temporels avec fenêtre glissante par trajectoire-individu

Avec la Figure 3.10, nous pouvons voir que les scores sont initialement accumulés que pour les *frames* disponibles. Les autres scores dans la somme à l'équation (3.4) y sont donc mis à zéro temporairement, jusqu'à ce que W scores soient atteints. Par la suite, la fenêtre glissante remplace graduellement le plus vieux score par le plus récent et réapplique la somme sur la nouvelle plage considérée à chaque nouvelle prédiction, soit sur le nouveau *frame* vidéo.

3.6 Optimisations pour le temps réel

Étant donné que l'un des critères primaires qui a été imposé pour le système de FRiVS conceptualisé, excepté bien évidemment la reconnaissance de visages performante, est d'assurer une efficacité opération rapide pour viser le temps réel, plusieurs considérations importantes d'implémentation ont été accomplies. Les quelques paragraphes qui suivent présentent quelques-uns des multiples développements accomplis pour optimiser davantage l'architecture du système.

3.6.1 Modification du code de référence pour le temps réel

Les bibliothèques de codes C++ OpenCV [22], OpenMP [97], et CUDA [90, 92, 93] ont largement été employées dans ce projet afin d'optimiser les procédés coûteux telle la triple détection VJ. Des besoins d'optimisation pour atteindre de meilleures rapidités d'exécution, une grande majorité du code de référence de FAST-DT, sinon dans son intégralité, a été remodélisé afin d'y introduire des opérations plus rapides ou des procédés en parallèle. Il en est de même pour les SVM qui ont été implémentés avec une version parallélisée des codes de LIBSVM [27] (*libsvm-multicore*) ainsi que des variations additionnelles de LIBLINEAR [52]. Ainsi, il ne serait pas surprenant qu'un lecteur qui tente de réaliser le code correspondant au procédé de FRiVS présenté dans ce projet, mais en employant les références originales des codes sources et algorithmes cités, n'atteignent pas les mêmes vitesses qu'atteintes ici. En effet, certaines parties du code a même été ajusté jusqu'au point de modifier des fonctions source d'OpenCV pour obtenir des variations différentes de `CascadeClassifier`[4]. Ainsi, dans un esprit de transfert de connaissances et d'ouverture, tout lecteur qui serait intéressé à poursuivre dans les mêmes voies ou à obtenir plus de détails est recommandé et le bienvenu de contacter l'auteur du rapport de projet[5].

3.6.2 Modèles binaires et temps d'entraînement

Étant donné que le système de FRiVS, dans une application réelle, serait employé dans des contextes où beaucoup d'individus seraient enregistrés dans la base de référence, il existe un besoin de performance pour le stockage mémoire et le temps de chargement des modèles. Or, l'approche par SVM est très compacte, ce qui fait que les modèles prennent $< 5\text{KB/SVM}$. L'entraînement à partir de plus de 20000 contre-exemples négatifs et 1 positif sont quant à eux effectués en quelques microsecondes, ce qui rend le système fort efficace.

---

[4] OpenCV, bibliothèque publique de code C++ pour la vision numérique (*Open Source Computer Vision*), Documentation de la classe « *CascadeClassifier* » https://docs.opencv.org/trunk/d1/de5/classcv_1_1CascadeClassifier.html

[5] Francis Charette Migneault, ing. jr. – francis.charette.migneault@gmail.com

3.6.3 Mémorisation et accès rapide des scores pour l'accumulation

Afin d'accomplir l'accumulation spatio-temporelle de scores, de sorte à regrouper les visages de mêmes trajectoires-individus ensemble, une mémoire circulaire (*circular buffer*), tel qu'illustré à la Figure 3.11, est employé avec une dimension maximale correspondant à la fenêtre glissante *W* des plus récents *frames*. Celui-ci permet d'automatiquement conserver uniquement les scores les plus récents pour assurer une analyse *on-line* et en temps réel, étant donné que sa structure interne à deux pointeurs écrase, par concept, les plus vieilles entrées par les nouvelles. Ce type d'approche est couramment employée pour le traitement de signaux *on-line* en raison de son efficacité et rapidité d'accès mémoire aux données sauvegardées, tel que discuté dans [119]. La rapidité de cette méthode est principalement associée à la dimension fixe du conteneur et des pointeurs de contrôle. Cela évite de réallouer continuellement des espaces mémoire à chaque nouveau *frame*, ce qui se produit très fréquemment, et réduirait grandement la vitesse du code autrement. Aussi, l'écrasement naturel des données fait qu'aucun déphasage (*copy-move* en mémoire) de l'intégralité du conteneur d'un index n'est requis à chaque itération. Ceci permet d'appliquer une simple somme sur une zone mémoire constamment préservée, pour une accumulation optimale en temps de calcul.

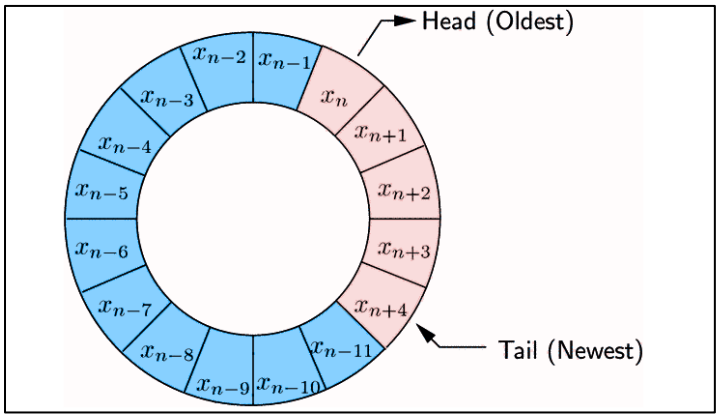

Figure 3.11 Représentation d'une mémoire circulaire (*circular buffer*) pour l'accumulation de scores de reconnaissance de visages[6]

[6] Tiré de « Lab 3: Real-Time Block Simulation in MATLAB », [Online] http://www.faculty.jacobs-university.de/jwallace/xwallace/courses/dsp/labs/03_ml_block/

CHAPITRE 4

SYSTÈME AVEC MULTIPLES REPRÉSENTATIONS DE VISAGE

Ce chapitre introduira les ajustements apportés au système initial de FRiVS (Chapitre 3), afin de rehausser davantage sa précision et sa robustesse. Les méthodes explorées et leurs justifications respectives seront introduites pour mettre en situation les évaluations expérimentales qui suivront au prochain chapitre (Chapitre 5). Chaque sous-section consiste en un bloc de tests qui ont été effectués en parallèle, pour y tirer des conclusions, avant de passer au prochain.

De manière plus spécifique, la problématique associée à la limitation imposée par l'image unique de référence sera présentée. Celle-ci donnera une meilleure mise en contexte de la complexité reliée à la réalisation spécifique d'un système complet de FRiVS basé sur le mode d'opération SV2 avec une seule représentation du visage par individu à entraîner dans le système. Ainsi, le problème associé à l'entraînement des ensembles de classificateurs qui sont exploités permettra de mettre plus d'emphase sur les besoins de représentations multiples robustes des visages, afin que le système demeure invariant aux conditions de l'environnement dynamique, malgré les variations brusques et imprévisibles pouvant être apportées par le suivi de trajectoire de visages des individus qui traversent la scène.

Ensuite, les approches explorées pour permettre d'adapter les visages produits par le module de suivi de trajectoire vers le module de classification seront présentées à proprement parler. Les méthodes considérées qui seront présentées regroupent de la génération de visages synthétiques, de la sélection aléatoire de caractéristiques et de l'adaptation au domaine des caméras. Dans tous les cas, les approches qui seront introduites viseront à adapter les visages, à l'aide d'une addition de multiples représentations de ceux-ci, pour venir aider au transfert de ceux-ci entre les modules mentionnés, de sorte à rendre le système plus invariant aux changements à travers les sévères conditions d'acquisition.

4.1 Problématique de représentation de visages limitée

Considérant que les Eo*e*-SVM ne possèdent qu'un seul exemple d'entraînement (l'image haute résolution de référence unique), la performance atteignable par ceux-ci est grandement limitée par cette unique représentation. À cause de l'unique échantillon qui leur est disponible, il est très complexe pour les *e*-SVM de modéliser la variation intra-classe à caractériser, de sorte à bien reconnaître les multiples combinaisons de visages possibles à partir des conditions d'acquisition variées dans la scène sans, toutefois, se tromper avec les visages des autres personnes. La variation intra-classe potentielle des représentations positives de l'individu est représentée à la Figure 4.1 par les multiples régions grisées. On peut donc y voir que les vecteurs de support en rouge établissent un hyperplan très clair, mais l'étendue de la classe positive demeure inconnue à partir d'un seul échantillon.

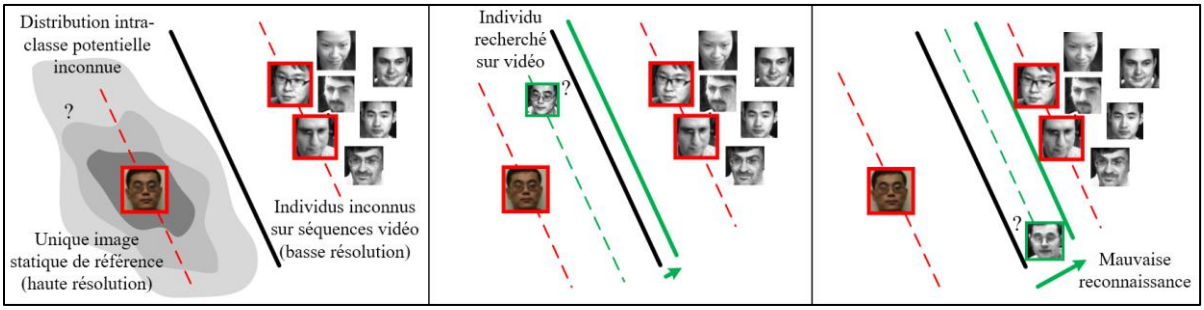

Figure 4.1 Représentation du manque de connaissance des Eo*e*-SVM pour définir la portée de la distribution intra-classe positive et classer les nouvelles apparences de test

Autrement dit, les *e*-SVM peuvent facilement déterminer la meilleure marge de séparation qui existe entre la classe de leur individu à représenter et les autres, car ils ont beaucoup de cas pour tester cette frontière de décision. Par contre, le manque d'exemples de la personne visée fait qu'ils n'ont aucune idée de l'étendue réelle de la distribution où les nouveaux visages potentiels du *target* peuvent se situer dans l'espace de haute dimensionnalité des vecteurs descripteurs. Ainsi, si l'on supposait que deux échantillons « probables » subvenaient, en vert à la Figure 4.1, le cas au centre serait bien classé, mais pas celui de droite. Les hyperplans de séparation qui seraient autrement obtenus selon ces nouvelles représentations seraient aussi très différents selon le cas. Cela démontre ainsi la sensibilité que les *e*-SVM ont aux vecteurs

d'entraînement positifs (aussi à cause du coût C^+ appliqué), et pourquoi il est si important de bien définir la distribution intra-classe des individus *target*. Donc, s'il venait possible de connaître à l'avance (à l'entraînement) approximativement jusqu'où ces échantillons pourraient s'étendre dans l'espace à haute dimensionnalité, cela permettrait aux *e*-SVM de corriger leurs poids pour obtenir un hyperplan plus adéquat aux attentes des représentations futures dans la scène, permettant une bonne reconnaissance de l'individu même dans les cas complexes d'acquisition.

L'ajout d'information permettant d'indiquer aux classificateurs jusqu'où s'étale la classe de visages de l'individu leur permettrait définitivement d'être plus robustes aux facteurs nuisibles d'acquisition vidéo. Par contre, le contexte du SSPP fait que nous ne possédons pas directement de tels exemples supplémentaires de la personne à reconnaître. Il devient alors nécessaire de se tourner vers d'autres approches pour trouver des méthodes qui fournissent davantage d'exemples à chaque ensemble de classificateurs, afin que ceux-ci puissent apprendre à supprimer les doutes de reconnaissance qui existent entre les divers individus qui sont captés dans la scène et les visages inconnus de la personne à reconnaître.

Une méthode très répandue dans la littérature permettant de remédier au manque de données pour des classes déséquilibrées à classifier est l'augmentation de données (*data augmentation*) [126, 134]. L'augmentation de données peut consister en diverses techniques, par exemple, en augmentant l'information par (1) *data warping* ou par (2) *over-sampling* [138]. Dans le contexte de l'imagerie par ordinateur, la première approche consiste à générer de nouvelles représentations d'images en appliquant habituellement des opérations géométriques traditionnelles, soit des translations, des rotations, du changement d'échelle ou du cisaillement. Ce type de transformation est donc appliqué au niveau des données initiales (*data-space*) pour générer des échantillons variés. La seconde méthode vise plutôt à générer des données directement au niveau des descripteurs de caractéristiques (*feature-space*), par exemple, avec l'algorithme SMOTE [30]. Une telle application est d'ailleurs présentée dans [42], pour la classification SVM de classes déséquilibrées. Ainsi, les techniques d'augmentation de données sont très adéquates pour les cas déséquilibrés comme le SSPP.

En ce qui concerne la génération de données dans l'espace des images, il est également possible de considérer des techniques plus avancées que les transformations géométriques de base. Entre autres, il y a la génération d'illumination, qui a été explorée par plusieurs auteurs spécifiquement pour la reconnaissance de visages [31, 86, 98, 133], ainsi que la génération de pose [18, 123, 128], ou encore l'alignement de pose du visage [153]. Aussi, des auto-encodeurs permettant l'apprentissage de représentations variées de visages ont été employés dans [101], ce qui peut être exploité pour obtenir davantage de visages des individus *targets* avec l'application de conditions d'acquisition variées. Malgré la variabilité des techniques, toutes démontrent un avantage net à employer de telles méthodes de génération de nouvelles représentations des visages, car il devient possible de transférer l'information obtenue dans le domaine des vidéo-caméras, et de l'exploiter pour obtenir des classificateurs plus robustes à l'environnement. Des études [28, 104] ont également spécifiquement évalué les bénéfices obtenus avec diverses approches d'additions de représentations synthétiques et d'illumination, afin de confirmer grand nombre de propositions existantes.

Ainsi, basées sur les justifications indiquées aux publications des méthodes faisant partie de l'état-de-l'art, la génération de représentations de visages et l'augmentation de données semblent appropriées pour les requis du projet. Certaines de ces techniques ont donc été évaluées avec l'architecture de FRiVS précédemment développée, afin de les employer si elles s'avéraient effectivement apporter un bénéfice. Les sections suivantes élaboreront sur ceci.

4.2 Système de FRiVS avec l'ajout des visages synthétiques

4.2.1 Génération synthétique par transformations géométriques

Le premier type de génération synthétique exploitée est basé sur des transformations géométriques, afin de produire les échantillons positifs additionnels requis pour permettre aux *e*-SVM de mieux représenter la distribution de la classe positive dans l'hyper-espace de caractéristiques, tel qu'introduit précédemment. Plus particulièrement, la notion de sélection spécifique d'exemplaires les plus compétents est introduite, en offrant des échantillons supplémentaires formés de variations à partir de l'unique image statique de référence. Les

travaux dans [83], qui utilisent également les classificateurs *e*-SVM, mais pour la classification de voitures, ont inspiré l'étape qui suit. L'idée générale dans ce travail est que, malgré que les voitures forment un groupe similaire sur certains points (formes et dimensions semblables par rapport aux autres objets), le groupe est fondamentalement très diversifié (couleurs, modèles, etc.). Ainsi, tenter de tout représenter le groupe avec un seul exemplaire est trop complexe mais, par la sélection d'un ensemble d'exemplaires les plus compétents, la tâche est réalisable. En contexte de FRiVS, le concept est aussi applicable puisque chaque individu à reconnaître est essentiellement représenté par plusieurs types d'apparitions dues à l'illumination, la pose, l'expression et ainsi de suite, malgré qu'il s'agisse de la même personne fondamentalement. Par contre, en SSPP, nous devons générer les exemplaires les plus compétents pour permettre leur sélection, car nous ne disposons pas de ceux-ci comme dans [83]. Nous procédons alors à une augmentation de données par transformations géométriques comme illustré à la Figure 4.2.

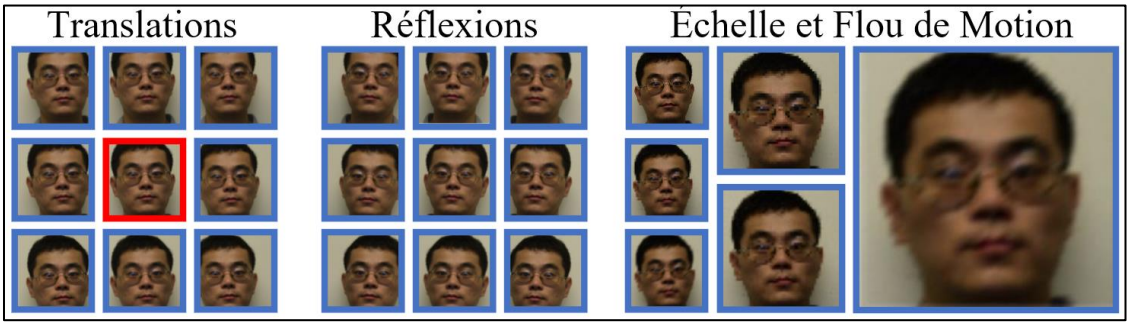

Figure 4.2 Exemples d'échantillons synthétiques (bleus) générés à partir d'opérations géométriques appliquées à l'image de référence unique (rouge)

La Figure 4.2 présente seulement quelques-unes des transformations qui ont été effectivement appliquées pour l'entraînement de chaque ensemble de classificateurs. En considérant toutes les combinaisons d'opérations, la quantité d'échantillons obtenus avoisinait au minimum 30 visages additionnels, tout dépendamment des dimensions originales de la ROI détectée. Étant donné que plus qu'un échantillon positif est désormais utilisé grâce à la génération synthétique proposée, le SVM employé n'est en fait plus un Eo*e*-SVM. À la place, un ensemble de classificateurs SVM à kernel linéaire, Eoℓ-SVM, vient le remplacer. Ceci n'est par contre qu'une formalité technique, car le reste des paramètres d'entraînement sont préservés

identiquement à l'exemplaire (*i.e.* : $C^+ = 1$ et $C^- = 0.01$) parce que, même avec l'ajout de ses échantillons positifs, les classes demeurent très déséquilibrées (milliers de négatifs).

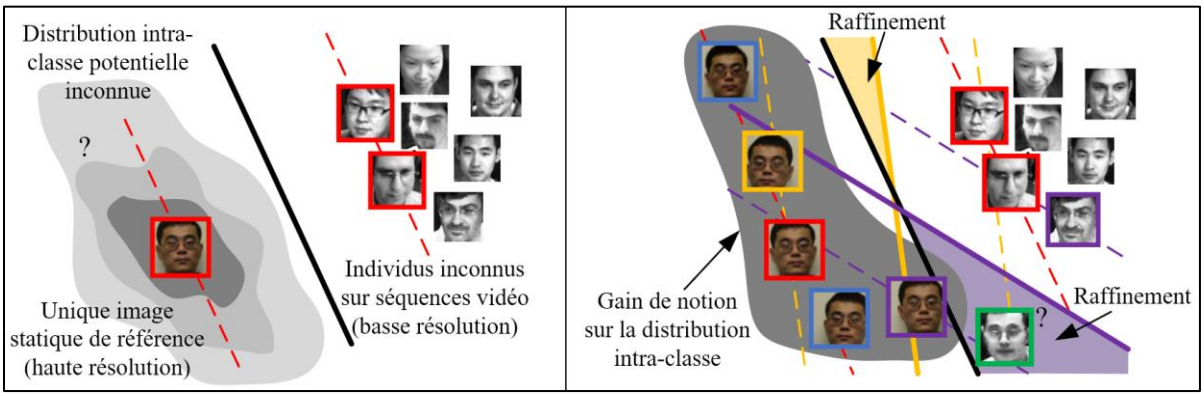

Figure 4.3 Représentation des Eoℓ-SVM mieux adaptés à la distribution des images synthétiques de l'individu à reconnaître (droite) par rapport à l'approche initiale (gauche)

Avec les nouvelles variations, il devenait alors possible d'entraîner un Eoℓ-SVM par individu plus robuste étant donné que les vecteurs de support résultants pour la marge de séparation maximale étaient beaucoup plus adaptés aux cas les plus complexes à séparer, comme démontré à la Figure 4.3. Les vecteurs de supports plus stricts peuvent être maintenant obtenus, car les SVM ont désormais la capacité de formuler une meilleure approximation de la distribution de données représentant la variation intra-classe positive (gris foncé à droite). Lors de la mise en application de cette génération synthétique, le nombre de vecteurs nécessaires pour accomplir la séparation venait souvent doubler. Cela démontrait que les ensembles accomplissaient effectivement une sélection plus éclairée des classificateurs les plus compétents pour résoudre le problème de classification ajusté, et avaient acquis une véritable compréhension de la distribution de données à modéliser, ce qui ne pouvait simplement pas être envisagé avec l'image unique de référence. La Figure 4.3 démontre également le bénéfice de cette adaptation aux cas complexes synthétiques, avec la même image de test (vert) qu'à la Figure 4.1 (droite), qui est dorénavant correctement reconnue comme l'individu recherché. Certains des raffinements procurés par les images positives synthétiques grâce à l'ajout de vecteurs de supports plus discriminants sont également représentés avec des couleurs correspondantes à la Figure 4.3.

4.2.2 Ajout de visages spécialisés supplémentaires

La seconde augmentation de données employée est basée sur le fait qu'il arrive que certains individus du *watchlist* se ressemblent plus entre eux que d'autres, ce qui vient réduire la performance des classificateurs spécifiques à ces gens semblables, mais pas pour les autres qui demeurent clairement différents d'apparence. Nous n'avons qu'à considérer par exemple des jumeaux, qui sont beaucoup plus difficiles à distinguer que deux individus quelconques à traits caractéristiques très distincts, que ce soit induit par la couleur de peau, la vieillesse, le port de lunettes, des taches de naissance sur le visage où toute autre particularité du visage qui pourraient être captées par les descripteurs. Ainsi, il est normal de voir que les E*o*e-SVM spécifiques à certaines personnes performent extrêmement bien, alors que d'autres produisent des prédictions décevantes en se basant seulement sur la référence unique (voir un exemple de résultat à l'ANNEXE V qui démontre les performances très variables entre différents *targets*).

Étant donné que l'extraction de caractéristiques offre un niveau d'abstraction des données dans un espace de haute dimensionnalité, il devient difficile de savoir quels traits spécifiques du visage deviennent plus pertinents pour discriminer les individus entre eux. Ainsi, plutôt qu'accomplir une analyse comme PCA ou LDA pour trouver des caractéristiques particulières qui distingueront les individus semblables, nous pouvons laisser cette tâche aux E*o*e-SVM qui sont spécifiquement construits pour le *hard negative mining*. Il ne suffit donc que de directement prendre les images à haute résolution des autres *targets* à enregistrer dans la *watchlist*, et de les ajouter à la base d'individus inconnus sur vidéos, pour entraîner les classificateurs de chaque *target* spécifiquement. Ceci est illustré à l'aide de la Figure 4.4, où l'on retrouve divers E*o*e-SVM spécifiquement entraînés avec des *non-targets* formés aussi des autres individus d'intérêt. À noter que les combinaisons de vecteurs de support ne sont pas limitées à un domaine particulier (voir image à gauche), et que ceux-ci changent selon les besoins spécifiques de chaque individu d'intérêt, tout dépendamment des positions constantes des représentations dans l'hyper-espace. L'approche par *stills* additionnels a été considérée dans [9], et l'on y note un gain en précision intéressant grâce à la spécialisation supplémentaire face aux particularités de visages sur images de référence, justifiant l'utilisation pour ce projet.

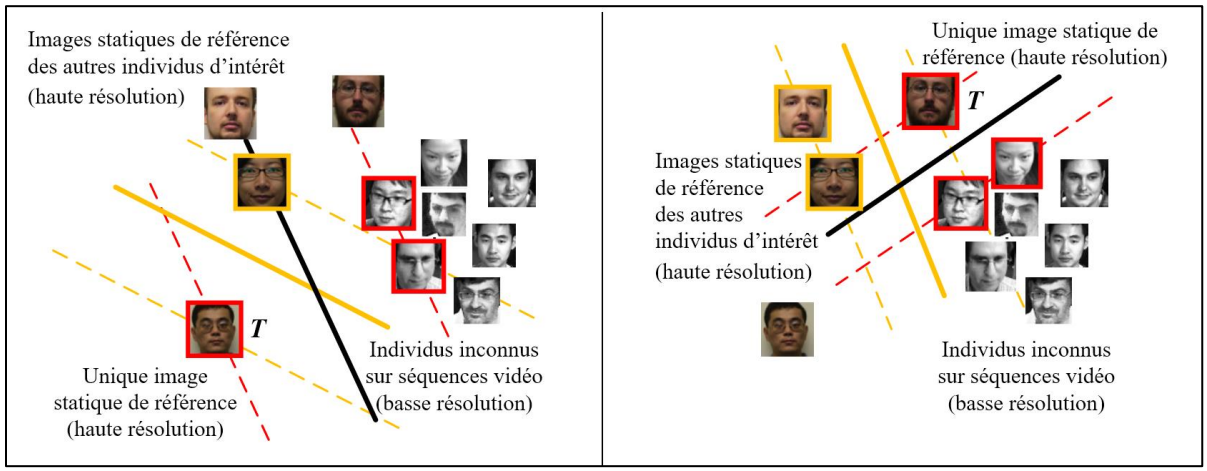

Figure 4.4 Représentation des Eo*e*-SVM spécialisés avec l'ajout des autres individus d'intérêts comme contre-exemples négatifs supplémentaires (*targets* annotés avec « ***T*** »)

Effectivement, en ajoutant les images statiques de référence, les classificateurs ont maintenant une façon de comparer de manière beaucoup plus approfondie les particularités de visage qui différencie les individus entre eux à travers le même domaine (l'environnement d'acquisition contrôlé). Ainsi, les Eo*e*-SVM trouvent habituellement de bien meilleures distinctions problématiques entre les individus dans cet espace commun, étant donné que les variations observées sur les visages ne sont attribuées qu'aux visages mêmes, plutôt qu'à une combinaison entre les traits du visage et des conditions d'acquisition nuisibles sur vidéo.

Cette dernière particularité, dont les traits spécifiques du visage sont plus facilement différenciés sur les images statiques comparables, est représentée à la Figure 4.5. Nous pouvons y voir, par exemple, que la définition de la bouche et de la mâchoire est nettement plus évidente dans l'environnement contrôlé que dans l'environnement semi-contrôlé des caméras de surveillance, étant donné les poses semblables comparées à celles sur vidéo.

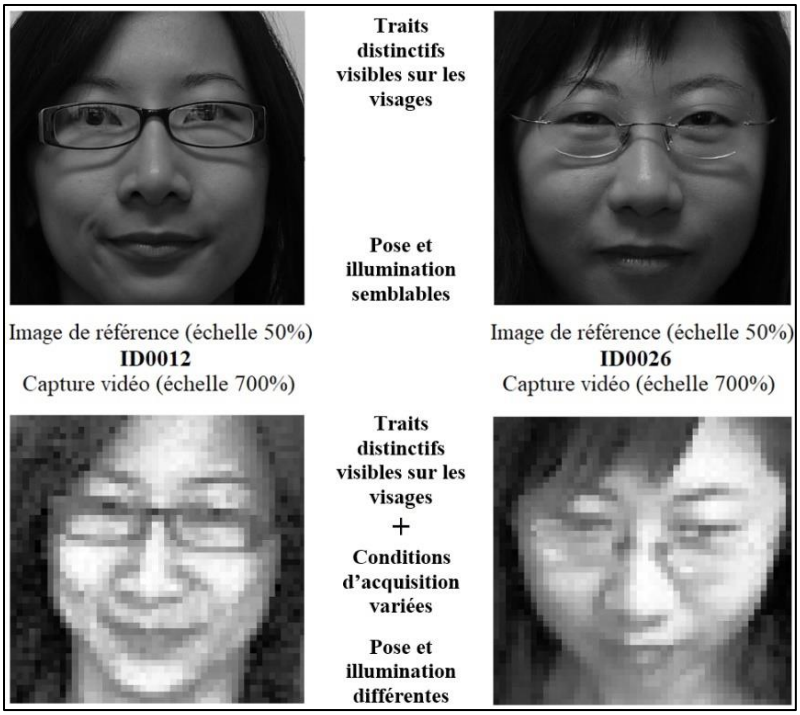

Figure 4.5 Comparaison des distinctions du visage de différents individus d'intérêt selon le type d'environnement

Finalement, considérant que les Eo*e*-SVM sont fondamentalement à la recherche des meilleurs contre-exemples (*hard negative mining*) pour mieux définir leur exemplaire à distinguer, il est intéressant d'ajouter encore davantage d'images négatives pouvant devenir l'un de ces cas complexes. Puisque les classes sont déjà extrêmement déséquilibrées, le fait d'ajouter encore plus de contre-exemples ne vient pas plus perturber l'entraînement des SVM. Ainsi, les images statiques de référence des autres individus d'intérêt ont également été traitées avec les opérations de transformations géométriques précédentes, de sorte à les ajouter comme représentations additionnelles à la base de contre-exemples négatifs. Ceci permet d'ajouter un autre niveau de spécialisation des classificateurs pour leur individu particulier à reconnaître.

4.2.3 Génération synthétique de pose

Étant donné les majeures variations de pose observées sur vidéo par rapport aux références, il est intéressant de tenter de mieux définir les contre-exemples à partir de représentations plus

justes des profils de visage possibles des individus. En effet, malgré que la génération synthétique par transformation géométrique apporte un grand nombre d'échantillons additionnels pour contrefaire le problème du SSPP et améliorer les performances, les images demeurent des représentations frontales du visage. De plus, des évaluations expérimentales présentées dans [28] semblent supporter que combiner les images 2D, ainsi que des représentations par modèles 3D, vient augmenter la performance globale d'un système de FR étant donné l'ajout de représentations apporté. La conclusion de ces travaux est que plus l'on ajoute de variations afin d'anticiper les observations futures de visages, plus le système possède de chances de réussir à correctement reconnaître les individus.

Ainsi, des approches de modélisation morphologiques ou de reconstructions 3D du visage sont considérables dans ce projet pour parvenir à générer les poses auxquelles l'on s'attend d'observer sur les caméras de surveillance. En procédant ainsi, il devient possible d'obtenir multiples représentations synthétiques plus similaires aux vrais profils de visages qui seront capturés dans la scène, ce qui apporte de la robustesse au système envers ces variations [71].

Pour accomplir une génération de pose à apparence tridimensionnelle du visage à partir d'une image frontale de visage, des approches basées sur 3DMM (3D *morphable models*) ont été proposées en littérature [17, 18, 113, 123]. En employant des modèles 3DMM, nous pouvons alors générer des poses plus fidèles à partir de l'unique image de référence à apparence frontale neutre, ce qui ne serait pas aussi finement obtenu avec, par exemple, des rotations 2D des images ou du cisaillement. Le principe à l'application de ces approches consiste très sommairement à générer le modèle 3D de visage sur lequel un *mapping* de la texture doit être effectué pour transférer l'information du visage (peau, rides, illumination, relief et ainsi de suite) contenue sur l'image 2D vers le modèle 3D, tout en demeurant fidèle à la référence.

Considérant la disponibilité de code pour implémenter ces méthodes, l'approche proposée par Mokhayeri et al. (2015) [85] a été employée pour l'étude actuelle, afin d'extraire les poses observées sur des séquences vidéo, et générer les modèles 3D à partir de 3DMM et composés de textures correspondantes aux visages de référence, afin obtenir des variations de visages en

pose réalistes et représentant les observations attendues dans la scène. Ainsi, un exemple de modèle 3D obtenu pour venir générer la pose du visage visible sur la séquence vidéo visée est représenté à l'aide de la Figure 4.6. La figure présente également de réelles captures obtenues du point de vue de la caméra correspondante, afin de comparer qualitativement les angles de poses du visage obtenus avec modélisation 3D par rapport à ceux extraient réellement des trajectoires de visage générées par le FT. Ainsi, selon les différentes poses observées à travers toute la séquence vidéo de cette caméra, les différentes poses les plus potentielles sont générées synthétiquement par 3DMM. Visuellement, les images obtenues semblent adéquates pour les ajouter à l'ensemble de représentations positives de l'individu d'intérêt, afin d'augmenter les représentations disponibles d'entraînement des ensembles de classificateurs.

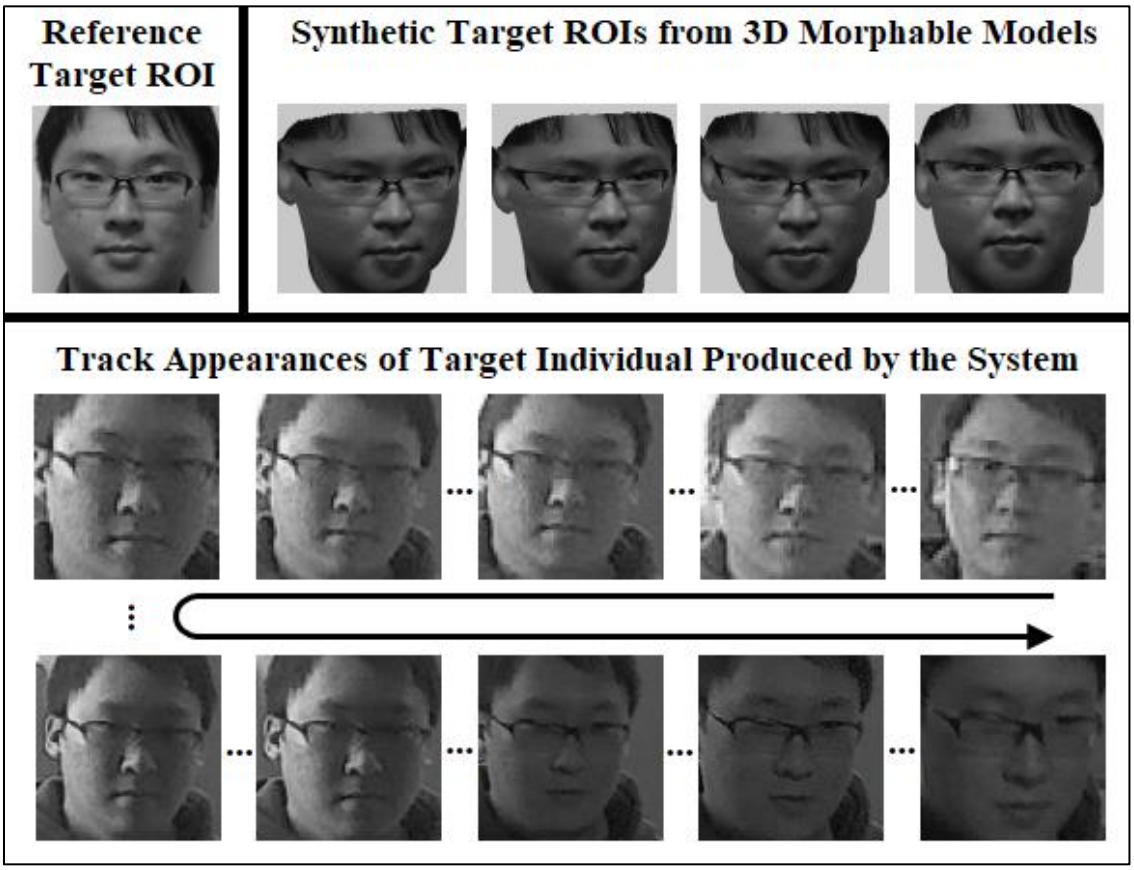

Figure 4.6 Comparaison d'un modèle 3D obtenu par génération synthétique avec des visages observés sur trajectoire vidéo dans la scène à tenter de représenter

4.3 Système de FRiVS avec sélection aléatoire de caractéristiques

Étant donné que les méthodes précédentes s'attardent principalement à des opérations directement appliquées dans le domaine des images de référence, celles-ci sont particulièrement propices à générer du bruit à travers les données d'entraînement. En effet, puisque n'importe quelle représentation inadéquate, ou insuffisamment fidèle aux vrais visages de référence et ceux observables dans la scène, viendrait complexifier le problème de classification que les ensembles de SVM doivent résoudre, la FR y serait fortement affectée.

De plus, les SVM sont particulièrement sensibles aux *outliers* positifs, étant donné la quantité limitée qui est disponible en SSPP afin d'établir les vecteurs de support pour la meilleure marge de séparation. Ainsi, l'approche RMS (*random subspace method*) est considérée ici afin d'appliquer de la génération de nouvelles représentations de visages directement dans l'hyper-espace des caractéristiques de visage extraites et contenues à travers les vecteurs descripteurs d'entraînement. En effet, si le système réussit à être relativement performant (FR suffisamment forte, mais pas idéale) à partir des descripteurs extraits pour caractériser les visages des individus d'intérêt, cela implique que les *features* sous-jacentes sont minimalement discriminantes pour ces personnes. Les caractéristiques sont donc « validées » comme étant efficaces pour la tâche de FR et, ainsi, la génération d'échantillons supplémentaires à partir de celles-ci devrait aussi produire des représentations de visages robustes et fidèles aux visages originaux transformés dans l'espace de très haute dimensionnalité.

L'utilisation de RSM pour ce projet est inspirée des travaux dans [29] qui ont démontré que cette approche, tel que supposé plus haut, permettrait en effet de grandement améliorer la robustesse du système pour une FR fiable et efficace. La méthode RSM consiste en une approche de type *over-sampling*, tel que présenté au début de ce chapitre, étant donné que les opérations de génération des nouvelles représentations de visages pour l'entraînement des classificateurs sont directement appliquées au *feature-space*.

Le principe de la méthode RSM est illustré à l'aide de la Figure 4.7, où l'on peut y voir qu'un vecteur descripteur originalement extrait à partir d'un visage, et qui est sous-divisé en multiples sous-ensembles de caractéristiques distinctes. Ainsi, à partir du vecteur initial formé de N_d caractéristiques, une sélection aléatoire de $N_d^{(rs)} < N_d$ *features* est effectuée, et ce, pour N_{rs} itérations qui viennent former les multiples représentations alternatives de *random subspaces*. À noter que les caractéristiques peuvent être dupliquées, afin d'obtenir une meilleure redondance de l'information discriminante contenue à travers celles-ci. De toute manière, pour une quantité de sous-ensembles de vecteurs élevée, il est nécessaire de dupliquer les informations considérant les dimensions employées durant ce projet. De plus, les *features* n'ont pas besoin de suivre un ordre particulier pour la génération des nouveaux vecteurs.

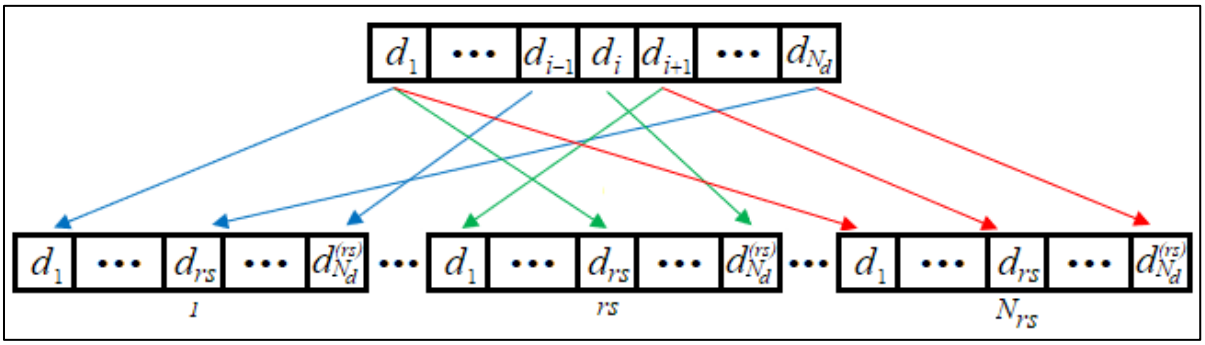

Figure 4.7 Représentation de l'application de RSM sur les vecteurs descripteurs de visage (haut) pour la génération de sous-ensemble de représentations (bas)

L'opération RSM est ainsi appliquée pour chacune des N_p *patches* extraits du visage. Un ensemble de $N_p \cdot N_{rs}$ SVM sont entraînés en utilisant respectivement pour chacun d'eux les descripteurs correspondants à obtenus pour chaque combinaison de *patch-random subspace.* Afin de conserver une correspondance entre les caractéristiques extraites durant la phase d'entraînement des classificateurs et la phase d'opération, pour permettre une comparaison de traits caractéristiques du visage équivalents, les indices des *features* sélectionnés aléatoirement par RSM sont employés pour effectuer le *mapping* des nouveaux visages sur séquence vidéo qui seront présentés à l'entrée du système. La notation $\mathrm{RSM}_{(Nrs,\, N_d)}$ sera employée par la suite afin de référer à cette méthode avec les paramètres associés.

4.4 Système de FRiVS avec adaptation de domaine

En plus de la génération de représentions synthétiques qui offrent, en soit, une adaptation au domaine (DA) de la scène sous surveillance, car la pose des visages potentiels est reflétée sur les modèles 3D générés, il est possible de profiter du principe de DA en utilisant les apparences de chaque caméra individuellement pour la FRiVS. En effet, il est possible de segmenter les bases de vidéos d'entraînement pour toute nouvelle caméra de sécurité, ce qui permet d'obtenir des modèles de visages uniques et bien spécifiques à chacune d'entre elles. Cette approche est démontrée dans [10], où l'on peut voir un gain considérable en performance de FR lorsque le domaine particulier à chaque point de vue de chaque caméra, et pour leur scène respective, est employé distinctivement pour l'entraînement des classificateurs. Les travaux dans [6] ont également montré l'avantage d'avoir une normalisation d'illumination spécifique au domaine où les individus doivent être reconnus. Ainsi, avoir des descripteurs de *unknown non-targets* normalisés selon la caméra spécifique est définitivement à considérer pour la FRiVS.

Comme on peut le voir avec la Figure 4.8, les scènes captées par les multiples caméras qui peuvent former un réseau de sécurité sont effectivement très différentes les unes des autres en termes de niveaux d'illumination et de pose, d'où la nécessité de l'adaptation de domaine.

Figure 4.8 Comparaison de conditions d'acquisition variées par scène captée avec différentes caméras de surveillance, base ChokePoint [139, 140]

Un autre raisonnement venant supporter le DA appliqué particulièrement aux SVM employés dans ce projet part du principe que ceux-ci sont conceptualisés à la base afin de se spécialiser à un individu bien particulier, en renforçant la distinction qui existe par rapport aux cas négatifs les plus complexes à distinguer. Les ensembles de SVM dans le cas du projet sont d'autant plus spécialisés alors qu'ils se concentrent sur des sections du visage (*patches*) bien particulières à chacun d'eux. Or, tant qu'à exploiter la spécialisation les classificateurs aux cas complexes pour des individus spécifiques, mieux vaut procéder à leur entraînement en exploitant une sélection extrêmement contrôlée de contre-exemples pour la scène particulièrement visée par chaque caméra de surveillance.

En retirant réellement les meilleurs *hard negatives* encore plus spécifiques au domaine visé, nous avons davantage de chances d'obtenir des classificateurs les plus compétents possible, pour qu'ils soient plus robustes et adaptés à l'environnement de la caméra et de ces problèmes d'acquisition spécifiques. Ainsi, plutôt que de tenter de résoudre les difficultés de classifications dues aux conditions de captures multiples à travers toutes les caméras du système de surveillance considérées comme un seul problème, nous divisons le tout en sous-problèmes par rapport à chaque environnement individuellement, ce qui simplifie énormément la complexité de représentations de visages dans l'hyper-espace de caractéristiques pour les classificateurs.

CHAPITRE 5

ÉVALUATIONS, ANALYSE DE RÉSULTATS ET DISCUSSIONS

Ce chapitre abordera les méthodologies qui ont été présentées lors du Chapitre 4, soit la génération synthétique d'images par transformations géométriques et de pose par *morphable models*, l'augmentation de données par la sélection aléatoire de caractéristiques extraites des visages, et l'adaptation de domaine spécifique aux caméras, afin de produire de multiples représentations de visage d'individus à reconnaître. À partir de celles-ci, ajoutées au système initial avec suivi de trajectoire et ensembles de SVM exemplaires, présenté au Chapitre 3, les évaluations de la performance du système de FRiVS réalisées durant le projet seront présentées à ce chapitre.

De plus, les détails mettant en situation la méthodologie expérimentale qui a permis d'accomplir les évaluations des variantes du système seront abordés. En outre, la procédure de test et la base de vidéos employées seront présentées avant de passer à une description des mesures de performances utilisées pour comparer les solutions envisagées. Les résultats d'évaluation seront présentés premièrement pour le système initial avec suivi de trajectoire, mais sans l'ajout des multiples représentations de visage supplémentaires pour entraîner les classificateurs. Ensuite les résultats des autres variations, avec l'ajout des représentations de visages, seront abordés un à un pour observer les gains graduels de chaque approche testée apporte, et ainsi y tirer des interprétations progressives jusqu'à la solution finale.

Finalement, les meilleurs résultats seront comparés à l'état-de-l'art en deux temps, (1) par rapport à multiples systèmes de S2V-FR de manière générale, soit plus explicitement par rapport à Eo*e*-SVM [11, 12], Eo*e*-SVM-DS [9, 11], ESRC-DA [91], CCM-CNN [99], CFR-CNN [101] et HaarNet [100], puis (2), pour des systèmes de S2V-FR particulièrement avec suivi de trajectoire, c'est-à-dire TM-FR [26], TM-SU [112], MFR [13], SVDL [141] et AAMT-FR [45]. Certaines méthodes de ce second groupe utilisent aussi de l'adaptation *on-line* du modèle de visage comme dans notre cas, ce qui les rendent intéressantes pour comparer.

5.1 Procédures de test

5.1.1 Détails de la base de données pour l'évaluation

La base de vidéos ChokePoint [139] est employée pour les l'évaluation des systèmes étant donné qu'elle offre une bonne source de vidéos représentant des captures par caméras de sécurité, tout en offrant également l'unique image de référence haute résolution nécessaire pour enregistrer les individus recherchés dans le système.

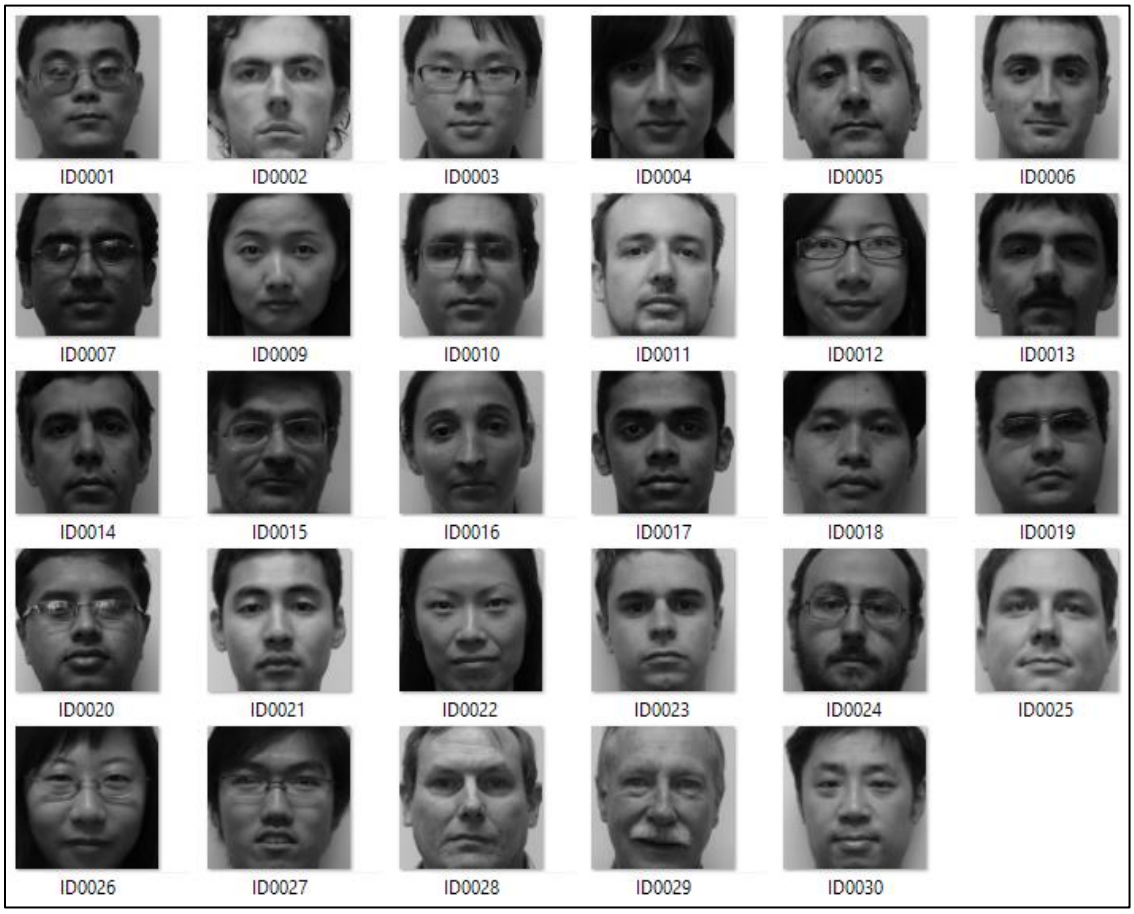

Figure 5.1 ROI de visage des individus formant la base de test ChokePoint

La base comporte 29 individus d'intérêt potentiels. L'ensemble des individus retrouvés dans la base de vidéos ChokePoint sont présentés à la Figure 5.1.

Les 29 individus de la base de vidéos passent à travers 2 différents portails, et selon deux directions (entrée et sortie), afin de générer des acquisitions en milieux distincts. De plus, chacun des portails est capturé à l'aide de 3 caméras simultanément, de sorte à obtenir divers points de vue et angles de pose du visage alors que les individus traversent les portails. Pour finir, 4 sessions de tests sont accomplies, afin d'obtenir des passages à trajectoires sensiblement variables de chacun des individus à chaque fois. Considérant toutes les combinaisons, un total de 72 séquences vidéo différentes sont obtenues. Sur ces dernières, selon l'apparition ou non de certains individus à travers les séquences, nous pouvons compter un grand total de 1281 passages pour former les trajectoires d'individu circulant à travers les multiples portails. Parmi ces séquences, la base de données indique également lesquelles correspondent aux apparitions de visages les plus « frontales », considérant la position des trois caméras à chaque fois. Nous pouvons donc voir que nous avons une grande diversité de membres, afin d'avoir des conditions de test réalistes.

Pour chacune des séquences vidéo disponibles dans la base de test, et pour chacun des *frames*, l'information de vérité (*ground truth*) indique l'individu qui est actuellement observé dans la scène, ainsi que la position de ses yeux. Ayant cette information en mains, il était possible de valider la position des ROI générées par l'algorithme de suivi de trajectoire en combinaison avec la triple détection VJ. Par contre, plusieurs erreurs dans la base ont été trouvées, tel qu'illustré avec la Figure 5.2 et la Figure 5.3, où plusieurs individus sont simultanément visibles sur la séquence vidéo, alors qu'il ne devrait qu'en avoir un seul selon les *ground truths*. Étant donné que ce projet évalue un système de FRiVS de bout-en-bout, les visages correctement détectés (il y a effectivement quelqu'un) venaient générer multiples « erreurs », car celui-ci ne figure pas dans les références pour valider sa présence en particulier.

Ainsi, la base a été nettoyée autant que possible lorsque d'autres cas de ce type étaient trouvés. Par contre, il demeure qu'il doit persister des situations qui n'aient pas été détectées considérant la taille des données à analyser. Donc, malgré que les performances finalement obtenues sont très satisfaisantes (voir section 5.3.2), il faut considérer que certains de ces cas sont probablement encore en cause d'une légère baisse de performance.

Figure 5.2 Exemple de séquence qui contient plusieurs individus non indiqués dans les tables de vérité, séquence P1E_S1_C1, individu ID0003, ChokePoint

Figure 5.3 Exemple de séquence qui contient plusieurs individus non indiqués dans les tables de vérité, séquence P1E_S2_C3, individu ID0027, ChokePoint

Figure 5.4 Exemple de séquence avec détections multiples dues à la réflexion sur la vitre de la porte du portail, séquence P2E_S3_C1.1, individu ID0030, ChokePoint

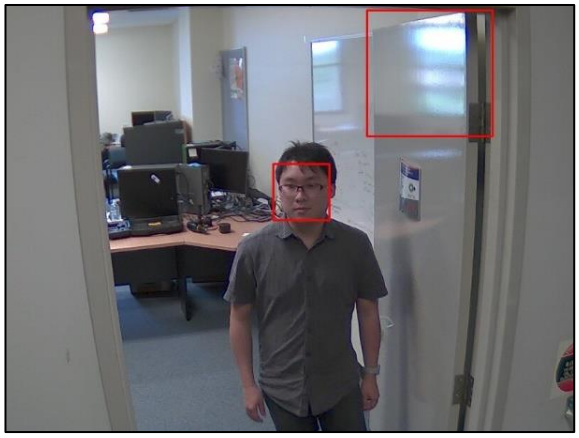

Figure 5.5 Exemple de réflexion constante d'illumination produisant de fausses détections répétées, séquence P1L_S3_C2, individu ID0003, ChokePoint

Deux autres situations problématiques de cette base de données sont illustrées avec la Figure 5.4 et la Figure 5.5. Dans le premier cas, de « fausses » détections sont obtenues dues aux réflexions dans la porte du portail et, dans le second, l'illumination en réflexion constante sur la porte produit une réelle fausse détection de manière répétée sur plusieurs *frames*.

Dans le premier cas, la réponse est problématique, car l'individu détecté peut correspondre en effet à une personne d'intérêt mais, étant une réflexion très floue, cela produit très possiblement des faux-négatifs de reconnaissance. Dans le second cas, ce genre d'erreur n'est pas répétable, car les systèmes de surveillances sont à l'habitude très soigneusement mis au point (*fine tuning*) lors de la configuration des caméras, afin d'éviter telle sorte de problèmes.

Or, certains cas d'études pourraient vouloir filtrer ces cas particuliers. Cela serait valable lorsque seulement les classificateurs de FR sont évalués pour établir leur performance spécifiquement à partir des ROI de visages validées (*ground truths* sans erreur) et dont ces problèmes ne devraient simplement pas exister. C'est le cas par exemple dans [99], qui considère que ChokePoint possède 64204 ROI de visages à partir des *ground truths*, alors qu'en fait, les séquences vidéo permettent de détecter jusqu'à 72329 *frames* possédant des visages. Dans la situation de ce projet avec le système FRiVS de bout-en-bout réalisé, les cas des deux figures du bas (Figure 5.4 et Figure 5.5) persistent dans l'étude, car il s'agit de situations réelles possibles en vidéosurveillance. Nous avons donc approximativement 8000 visages potentiels supplémentaires à évaluer, ce qui inclut des cas plus difficiles à classer, sans compter en plus de cela les erreurs de fausses détections additionnelles.

À noter, seulement les deux cas du haut (Figure 5.2 et Figure 5.3) sont filtrés pendant cette étude, étant donné qu'il s'agit de séquences à individus multiples, ce qui crée des trajectoires de visages complètes et correctement gérées par le système proposé, mais qui ne peuvent pas être confirmées sans *ground truth*. En prenant ces considérations en compte, la performance atteinte par les classificateurs serait possiblement encore meilleure à la section 5.3.2 des résultats finaux si ces cas complexes étaient filtrés. Toutefois, la robustesse des trajectoires et la correction des ROI par multiples contre-validations FD+FT démontrent avec les résultats que ces cas sont tout de même correctement gérés par le système de FRiVS proposé.

5.1.2 Segmentation de la base de données

Afin d'accomplir des évaluations de performances justes du système et statistiquement répétables, plusieurs réplications de test sont nécessaires. Donc, différents individus ont été interchangés 5 fois afin de réévaluer la performance de la FR indépendamment des personnes employées pour l'entraînement (tant positifs que négatifs), ainsi que pour tester en opération. Les combinaisons de réplications qui ont été employées sont présentées au Tableau A.3 situé à l'ANNEXE II. Nous pouvons y voir qu'une combinaison aléatoire d'individus pour chaque catégorie est employée à chaque réplication. La méthodologie applique pour chaque test quant à elle est basée sur [9, 10, 11, 12, 101], soit l'utilisation de $N_{nt} = 10$ individus inconnus pour recueillir des visages sur vidéos pour employer comme contre-exemples d'entraînement des modèles et $N_{wl} = 5$ individus d'intérêt enregistrés dans la liste de contrôle du système de FRiVS à partir de leur image de référence. Le reste des vidéos inutilisés pendant l'entraînement, soit $N_{ut} = 14$ individus inconnus (*negative probes*), combinés aux vidéos correspondants aux N_{wl} individus enregistrés (*positive probes*), sont employés pour évaluer les réponses de reconnaissance lors de la phase d'opération. Puisque le nombre d'individus inconnus au système est approximativement trois fois plus élevé que le nombre d'individus enregistrés, l'évaluation utilisée considère le déséquilibre qui serait observé en mode d'opération sur le terrain, dans un système de sécurité sur vidéo-caméra déployé pour la surveillance. Évidemment, il serait préférable d'avoir encore plus d'exemples de vidéos inconnus mieux représenter le réel niveau de déséquilibre encore plus prononcé qui serait observé sur le terrain, mais la limitation de la base de données nous pose problème, car nous voulons conserver une quantité raisonnable d'individus dans la *watchlist* simultanément pour évaluer l'impact de reconnaissance avec plusieurs personnes à reconnaître en même temps.

5.1.3 Variations d'entraînement par méthode

Considérant que plusieurs méthodologies aient été envisagées pour l'évaluation du système optimal de FRiVS à partir de trajectoire de visage, de modèles synthétiques et d'augmentation de données, diverses combinaisons d'entraînement ont été effectuées. Les acronymes et explications du Tableau 5.1 seront employés pour la présentation des résultats qui suivront.

Tableau 5.1 Description des combinaisons d'entraînement évaluées par bloc de tests

Bloc de test	Combinaison	Méthodes d'entraînement						Description	Classifieur
		ST_{nt}	GT_{nt}	GT_{wl}	3D	RSM	DA		
N/A	Système initial							L'individu d'intérêt est enregistré au système à partir de son image de référence unique à haute résolution, et est comparé aux individus inconnus provenant de vidéos en mode opérationnel.	Eo*e*- SVM
Génération synthétique par transformations géométriques + Ajout d'images de référence des contre-exemples négatifs	ST_{nt}	✓						Les autres individus d'intérêt faisant partie de la cohorte d'entraînement sont ajoutés comme contre-exemples négatifs additionnels au système initial.	Eo*e*- SVM
	ST_{nt}+GT_{nt}	✓	✓					Les autres individus d'intérêt faisant partie de la cohorte d'entraînement sont ajoutés, avec leurs visages synthétiques générés par transformations géométriques, comme contre-exemples négatifs additionnels au système initial.	Eo*e*- SVM
	GT_{wl}+ST_{nt}	✓		✓				Des représentations de visages additionnelles de l'individu d'intérêt générées synthétiquement à partir de transformations géométriques sont ajoutées au système initial.	Eo*ℓ*-SVM
	GT_{wl}+ST_{nt}+GT_{nt}	✓	✓	✓				Les autres individus d'intérêt de la cohorte, leur visages synthétiques par transformations géométriques, ainsi que celles de l'individu d'intérêt, sont tous ajoutés au système intial, selon leurs classes respectives.	Eo*ℓ*-SVM
Génération synthétique de pose par modèles 3D	3DMM				✓			Des représentations de poses du visage additionnelles de l'individu d'intérêt générées synthétiquement à partir de modèles morphologiques 3D sont ajoutés au système initial.	Eo*ℓ*-SVM
	3DMM+ GT_{wl}+ST_{nt}+GT_{nt}	✓	✓	✓	✓			Tous les visages générés par méthodes de générations synthétiques avec transformations géométriques ainsi qu'avec modèles morphologiques 3D sont ajoutés au système initial selon leurs classes respectives.	Eo*ℓ*-SVM
Génération de représentations de visage par sélection alétaoire des caractéristiques	$RSM_{(20,128)}$					✓		Une sélection aléatoire de 128 caractéristiques extraites de l'image de référence unique de l'individu d'intérêt enregistré au système, ainsi que des contre-exemples, est utilisée pour générer 20 nouvelles représentations du visage.	Eo*e*- SVM
	$RSM_{(20,128)}$+ GT_{wl}+ST_{nt}+GT_{nt}	✓	✓	✓		✓		Tous les visages générés synthétiquement par transformations géométriques sont utilisés et une sélection aléatoire de 128 caractéristiques est employé sur ceux-ci pour générer 20 nouvelles représentations de visage pour chaque image obtenue.	Eo*ℓ*-SVM
Adaptation de domaine des caméras de surveillance	DA+ GT_{wl}+ST_{nt}+GT_{nt}	✓	✓	✓			✓	Les autres individus d'intérêt de la cohorte, leur visages synthétiques par transformations géométriques, ainsi que celles de l'individu d'intérêt, sont sélectionnés des scènes de caméra spécifiques pour comparer avec les vidéos correspondants.	Eo*ℓ*-SVM

Ainsi, les transformations géométriques appliquées à l'image de référence unique de chaque individu d'intérêt afin de générer des variations synthétiques de son visage est annoté par GT_{wl}. Ces transformations y sont ensuite ajoutées à chaque ensemble de classificateurs respectif à chacun des individus de la cohorte à inscrire à la *watchlist* du système de FRiVS, et ce, à tour de rôle. L'application des mêmes transformations géométriques aux autres individus d'intérêt que celui actuellement entraîné au système sont quant à elles représentées par GT_{nt}, pour être ajoutés comme visages supplémentaires de contre-exemples aux vidéos de caméras déjà employés. À noter que l'insertion des négatifs supplémentaires par GT_{nt} implique aussi nécessairement l'utilisation de leur image de référence unique ST_{nt} comme contre-exemples, ce qui fait que la combinaison GT_{nt} est toujours accompagné de ST_{nt}. Par contre, ST_{nt} peut être employé à lui seul, c'est-à-dire que les contre-exemples *stills* sont utilisés directement, sans transformations géométriques additionnelles. L'obtention des poses de visages par génération tridimensionnelle des visages avec *morphable models* est représentée par 3DMM. La génération de vecteurs descripteurs à partir de sélection aléatoire des caractéristiques est indiquée par $\mathrm{RSM}_{(20,128)}$. Pour finir, l'adaptation de domaine par caméra spécifique est représentée par DA au Tableau 5.1.

Nous pouvons aussi voir au Tableau 5.1 lequel des classificateurs (Eo*e*-SVM ou Eoℓ-SVM) est utilisé à chaque variation d'entraînement visée, en fonction de l'utilisation de plus d'échantillons positifs ou non. Rappelons ici que les SVM à *kernel* linéaire doivent être utilisés au lieu des exemplaires à une seule représentation du visage lorsque la classe caractérisant l'individu d'intérêt est augmentée par données synthétiques, puisque plus d'un échantillon y est alors employé.

5.2 Évaluation de la performance de la reconnaissance de visages

Les sous-sections qui suivront par la suite présentent les mesures de performances qui sont employées afin d'évaluer la qualité de la reconnaissance de visages des systèmes à travers les multiples combinaisons d'entraînement. Les différentes manières d'application de ces mesures y sont également discutées.

5.2.1 Mesures de performance

Pour chaque combinaison de test évaluée, la courbe ROC (*Receiver Operating Characteristic curve*) respective à chacun des individus enregistrés dans la *watchlist* est tracée afin d'évaluer le compromis entre le FAR (*False Accept Rate*) et le FRR (*False Reject Rate*). Ceci permet d'évaluer respectivement la probabilité d'erreurs de type I, soit lorsque des visages sont incorrectement reconnus comme des personnes d'intérêt sans pour autant en êtres (FP), par rapport à la probabilité erreurs de type II, lorsqu'un visage dans la scène n'est pas reconnu comme appartenant à un individu d'intérêt, alors qu'il y en a effectivement un (FN) [19]. Or, en accumulant le nombre de fausses alarmes (FP) par rapport au total d'essais de reconnaissance correspondant à des cas souhaités comme négatifs (FP+TN), nous obtenons le ratio de faux-positifs $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$. En accumulant à l'inverse le nombre de bonnes reconnaissances d'individu d'intérêt (TP), par rapport aux essais de reconnaissances qui s'attribuent aux cas espérés comme étant positifs (TP+FN), nous obtenons le ratio de vrais-positifs $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$. Ces quantités et ratios découlant sont déterminés en fonction des scores de reconnaissances de tous les individus de test par rapport à chacun des classificateurs entraînés, et du seuil de décision spécifique à chaque individu (γ_s dans le système de FRiVS) afin de déterminer un point sur la courbe ROC. En faisant varier ce seuil de décision entre $[0,1]$ pour des scores normalisés sur chaque *frame*, ou entre $[0, W]$ pour une accumulation de scores par trajectoire, nous obtenons le tracé de la courbe ROC par individu respectif.

La performance du système à accomplir la FR peut être évaluée selon un point optimal de compromis entre ces deux erreurs modélisées par la courbe ROC. Par contre, étant donné que le niveau de compromis à allouer au système, c'est-à-dire à quel point nous désirons être robuste à un certain type d'erreurs et à quel point nous sommes prêts à accepter l'autre, dépend de l'application même et aussi des besoins de l'utilisateur final. Ainsi, afin de considérer les multiples points de compromis existant sur la courbe ROC simultanément, la mesure de surface sous la courbe ROC (AUC) [59] est employée, ce qui fournit une valeur indiquant un aperçu d'ensemble de la robustesse aux deux types d'erreurs en fonction de l'allure de la courbe. Ceci

est dû au fait qu'une courbe ROC d'un système optimal visera à tendre vers le coin supérieur gauche étant donné qu'un FPR minimal ainsi qu'un TPR maximal sont obtenus.

Par contre, étant donné que l'erreur de type II est beaucoup grave que l'erreur de type I dans le contexte de la FRiVS, soit qu'il est bien pire de manquer la reconnaissance d'un individu d'intérêt du *watchlist* que de générer une fausse alarme pour un individu quelconque observé, la mesure de surface partielle sous la courbe ROC (pAUC) [50, 81, 132, 144] est couramment employée dans le domaine FR. Cette mesure est utilisée afin de ne considérer que les cas de reconnaissances situés entre un taux de faux-positif (FPR) allant de 0% jusqu'à un taux relativement bas (ex. : 10% ou 20%), de sorte à ne considérer que la performance du système dans des cas réalistes d'application, là où le taux de fausses alarmes est raisonnablement faible. Ceci est représenté à l'aide de la Figure 5.6. En effet, un système de FRiVS ne serait pas très utile s'il générait une abondance de fausses alarmes, même s'il reconnaissait parfaitement les individus d'intérêts, car cela nous ferait perdre confiance en sa précision. De plus, pour une application réelle, le but d'un tel système vise souvent à reconnaître les individus efficacement à travers de multiples séquences vidéo, de sorte à alléger la tâche du vidéo-analyste qui ne peut vérifier l'intégralité des caméras et observations captées. Il serait donc inutile d'avoir un système où cet analyste devrait repasser continuellement après l'algorithme pour valider. Le mieux est donc d'avoir un nombre minime de fausses alarmes, habituellement représentant des cas extrêmement complexes à distinguer même pour l'humain, où l'assistance du vidéo-interprète viendrait confirmer la décision finale.

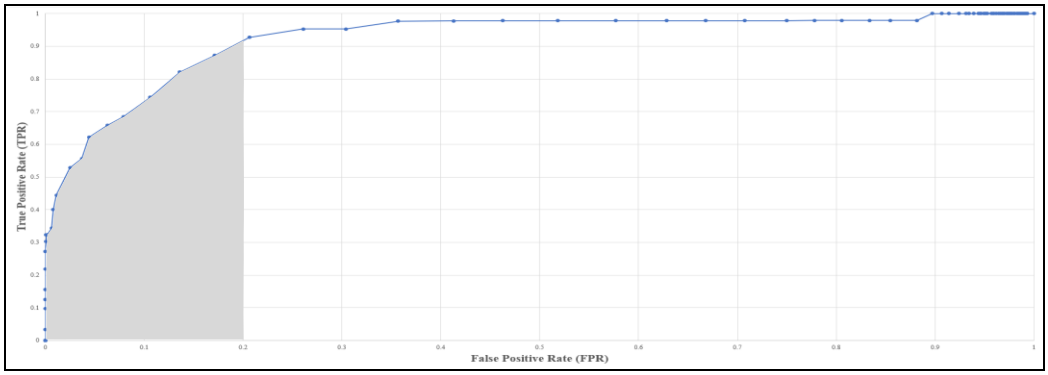

Figure 5.6 Démonstration du principe de la surface partielle sous la courbe ROC

En plus de la mesure pAUC provenant de la courbe ROC, la mesure AUPR de surface sous la courbe PR (précision-rappel ou *precision-recall*) est également employée pour évaluer le système de FRiVS. Cette courbe est basée sur la sensibilité (*recall*) du système à faire de bonnes prédictions (TP) malgré les possibilités de manquer la reconnaissance d'un individu d'intérêt circulant en effet dans la scène (FN), en fonction de la précision, ou fiabilité, de ces prédictions. La précision (PPV) est donc définie selon le taux de vraies reconnaissances des individus recherchés (TP) selon la quantité de réponses marquées comme reconnues, c'est-à-dire $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$. Par définition, le rappel est équivalent au TPR, ce qui permet de tracer une courbe PR inversée côte à côte à la courbe ROC (exemples d'une telle disposition employée à l'ANNEXE III et à l'ANNEXE IV), pour des comparaisons visuelles plus efficaces.

Un avantage marqué de la mesure AUPR par rapport à AUC est que le fait que la mesure de précision ne considère que les réponses positives (voir dénominateur de PPV), ce qui fait que cette mesure de performance n'est pas affectée par le majeur déséquilibre existant entre les classes positive et négative [37]. Effectivement, la mesure de FPR de la courbe ROC considère le grand nombre de TN potentiel dans son équation, ce qui fait qu'elle a tendance à grandement surestimer la performance des classificateurs dans les cas aussi déséquilibrés que le SSPP, car elle demeure insensible aux changements causés par la distribution inégale des classes [54]. D'ailleurs, cela explique aussi en partie pourquoi une mesure pAUC est requise. La valeur de AUPR offre toutefois une représentation nettement plus équitable des cas recherchés en FRiVS à partir de SSPP, soit de savoir si les quelques individus d'intérêt à reconnaître sont régulièrement détectés, et ceci, avec autant de précision que possible sans les confondre aux autres personnes, quiconque que ce soit.

Pour évaluer la performance optimale des classificateurs formés par le système de FRiVS en considérant le déséquilibre inévitable des classes, la mesure F_1-score est souvent utilisée [120]. Celle-ci est définie selon l'équation (5.1).

$$F_1\text{-score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{5.1}$$

La mesure F_1-score a pour avantage de ne pas considérer les classifications de type TN, étant donné la formulation des mesures PPV et TPR quelle utilise. Ainsi, il devient possible d'ignorer l'abondance de « bonnes réponses » obtenues en FRiVS lorsque les multiples individus inconnus au système sont identifiés comme non reconnus. Dû à la nature du problème, nous avons toujours naturellement plus de TN en FR, car il y a plus de personnes inconnues probables qu'enregistrées au système, mais les cas de bonnes classifications qui nous intéressent davantage sont les TP, ou les erreurs FP et FN le cas échéant. Les deux paramètres PPV et TPR du F_1-score ont pour numérateur les TP, ce qui permet de mesurer efficacement la qualité spécifique aux individus du *watchlist* reconnus à travers le réseau de sécurité par caméras.

La mesure F_1-score correspond en fait au cas spécifique de l'équation (5.2) lorsque $\beta = 1$, soit lorsque le score est balancé équitablement entre les deux mesures sous-jacentes. Ainsi, dans un cas où l'on voudrait davantage un système performant du point de vue du rappel ou de la précision, il est possible d'ajuster le compromis entre les deux mesures. Réduire à $\beta = 0.5$ produit une évaluation qui favorise la précision au détriment du rappel, ce qui vient atténuer l'influence des classifications FN, alors qu'appliquer $\beta = 2$ rehausse la pondération du rappel par rapport à la précision, ce qui leurs donne plus d'importance.

$$F_\beta = \frac{\left(1+\beta^2\right)\cdot precision \cdot recall}{\beta^2 \cdot precision + recall} = \frac{\left(1+\beta^2\right)\cdot \mathrm{PPV}\cdot \mathrm{TPR}}{\beta^2 \cdot \mathrm{PPV} + \mathrm{TPR}} \qquad (5.2)$$

Considérant que les cas FN correspondent à des individus d'intérêt du *watchlist* qui auraient incorrectement été classés comme des inconnus quelconques, c'est-à-dire qu'ils n'auraient pas été reconnus par le système, il semble logique d'appliquer une pondération $\beta = 2$ plus élevée. Les mesures de F_1-score et de F_2-score sont donc rapportées pour le système le plus performant qui sera retenu à la section 5.3.

Les dernières mesures d'évaluation considérées sont davantage à caractère applicatif au niveau de la rapidité et efficacité du système en phase opérationnelle pour accomplir les prédictions. Spécifiquement, la mesure de complexité computationnelle par rapport au nombre d'opérations, de paramètres, et de la vitesse d'exécution, permettent respectivement de prendre en considération le nombre d'opérations par produits scalaires, le nombre de valeurs du modèle à stocker en mémoire ainsi que la rapidité des calculs de l'algorithme de classification afin de produire un unique score de reconnaissance d'un individu. L'équation (5.3) définie la complexité des opérations pour faire une prédiction par chaque SVM individuellement, selon les définitions retirées de [10]. Étant donné les modes d'opérations des SVM dans le contexte de ce projet, le nombre de paramètres est directement équivalent au nombre d'opérations. Le nombre de paramètres définissant les modèles est donc directement le même que la quantité d'opérations employée pour évaluer les nouveaux visages. Les quantités utilisées dépendent des paramètres employés par le système, selon chaque variation testée, surtout par rapport au nombre de caractéristiques par vecteur de support (N_d ou $N_d^{(rs)}$) et l'utilisation de RSM ou non ($N_{rs} = 1$ si inutilisé, car seulement 1 descripteur HOG), qui varient pour différents tests.

$$\#opérations = N_p \cdot N_{rs} \cdot N_{sv} \cdot N_d \tag{5.3}$$

5.2.2 Évaluations par reconnaissance et par identification

Étant donné que l'on a un suivi des ROI de visage avec leurs scores associés le long des séquences vidéo, il est possible d'évaluer le système selon le mode par transaction (chaque *frame* indépendamment) ou par trajectoire (avec l'accumulation de scores). Par contre, les travaux qui utilisent des modules de suivi de visages par trajectoires montrent que les trajectoires produisent de bien meilleurs résultats [48], et c'est également ce qui a été observé lors des tests expérimentaux. Les performances présentées aux sections qui suivront seront donc spécifiquement par rapport à la FR par trajectoire de visage, profitant ainsi de la redondance d'information pour des prédictions plus fiables et robustes aux conditions de capture variées dans l'environnement et à travers les multiples caméras.

Finalement, deux types d'évaluations sont considérées, afin de prendre en compte les différents modes d'opération possibles du système de FRiVS, soit par reconnaissance de surveillance et par identification (ou contrôle d'accès) [19, 20]. Dans le premier cas, il s'agit simplement de la réponse de scores accumulés, par ensemble de classificateurs respectif à chaque individu d'intérêt, qui sont utilisés pour obtenir les mesures de performance précédemment décrites (pAUC et AUPR par individu), pour ensuite être moyennées selon la réplication donnée. Dans le cas de l'identification, les scores accumulés par trajectoire sont premièrement comparés, selon l'équation (5.4), afin de déterminer lequel des individus produit la réponse la plus forte (plus haute confidence), à condition que les réponses respectives dépassent le seuil de reconnaissance spécifique à chaque individu. Si le seuil n'est pas atteint, le score est mis à zéro, et si l'ensemble des scores de tous les individus dans la *watchlist* résultent à une valeur nulle, on considère qu'aucun des individus d'intérêt n'est présent dans la scène. Ainsi, cette approche fonctionne de manière similaire aux détecteurs CNN à multiples classes, où un vecteur de scores des classes possibles (nos individus du *watchlist*) est produit avec leurs confidences de détection respectives et, si tous les scores sont plus bas que le seuil de détection, l'objet n'est pas reconnu (dans notre cas, c'est un *unknown non-target*).

$$S_I^{acc*} = \max \mathbf{S}_s^{acc*} \Big|_{\mathbf{S}_s^{acc*} \geq \gamma_s} \tag{5.4}$$

Des analyses expérimentales *on-line* avec le système de bout-en-bout de FRiVS ont permis de voir l'efficacité de cette approche, ce qui est reflété dans les résultats de performances obtenus. Le choix de la fonction « max » est ainsi basé sur les observations expérimentales effectuées. Avec l'individu identifié (ou l'absence de tous les individus du *watchlist*), les mesures de performances sont ensuite calculées de la même manière que dans le mode d'évaluation en reconnaissance de surveillance. Les deux modes ont été appliqués de manière itérative pour chacune des réplications considérées, en interchangeant adéquatement les individus d'entraînement et de test, puis itérativement pour chacune des variations d'entraînement présentées au Tableau 5.1.

5.3 Résultats de la reconnaissance de visage

Cette section présentera les résultats obtenus en deux temps, soit avec le système initial n'employant que l'accumulation par trajectoire de visages d'individus captés par les caméras de sécurité, puis avec les différentes combinaisons de méthodes d'augmentation de données et de génération synthétique testées. Afin de clarifier et simplifier l'ensemble des résultats obtenus, seulement les performances sommaires des séquences à partir des caméras affichant des visages « frontaux », ainsi que pour l'évaluation sur l'intégralité de la base ChokePoint, seront présentées ici. Les résultats détaillés seront présentés à l'ANNEXE I pour référence.

5.3.1 Système initial avec suivi de trajectoire

Les résultats suivants correspondent au système initial. Celui-ci est donc employé comme référence de base pour la comparaison des autres variations qui suivront à la section suivante.

Tableau 5.2 Résultats moyens de 5 réplications de test du système initial de FRiVS avec suivi de trajectoires de visages, pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint

Sequences de test		Caméras frontales		Toute la base	
Variation d'entraînement	Mesure de performance	Surveillance	Identification	Surveillance	Identification
Système initial	pAUC(20%)	86.75% ± 0.66%	97.36% ± 2.32%	77.30% ± 0.31%	69.98% ± 0.70%
	AUPR	67.69% ± 4.55%	97.71% ± 1.57%	61.23% ± 5.66%	61.30% ± 0.09%

Il est possible de voir avec le Tableau 5.2 que les performances du système initial sont déjà extrêmement fortes pour le cas des caméras frontales en mode d'identification. Les autres cas sont toutefois loin d'atteindre les niveaux de performance faisant l'état-de-l'art. Ainsi, on peut fois que la pose à un impact majeur sur les résultats, justifiant pourquoi certaines des techniques de génération synthétique des visages ont été prévues pour ce problème spécifiquement. De plus, en mode de surveillance, les performances ne sont pas suffisantes dans tous les cas.

5.3.2 Système avec génération synthétique et adaptation de domaine

Les résultats des quatre premières combinaisons de générations synthétiques explorées, formant le premier bloc de test qui avait été considéré lors de la conception du système, sont tous présentés au Tableau 5.3. Ces combinaisons utilisent seulement de la génération synthétique sous forme des transformations géométriques et l'ajout des images de référence uniques des autres individus de la cohorte comme contre-exemples additionnels.

Tableau 5.3 Résultats moyens de 5 réplications de test du système avec les combinaisons ST_{nt}, ST_{nt}+GT_{nt}, GT_{wl}+ST_{nt} et GT_{wl}+ST_{nt}+GT_{nt} pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint (résultats optimaux en gras, tant pour ces combinaisons que par rapport au système initial)

Sequences de test		**Caméras frontales**		**Toute la base**	
Variation d'entraînement	**Mesure de performance**	**Surveillance**	**Identification**	**Surveillance**	**Identification**
ST_{nt}	pAUC(20%)	91.51% ± 0.24%	96.06% ± 3.09%	79.99% ± 0.16%	92.12% ± 0.01%
	AUPR	77.81% ± 3.07%	96.03% ± 1.00%	65.56% ± 4.69%	91.40% ± 0.55%
ST_{nt}+GT_{nt}	pAUC(20%)	91.32% ± 1.53%	97.35% ± 1.76%	81.12% ± 0.56%	92.38% ± 1.86%
	AUPR	81.66% ± 2.25%	96.76% ± 0.02%	66.96% ± 5.36%	91.15% ± 3.00%
GT_{wl}+ST_{nt}	pAUC(20%)	94.94% ± 0.56%	98.33% ± 1.67%	82.92% ± 1.04%	93.84% ± 0.79%
	AUPR	89.68% ± 0.19%	98.22% ± 1.78%	72.35% ± 0.79%	93.46% ± 1.11%
GT_{wl}+ST_{nt}+GT_{nt}	pAUC(20%)	**95.51% ± 0.26%**	**98.69% ± 0.86%**	**83.96% ± 1.07%**	**94.56% ± 1.40%**
	AUPR	**89.95% ± 0.23%**	**98.25% ± 0.75%**	**73.01% ± 1.70%**	**93.79% ± 1.79%**

Nous pouvons remarquer très rapidement à partir des résultats du Tableau 5.3 que les meilleures performances sont toutes atteintes avec la combinaison de test GT_{wl}+ST_{nt}+GT_{nt}, soit lorsque le maximum de nouvelles représentations de visages, indépendamment des classes, est ajouté pour l'entraînement très spécialisé de Eoℓ-SVM par individu du *watchlist*.

Étant donné les résultats obtenus au premier bloc de test effectué (évaluation des transformations géométriques), il a été possible de déterminer que le système avait été amélioré avec succès, comparativement aux résultats initiaux du Tableau 5.2. Par contre, alors que les résultats sur les séquences « frontales » étaient nettement meilleurs, autant en surveillance qu'en identification, les performances sur l'ensemble de la base de données demeuraient relativement basses (par rapport à l'état-de-l'art) dans le cas du mode en surveillance. Ceci laissait envisager que les modèles obtenus étaient très robustes à de multiples conditions d'acquisition, telle l'illumination ou le flou de motion, mais semblaient encore souffrir des poses importantes du visage sur les séquences vidéo, car il s'agit de la distinction principale observable entre les points de vue des vidéos de séquences « frontales » et le reste.

Ainsi, cette constatation a mené au second bloc de tests qui, lui, tente de s'attaquer très particulièrement au problème de poses par l'utilisation de modèles 3D pour générer ces variations probables du visage selon les observations en profil réellement capturées de la scène. Les résultats de ce second bloc sont présentés au Tableau 5.4, où tant l'utilisation de 3DMM à elle seule, ainsi que combinée avec la meilleure variation retenue du bloc de test précédent, permet de générer synthétiquement les modèles de visages avec poses tridimensionnelles.

Tableau 5.4 Résultats moyens de 5 réplications de test du système avec les combinaisons 3DMM à elle seule et 3DMM+GT_{wl}+ST_{nt}+GT_{nt} pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint

Sequences de test		**Caméras frontales**		**Toute la base**	
Variation d'entraînement	**Mesure de performance**	**Surveillance**	**Identification**	**Surveillance**	**Identification**
3DMM	pAUC(20%)	78.05% ± 2.82%	91.14% ± 7.14%	72.30% ± 0.93%	87.49% ± 4.52%
	AUPR	62.15% ± 3.17%	91.73% ± 4.96%	54.04% ± 2.63%	86.22% ± 4.34%
3DMM + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(20%)	90.29% ± 0.59%	95.54% ± 0.24%	85.04% ± 1.10%	93.01% ± 0.73%
	AUPR	83.11% ± 3.46%	95.75% ± 0.36%	75.02% ± 1.39%	92.55% ± 0.05%

Malheureusement, les performances obtenues lors de ce second bloc de test n'étaient pas aussi satisfaisantes que pour le bloc précédent, même lorsque les deux approches (générations synthétiques par translations géométriques et par modèles 3DMM) étaient employées ensemble. En comparaison aux résultats du système initial au Tableau 5.2, l'ajout des visages de poses synthétiques avec 3DMM uniquement venait détériorer le système de FRiVS. Lorsque les approches de transformations géométriques étaient combinées aux 3DMM, les performances étaient maintenant meilleures que le système initial, mais toujours plus faibles que la combinaison GT_{wl}+ST_{nt}+GT_{nt} à elle seule. Ainsi, il est fort probable que la seule raison en cause du gain en performance de 3DMM+GT_{wl}+ST_{nt}+GT_{nt} par rapport au système initial soit induite par les générations synthétiques de transformations géométriques qui ont démontré une nette amélioration.

Or, il a été considéré, basé sur ces résultats expérimentaux, que les modèles 3D qui ont été générés pour représenter les visages des individus d'intérêt ne soient probablement pas suffisamment représentatifs ou fidèles aux observations attendues sur les vidéo-caméras. Étant donné que les Eoℓ-SVM qui sont employés tentent de très fidèlement représenter les individus avec un nombre limité d'exemples, le fait d'ajouter des représentations insuffisamment discriminatives pour différencier l'individu d'intérêt des autres ne fait qu'en réalité bruiter les modèles déjà extrêmement sensibles aux faibles variations présentent dans la classe positive. Effectivement, étant donné la quantité très limitée de vecteurs de support disponibles du côté de la classe positive, les Eoℓ-SVM sont rapidement affectés par n'importe quels *outliers* ajoutés durant l'entraînement. C'est pourquoi il est bien nécessaire d'avoir des représentations d'extrêmement bonne qualité par rapport aux observations attendues dans la scène.

Pour le cas des transformations géométriques, puisque les modèles de visages additionnels introduits n'étaient que de simples variations de la position des pixels, mise à l'échelle ou de flou, les distinctions par rapport aux images de référence demeuraient relativement contrôlées et rapprochées dans l'hyper-espace des vecteurs. Cela ne paraît pas être le cas pour les modèles générés par 3DMM, car trop de variation intra-classe semble être introduite, et les classificateurs deviennent confondus par ces mauvais exemples de nouveaux visages.

Les conclusions qui avaient été tirées par les résultats du second bloc de test ont motivé la suppression de la génération de pose synthétique par les modèles 3D du système final, ce qui a nécessité une recherche de méthode alternative pour améliorer la performance de FR, tout en demeurant fidèles aux représentations de visages attendues sur vidéos. Ceci explique l'introduction du troisième bloc de test, qui exploite la génération de multiples représentations à partir des vecteurs de support variés, directement par la sélection des caractéristiques qui avaient été précédemment extraites plutôt que par représentations à partir de l'image.

Considérant que la combinaison GT_{wl}+ST_{nt}+GT_{nt} avait démontré par preuve empirique que les vecteurs obtenus correspondaient à des représentations valides et robustes de visages attendus dans la scène pour les individus enregistrés à la *watchlist*, le raisonnement était alors que les caractéristiques sous-jacentes qui les composent étaient également adéquates pour discriminer les visages. Ainsi, la méthode RSM, qui ne fait que sélectionner aléatoirement certaines de ces caractéristiques, afin de les recombiner en sous-ensembles distincts, ne fait qu'exploiter les *features* confirmées comme robustes aux conditions d'acquisition des séquences vidéo. L'application de RSM, tant à elle seule que combinée avec les générations GT_{wl}+ST_{nt}+GT_{nt}, a produit les résultats du Tableau 5.5.

Tableau 5.5 Résultats moyens de 5 réplications de test du système avec les combinaisons $RSM_{(20,128)}$ à elle seule et $RSM_{(20,128)}$+GT_{wl}+ST_{nt}+GT_{nt} pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint (résultats optimaux en gras, tant pour ces combinaisons que par rapport à GT_{wl}+ST_{nt}+GT_{nt})

Sequences de test		**Caméras frontales**		**Toute la base**	
Variation d'entraînement	**Mesure de performance**	**Surveillance**	**Identification**	**Surveillance**	**Identification**
$RSM_{(20,128)}$	pAUC(20%)	94.15% ± 0.70%	99.22% ± 0.44%	83.53% ± 0.86%	94.52% ± 1.19%
	AUPR	89.66% ± 1.49%	96.87% ± 2.16%	**77.06% ± 1.86%**	**95.47% ± 1.70%**
$RSM_{(20,128)}$ + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(20%)	94.40% ± 1.03%	**99.70% ± 0.11%**	82.61% ± 0.09%	94.09% ± 0.24%
	AUPR	89.09% ± 3.42%	**98.32% ± 1.28%**	75.84% ± 2.87%	95.22% ± 0.81%

Lorsqu'utilisée simplement, l'approche RSM est appliquée aux descripteurs HOG_{588} originaux provenant des *patches* de la ROI d'images uniques statiques de référence des *targets*, ainsi que pour des *patches* des ROI d'inconnus provenant des séquences vidéo. Dans le cas où RSM est combinée avec les transformations géométriques ($RSM_{(20,128)}$+GT_{wl}+ST_{nt}+GT_{nt}), ces dernières sont tout d'abord appliquées afin d'obtenir l'ensemble de variations de visages synthétiques, afin d'y employer RSM par la suite. Les visages obtenus sur séquences vidéo en mode opérationnel sont directement traités avec les mêmes opérations qu'à l'entraînement pour les deux combinaisons avec ou sans images synthétiques.

Ainsi, plutôt que d'obtenir des ensembles de N_p SVM à base de vecteurs de dimensionnalité $N_d = 588$, nous obtenons des ensembles de $N_p \cdot N_{rs}$ SVM à base de vecteurs à $N_d = 128$ *features*. Étant donné la quantité de vecteurs générés aléatoirement et la quantité de caractéristiques utilisées par chacun, plusieurs des caractéristiques originales sont transférées en redondance, ce qui fait que certaines combinaisons de relations entre celles-ci sont préservées, tout en simplifiant la complexité d'ensemble des descripteurs en réduisant leur dimension. Aussi, en ajoutant de multiples variations de sous-ensembles de données dans l'hyper-espace de caractéristiques, nous obtenons des ensembles de classificateurs potentiellement plus compétents pour classer certaines particularités bien distinctes du visage ou pour certains cas très spécifiques de conditions ardues d'acquisition sur les vidéos. En d'autres termes, certains des classificateurs dans l'ensemble peuvent devenir encore plus spécialisés pour effectuer la classification de davantage de combinaisons différentes, et les meilleurs font globalement améliorer la prédiction finale de l'individu à reconnaître.

Les résultats présentés au Tableau 5.5 semblent effectivement montrer que l'approche RSM a permis d'améliorer légèrement certains des cas d'évaluation, ce qui vient supporter l'idée que des classificateurs compétents sont sûrement obtenus avec cette nouvelle génération de représentations des vecteurs descripteurs extraits des visages. Par contre, nous remarquons également que les améliorations apportées principalement pour l'identification viennent au détriment de quelques baisses de performance du côté de la reconnaissance en surveillance.

Toujours en quête d'amélioration du système de FRiVS, un dernier ajustement a été apporté, soit de considérer la sélection très spécifique des individus inconnus à partir de chacune des caméras individuellement. Ceci permet d'obtenir des modèles de Eoℓ-SVM extrêmement spécialisés par un entraînement adapté au domaine particulier à chaque scène observée. Ce dernier bloc de test emploie la combinaison la plus optimale jusqu'à présent (GT_{wl}+ST_{nt}+GT_{nt}) en tenant compte de l'ensemble des scénarios de test évalués (séquences frontales/complètes et en mode surveillance/identification). Malgré que la quantité de vidéos de contre-exemples soit grandement diminuée à cause de la sélection spécifique par caméra (nous passons facilement de 20000 visages répartis à travers toutes les séquences à approximativement 500 par caméra), la génération synthétique par transformations géométriques n'est pas affectée, car celle-ci n'emploie que les images uniques de référence pour enregistrer les individus d'intérêt à la liste de contrôle. Les résultats obtenus pour cette évaluation sont présentés au Tableau 5.6.

Tableau 5.6 Résultats moyens de 5 réplications de test du système avec la combinaison DA+GT_{wl}+ST_{nt}+GT_{nt} pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint
(résultats optimaux en gras par rapport à tous les tests précédents)

Sequences de test		**Caméras frontales**		**Toute la base**	
Variation d'entraînement	**Mesure de performance**	**Surveillance**	**Identification**	**Surveillance**	**Identification**
DA + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(20%)	**97.00% ± 1.09%**	96.75% ± 0.69%	**90.48% ± 1.87%**	**95.45% ± 0.05%**
	AUPR	88.94% ± 1.20%	92.95% ± 2.61%	**85.87% ± 1.53%**	**95.92% ± 0.40%**

Nous pouvons voir à partir des résultats obtenus que l'adaptation de domaine vient définitivement bénéficier à la performance du système, principalement lorsque l'on compare les écarts obtenus pour le mode de surveillance avec tous les autres tests. Ainsi, non seulement la spécialisation des Eoℓ-SVM au domaine spécifique est efficace, mais il est également possible de considérer que la sélection réduite des contre-exemples fait que les classes résultantes soient grandement moins déséquilibrées qu'auparavant. Ceci a très possiblement aidé les classificateurs à mieux définir les marges de séparations.

5.3.3 Sommaire des résultats et comparaison à l'état-de-l'art

Les meilleures combinaisons de test obtenues pour l'intégralité des évaluations effectuées lors de ce projet sont présentées au Tableau 5.7 et au Tableau 5.8. Ainsi, nous pouvons voir que, selon le type de mode d'opération du système de FRiVS désiré, différentes approches devraient être considérées parce que les performances varient de manière notable entre chaque cas. Ceci reflète d'ailleurs qu'il existe rarement une solution unique à toutes les situations.

Tableau 5.7 Résultats moyens de 5 réplications de test du système avec les meilleures combinaisons retenues pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint
(résultats optimaux en gras par rapport à tous les tests effectués)

Sequences de test		**Caméras frontales**		**Toute la base**	
Variation d'entraînement	**Mesure de performance**	**Surveillance**	**Identification**	**Surveillance**	**Identification**
$GT_{wl}+ST_{nt}+GT_{nt}$	pAUC(20%)	95.51% ± 0.26%	98.69% ± 0.86%	83.96% ± 1.07%	94.56% ± 1.40%
	AUPR	**89.95% ± 0.23%**	98.25% ± 0.75%	73.01% ± 1.70%	93.79% ± 1.79%
$RSM_{(20,128)}$ + $GT_{wl}+ST_{nt}+GT_{nt}$	pAUC(20%)	94.40% ± 1.03%	**99.70% ± 0.11%**	82.61% ± 0.09%	94.09% ± 0.24%
	AUPR	89.09% ± 3.42%	**98.32% ± 1.28%**	75.84% ± 2.87%	95.22% ± 0.81%
DA + $GT_{wl}+ST_{nt}+GT_{nt}$	pAUC(20%)	**97.00% ± 1.09%**	96.75% ± 0.69%	**90.48% ± 1.87%**	**95.45% ± 0.05%**
	AUPR	88.94% ± 1.20%	92.95% ± 2.61%	**85.87% ± 1.53%**	**95.92% ± 0.40%**

Tableau 5.8 Résultats moyens F_β de 5 réplications de test du système avec les meilleures combinaisons retenues pour les modes de surveillance et d'identification, sur les séquences à caméras frontales et l'intégralité de ChokePoint

Sequences de test		**Caméras frontales**		**Toute la base**	
Variation d'entraînement	**Mesure de performance**	**Surveillance**	**Identification**	**Surveillance**	**Identification**
$RSM_{(20,128)}$ + $GT_{wl}+ST_{nt}+GT_{nt}$	F_1	**56.81% ± 9.93%**	**96.85% ± 0.30%**	**54.77% ± 11.96%**	91.57% ± 0.03%
	F_2	**65.19% ± 4.96%**	**98.36% ± 0.23%**	62.60% ± 7.55%	**89.97% ± 0.10%**
DA + $GT_{wl}+ST_{nt}+GT_{nt}$	F_1	48.44% ± 8.27%	89.37% ± 3.65%	52.91% ± 9.72%	89.77% ± 0.49%
	F_2	56.28% ± 7.18%	92.62% ± 1.72%	**63.23% ± 5.19%**	**91.62% ± 0.08%**

Afin d'évaluer les résultats obtenus selon des méthodes faisant partie de l'état-de-l'art, deux catégories d'algorithmes sont considérées. Dans le premier cas, il s'agit de certains des meilleurs classificateurs retrouvés et employés spécifiquement pour la FRiVS en mode S2V indépendamment des techniques sous-jacentes. La seconde catégorie implique quant à elle seulement des méthodes faisant explicitement la mise en application d'une approche par suivi de trajectoire de visages pour une reconnaissance également à base de S2V et d'entraînement en condition de SSPP, ce qui inclus les complexités supplémentaires de FT, comme les erreurs de *fail-to-acquire*, qui ne sont simplement pas considérées dans la première catégorie. Ainsi, la comparaison avec la première catégorie se veut un peu plus pessimiste, afin d'avoir un aperçu des améliorations supplémentaires souhaitables pour obtenir un système minimalement équivalent à l'état-de-l'art, advenant aucune erreur de FD ou de FT rencontrée. Par contre, la seconde comparaison se veut plus raisonnable, car les systèmes comparés fonctionnent sur les mêmes bases et, donc, sont définitivement plus équivalents pour comparer selon les mêmes difficultés rencontrées, incluant le désalignement des pixels et le *drift* de trajectoire.

Afin de simplifier l'interface des tableaux, la meilleure solution qui utilise la méthode RSM est annoté par Eoℓ-SVM-RSM (au lieu de $RSM_{(20,128)}$+GT_{wl}+ST_{nt}+GT_{nt}) alors que celle qui applique la DA est annotée Eoℓ-SVM-DA (au lieu de DA+GT_{wl}+ST_{nt}+GT_{nt}). Les deux utilisent leur méthode en combinaison avec les multiples générations synthétiques par transformations géométriques des images de visages de référence des individus d'intérêt. La complexité computationnelle en termes d'opérations et de paramètres sont respectivement déterminées par $(9 \cdot 20 \cdot 46 \cdot 128)$ et $(9 \cdot 1 \cdot 46 \cdot 588)$ pour Eoℓ-SVM-RSM et Eoℓ-SVM-DA étant donné qu'en moyenne $N_{sv} = 46$ vecteurs de support étaient générés par chacun des sous-SVM. La vitesse de calcul des deux méthodes est approximée à environ 0.0598±0.01 s/frame (~16.72 FPS) puisque les opérations de prédiction des ensembles de SVM sont accomplies en parallèle avec plusieurs niveaux d'optimisation, ce qui fait que la partie la plus lente et limitante du système est en fait la triple détection de Viola-Jones. Le temps requis pour la classification même des visages par les Eoℓ-SVM est donc largement négligeable par rapport à la FD globale sur l'ensemble du *frame* du vidéo.

Les résultats comparatifs des différentes catégories de systèmes de FRiVS, telles que tout juste décrites, sont présentés respectivement au Tableau 5.9 et au Tableau 5.10. Nous pouvons ainsi voir que les variantes optimales de Eoℓ-SVM retenues permettent d'obtenir des résultats qui s'approchent des meilleurs systèmes de FRiVS en S2V faisant l'état-de-l'art pour toutes les conditions et tous les modes d'opération confondus, malgré les difficultés de FT ajoutées.

Tableau 5.9 Comparaison des performances des meilleures variantes du système avec d'autres méthodes en S2V faisant l'état-de-l'art, toutes situations confondues, sur l'intégralité des séquences de ChokePoint

Système de FRiVS en mode S2V		**pAUC(20%)**	**AUPR**	**Complexité**	
				#opérations	**# paramètres**
Eo*e*-SVM	[11, 12]	100.0±0.00	99.24±0.38	2.23M	230K
Eo*e*-SVM-DS(1)	[9]	97.52±0.50	96.86±0.72	114K	99K
Eo*e*-SVM-DS(2)	[11]	100.0±0.00	99.31±0.46	N/A	N/A
ESRC-DA	[91]	N/A	76.97±0.07	228M	41.5M
CCM-CNN	[99]	N/A	98.87±0.63	33.3M	2.40M
CFR-CNN	[101]	N/A	96.47±0.86	3.75M	1.20M
HaarNet	[100]	N/A	99.36±0.59	3.50B	13.1M
Eoℓ-SVM-RSM	**nôtre**	94.09±0.24	95.22±0.81	1.06M	1.06M
Eoℓ-SVM-DA	**nôtre**	95.45±0.05	95.92±0.40	243K	243K

Tableau 5.10 Comparaison des performances des meilleures variantes du système avec d'autres méthodes en S2V à partir de SSPP et avec suivi de trajectoire faisant l'état-de-l'art, sur l'intégralité des séquences de ChokePoint, en modes transaction et trajectoire

Système de FRiVS en mode S2V et à partir de SSPP et suivi de trajectoire		**Transaction**		**Trajectoire**		**Temps de calcul (s/frame)**	
		pAUC(5%)	**AUPR**	**pAUC(5%)**	**AUPR**		
TM-FR	[26]	0.249±0.02	0.457±0.04	0.318±0.02	0.423±0.03	0.1910±0.05	[7]
TM-SU	[112]	0.347±0.03	0.452±0.05	0.391±0.02	0.478±0.02	0.2920±0.07	[7]
MFR	[13]	0.417±0.03	0.512±0.05	0.472±0.02	0.532±0.02	0.2500±0.08	[7]
SVDL	[141]	0.439±0.03	0.533±0.05	0.583±0.02	0.631±0.02	0.2370±0.05	[7]
AAMT-FR	[45, 47]	**0.494±0.02**	**0.588±0.04**	0.649±0.07	0.793±0.03	0.2170±0.06	[7]
Eoℓ-SVM-RSM	**nôtre**	0.406±0.03	0.495±0.04	0.711±0.03	0.758±0.03	0.0598±0.01	[8]
Eoℓ-SVM-DA	**nôtre**	0.469±0.03	0.565±0.03	**0.823±0.02**	**0.859±0.02**	**0.0598±0.01**	[8]

[7] Temps de calcul reporté par l'article correspondant en référence.

[8] Évalué par une machine équipée de carte Intel® Core™ i7-5500U (2.40Ghz) 16GB RAM, et d'un GPU NVIDIA GeForce GTX 950M, avec algorithmes optimisés par les bibliothèques OpenCV, OpenMP et CUDA.

Dans le cas des méthodes employant spécifiquement du *tracking* pendant la phase d'opération (Tableau 5.10), et conditionnées par le SSPP pour l'entraînement et le mode d'opération S2V-FR, l'augmentation du nombre de représentations de visage permet de nettement surpasser les systèmes faisant l'état-de-l'art, surtout lorsque l'information spatio-temporelle de la trajectoire est employée. Lorsque seulement les prédictions par transaction sont employées, soit chaque *frame* individuellement, nous perdons beaucoup en performances. Malgré tout, ce dernier mode produit tout de même des reconnaissances de visage qui se rapprochent à des méthodes existantes, ou les surpassent tout simplement. Le cas du AAMT-FR est particulièrement intéressant sur ce point étant donné que le modèle interne de visage du *tracker* y est adapté en-ligne selon les nouvelles apparences de visages captées dans la scène, c'est-à-dire selon le même principe que le module de suivi de trajectoire employé lors de ce projet. Ceci fait que ce système est probablement la meilleure base de comparaison pour démontrer les gains de précision de la reconnaissance de visages atteints.

Nous pouvons voir aussi avec le Tableau 5.9 que, bien que les performances des approches proposées par Eoℓ-SVM soient un peu plus faibles que les systèmes existants en littérature, soit principalement de gros modèles à base de CNN, les solutions de FRiVS proposées pour le S2V sont clairement plus compactes en mémoire, surtout pour le cas de Eoℓ-SVM-DA, étant donné la quantité réduite de paramètres nécessaires par les vecteurs de support pour représenter un modèle d'un individu d'intérêt dans la *watchlist*. Les méthodes Eo*e*-SVM et Eo*e*-SVM-DS sont quant à elles de bons points de comparaison pour définir les futurs objectifs à viser, ainsi que des pistes de développements futurs, car il s'agit à la base des mêmes classificateurs que le Eoℓ-SVM employé ici.

Par contre, comme il l'a déjà été mentionné, les résultats de ces deux techniques ne considèrent pas la complexité ajoutée du suivi de trajectoire qui est nécessaire afin de distinguer divers individus simultanément dans la scène pour obtenir un système complet de FRiVS. Ainsi, ces deux méthodes sont alors de bons points de référence dans le cas où nous réussissions à supprimer les problèmes observés en lien au suivi de trajectoire.

Les performances du Tableau 5.10 permettent quant à elles de voir les gains en performances majeurs par l'utilisation de Eoℓ-SVM très spécialisés pour la reconnaissance d'individus entraînés à base de SSPP. Cette robustesse n'est que davantage renforcée avec l'application du suivi de trajectoire d'apparences de visages dans la scène des caméras de surveillance pour une accumulation spatio-temporelle de scores de FR des individus recherchés. De plus, l'utilisation de la génération synthétique à partir de transformations géométriques sur les images de références d'individus d'intérêt, combinée à la sélection aléatoire de caractéristiques pour encore davantage de représentations des visages, ou avec l'adaptation de domaine très spécifique de chaque caméra, permet toujours d'obtenir un système de FRiVS encore plus robuste. Ceci rend la S2V-FR efficace sur les séquences vidéo même si elles sont marquées de grandes nuisances de conditions de capture en environnement semi-contrôlé.

Il est également important de mettre davantage d'emphase sur la rapidité du système obtenue qui est indiquée au Tableau 5.10. Cette vitesse d'exécution en mode d'opération qui approche environ 16.72FPS est obtenue à l'aide d'une machine relativement simple, comparativement à plusieurs travaux en littérature qui se dotent souvent des cartes graphiques GTX1080 ou TITAN beaucoup plus performantes. Ceci indique alors que l'architecture proposée du système de FRiVS (Figure 3.1) est extrêmement optimale pour une application légère et propice à des performances encore plus accrues si elle était déployée sur une machine dédiée de plus haut calibre ou, encore plus, sur un serveur distribué spécialement pour la vidéosurveillance.

CONCLUSION

Ce rapport a présenté une nouvelle architecture qui vient combiner les avantages de l'accumulation spatio-temporelle de scores de reconnaissance de visages par suivi de trajectoire avec diverses variantes d'ensemble de classificateurs SVM exemplaires très spécialisé par un entraînement à partir de représentations du visage des individus d'intérêt générés à l'aide de plusieurs méthodes.

Les différentes approches considérées pour la conception initiale du système de FRiVS ainsi que les bases qui s'y rattachent ont également été discutées. Les méthodes retenues ont été justifiées selon les résultats expérimentaux, ainsi que des références en littérature scientifique, qui appuient leur utilisation répandue et qui démontrent que celles-ci apportent de meilleures performances dans le domaine de la FR spécifiquement. Les variantes du système amélioré par la génération des représentations multiples de visages, fondées sur diverses approches et principes distincts, ont ensuite été présentées en détail. Les variantes comparées ont été évaluées avec la base de vidéos ChokePoint, permettant un apprentissage en SSPP d'individus d'intérêt à partir de leur image de référence unique, et sont testées contre plusieurs séquences vidéo correspondant à l'environnement d'acquisition de différentes caméras de sécurité.

Les résultats obtenus et comparés à plusieurs systèmes formant l'état-de-l'art en reconnaissance de visage démontrent deux conclusions selon la situation. Pour une évaluation comparative générale des systèmes de FRiVS opérant selon divers modes d'opération, le système actuel demeure sous les requis de performances à atteindre. Par contre, cette évaluation sous-estime l'importance des problématiques additionnelles reliées au suivi de trajectoire de visages. Malgré tout, le système proposé offre des améliorations en termes de complexité des modèles par rapport à ces approches. L'autre évaluation, qui compare des systèmes de S2V similairement entraînés à partir de SSPP et fondés sur l'évaluation de la FR à partir de suivi de trajectoires des visages détectés, démontre qu'une amélioration majeure est observée grâce à l'utilisation de la génération synthétique combinée aux classificateurs hautement spécialisés par ces représentations multiples du visage de référence.

Contributions

Le présent projet apporte plusieurs contributions en lien avec les différents aspects qui y sont traités. Tout d'abord, une évaluation de plusieurs méthodes de génération de représentations du visage reliées à différentes méthodologies est accomplie pour les comparer de manière exhaustive. Ceci a permis de produire des résultats expérimentaux fondés sur des mesures de performances courantes en apprentissage machine qui peuvent être employées comme référence pour justifier leur sélection ou leurs applications futures dans un autre système.

Ensuite, diverses architectures de systèmes de FRiVS de bout-en-bout sont proposées afin de produire une reconnaissance de visage robuste considérant les réalités d'application pour les systèmes de sécurité. Ainsi, les solutions proposées ne font pas abstraction d'un grand ensemble de problématiques rencontrées dans ce domaine, qui sont malheureusement souvent simplifiées ou très peu discutées en littérature. Parmi les multiples problèmes existants, le système de FRiVS proposé considère les cas suivants sans s'y limiter :

- Les problèmes produisant des manquements ou de fausses détections de visages sur vidéo et induisant des erreurs de reconnaissance, n'ayant aucun visage à comparer;
- Les problèmes de perte et de dérive de trajectoire, ainsi que de désalignement des régions d'intérêt des visages, réduisant les possibilités de fortes prédictions de FR;
- Les complexités associées à la combinaison en parallèle de la détection de visages, du suivi de trajectoires, ainsi que les manières pour produire des ROI robustes;
- Les conditions ardues et distinctes d'acquisition sur séquences vidéo provenant de divers environnements à partir de multiples caméras de sécurité;
- Le manque de données pour l'entraînement de modèles de visages discriminants pour reconnaître les individus dans la scène, à travers des domaines d'acquisition différents;
- Une considération pour le fonctionnement en temps réel et les optimisations associées à appliquer pour réduire la complexité et l'espace mémoire d'un système dynamique;
- L'élaboration de modèles légers, rapides à entraîner, et selon une architecture particulière permettant d'ajouter facilement de nouveaux individus d'intérêt *on-line*.

Ainsi, les points précédents font que les systèmes proposés sont réalistes par rapport aux réalités qui seraient rencontrées en milieu industriel, dans les lieux publics ou pour les points de contrôle par les agences gouvernementales, qui pourraient vouloir employer des systèmes de FRiVS de la sorte. Ceci apporte à tous ces utilisateurs potentiels des méthodologies applicables rapidement tout en tenant de limiter autant que possible les mauvaises surprises pouvant subvenir autrement lors de la mise en application de techniques proposées.

Une autre contribution importante est l'ajout des modèles par représentations multiples employées à travers le nouveau classificateur proposé, Eoℓ-SVM, fondé sur des ensembles de SVM exemplaire robustes, encore plus spécialisés et adaptés aux conditions problématiques rencontrées en FRiVS. Le modèle du classificateur proposé ainsi que ses méthodes sous-jacentes pourraient alors être réutilisés pour des cas similaires ou divers selon les besoins de futurs chercheurs en source de solutions aux mêmes genres de problématiques rencontrées durant ce projet. Les évaluations accomplies pourraient leur venir en aide pour rapidement déterminer de nouvelles voies d'explorations ou d'améliorations du domaine de la FR.

En plus des contributions précédentes, les évaluations sont accomplies en considérant plusieurs des modes d'opération possibles des systèmes de FRiVS (en transaction vs trajectoire, et surveillance vs identification). Or, en plus de permettre davantage d'opportunités d'application diverses des solutions introduites, cela pourrait inspirer plusieurs à considérer les différentes manières de percevoir les systèmes de FR. En effet, les résultats démontraient que certains cas étaient très performants, alors que d'autres non, selon une même variante testée. Ainsi, il se pourrait que de futurs lecteurs réalisent selon la démonstration appliquée et détaillée du rapport que ceux-ci aient omis certaines de ces considérations ou interprétations pour leurs travaux.

Finalement, afin de soulever encore l'importance de ce point, une contribution majeure est que le système de FRiVS proposé est une solution de bout-en-bout en combinant judicieusement plusieurs méthodes de l'état-de-l'art, sans simplification de la grande problématique de la FR. Ceci est donc un rappel à ceux qui s'aventureront dans ce domaine complexe que la problématique n'est pas dissociable, il s'agit d'un grand tout à considérer dans son ensemble.

RECOMMANDATIONS ET PERSPECTIVES

Basées sur les résultats de performance obtenus, les opérations de transformation géométriques synthétiques pour augmenter les données devraient être employées dans tous les cas, car elles viennent grandement combattre les difficultés associées au déséquilibre des classes et elles fournissent de la robustesse par rapport à la variation intra-classe des individus d'intérêt positifs simultanément. Pour la génération de représentations de visages supplémentaires, il semble que Eoℓ-SVM-DA soit plus efficace que Eoℓ-SVM-RSM, tant sur le point de la précision de la reconnaissance que la complexité du problème, ce qui fait que cette variante serait à prioriser. Par contre, l'utilisation de RSM augmente tout de même les performances, par rapport à l'application à elle seule de la génération synthétique des visages par transformations géométriques. Ainsi, cette méthode ne devrait pas être simplement rejetée.

Il est donc proposé, probablement pour le contexte de travaux futurs, de considérer une approche qui combine les deux meilleures variantes. Ainsi, en quelque sorte, une forme Eoℓ-SVM-DA-RSM qui accomplirait premièrement une adaptation de domaine selon la caméra spécifique sur laquelle chaque classificateur doit opérer respectivement puis, à partir des échantillons négatifs retirés de ces vidéos respectifs, d'effectuer une sélection aléatoire RSM pour générer davantage de variations pour représenter le visage et agrandir l'ensemble de SVM. Ceci n'a pas pu être testé faute de temps, mais semble être définitivement une approche prometteuse, en plus de ne nécessiter que très peu d'ajustement. Il s'agirait donc d'une recommandation à court terme rapide et facile à évaluer.

Ensuite, une des majeures considérations observées par rapport à l'application du système complet était que celui-ci était très grandement limité selon plusieurs points de vue par la détection de visages. En effet, le plus long délai de computation est relié à la triple détection VJ, et il demeure que plusieurs cas de faux-positifs ou *fail-to-acquire* survient. Ainsi, une recommandation plus à moyen terme serait de remplacer ce détecteur par une approche plus récente de FD, comme un des multiples CNN de détection d'objet présentés dans le rapport.

Par rapport aux recommandations à plus long terme, il serait intéressant d'appliquer des descripteurs en parallèle. Un candidat potentiel serait LPQ considérant qu'il gère le problème de flou de motion, qui est le seul cas auquel HOG semble demeurer variant. Cet ajustement serait normalement rapide à accomplir, mais le nombre limité d'implémentations existantes pour fonctionner dans un environnement optimisé en C++ fait que celui devrait probablement être réimplémenté, ce qui prend beaucoup de temps pour le valider. Aussi, toute l'architecture du système de FRiVS serait affectée étant donné l'ajout du descripteur supplémentaire, ce qui nécessiterait un certain temps pour l'adapter à cette modification.

Une autre recommandation à plus long terme serait d'explorer des alternatives aux SVM actuels, comme exploiter les approches de sélection dynamique des SVM les plus compétents, ainsi que l'utilisation de pondérations dynamiques, proposées dans [9, 10, 25], ou même, peut-être essayer d'appliquer un One-Class SVM (OC-SVM) [12]. Ceci requiert toutefois beaucoup de changements ainsi que d'évaluations afin de déterminer les performances atteignables lorsque ces approches sont combinées à un système de bout-en-bout avec suivi de trajectoire.

La dernière recommandation est plus à titre indicatif, et fait référence aux besoins de bases de données alternatives de plus grande envergure pour évaluer les systèmes sur une quantité très grande d'individus simultanément. Effectivement, plusieurs des systèmes formant l'état-de-l'art en FRiVS semblent avoir atteint un plateau, où les performances se jouent dans les fractions du dernier pourcentage avant la perfection. Or, beaucoup des résultats sont obtenus sur des bases de vidéos comme ChokePoint [139, 140] ou COX-S2V [64, 65], qui s'attribuent davantage aux environnements semi-contrôlés que sans-contraintes. Ainsi, à titre d'ouverture, il serait intéressant d'évaluer les méthodes proposées en littérature à travers des mises en application qui les exposeraient à des conditions dans des environnements encore plus sévères. Il faudrait alors voir comment ceux-ci performent avec des milliers d'individus enregistrés dans la *watchlist* en même temps, en employant des caméras avec de pires angles de poses et davantage d'occlusions. Sommairement, il serait intéressant de tester sur les conditions réelles auxquels les systèmes de FRiVS devront faire face, mais cela requiert d'effectuer une tâche colossale d'acquisition de données vidéo annotées pour le cas spécifique de S2V.

ANNEXE I

RÉSULTATS DÉTAILLÉS DES VARIANTES DE TEST ÉVALUÉES

Le Tableau A.1 présente toutes les combinaisons d'évaluations ayant fait objet de l'étude effectuée pendant ce projet. Une échelle comparative très simple (à droite) est employée afin de compter les occurrences de meilleures performances correspondantes, indiquées en gras, pour les combinaisons de test respectives et l'ensemble des séquences. L'échelle fait ressortir clairement qu'avec l'adaptation de domaine par caméra et l'ensemble des techniques de génération d'images synthétiques à partir de transformations géométriques, les meilleures performances sont atteintes. Un aperçu agrandi est aussi présenté au Tableau A.2 pour certaines des colonnes d'intérêt plus particulièrement employées dans le rapport.

Tableau A.1 Moyennes sur 5 réplications de l'ensemble de tous les résultats, pour toutes les combinaisons d'évaluation considérées du système de FRiVS, pour diverses combinaisons de séquences de ChokePoint, et comparées par performances correspondantes (les valeurs en gras représentent les performances optimales correspondantes)

Séquences de test		Caméras frontales				Toute la base				Séquence S1				Séquence S2				Séquence S3				Séquence S4				1st	SCORE	
Variation d'entraînement	Mesure de performance	Recognition		Identification		Recognition		Identification		Recognition		Identification		Recognition		Identification		Recognition		Identification		Recognition		Identification		2nd		
		AVG	SE	AVG	SE	AVG	SE	AVG	SE	AVG	SE	AVG	SE	AVG	SE	AVG	SE	AVG	SE	AVG	SE	AVG	SE	AVG	SE	3rd		
Système initial	pAUC(10%)	78.91%	±0.96%	96.75%	±2.60%	70.43%	±1.86%	65.20%	±1.28%	61.98%	±0.67%	77.26%	±1.90%	79.83%	±0.25%	91.98%	±2.73%	79.43%	±7.02%	90.18%	±0.22%	68.62%	±1.31%	88.96%	±6.35%	>0	0.00	0.00
	pAUC(20%)	86.75%	±0.66%	97.36%	±2.32%	77.30%	±0.31%	69.98%	±0.70%	68.39%	±1.33%	82.25%	±2.75%	84.74%	±2.97%	94.78%	±2.57%	85.88%	±4.69%	93.45%	±0.12%	76.04%	±0.02%	91.64%	±4.24%		0.00	
	AUC	95.23%	±1.73%	99.44%	±0.50%	91.39%	±1.13%	79.74%	±0.18%	85.54%	±2.37%	91.49%	±2.66%	94.99%	±2.13%	98.85%	±0.62%	96.39%	±0.93%	97.13%	±0.03%	90.40%	±1.35%	96.15%	±1.88%		0.00	
	AUPR	67.69%	±4.55%	97.71%	±1.57%	61.23%	±5.66%	61.30%	±0.09%	53.17%	±2.40%	82.02%	±1.89%	70.10%	±5.87%	94.02%	±0.79%	66.99%	±11.61%	91.02%	±1.10%	62.15%	±4.46%	91.88%	±4.34%		0.00	
ST_{nt}	pAUC(10%)	85.12%	±0.88%	94.71%	±3.60%	73.71%	±1.32%	89.32%	±0.66%	65.62%	±0.50%	82.01%	±3.20%	84.79%	±1.06%	95.67%	±1.21%	81.31%	±6.28%	92.49%	±2.10%	71.03%	±1.82%	88.44%	±7.33%		0.00	0.00
	pAUC(20%)	91.51%	±0.24%	96.06%	±3.09%	79.99%	±0.16%	92.12%	±0.01%	71.86%	±0.01%	85.68%	±4.21%	88.54%	±1.39%	97.69%	±0.75%	87.43%	±4.17%	95.41%	±1.89%	78.16%	±2.84%	90.61%	±5.57%		0.00	
	AUC	97.30%	±0.95%	99.20%	±0.63%	92.34%	±1.21%	96.61%	±0.56%	87.07%	±1.49%	93.34%	±4.08%	96.86%	±0.65%	99.54%	±0.15%	97.01%	±0.70%	98.34%	±1.12%	90.24%	±3.60%	95.96%	±3.16%		0.00	
	AUPR	77.81%	±3.07%	96.03%	±1.00%	65.56%	±4.69%	91.40%	±0.55%	59.49%	±2.19%	85.51%	±4.16%	76.41%	±6.96%	96.34%	±0.45%	72.31%	±9.96%	92.05%	±3.47%	63.68%	±1.17%	91.47%	±5.50%		0.00	
ST_{nt}+GT_{nt}	pAUC(10%)	85.46%	±0.89%	96.60%	±1.63%	74.64%	±1.91%	89.62%	±2.18%	65.03%	±0.50%	82.60%	±4.61%	84.16%	±1.64%	**97.73%**	±1.17%	81.25%	±6.89%	89.35%	±5.48%	75.05%	±0.02%	90.44%	±3.72%		1.00	**3.00**
	pAUC(20%)	91.32%	±1.53%	97.35%	±1.76%	81.12%	±0.56%	92.38%	±1.86%	71.70%	±0.62%	86.18%	±4.69%	88.12%	±3.16%	**98.86%**	±0.58%	88.27%	±4.33%	92.93%	±4.24%	82.05%	±1.70%	92.91%	±2.45%		1.00	
	AUC	96.31%	±2.26%	99.39%	±0.43%	92.99%	±1.01%	96.74%	±1.19%	86.79%	±0.94%	93.16%	±4.60%	96.13%	±1.65%	**99.77%**	±0.12%	97.33%	±0.68%	97.79%	±1.65%	93.23%	±2.70%	96.91%	±1.96%		1.00	
	AUPR	81.66%	±2.25%	96.76%	±0.02%	66.96%	±5.36%	91.15%	±3.00%	60.06%	±1.99%	85.21%	±5.83%	75.24%	±5.61%	96.96%	±2.26%	71.95%	±8.98%	88.88%	±6.95%	68.67%	±2.72%	92.74%	±2.33%		0.00	
GT_{wl}+ST_{nt}	pAUC(10%)	92.65%	±0.84%	96.66%	±3.34%	77.46%	±0.86%	91.77%	±0.31%	71.82%	±3.05%	83.60%	±0.15%	86.23%	±3.94%	97.22%	±0.06%	81.25%	±4.41%	92.95%	±3.50%	77.47%	±1.44%	93.60%	±0.52%		0.00	0.00
	pAUC(20%)	94.94%	±0.56%	98.33%	±1.67%	82.92%	±1.04%	93.84%	±0.79%	78.64%	±2.47%	87.01%	±1.23%	88.36%	±4.73%	98.61%	±0.03%	86.87%	±2.60%	94.86%	±3.37%	82.40%	±1.00%	95.59%	±0.47%		0.00	
	AUC	97.92%	±0.82%	99.67%	±0.33%	93.59%	±1.60%	97.76%	±0.70%	90.55%	±2.93%	95.13%	±1.50%	96.61%	±1.82%	99.72%	±0.01%	96.15%	±0.52%	98.03%	±1.61%	92.41%	±1.53%	98.97%	±0.25%		0.00	
	AUPR	89.68%	±0.19%	98.22%	±1.78%	72.35%	±0.79%	93.46%	±1.11%	66.56%	±2.23%	86.93%	±2.25%	81.40%	±2.02%	97.32%	±0.71%	77.11%	±5.88%	94.10%	±2.97%	73.68%	±1.27%	95.19%	±0.25%		0.00	
GT_{wl}+ST_{nt}+GT_{nt}	pAUC(10%)	93.52%	±0.53%	97.38%	±1.72%	78.57%	±0.58%	92.22%	±1.31%	74.87%	±2.48%	**87.08%**	±0.49%	86.56%	±3.95%	96.36%	±0.00%	81.26%	±5.11%	91.41%	±4.83%	77.36%	±2.69%	93.31%	±1.82%		1.00	**2.00**
	pAUC(20%)	95.51%	±0.26%	98.69%	±0.86%	83.96%	±1.07%	94.56%	±1.40%	81.59%	±1.95%	89.57%	±1.19%	89.72%	±3.60%	98.18%	±0.00%	86.90%	±3.75%	94.37%	±3.75%	81.83%	±3.67%	95.82%	±1.74%		0.00	
	AUC	98.04%	±0.88%	99.74%	±0.17%	93.79%	±1.75%	98.07%	±0.81%	92.34%	±1.67%	95.91%	±1.32%	96.83%	±1.60%	99.64%	±0.00%	96.28%	±0.32%	97.88%	±1.75%	90.73%	±3.64%	99.16%	±0.35%		0.00	
	AUPR	**89.95%**	±0.23%	98.25%	±0.75%	73.01%	±1.70%	93.79%	±1.79%	68.48%	±0.74%	89.16%	±2.20%	82.13%	±0.54%	96.65%	±1.27%	77.17%	±7.41%	93.09%	±4.04%	71.65%	±1.17%	95.46%	±1.03%		1.00	
3DMM	pAUC(10%)	70.55%	±5.60%	88.06%	±8.51%	64.60%	±1.70%	83.20%	±6.20%	57.25%	±1.56%	78.35%	±3.23%	71.23%	±1.57%	88.92%	±10.68%	71.18%	±0.54%	82.86%	±7.21%	66.33%	±0.99%	87.61%	±5.28%		0.00	0.00
	pAUC(20%)	78.05%	±2.82%	91.14%	±7.14%	72.30%	±0.93%	87.49%	±4.52%	64.63%	±0.40%	84.18%	±1.23%	79.13%	±0.23%	91.14%	±8.67%	79.51%	±1.26%	87.29%	±5.74%	71.59%	±0.55%	90.82%	±3.56%		0.00	
	AUC	91.64%	±1.55%	97.38%	±2.27%	89.50%	±0.72%	95.67%	±0.86%	85.01%	±1.49%	93.47%	±1.03%	93.83%	±0.45%	97.41%	±2.55%	92.95%	±0.18%	96.22%	±2.02%	88.53%	±1.42%	96.70%	±0.78%		0.00	
	AUPR	62.15%	±3.17%	91.73%	±4.96%	54.04%	±2.63%	86.22%	±4.34%	51.86%	±2.02%	81.78%	±1.07%	57.95%	±5.20%	91.54%	±7.60%	58.38%	±0.12%	85.81%	±5.90%	58.90%	±1.13%	90.36%	±2.97%		0.00	
3DMM + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(10%)	86.64%	±2.24%	93.67%	±0.28%	80.03%	±1.48%	90.11%	±0.96%	69.50%	±0.25%	84.06%	±5.86%	88.89%	±0.06%	95.09%	±2.75%	85.50%	±2.08%	90.83%	±0.37%	80.39%	±0.81%	88.96%	±0.35%		0.00	0.00
	pAUC(20%)	90.29%	±0.59%	95.54%	±0.24%	85.04%	±1.10%	93.01%	±0.73%	75.89%	±0.81%	88.36%	±2.91%	92.66%	±0.17%	97.46%	±1.45%	90.76%	±1.30%	94.28%	±0.65%	84.99%	±0.18%	91.43%	±1.33%		0.00	
	AUC	96.02%	±1.93%	98.97%	±0.18%	94.29%	±0.62%	97.59%	±0.50%	90.31%	±1.65%	95.00%	±1.30%	97.55%	±0.55%	99.49%	±0.29%	96.99%	±0.03%	98.86%	±0.13%	93.98%	±0.96%	97.18%	±1.13%		0.00	
	AUPR	83.11%	±3.46%	95.75%	±0.36%	75.02%	±1.39%	92.55%	±0.05%	64.66%	±4.00%	86.68%	±2.82%	85.31%	±1.50%	96.19%	±2.48%	78.69%	±1.49%	93.60%	±1.11%	77.55%	±2.75%	92.24%	±0.03%		0.00	
$RSM_{(20,128)}$	pAUC(10%)	92.05%	±0.14%	98.43%	±0.88%	78.08%	±0.07%	92.32%	±1.27%	72.28%	±0.19%	84.71%	±3.59%	85.87%	±3.11%	97.60%	±0.37%	82.15%	±4.74%	**96.76%**	±0.06%	79.75%	±1.00%	92.73%	±3.91%		1.00	**7.00**
	pAUC(20%)	94.15%	±0.70%	99.22%	±0.44%	83.53%	±0.86%	94.52%	±1.19%	78.52%	±0.79%	87.10%	±4.11%	89.67%	±2.82%	98.80%	±0.19%	87.51%	±3.38%	**98.38%**	±0.03%	84.30%	±1.05%	95.46%	±2.86%		1.00	
	AUC	96.71%	±2.01%	99.84%	±0.09%	93.39%	±2.12%	98.07%	±0.77%	89.99%	±2.87%	95.09%	±2.59%	96.77%	±1.39%	99.76%	±0.04%	96.35%	±0.42%	**99.68%**	±0.01%	91.90%	±3.31%	99.09%	±0.57%		1.00	
	AUPR	89.66%	±1.49%	96.87%	±2.16%	77.06%	±1.86%	95.47%	±1.70%	70.34%	±1.18%	90.79%	±3.61%	83.23%	±1.78%	**97.82%**	±1.19%	**81.48%**	±6.94%	**95.89%**	±2.77%	76.93%	±3.04%	**95.84%**	±2.89%		4.00	
$RSM_{(20,128)}$ + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(10%)	91.79%	±1.58%	**99.41%**	±0.23%	76.91%	±1.21%	91.65%	±0.18%	71.07%	±0.45%	85.07%	±1.04%	78.70%	±8.08%	91.51%	±7.49%	71.66%	±1.04%	88.47%	±4.44%	72.93%	±2.31%	86.23%	±2.21%		1.00	**5.00**
	pAUC(20%)	94.40%	±1.03%	**99.70%**	±0.11%	82.61%	±0.09%	94.09%	±0.24%	77.95%	±0.59%	88.62%	±0.60%	84.37%	±7.01%	94.36%	±5.14%	78.35%	±0.99%	92.54%	±3.32%	79.95%	±2.59%	90.46%	±1.24%		1.00	
	AUC	96.44%	±2.03%	**99.94%**	±0.02%	92.07%	±2.64%	98.00%	±0.05%	90.69%	±3.15%	96.22%	±1.30%	92.84%	±5.30%	98.71%	±1.19%	90.65%	±3.11%	98.35%	±0.83%	90.87%	±3.33%	97.93%	±0.41%		1.00	
	AUPR	89.09%	±3.42%	**98.32%**	±1.28%	75.84%	±2.87%	95.22%	±0.81%	69.46%	±1.13%	**91.24%**	±1.96%	76.50%	±5.91%	95.46%	±2.26%	70.77%	±0.17%	93.35%	±0.15%	69.72%	±0.87%	92.85%	±0.36%		2.00	
DA + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(10%)	**95.26%**	±0.92%	93.49%	±1.37%	**86.62%**	±1.60%	**92.40%**	±0.50%	**82.80%**	±6.06%	86.15%	±2.51%	**93.19%**	±2.90%	95.95%	±1.96%	**87.40%**	±2.69%	95.72%	±1.91%	**84.28%**	±0.51%	**93.91%**	±0.03%		8.00	**31.00**
	pAUC(20%)	**97.00%**	±1.09%	96.75%	±0.69%	**90.48%**	±1.87%	**95.45%**	±0.05%	**87.47%**	±5.47%	**91.59%**	±1.18%	**96.16%**	±1.80%	97.97%	±0.98%	**91.56%**	±1.96%	96.90%	±1.91%	**87.63%**	±1.89%	**96.95%**	±0.02%		9.00	
	AUC	**98.45%**	±1.17%	99.35%	±0.14%	**96.54%**	±1.31%	**98.58%**	±0.23%	**94.55%**	±2.95%	**97.38%**	±0.24%	**99.23%**	±0.36%	99.59%	±0.20%	**97.95%**	±0.21%	98.38%	±1.38%	**94.94%**	±1.91%	**99.39%**	±0.00%		9.00	
	AUPR	88.94%	±1.20%	92.95%	±2.61%	**85.87%**	±1.53%	**95.92%**	±0.40%	**75.84%**	±5.62%	87.51%	±0.72%	**87.76%**	±2.41%	95.48%	±1.56%	79.52%	±2.46%	92.64%	±3.00%	**78.40%**	±0.44%	94.18%	±0.99%		5.00	

Tableau A.2 Agrandissement des colonnes de séquences « Caméras frontales » et « Toute la base » du Tableau A.1 pour les évaluations du système de FRiVS avec échelle comparative

Séquences de test		Caméras frontales				Toute la base				SCORE	
Variation d'entraînement	Mesure de performance	Recognition AVG	Recognition SE	Identification AVG	Identification SE	Recognition AVG	Recognition SE	Identification AVG	Identification SE		
Système initial	pAUC(10%)	78.91%	±0.96%	96.75%	±2.60%	70.43%	±1.86%	65.20%	±1.28%	0.00	0.00
	pAUC(20%)	86.75%	±0.66%	97.36%	±2.32%	77.30%	±0.31%	69.98%	±0.70%	0.00	
	AUC	95.23%	±1.73%	99.44%	±0.50%	91.39%	±1.13%	79.74%	±0.18%	0.00	
	AUPR	67.69%	±4.55%	97.71%	±1.57%	61.23%	±5.66%	61.30%	±0.09%	0.00	
ST_{nt}	pAUC(10%)	85.12%	±0.88%	94.71%	±3.60%	73.71%	±1.32%	89.32%	±0.66%	0.00	0.00
	pAUC(20%)	91.51%	±0.24%	96.06%	±3.09%	79.99%	±0.16%	92.12%	±0.01%	0.00	
	AUC	97.30%	±0.95%	99.20%	±0.63%	92.34%	±1.21%	96.61%	±0.56%	0.00	
	AUPR	77.81%	±3.07%	96.03%	±1.00%	65.56%	±4.69%	91.40%	±0.55%	0.00	
ST_{nt}+GT_{nt}	pAUC(10%)	85.46%	±0.89%	96.60%	±1.63%	74.64%	±1.91%	89.62%	±2.18%	1.00	**3.00**
	pAUC(20%)	91.32%	±1.53%	97.35%	±1.76%	81.12%	±0.56%	92.38%	±1.86%	1.00	
	AUC	96.31%	±2.26%	99.39%	±0.43%	92.99%	±1.01%	96.74%	±1.19%	1.00	
	AUPR	81.66%	±2.25%	96.76%	±0.02%	66.96%	±5.36%	91.15%	±3.00%	0.00	
GT_{wl}+ST_{nt}	pAUC(10%)	92.65%	±0.84%	96.66%	±3.34%	77.46%	±0.86%	91.77%	±0.31%	0.00	0.00
	pAUC(20%)	94.94%	±0.56%	98.33%	±1.67%	82.92%	±1.04%	93.84%	±0.79%	0.00	
	AUC	97.92%	±0.82%	99.67%	±0.33%	93.59%	±1.60%	97.76%	±0.70%	0.00	
	AUPR	89.68%	±0.19%	98.22%	±1.78%	72.35%	±0.79%	93.46%	±1.11%	0.00	
GT_{wl}+ST_{nt}+GT_{nt}	pAUC(10%)	93.52%	±0.53%	97.38%	±1.72%	78.57%	±0.58%	92.22%	±1.31%	1.00	**2.00**
	pAUC(20%)	95.51%	±0.26%	98.69%	±0.86%	83.96%	±1.07%	94.56%	±1.40%	0.00	
	AUC	98.04%	±0.88%	99.74%	±0.17%	93.79%	±1.75%	98.07%	±0.81%	0.00	
	AUPR	**89.95%**	±0.23%	98.25%	±0.75%	73.01%	±1.70%	93.79%	±1.79%	1.00	
3DMM	pAUC(10%)	70.55%	±5.60%	88.06%	±8.51%	64.60%	±1.70%	83.20%	±6.20%	0.00	0.00
	pAUC(20%)	78.05%	±2.82%	91.14%	±7.14%	72.30%	±0.93%	87.49%	±4.52%	0.00	
	AUC	91.64%	±1.55%	97.38%	±2.27%	89.50%	±0.72%	95.67%	±0.86%	0.00	
	AUPR	62.15%	±3.17%	91.73%	±4.96%	54.04%	±2.63%	86.22%	±4.34%	0.00	
3DMM + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(10%)	86.64%	±2.24%	93.67%	±0.28%	80.03%	±1.48%	90.11%	±0.96%	0.00	0.00
	pAUC(20%)	90.29%	±0.59%	95.54%	±0.24%	85.04%	±1.10%	93.01%	±0.73%	0.00	
	AUC	96.02%	±1.93%	98.97%	±0.18%	94.29%	±0.62%	97.59%	±0.50%	0.00	
	AUPR	83.11%	±3.46%	95.75%	±0.36%	75.02%	±1.39%	92.55%	±0.05%	0.00	
$RSM_{(20,128)}$	pAUC(10%)	92.05%	±0.14%	98.43%	±0.88%	78.08%	±0.07%	92.32%	±1.27%	1.00	**7.00**
	pAUC(20%)	94.15%	±0.70%	99.22%	±0.44%	83.53%	±0.86%	94.52%	±1.19%	1.00	
	AUC	96.71%	±2.01%	99.84%	±0.09%	93.39%	±2.12%	98.07%	±0.77%	1.00	
	AUPR	89.66%	±1.49%	96.87%	±2.16%	77.06%	±1.86%	95.47%	±1.70%	4.00	
$RSM_{(20,128)}$ + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(10%)	91.79%	±1.58%	**99.41%**	±0.23%	76.91%	±1.21%	91.65%	±0.18%	1.00	**5.00**
	pAUC(20%)	94.40%	±1.03%	**99.70%**	±0.11%	82.61%	±0.09%	94.09%	±0.24%	1.00	
	AUC	96.44%	±2.03%	**99.94%**	±0.02%	92.07%	±2.64%	98.00%	±0.05%	1.00	
	AUPR	89.09%	±3.42%	**98.32%**	±1.28%	75.84%	±2.87%	95.22%	±0.81%	2.00	
DA + GT_{wl}+ST_{nt}+GT_{nt}	pAUC(10%)	**95.26%**	±0.92%	93.49%	±1.37%	**86.62%**	±1.60%	**92.40%**	±0.50%	8.00	**31.00**
	pAUC(20%)	**97.00%**	±1.09%	96.75%	±0.69%	**90.48%**	±1.87%	**95.45%**	±0.05%	9.00	
	AUC	**98.45%**	±1.17%	99.35%	±0.14%	**96.54%**	±1.31%	**98.58%**	±0.23%	9.00	
	AUPR	88.94%	±1.20%	92.95%	±2.61%	**85.87%**	±1.53%	**95.92%**	±0.40%	5.00	

Légende de scores: **1er rang** | **2e rang** | **3e rang** | **plus grand que zéro**

ANNEXE II

RÉPARTITION DES INDIVIDUS POUR LES RÉPLICATIONS DE TEST

Le Tableau A.3 démontre la répartition des individus de la base ChokePoint à travers les réplications de tests et les diverses catégories, tel que décrit à la section 5.1.2.

Tableau A.3 Segmentation de la base ChokePoint pour l'évaluation du système de FRiVS

Type d'échantillons	**Mode des échantillons**	**#**	**ID de réplications de test dans ChokePoint**				
			1	**2**	**3**	**4**	**5**
$[N_{wl} = 5]$ **Individus d'intérêt enregistrés au système (*watchlist targets*)**	Phase d'entraînement (images uniques) Phase d'opération (vidéos)	1	ID0003	ID0010	ID0001	ID0011	ID0004
		2	ID0005	ID0011	ID0003	ID0012	ID0016
		3	ID0006	ID0015	ID0013	ID0017	ID0020
		4	ID0010	ID0025	ID0018	ID0019	ID0021
		5	ID0024	ID0026	ID0019	ID0024	ID0023
$[N_{nt} = 10]$ **Individus de contre-exemples négatifs (*unknown non-targets*)**	Phase d'entraînement (vidéos)	1	ID0001	ID0001	ID0004	ID0006	ID0003
		2	ID0002	ID0004	ID0007	ID0007	ID0005
		3	ID0007	ID0005	ID0009	ID0013	ID0006
		4	ID0009	ID0012	ID0011	ID0014	ID0009
		5	ID0011	ID0014	ID0016	ID0020	ID0010
		6	ID0013	ID0016	ID0021	ID0021	ID0011
		7	ID0014	ID0018	ID0024	ID0022	ID0017
		8	ID0016	ID0019	ID0026	ID0023	ID0022
		9	ID0017	ID0023	ID0027	ID0026	ID0027
		10	ID0018	ID0027	ID0030	ID0027	ID0029
$[N_{ut} = 14]$ (reste) **Individus inconnus de test *(unknown test probes)***	Phase d'opération (vidéos)	1	ID0004	ID0002	ID0002	ID0001	ID0001
		2	ID0012	ID0003	ID0005	ID0002	ID0002
		3	ID0015	ID0006	ID0006	ID0003	ID0007
		4	ID0019	ID0007	ID0010	ID0004	ID0012
		5	ID0020	ID0009	ID0012	ID0005	ID0013
		6	ID0021	ID0013	ID0014	ID0009	ID0014
		7	ID0022	ID0017	ID0015	ID0010	ID0015
		8	ID0023	ID0020	ID0017	ID0015	ID0018
		9	ID0025	ID0021	ID0020	ID0016	ID0019
		10	ID0026	ID0022	ID0022	ID0018	ID0024
		11	ID0027	ID0024	ID0023	ID0025	ID0025
		12	ID0028	ID0028	ID0025	ID0028	ID0026
		13	ID0029	ID0029	ID0028	ID0029	ID0028
		14	ID0030	ID0030	ID0029	ID0030	ID0030

ANNEXE III

RÉSULTATS D'ANALYSE EMPIRIQUE DE CERTAINS DESCRIPTEURS POUR LA RECONNAISSANCE DE VISAGES

L'évaluation empirique est basée sur une étude de trois descripteurs potentiels pour la caractérisation de visages employés en reconnaissance de visages. Les descripteurs sont LBP_{59}, HOG_{144} et CNN_{259} avec pour indice la dimensionnalité des vecteurs générés. Les deux premières approches sont les méthodes traditionnelles (sans variation spécifique) et l'extraction par CNN est basée sur [102] afin de faire du *feature learning* et dont le descripteur est formé des valeurs obtenues d'une couche intermédiaire à la sortie du réseau. Le classificateur est un SVM à kernel linéaire traditionnel dans tous les cas, entraîné avec 100 individus positifs contre 100 négatifs, et testés contre 200 autres, tous provenant de visages frontaux de la base d'images Color FERET [89, 105, 106].

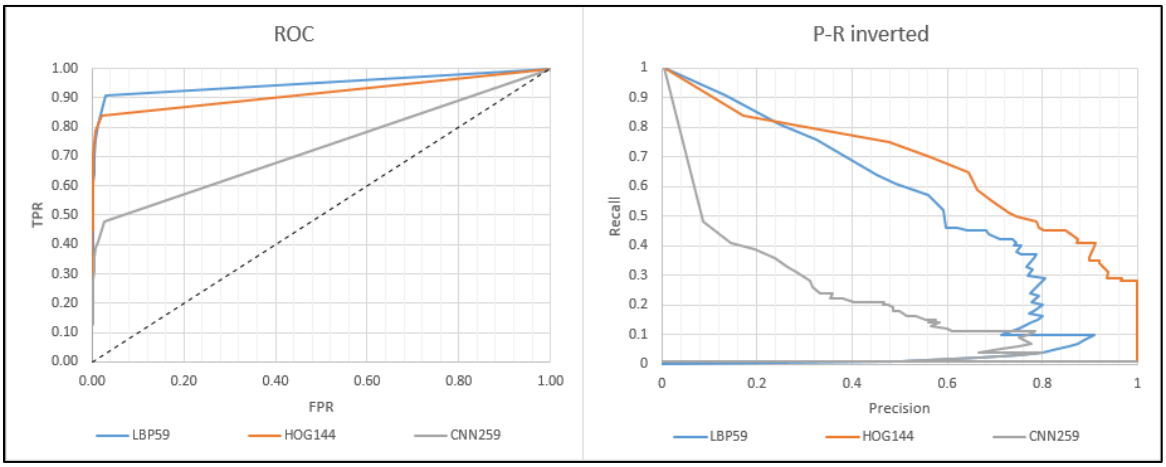

Figure A.1 Courbes ROC et PR en fonction des descripteurs LBP_{59}, HOG_{144} et CNN_{259}

La Figure A.1 indique que LBP et HOG ont des performances assez proches en termes de prédiction des TP par rapport aux FP, mais la prédiction en fonction du *recall* donne un net avantage vers l'approche HOG. Ceci est une considération importante étant donné le déséquilibre majeur entre les positifs et négatifs habituellement observé en SSPP. Ainsi, pour

les cas lorsque les classes ne seront pas 100P/100N, l'impact sera encore plus marqué. L'approche CNN semble ici inadéquate dans tous les cas.

La Figure A.2 démontre que, pour une reconnaissance de visages à partir de différentes dimensions de ROI, les performances varient légèrement pour le descripteur LBP, mais changent brusquement pour HOG. Les ROI englobant les visages employés pour l'entraînement des classificateurs sont obtenues à partir des multiples réponses de Viola-Jones lorsque la suppression des détections redondantes est désactivée.

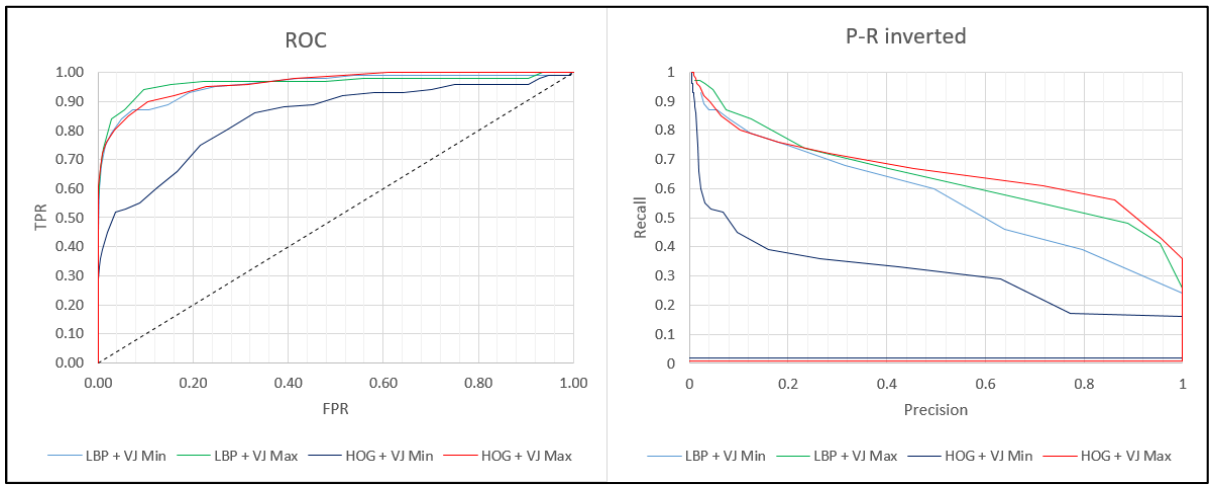

Figure A.2 Courbes ROC et PR en fonction des descripteurs LBP_{59} et HOG_{144} appliqués aux ROI de visages minimales et maximales obtenues par Viola-Jones

Ainsi, malgré que HOG offre de hauts potentiels de reconnaissance de visages à partir de la Figure A.1 et de la Figure A.2, une détection de ROI constantes à travers les visages afin de normaliser la position des caractéristiques discriminantes semble être un prérequis.

ANNEXE IV

RÉSULTATS DE PRÉDICTION DE SVM TRADITIONNELS

Les SVM traditionnels à *kernel* linéaire et sortie binaire, entraînés à partir d'un unique échantillon de référence de l'individu d'intérêt, contre une multitude de contre-exemples. Les performances évaluées sur Color FERET [106], premièrement, pour des images frontales et ensuite pour des angles de pose variés, semblent démontrer que la méthode n'est pas suffisante à elle seule pour le SSPP. D'autant plus, Color FERET correspond à des images en environnement contrôlé avec des conditions d'acquisition très similaires entre les différents individus et visages observés, ce qui indique que les performances seraient probablement encore pires dans un contexte d'entraînement S2V qui serait ensuite testé contre un environnement affichant des conditions d'acquisition nuisibles très sévères.

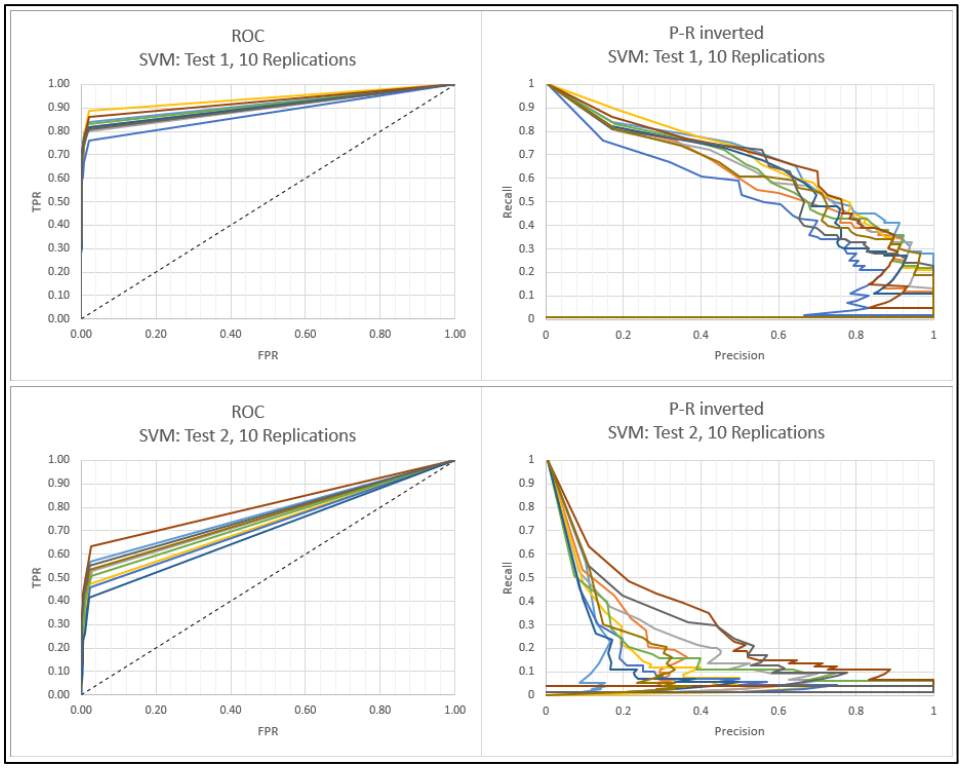

Figure A.3 Courbes ROC et PR en fonction de visages frontaux (haut) et de visages avec poses (bas), à l'aide de SVM à kernel linéaire entraînés et testés sur la base de visages Color FERET, pour 10 réplications de test avec 100 individus entraînés et testés contre 200 autres

ANNEXE V

RÉSULTATS DE PRÉDICTION DE DIFFÉRENTS ENSEMBLES DE CLASSIFICATEURS POUR DIVERS INDIVIDUS DANS LA WATCHLIST

Le Tableau A.4 présente les résultats de FR par *target* individuellement, avec leur moyenne et erreur standard correspondantes, et par combinaison de prédiction selon la confidence maximale (voir la section 5.2.2 pour les détails), à travers les multiples paramètres d'évaluation de performance.

Tableau A.4 Résultats de prédiction de 5 individus d'intérêt enregistrés dans la *watchlist* à l'aide du système de base (*baseline*) présenté au Chapitre 3, et testé sur ChokePoint

Baseline		MaxConf	ID0003	ID0005	ID0006	ID0010	ID0024	AVG(ID)	SE(ID)	
Frontal	pAUC(10%)	94.2%	86.5%	74.8%	51.3%	86.7%	100.0%	79.9%	±8.18%	79.9%
	pAUC(20%)	95.0%	87.0%	81.2%	75.5%	93.4%	100.0%	87.4%	±4.33%	87.4%
	AUC	98.9%	87.5%	86.3%	95.1%	98.7%	100.0%	93.5%	±2.83%	93.5%
	AUPR	96.1%	85.7%	60.2%	41.9%	73.5%	100.0%	72.2%	±10.05%	72.2%
ALL	pAUC(10%)	85.1%	59.6%	84.9%	56.4%	66.6%	94.0%	72.3%	±7.35%	72.3%
	pAUC(20%)	89.5%	66.6%	88.7%	65.3%	71.4%	96.0%	77.6%	±6.22%	77.6%
	AUC	95.9%	84.6%	93.0%	85.4%	89.1%	99.2%	90.3%	±2.69%	90.3%
	AUPR	89.2%	58.8%	78.3%	49.6%	55.9%	91.8%	66.9%	±7.86%	66.9%
S1	pAUC(10%)	79.2%	39.1%	76.6%	53.0%	43.6%	94.2%	61.3%	±10.46%	61.3%
	pAUC(20%)	85.0%	48.2%	79.2%	61.0%	49.8%	97.1%	67.1%	±9.33%	67.1%
	AUC	94.2%	76.5%	84.5%	79.4%	76.1%	99.4%	83.2%	±4.33%	83.2%
	AUPR	83.9%	41.5%	74.1%	50.3%	24.3%	87.7%	55.6%	±11.36%	55.6%
S2	pAUC(10%)	89.2%	71.9%	85.5%	57.9%	84.6%	98.0%	79.6%	±6.81%	79.6%
	pAUC(20%)	92.2%	75.8%	87.8%	59.0%	87.3%	99.0%	81.8%	±6.78%	81.8%
	AUC	98.2%	91.8%	90.9%	84.6%	97.1%	99.8%	92.9%	±2.64%	92.9%
	AUPR	93.2%	73.7%	76.7%	57.2%	76.5%	95.7%	76.0%	±6.11%	76.0%
S3	pAUC(10%)	90.4%	76.0%	96.8%	72.0%	87.9%	99.5%	86.5%	±5.47%	86.5%
	pAUC(20%)	93.6%	81.1%	98.4%	83.0%	90.5%	99.8%	90.6%	±3.82%	90.6%
	AUC	97.2%	92.2%	99.7%	96.6%	98.1%	100.0%	97.3%	±1.41%	97.3%
	AUPR	92.1%	71.5%	90.7%	45.3%	87.4%	98.1%	78.6%	±9.40%	78.6%
S4	pAUC(10%)	82.6%	54.0%	87.6%	58.4%	60.4%	89.3%	69.9%	±7.63%	69.9%
	pAUC(20%)	87.4%	63.3%	92.7%	68.1%	65.2%	91.0%	76.1%	±6.50%	76.1%
	AUC	94.3%	78.5%	98.5%	81.2%	88.8%	98.2%	89.1%	±4.16%	89.1%
	AUPR	87.5%	51.9%	86.6%	55.7%	50.1%	88.7%	66.6%	±8.64%	66.6%
AVG(S#)	pAUC(10%)	85.4%	60.3%	86.6%	60.3%	69.1%	95.3%			
	pAUC(20%)	89.5%	67.1%	89.5%	67.8%	73.2%	96.7%			
	AUC	96.0%	84.8%	93.4%	85.5%	90.0%	99.3%			
	AUPR	89.2%	59.7%	82.0%	52.1%	59.6%	92.6%			
SE(S#)	pAUC(10%)	±2.68%	±8.51%	±4.15%	±4.07%	±10.48%	±2.28%			
	pAUC(20%)	±2.01%	±7.32%	±4.07%	±5.45%	±9.62%	±1.99%			
	AUC	±1.03%	±4.22%	±3.56%	±3.87%	±5.10%	±0.40%			
	AUPR	±2.15%	±7.79%	±3.95%	±2.73%	±14.13%	±2.56%			

Les résultats du Tableau A.4 proviennent du système de base avec Eo*e*-SVM entraîné avec la seule image de référence des individus, et utilisant le suivi de trajectoire, sans l'application d'aucune des améliorations présentées au Chapitre 4. Il est intéressant d'observer dans ce tableau que les prédictions respectives par individu sont extrêmement différentes en termes de précision, alors que les ensembles de classificateurs sont testés sur les mêmes séquences vidéo. Par exemple, l'individu ID0024 est généralement bien reconnu à travers l'ensemble de la base de données ChokePoint. Par contre, l'individu ID0006 est à l'inverse très peu souvent reconnu correctement. Les trois autres individus sont quant à eux reconnus plus ou moins bien, tout dépendamment de la séquence d'apparition vidéo spécifique sur laquelle ils sont évalués.

La grande variabilité des prédictions de FR présentée au Tableau A.4 justifie les besoins d'amélioration du système initial qui sont explorés au Chapitre 4, afin d'atteindre des prédictions robustes indépendamment des séquences et individus particulièrement évalués.

BIBLIOGRAPHIE

[1] *Visual Object Tracking Challenge, VOT Challenge* [Online]. Available: http://www.votchallenge.net/

[2] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28(12), pp. 2037-2041, 2006.

[3] T. Ahonen, J. Matas, C. He, and M. Pietikäinen, "Rotation invariant image description with local binary pattern histogram fourier features," in *Scandinavian Conference on Image Analysis*, 2009, pp. 61-70.

[4] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila, "Recognition of blurred faces using local phase quantization," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1-4.

[5] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol, "Face recognition using HOG–EBGM," *Pattern Recognition Letters*, vol. 29(10), pp. 1537-1543, 2008.

[6] I. Amara, E. Granger, and A. Hadid, "On the Effects of Illumination Normalization with LBP-Based Watchlist Screening," in *Computer Vision-ECCV 2014 Workshops*, 2014, pp. 173-188.

[7] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3444-3451.

[8] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video: A review," Article, *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26(5), p. 53, Aug 2012.

[9] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, "Dynamic Ensembles of Exemplar-SVMs for Still-to-Video Face Recognition," *Pattern Recognition,* 2017-02-12 2017.

[10] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, "Dynamic Selection of Exemplar-SVMs for Watch-List Screening Through Domain Adaptation," presented at the International Conference on Pattern Recognition Applications and Methods (ICPRAM), Porto, Portugal, 2017.

[11] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, "Ensembles of exemplar-SVMs for video face recognition from a single sample per person," in *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, 2015, pp. 1-6.

[12] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, "Robust watch-list screening using dynamic ensembles of SVMs based on multiple face representations," *Machine Vision and Applications*, pp. 1-23, 2017.

[13] S. Bashbaghi, E. Granger, R. Sabourin, and G. A. Bilodeau, "Watch-List Screening Using Ensembles Based on Multiple Face Representations," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 2014, pp. 4489-4494.

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

[15] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, "On the use of SIFT features for face authentication," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 2006, pp. 35-35.

[16] C. M. Bishop, "Pattern Recognition," *Machine Learning*, 2006.

[17] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25(9), pp. 1063-1074, 2003.

[18] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187-194.

[19] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. (2013). *Guide to Biometrics*. [Online] Available: https://books.google.ca/books?id=DLLbBwAAQBAJ

[20] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "The Relation Between the ROC Curve and the CMC," in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, 2005, pp. 15-20.

[21] D. Bradley. Haar classifier for profile faces (Contribution of "haarcascade_profileface.xml" to OpenCV), [Online]. Available: https://github.com/opencv/opencv/wiki/Contributors.

[22] G. Bradski. The OpenCV Library (The OpenCV Library), [Online]. Available: https://github.com/opencv/opencv.

[23] G. R. Bradski, "Computer Vision Face Tracking for Use in a Perceptual User Interface," 1998.

[24] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust Tracking-by-Detection using a Detector Confidence Particle Filter," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 1515-1522.

[25] A. S. Britto, R. Sabourin, and L. E. Oliveira, "Dynamic selection of classifiers—A comprehensive review," *Pattern Recognition,* vol. 47(11), pp. 3665-3680, 2014.

[26] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* (10), pp. 1042-1052, 1993.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2(3), p. 27, 2011.

[28] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "An evaluation of multimodal 2D+3D face biometrics," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27(4), pp. 619-624, 2005.

[29] N. V. Chawla and K. W. Bowyer, "Random Subspaces And Subsampling For 2-D Face Recognition," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 582-589.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of artificial intelligence research,* vol. 16 pp. 321-357, 2002.

[31] T. Chen, W. Yin, Z. Xiang Sean, D. Comaniciu, and T. S. Huang, "Illumination Normalization for Face Recognition and Uneven Background Correction Using Total Variation Based Image Models," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 532-539 vol. 2.

[32] L. Chi, H. Zhang, and M. Chen, "End-To-End Face Detection and Recognition," *arXiv preprint arXiv:1703.10818,* 2017.

[33] J.-T. Chien and C.-P. Liao, "Maximum Confidence Hidden Markov Modeling for Face Recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 30(4), pp. 606-616, 2008.

[34] F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal, "Online Multi-Face Detection and Tracking using Detector Confidence and Structured SVMs."

[35] F. Comaschi, S. Stuijk, T. Basten, and H. Corporaal, "Robust Online Face Tracking-by-Detection," in *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, 2016, pp. 1-6.

[36] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886-893.

[37] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233-240.

[38] M. De-la-Torre, E. Granger, P. Radtke, R. Sabourin, and D. O. Gorodnichy, "Self-updating with facial trajectories for video-to-video face recognition," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 2014, pp. 1669-1674.

[39] M. De-la-Torre, E. Granger, P. V. W. Radtke, R. Sabourin, and D. O. Gorodnichy, "Partially-supervised learning from facial trajectories for face recognition in video surveillance," *Information Fusion,* vol. 24 pp. 31-53, 2015/07/01/ 2015.

[40] P. F. De Carrera and I. Marques, "Face Recognition Algorithms," *Master's thesis in Computer Science, Universidad Euskal Herriko,* 2010.

[41] M. De Marsico, *Face Recognition in Adverse Conditions*. Hershey: Information Science Reference, 2014.

[42] L. Demidova and I. Klyueva, "Improving the Classification Quality of the SVM Classifier for the Imbalanced Datasets on the Base of Ideas the SMOTE Algorithm," in *ITM Web of Conferences*, 2017, p. 02002.

[43] H. Demirel, T. J. Clarke, and P. Y. Cheung, "Adaptive automatic facial feature segmentation," in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, 1996, pp. 277-282.

[44] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using Histograms of Oriented Gradients," *Pattern Recognition Letters,* vol. 32(12), pp. 1598-1603, 2011.

[45] M. A. A. Dewan, E. Granger, G. L. Marcialis, R. Sabourin, and F. Roli, "Adaptive appearance model tracking for still-to-video face recognition," *Pattern Recognition,* vol. 49 pp. 129-151, 2016/01/01/ 2016.

[46] M. A. A. Dewan, E. Granger, F. Roli, R. Sabourin, and G. Marcialis, "A Comparison of Adaptive Appearance Methods for Tracking Faces in Video Surveillance," 2013.

[47] M. A. A. Dewan, E. Granger, R. Sabourin, G. L. Marcialis, and F. Roli, "Video Face Recognition From A Single Still Image Using an Adaptive Appearance Model Tracker," in *2015 IEEE Symposium Series on Computational Intelligence*, 2015, pp. 196-202.

[48] M. A. A. Dewan, D. Qiao, F. Lin, and D. Wen, "An Approach to Improving Single Sample Face Recognition Using High Confident Tracking Trajectories," in *Canadian Conference on Artificial Intelligence*, 2016, pp. 115-121.

[49] C. Ding and D. Tao, "Trunk-Branch Ensemble Convolutional Neural Networks for Video-based Face Recognition," *arXiv preprint arXiv:1607.05427,* 2016.

[50] L. E. Dodd and M. S. Pepe, "Partial AUC Estimation and Regression," *Biometrics,* vol. 59(3), pp. 614-623, 2003.

[51] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*: John Wiley & Sons, 2012.

[52] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *The Journal of Machine Learning Research,* vol. 9 pp. 1871-1874, 2008.

[53] M. R. Faraji and X. Qi, "Face recognition under varying illumination based on adaptive homomorphic eight local directional patterns," *IET Computer Vision,* vol. 9(3), pp. 390-399, 2014.

[54] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters,* vol. 27(8), pp. 861-874, 2006.

[55] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440-1448.

[56] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[57] R. Gopalan, S. Taheri, P. Turaga, and R. Chellappa, "A Blur-Robust Descriptor with Applications to Face Recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 34(6), pp. 1220-1226, 2012.

[58] M. Grgic, K. Delac, and S. Grgic, "SCface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51(3), pp. 863-879, 2011.

[59] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* vol. 143(1), pp. 29-36, 1982.

[60] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured Output Tracking with Kernels."

[61] D. Held, S. Thrun, and S. Savarese, "Learning to Track at 100 FPS with Deep Regression Networks," in *European Conference Computer Vision (ECCV)*, 2016.

[62] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.

[63] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst2007.

[64] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, *et al.*, "A benchmark and comparative study of video-based face recognition on COX face database," *IEEE Transactions on Image Processing,* vol. 24(12), pp. 5967-5981, 2015.

[65] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "Benchmarking Still-to-Video Face Recognition via Partial and Local Linear Discriminant Analysis on COX-S2V Dataset," in *Computer Vision – ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part II*, K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 589-600.

[66] S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *British machive vision conference*, 2012, p. 11 pages.

[67] A. K. Jain, R. P. Duin, and J. Mao, "Statistical pattern recognition: A review," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 22(1), pp. 4-37, 2000.

[68] H. Jiang and E. Learned-Miller, "Face Detection with the Faster R-CNN," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, 2017, pp. 650-657.

[69] Z. Jun, Y. Yong, and M. Lades, "Face Recognition: Eigenface, Elastic Matching, and Neural Nets," *Proceedings of the IEEE,* vol. 85(9), pp. 1423-1435, 1997.

[70] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection."

[71] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Toward development of a face recognition system for watchlist surveillance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 33(10), pp. 1925-1937, 2011.

[72] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 12(1), pp. 103-108, 1990.

[73] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin, *et al.*, "The Visual Object Tracking VOT2016 Challenge Results," ed: Springer Nature.

[74] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernández, *et al.*, "The Visual Object Tracking VOT2015 Challenge Results," in *Visual Object Tracking, Single Object Tracking Dataset*, 2015, pp. 1-23.

[75] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, *et al.*, "The Visual Object Tracking VOT2014 Challenge Results," in *Computer Vision - ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, L. Agapito, M. M. Bronstein, and C. Rother, Eds., ed Cham: Springer International Publishing, 2015, pp. 191-217.

[76] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, *et al.*, "The Visual Object Tracking VOT2013 Challenge Results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 98-111.

[77] B. Leibe, K. Schindler, and L. Van Gool, "Coupled Detection and Trajectory Estimation for Multi-Object Tracking," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1-8.

[78] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A Convolutional Neural Network Cascade for Face Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5325-5334.

[79] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of Exemplar-SVMs for Object Detection and Beyond," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 89-96.

[80] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Computer Vision–ECCV 2014*, ed: Springer, 2014, pp. 720-735.

[81] D. K. McClish, "Analyzing a portion of the ROC curve," *Medical Decision Making*, vol. 9(3), pp. 190-195, 1989.

[82] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[83] I. Misra, A. Shrivastava, and M. Hebert, "Data-driven Exemplar Model Selection," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, 2014, pp. 339-346.

[84] T. Mita, T. Kaneko, and O. Hori, "Joint Haar-like features for face detection," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 2005, pp. 1619-1626 Vol. 2.

[85] F. Mokhayeri, E. Granger, and G.-A. Bilodeau, Synthetic Face Generation for Still-to-Video Face Recognition (Synthetic Face Generation for Still-to-Video Face Recognition). IEEE, 2015, unpublished.

[86] F. Mokhayeri, E. Granger, and G.-A. Bilodeau, "Synthetic face generation under various operational conditions in video surveillance," in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 4052-4056.

[87] J. Mu, F. Chen, and C. Wang, "Research on Emotion Recognition for Facial Expression Images Based on Hidden Markov Model," *International Journal of Modern Trends in Engineering and Research*, 2014.

[88] M. Murtaza, M. Sharif, M. Raza, and J. Shah, "Face recognition using adaptive margin fisher's criterion and linear discriminant analysis," *International Arab Journal of Information Technology*, vol. 11(2), pp. 1-11, 2014.

[89] National Institute of Standards and Technology. (2003) The Color FERET Database [Online]. Available: http://www.itl.nist.gov/iad/humanid/colorferet/home.html

[90] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with CUDA," *Queue*, vol. 6(2), pp. 40-53, 2008.

[91] F. Nourbakhsh, E. Granger, and G. Fumera, "An Extended Sparse Classification Framework for Domain Adaptation in Video Surveillance," in *Asian Conference on Computer Vision*, 2016, pp. 360-376.

[92] NVIDIA CUDA. Compute unified device architecture programming guide (Compute unified device architecture programming guide), [Online].

[93] NVIDIA CUDA. Programming guide (Programming guide), [Online]. Available: http://docs.nvidia.com/cuda/index.html.

[94] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24(7), pp. 971-987, 2002.

[95] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International conference on image and signal processing*, 2008, pp. 236-243.

[96] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe, "A Boosted Particle Filter: Multitarget Detection and Tracking," in *European conference on computer vision*, 2004, pp. 28-39.

[97] A. OpenMP. OpenMP application program interface, v. 3.0 (OpenMP application program interface, v. 3.0), [Online]. Available: http://www.openmp.org/.

[98] C. Pagano, E. Granger, R. Sabourin, A. Rattani, G.-L. Marcialis, and F. Roli, "Efficient adaptive face recognition systems based on capture conditions," in *Computational Intelligence in Biometrics and Identity Management (CIBIM), 2014 IEEE Symposium on*, 2014, pp. 60-67.

[99] M. Parchami, S. Bashbaghi, and E. Granger, "CNNs with Cross-Correlation Matching for Face Recognition in Video Surveillance Using a Single Training Sample Per Person," presented at the Advanced Video and Signal based Surveillance, Lecce, Italy, 2017.

[100] M. Parchami, S. Bashbaghi, and É. Granger, "Video-Based Face Recognition Using Ensemble of Haar-Like Deep Convolutional Neural Networks," presented at the International Joint Conference on Neural Networks (IJCNN'2017), Anchorage, Alaska, USA, 2017.

[101] M. Parchami, S. Bashbaghi, E. Granger, and S. Iftekar Sayed, "Using Deep Autoencoders to Learn Robust Domain-Invariant Representations for Still-to-Video Face Recognition," presented at the Advanced Video and Signal based Surveillance, Lecce, Italy, 2017.

[102] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," *Proceedings of the British Machine Vision,* vol. 1(3), p. 6, 2015.

[103] O. M. Parkhi, A. Vedaldi, and A. Zisserman. (2015, Accessed: 2016-07-27). *VGG Face Descriptor, VGG Face Descriptor* [Online]. Available: http://www.robots.ox.ac.uk/~vgg/software/vgg_face/

[104] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, *et al.*, "Overview of the face recognition grand challenge," in *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, 2005, pp. 947-954.

[105] P. J. Phillips, H. Moon, S. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 22(10), pp. 1090-1104, 2000.

[106] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing,* vol. 16(5), pp. 295-306, 4/27/ 1998.

[107] S. Puttemans, E. Can, and T. Goedemé, "Improving Open Source Face Detection by Combining an Adapted Cascade Classification Pipeline and Active Learning," in *Proceedings of VISAPP 2017*, 2017.

[108] E. Rahtu, J. Heikkilä, V. Ojansivu, and T. Ahonen, "Local phase quantization for blur-insensitive image analysis," *Image and Vision Computing,* vol. 30(8), pp. 501-512, 2012.

[109] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779-788.

[110] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *arXiv preprint arXiv:1612.08242,* 2016.

[111] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in neural information processing systems*, 2015, pp. 91-99.

[112] F. Roli and G. L. Marcialis, "Semi-Supervised PCA-based Face Recognition Using Self-Training," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 2006, pp. 560-568.

[113] S. Romdhani, J. Ho, T. Vetter, and D. J. Kriegman, "Face Recognition Using 3-D Models: Pose and Illumination," *Proceedings of the IEEE,* vol. 94(11), pp. 1977-1999, 2006.

[114] A. Rosebrock. (2016, Accessed: 2017-12-05). *Intersection over Union (IoU) for object detection, Intersection over Union (IoU) for object detection* [Online]. Available: https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

[115] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable Model Fitting by Regularized Landmark Mean-Shift," *International Journal of Computer Vision,* vol. 91(2), pp. 200-215, 2011.

[116] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815-823.

[117] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features off-the-shelf: an Astounding Baseline for Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops,* 2014, pp. 806-813.

[118] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Josa a,* vol. 4(3), pp. 519-524, 1987.

[119] S. W. Smith, "Circular Buffering," in *The Scientist and Engineer's Guide to Digital Signal Processing*. vol. The Scientist and Engineer's Guide to Digital Signal Processing Chapter 28: Digital Signal Processors, ed California, USA: Newnes, 1997, [Online]. Available: http://www.dspguide.com/ch28/2.htm

[120] R. Soleymani, E. Granger, and G. Fumera, "Progressive Boosting for Class Imbalance," *arXiv preprint arXiv:1706.01531,* 2017.

[121] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701-1708.

[122] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern recognition,* vol. 39(9), pp. 1725-1745, 2006.

[123] A. T. Tran, T. Hassner, I. Masi, and G. Medioni, "Regressing Robust and Discriminative 3D Morphable Models with a very Deep Neural Network," *arXiv preprint arXiv:1612.04904,* 2016.

[124] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience,* vol. 3(1), pp. 71-86, 1991.

[125] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, 1991, pp. 586-591.

[126] C. N. Vasconcelos and B. N. Vasconcelos, "Increasing Deep Learning Melanoma Classification by Classical And Expert Knowledge Based Image Transforms," *arXiv preprint arXiv:1702.07025*, 2017.

[127] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the Sensitivity of Support Vector Machines," in *Proceedings of the international joint conference on AI*, 1999, pp. 55-60.

[128] T. Vetter and T. Poggio, "Linear Object Classes and Image Synthesis From a Single Example Image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19(7), pp. 733-742, 1997.

[129] P. Viola and M. Jones, "Rapid Object Detection Using a BoostedCascade of Simple Features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, pp. I-I.

[130] P. Viola and M. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57(2), pp. 137-154, 2004/05/01 2004.

[131] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24(2), pp. 137-154, 1997.

[132] S. Walter, "The partial area under the summary ROC curve," *Statistics in medicine*, vol. 24(13), pp. 2025-2040, 2005.

[133] H. Wang, S. Z. Li, Y. Wang, and W. Zhang, "Illumination Modeling and Normalization for Face Recognition," in *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, 2003, pp. 104-111.

[134] J. Wang and L. Perez, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning."

[135] J. Wang, K. N. Plataniotis, and A. N. Venetsanopoulos, "Selecting discriminant eigenfaces for face recognition," *Pattern Recognition Letters*, vol. 26(10), pp. 1470-1482, 2005.

[136] L. Wolf, T. Hassner, and I. Maoz, "Face Recognition in Unconstrained Videos with Matched Background Similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 529-534.

[137] L. Wolf, T. Hassner, and I. Maoz. (2003) YouTube Faces Database [Online]. Available: https://www.cs.tau.ac.il/~wolf/ytfaces/

[138] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?," in *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, 2016, pp. 1-6.

[139] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. (2011) ChokePoint Dataset [Online]. Available: http://arma.sourceforge.net/chokepoint/

[140] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, 2011, pp. 74-81.

[141] M. Yang, L. Van Gool, and L. Zhang, "Sparse Variation Dictionary Learning for Face Recognition with A Single Training Sample Per Person," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 689-696.

[142] X. Yang, Q. Song, and Y. Wang, "A weighted support vector machine for data classification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21(05), pp. 961-976, 2007.

[143] L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data," *Knowledge-Based Systems*, vol. 94 pp. 88-104, 2016.

[144] W. A. Yousef, "Assessing classifiers in terms of the partial area under the ROC curve," *Computational Statistics & Data Analysis*, vol. 64 pp. 51-70, 8// 2013.

[145] K. Zhang, L. Zhang, and M.-H. Yang, "Real-Time Compressive Tracking," in *European conference on computer vision*, 2012, pp. 864-877.

[146] X. Zhang and Y. Gao, "Face Recognition across Pose: A Review," *Pattern Recognition*, vol. 42(11), pp. 2876-2896, 2009.

[147] Y. Zhang and D. Wang, "A Cost-Sensitive Ensemble Method for Class-Imbalanced Datasets," in *Abstract and applied analysis*, 2013.

[148] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29(6), pp. 915-928, 2007.

[149] G. Zhao and M. Pietikäinen, "Local binary pattern descriptors for dynamic texture recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, pp. 211-214.

[150] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35(4), pp. 399-458, 2003.

[151] G. Zhenhua, D. Zhang, and D. Zhang, "A Completed Modeling of Local Binary Pattern Operator for Texture Classification," *Image Processing, IEEE Transactions on*, vol. 19(6), pp. 1657-1663, 2010.

[152] F. J. Zhong and J. S. Zhang, "Face recognition with enhanced local directional patterns," *Neurocomputing*, vol. 119 pp. 375-384, Nov 7 2013.

[153] Y. Zhong, J. Chen, and B. Huang, "Towards End-to-End Face Recognition through Alignment Learning," *arXiv preprint arXiv:1701.07174*, 2017.