Analyse faciale dans les flux vidéo

Ioan Marius Bilasco

HABILITATION À DIRIGER DES RECHERCHES

delivrée par l'Université de Lille

Discipline : Informatique

Analyse faciale dans les flux vidéo

Ioan Marius BILASCO

8 février 2019

Jury

Mme Jenny Benois-Pineau	Professeur - Université de Bordeaux	Rapporteur
M. Farzin Deravi	Professeur - Kent University	Rapporteur
M. Mohammed Bennamoun	Professeur - West Australia University	Rapporteur
Mme Alice Caplier	Professeur - Grenoble-INP	Examinateur
M. Atilla Baskurt	Professeur - INSA Lyon	Examinateur
Mme Laurence Duchien	Professeur - Université de Lille	Examinateur
M. Chaabane Djeraba	Professeur - Université de Lille	Garant

Résumé

Les flux vidéo sont omniprésents dans notre quotidien. L'exploitation efficace de l'information visuelle contenue dans ces flux laisse présager l'avènement de systèmes intelligents accompagnant l'activité humaine au quotidien.

Dans ce contexte général, nos travaux de recherche s'articulent autour de la compréhension des comportements humains dans un environnement personnel sur la base de l'analyse vidéo. En particulier, nous nous intéressons à certains aspects de l'analyse faciale tels que l'orientation de la tête, la reconnaissance du genre et la reconnaissance des expressions. L'ensemble de ces aspects permet de développer des applications qui dressent de manière précise le profil, les intérêts ou les besoins de l'utilisateur.

Plus spécifiquement, nous exposons des approches non-intrusives pour estimer l'orientation de la tête d'une personne en s'appuyant sur la symétrie du visage et sur l'identification des transformations permettant de ramener un visage dans une situation quasi-frontale. Nous développons aussi des techniques de normalisation géométrique et photométrique pour reconnaître le genre de la personne en partant de descripteurs simples tels que les intensités des pixels. Les attributs intra- et extra-faciaux tels que la chevelure, la barbe ou la moustache viennent alimenter un système d'inférence floue pour la reconnaissance du genre. Enfin, nous nous intéressons à l'étude unifiée de micro- et macro-expressions. Nous introduisons la notion de motifs locaux de mouvements afin de répondre au large panel d'intensités d'expression que l'on peut rencontrer dans un cadre d'interaction naturelle.

Mots-clés : Vision par ordinateur, Analyse faciale, Estimation de l'orientation de la tête, Symétrie bilatérale, Reconnaissance du genre, Normalisation faciale, Spécification d'histogrammes, Systèmes d'inférence floue, Reconnaissance des expressions, Micro expression, Macro expression, Flux optique, Mouvement cohérent, Motifs locaux de mouvement, Modèle de segmentation faciale.

Abstract

Video streams are ubiquitous in our daily lives. The effective use of the visual information embedded in video streams forsees the advent of intelligents systems for assisting humans in their daily life activities. In this general context, our research addresses human behavior understanding in an individual setting based on video analysis. More specifically, we explore aspects related to facial video analysis such as : head pose estimation, gender recognition and expression recognition. All these aspects together sustain the developpement of intelligent systems having access to accurate user profiles in terms of characteristics, interests and needs.

In this work, we present non-intrusive approaches for estimating head pose by using the characteristics of facial symmetry and by identifying transformations that can bring the face in a near-frontal configuration. We also employ geometric and photometric normalisation techniques for gender recognition using simple features like pixel intesities. The intra- and extra-facial features like hair, beard or mustache are used jointly with the normalized pixel intensities within a fuzzy inference system to make gender recognition more effective. Finally, we propose a unified approach for micro- and macro-expression recognition. We introduce local motion patterns in order to deal with a large panel of expression intensities as one can meet in uncontrolled settings.

Keywords : Computer vision, Facial analysis, Head orientation, Bilateral symmetry, Gender recognition, Facial normalisation, Histogram specification, Fuzzy inference system, Expression recognition, Micro expression, Macro expression, Optical flow, Coherent movement, Local motion patterns, Facial segmentation models.

Je tiens à remercier :

Madame Jenny Benois-Pineau (Professeur à l'université de Bordeaux), Monsieur Farzin Deravi (Professeur à l'université de Kent), Monsieur Mohammed Bennamoun (Professeur à l'université de West Australia) qui m'ont fait l'honneur de rapporter mon travail et dont les remarques m'ont permis d'en améliorer certains aspects.

Madame Alice Caplier (Professeur à Grenoble-INPG), Monsieur Atilla Baskurt (Professeur à l'INSA de Lyon) et Madame Laurence Duchien (Professeur à l'université de Lille) qui ont accepté d'examiner mon travail.

Monsieur Chaabane Djeraba (Professeur à l'université de Lille), mon garant, pour son encadrement, ses nombreux conseils et son soutien. La confiance qu'il m'a accordée depuis mes débuts à Lille, m'a permis de m'épanouir en tant que chercheur et de mener à bon port les différents travaux et co-encadrements qui abouttissent aujourd'hui à l'écriture de ce manuscrit.

Monsieur Jean Martinet, Pierre Tirilly, Maîtres de Conférence à l'université de Lille, José Mennesson, Maître Assistant à l'IMT Lille Douai, mes collègues de l'équipe FOX, pour les nombreuses discussions qui ont beaucoup contribué à l'avancement des travaux que nous avons l'occasion de co-encadrer.

Madame Afifa Dahmane (Maître Assistant à l'université de Science et Technologie Houari-Boumédiène, Algéria), Monsieur Samir Amir (Directeur de recherche à Press'Innov), Benjamin Allaert (Ingénieur de recherche à l'université de Lille), Monsieur Taner Danisman (Enseignant-chercheur à l'université d'Akendiz, Turquie), Monsieur Yassine Kazi-Tani (Maître Assistant à l'Ecole Supérieure en Informatique Sidi-Bel-Abbes, Algérie) qui ont su me challenger et qui ont su accepter mes nombreux suggestions, remarques et critiques émises au cours de leur thèse. C'est grâce à l'investissement de ces anciens doctorants que j'ai co-encadrés, que les résultats que je vous présente aujourd'hui ont été rendus possibles.

Madame Delphine Poux, Monsieur Pierre Falez, Monsieur Romain Belmonte, Monsieur Cagan Arlsan pour leur tenacité et leur envie de poursuivre leurs travaux de thèse, en partie à mes côtés.

Tous ceux qui m'ont soutenu et encouragé tout au long de ces années passées à Lille et je pense notamment à mes collègues de l'équipe FOX, du groupe IMAGE, de l'IRCICA et du département Informatique de la faculté de Sciences et Technologies de l'université de Lille.

Mes anciens encadrants de thèse de l'université Joseph Fourier : Hervé Martin, Marlène Villanova et Jérôme Gensel qui ont su ouvrir mon appétit pour la recherche et ont rendu possible le démarrage de ma carrière.

Mes parents, mes grands parents, mon frère qui malgré leur éloignement géographique ont toujours su être près de moi. Ma belle famille qui depuis longtemps déjà m'offre beaucoup d'amour et de réconfort.

En dernier, je tiens à remercier tout particulièrement ma femme, Céline, et mes enfants, Olivia et Roman qui ont su m'offrir toute l'envie, la compréhension et l'amour dont j'ai eu besoin pour mener à bien ce travail.

Sommaire

Première partie

Introduction

1 Résumé de mon parcours recherche

Depuis ma nomination au poste de Maître de Conférences en septembre 2009, je me suis consacré principalement au domaine de la vision par ordinateur sur des thématiques portant sur l'analyse faciale. En particulier, je me suis focalisé sur l'estimation de l'orientation de la tête, la caractérisation du genre et la reconnaissance des expressions faciales. Parallèlement, j'ai eu l'occasion de poursuivre des travaux en relation avec ma thèse sur les métadonnées et la sémantique. En effet, lorsque l'on conçoit un système intelligent capable de réagir en adéquation avec les événements ou les comportements identifiés, on rencontre des problèmes liés à l'intégration des informations issues de sources hétérogènes (analyse vidéo ou autres). Ainsi, il est nécessaire de disposer de méthodologies et d'outils qui facilitent l'interopérabilité et l'intégration de données semi-structurées issues de différents systèmes.

Je donne aujourd'hui de nouvelles orientations à mes recherches autour de l'analyse faciale en m'intéressant à l'apprentissage évolué (en profondeur et neuromorphique). Ces travaux récents ont été rendus possibles grâce aux interactions avec l'équipe Embedded Real-Time Adaptative System Design (Émeraude) du Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) et avec l'équipe Circuits, Systèmes et Applications des Micro-ondes (CSAM) de l'Institut d'Électronique, de Microélectronique et Nanotechnologies (IEMN) de Lille, toutes hébergées au sein de l'Institut de Recherche sur les Composants logiciels et matériels pour l'Information et la Communication Avancée (IRCICA, USR CNRS 3380).

La Figure I.1 illustre la structuration temporelle de mes activités de recherche à travers les travaux de thèses et les post-doctorants ou assimilés que j'ai pu co-encadrer.

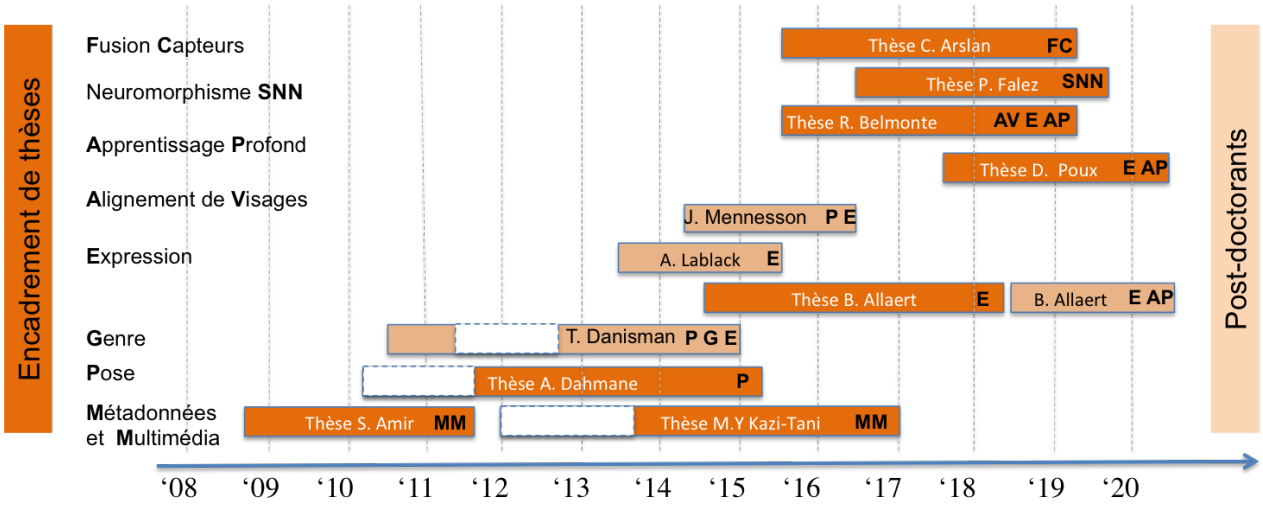


FIGURE I.1 – *Co-encadrements de doctorants (8) et de post-docs ou assimilés (3) selon les principaux axes de recherche.*

Après ma thèse soutenue dans le domaine de la gestion de métadonnées et de documents, j'ai intégré l'équipe Fouille de dOnnées CompleXes (FOX), fondée par Chaabane Djeraba (PR, lab. CRIStAL, Université de Lille) en 2004. Mes travaux ont débuté avec l'encadrement de la thèse de Sa-

mir Amir [1] qui porte sur la mise en correspondance d'ontologies. Graduellement, j'ai commencé à diriger mes recherches vers l'analyse faciale dans le cadre des activités d'estimation de l'orientation de la tête, notamment en co-encadrant les travaux d'Afifa Dahmane [2] et coordonnant ceux de Taner Danisman [3] et de José Mennesson [4]. La collaboration avec Taner Danisman s'est poursuivie avec des travaux autour de la reconnaissance du genre et des expressions à partir d'images statiques. Dans ce domaine, j'ai également coordonné les travaux d'Adel Lablack [5]. J'ai ensuite souhaité donner une nouvelle orientation aux travaux autour de l'expression en m'intéressant, dans la thèse de Benjamin Allaert [6], à la reconnaissance des expressions dans un contexte dynamique. Avec l'arrivée de l'apprentissage profond, l'équipe a décidé d'étudier l'impact de ces nouveaux outils pour l'analyse faciale. Ainsi, j'ai pu participer à l'encadrement des travaux de thèse de Romain Belmonte [7] qui étudie l'alignement facial en se basant sur l'apprentissage profond. Dans ce même contexte, j'ai initié les travaux de la thèse de Delphine Poux [8] sur l'étude de l'impact des occultations sur les expressions faciales en prenant le parti de privilégier les approches d'apprentissage profond. En parallèle de ces travaux liés à l'apprentissage profond, je m'intéresse également à d'autres formes d'apprentissage évoluées dans un contexte de vision, notamment aux réseaux de neurones à spikes, dans le cadre de la thèse de Pierre Falez [9]. Au-delà des travaux évoqués ci-dessus qui constituent le coeur de mes activités de recherche, j'ai pu également participer aux co-encadrements de la thèse de Cagan Arlsan [10] et de celle de Mohamed Yassine Kazi Tani [11]. La thèse de Cagan Arlsan s'intéresse à la fusion de capteurs dans un contexte d'interaction. Celle de Mohamed Yassine Kazi Tani explore l'utilisation de raisonnements sémantiques dans un contexte de vision pour la caractérisation de mouvements de foule.

1. Samir Amir, doctorant de septembre 2008 à décembre 2011, équipe FOX, lab. CRIStAL, actuellement chef d'équipe R&D à Press'Inov, Lyon.
2. Afifa Dahmane, doctorante de mai 2010 à février 2015, équipe FOX, lab. CRIStAL en co-tutelle avec l'Université de Sciences et Technologies Houari Boumediene (USTHB), Algérie, actuellement Maître de Conférences à l'USTHB, Algérie.
3. Taner Danisman, post-doctorant septembre 2010 à juin 2011 et ingénieur de recherche de juin 2012 à octobre 2014, équipe FOX, lab. CRIStAL, actuellement enseignant-chercheur à Akdeniz Üniversitesi, Turquie.
4. José Mennesson, ingénieur de recherche de novembre 2014 à décembre 2015, équipe FOX, lab. CRIStAL, actuellement Maître Assistant à l'Institut Mines-Telecom Lille Douai, équipe FOX, lab. CRIStAL.
5. Adel Lablack, ingénieur de recherche de juin 2013 à août 2015, équipe FOX, lab. CRIStAL, actuellement ingénieur R&D chez FLIR, Courtrai, Belgique.
6. Benjamin Allaert, doctorant de octobre 2014 à juin 2018, équipe FOX, lab. CRIStAL, actuellement ingénieur de recherche équipe FOX, lab. CRIStAL.
7. Romain Belmonte, doctorant depuis octobre 2015, équipe FOX, lab. CRIStAL cofinancement par Ecole d'ingénieurs des Hautes Technologies et du Numérique - ISEN YNCREA et la Métropole Européenne de Lille.
8. Delphine Poux, doctorante depuis octobre 2015, équipe FOX, lab. CRIStAL cofinancement par l'Ecole d'ingénieurs des Hautes Technologies et du Numérique - ISEN YNCREA.
9. Pierre Falez, doctorant depuis octobre 2016, équipe FOX, lab. CRIStAL en co-direction avec l'équipe Émeraude, lab. CRIStAL.
10. Cagan Arslan, doctorant depuis octobre 2015, équipe FOX, lab. CRIStAL en co-direction avec l'équipe MINT, lab. CRIStAL.
11. MY. Kazi Tani, doctorant de janvier 2012 à mars 2018, inscrit à l'Université d'Oran Es Sénia, Algérie et accueilli dans l'équipe FOX, lab. CRIStAL de septembre à octobre 2013, d'avril à mai 2014 et en novembre 2014, actuellement Maître Assistant à l'École Supérieure en Informatique 8 mai 1945 - Sidi Bel Abbes, Algérie.

2 Thèmes de recherche

Dans ce qui suit, je donne un bref aperçu de l'ensemble des activités évoquées précédemment en listant ci-dessous les résultats notables et les encadrements et collaborations que ces activités ont occasionnés :

Estimation de l'orientation de la tête Afin d'apporter des solutions aux problèmes de détection de l'orientation de la tête, dans le cadre de la thèse d'Afifa Dahmane, nous avons proposé une approche basée sur la symétrie du visage (Dahmane et al. 2015) [12]. D'autres travaux sur le même sujet ont été menés en collaboration avec les post-doctorants Taner Danisman et José Mennesson et valorisés dans (Danisman et Bilasco 2016) [13] et (Mennesson et al. 2016).

Reconnaissance du genre En poursuivant la collaboration avec Taner Danisman (post-doctorant, équipe FOX, lab. CRIStAL), nous nous sommes intéressés à la biométrie douce. La biométrie douce permet de différencier les individus sur la base de leurs traits caractéristiques (par exemple, la couleur des yeux, la forme du visage, le genre), sans toutefois les identifier précisément. Nous avons exploré la reconnaissance du genre des personnes en proposant une méthode globale fonctionnant dans des conditions (taille d'images très petite, faible illumination, etc.) difficilement abordables avec des méthodes classiques. Les travaux menés ont souligné l'importance du processus de normalisation du visage avant traitement lorsque la pose n'est pas contrainte. Des résultats intéressants ont pu être obtenus dans un contexte de validation croisée entre différentes bases de données (Danisman et al. 2014). Toutefois, la caractérisation du genre en se basant uniquement sur les éléments du visage ne peut pas répondre à la variété des situations que nous pouvons rencontrer dans la vie courante. Certains individus ont des traits de visage plus proches du sexe opposé que du leur. Ainsi, nous avons poursuivi ces travaux en les enrichissant avec des informations caractérisant à la fois, des éléments précis à l'intérieur du visage (moustache, barbe) et des éléments entourant le visage (comme les cheveux) (Danisman et al. 2015) [14].

Reconnaissance des expressions En partant du constat que la normalisation du visage permet d'améliorer les processus de caractérisation faciale, nous avons exploré la reconnaissance de la joie, en nous appuyant sur l'approche de normalisation mise en place pour la reconnaissance du genre. Ces travaux ont permis de construire des masques de pixels sur les visages, capables d'améliorer la reconnaissance de l'expression de joie à partir de l'intensité de pixels sans passer par des descripteurs complexes (Danisman et al. 2013) [15]. Les travaux conduits sur des données statiques (images

12. A. Dahmane ; S. Larabi ; I.M. Bilasco ; C. Djeraba - Head pose estimation based on face symmetry analysis - Signal, Image and Video Processing (SIViP), Springer, 2015, 9 (8), pp 1871-1880 (Facteur d'impact : 1,643 selon Journal Citations Reports 2018).

13. T. Danisman ; I.M. Bilasco - In-plane face orientation estimation in still images - Multimedia Tools and Applications, Springer Verlag, 2016, 75 (13), pp.7799-7829 (Facteur d'impact : 1,541 selon JCR 2018).

14. T. Danisman ; I.M. Bilasco ; J. Martinet - Boosting gender recognition performance with a fuzzy inference system - Expert Systems with Applications, Volume 42, Issue 5, 1 April 2015, pp. 2772-2784 (Facteur d'impact : 3,768 selon JCR 2018).

15. T. Danisman ; I.M. Bilasco ; J. Martinet ; C. Djeraba - Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron - Signal Processing, Elsevier, 2013, Special issue on Machine Learning in Intelligent Image Processing, 93 (6), pp. 1547-1556 (Facteur d'impact : 3,470 selon JCR 2017).

ou trames d'une vidéo) ont été poursuivis par les recherches menées dans le cadre de la thèse de doctorat de Benjamin Allaert. Ces travaux visent à caractériser dans le cadre d'une interaction naturelle l'état émotionnel des individus. L'interaction naturelle complexifie beaucoup le processus de reconnaissance car l'intensité des expressions peut varier (de micro- à macro-expressions). De plus, le mouvement global de la personne influe fortement sur les mouvements intra-faciaux qui matérialisent l'apparition et la disparition des expressions. En s'intéressant au défi posé par la variabilité de l'intensité, nous étudions une approche tenant compte des caractéristiques physiques de mouvements faciaux, permettant ainsi de filtrer le bruit sans nuire à l'information extraite (Allaert et al. 2017). Afin de contourner le problème relatif aux mouvements de tête, la plupart des méthodes proposent d'employer des techniques de normalisation du visage, mais cette normalisation induit des déformations de texture qui interfèrent avec les expressions. Dans (Allaert et al. 2018b) [16], nous avons proposé une méthodologie et un corpus de données, Synchronous Natural and Posed 2D Facial Expressions (SNaP-2DFE), qui permettent de quantifier l'impact de la normalisation sur la reconnaissance des expressions faciales. La ligne directrice confortée par les travaux initiés avec Benjamin Allaert se poursuit avec la thèse de Delphine Poux qui a démarré en octobre 2017. Les premiers résultats de Delphine Poux montrent que l'exploitation de l'information du mouvement trouve également sa place dans la reconnaissance des expressions en présence d'occultations dans (Poux et al. 2018).

Au-delà de ces travaux fortement liés au domaine de l'analyse faciale, j'ai pu, d'une part, poursuivre les travaux menés dans le cadre de mon doctorat, et d'autre part construire de nouvelles collaborations autour :

— des métadonnées et sémantique ;
— des approches profondes pour l'analyse faciale ;
— des approches neuromorphiques pour l'analyse faciale.

Les travaux relatifs à ces thématiques sont brièvement introduits ci-dessous et ne sont pas détaillés dans ce manuscrit. Les travaux liés à l'apprentissage profond et neuromorphique sont toutefois évoqués plus largement dans la Partie III de ce document lors de la présentation de mon futur projet scientifique.

Métadonnées et sémantique Dans un processus classique de vision par ordinateur, nous partons des données brutes, caractérisées par des descripteurs. Ensuite, les descripteurs alimentent des processus implicites (classification et apprentissage) ou explicites (systèmes de règles) de construction de nouveaux descripteurs. Dans la plupart des réalisations évoquées jusqu'à présent, nous nous appuyons sur des processus implicites. Dans le cadre des travaux de Mohamed Yassine Kazi Tani j'ai suggéré d'aborder les apports des raisonneurs sémantiques dans la résolution de certains problèmes de vision, à l'instar du suivi de personnes ou de groupes de personnes dans (Kazi-Tani et al. 2017). Ensemble, nous avons proposé une ontologie ainsi que des règles spécifiques qui permettent de déduire de l'évolution des blobs en mouvement (individu ou groupe) leur nature et leur comportement. Au-delà des problématiques propres à la vision, de par les activités menées pendant ma thèse et à l'issue de celle-ci, j'ai travaillé dans le domaine de l'information sémantique

16. B. Allaert ; J. Mennesson ; I.M. Bilasco ; C. Djeraba - Impact of the face registration techniques on facial expressions recognition - Signal Processing : Image Communication, EURASIP, Elsevier, 2017, 61, pp. 44-53 (Facteur d'impact : 2,073 selon Journal Citations Reports 2018).

extraite depuis les données vers des processus intelligents capables d'enrichir l'expérience utilisateur. Lorsque ces informations sont échangées et traitées par des tiers, du fait qu'elles sont issues de processus différents, il est nécessaire de disposer d'approches capables d'intégrer des données semi-structurées de nature variée (XML, RDF). Ainsi, dans les travaux de thèse de Samir Amir, nous avons abordé les problèmes d'intégration de données et proposé un cadre formel qui permet d'aligner des informations issues de sources hétérogènes dans (Amir et al. 2011) et (Amir et al. 2013).

Approches profondes pour l'analyse faciale L'interprétation d'un flux vidéo capté dans des environnements peu contrôlés doit tenir compte et s'adapter aux évolutions du contexte visuel (changement de luminosité, perte de netteté, etc.) et du comportement de l'individu (changements de pose, occultations, etc.). Ces changements influencent grandement les divers processus sous-jacents d'analyse. La prise en compte, a priori, de ces évolutions permettrait d'améliorer le processus d'analyse en choisissant les techniques les plus adaptées à une situation donnée. La caractérisation de l'environnement peut se révéler tout aussi compliquée que le traitement principal. Ainsi, nous abordons des approches capables d'intégrer la prise en compte de ces évolutions. Dans la thèse de Romain Belmonte, en collaboration avec l'Ecole d'ingénieurs des Hautes Technologies et du Numérique (ISEN - YNCREA), nous explorons les processus d'apprentissage profond pour la prise en compte de cette multitude de facteurs en multipliant les situations présentées (augmentation des données). Les premiers travaux qui en émergent concernent la prise en compte de l'évolution dans le temps du suivi des points caractéristiques du visage pour améliorer la précision des détections dans (Belmonte et al. 2019).

Approches neuromorphiques pour l'analyse faciale Les processus d'apprentissage se trouvent au coeur de l'ensemble des travaux évoqués jusqu'ici. Depuis 2014, nous nous intéressons à des processus d'apprentissage non-supervisés sur la base de réseaux de neurones impulsionnels. Entourés par des collègues intéressés par la fabrication (équipe CSAM, lab. IEMN) et la simulation (équipe Émeraude, lab. CRIStAL) de ces réseaux, nous explorons des problématiques liées aux besoins de la vision (réseaux de taille importante, encodage de l'information, etc.). Les travaux que nous avons réalisés conjointement nous ont permis en 2016 de bénéficier d'un contrat doctoral pour étudier les architectures neuromorphiques à base de neurones impulsionnels pour la vision. En étroite collaboration avec Pierre Tirilly (MCF, équipe FOX, lab. CRIStAL), Philippe Devienne (CR CNRS, équipe Émeraude, lab. CRIStAL) et Pierre Boulet (PR, équipe Émeraude, lab. CRIStAL), nous encadrons les travaux de Pierre Falez autour de l'apprentissage non-supervisé de descripteurs à base de réseaux de neurones impulsionnels et de mécanismes bio-inspirés de type *spike-timing dependent plasticity* (STDP). Les résultats obtenus ont été publiés dans (Falez et al. 2017) et (Falez et al. 2018) [17].

17. P. Falez ; P. Tirilly ; I. M. Bilasco ; Ph Devienne ; P. Boulet - Mastering the output frequency in spiking neural networks - Proc. of the International Joint Conference on Neural Networks (IJCNN), Jul. 2018, Rio de Janeiro, Brazil.

3 Plan

Dans la deuxième partie de ce document, je détaille les principales contributions apportées dans trois domaines de l'analyse faciale :

— l'estimation de l'orientation de la tête - chapitre 1 ;

— la reconnaissance du genre - chapitre 2 ;

— la reconnaissance des expressions faciales - chapitre 3.

Le document se poursuit avec la présentation de mes perspectives de recherche, en Partie III. Une présentation synthétique de mes activités de recherche incluant les encadrements et les montages de projets, ainsi que mes responsabilités pédagogiques et administratives est réalisée en Partie IV, Chapitre 1. Le détail des publications réalisées est présenté dans la Partie IV, Chapitre 2.

Deuxième partie

Analyse faciale dans les flux vidéo

L'omniprésence de flux vidéo dans notre quotidien ouvre de nouvelles perspectives pour enrichir l'expérience utilisateur. Souvent associée à la télé-protection, la capacité de traiter, extraire et interpréter les comportements des individus trouve des applications dans des domaines tels que : l'e-learning, l'e-health, le commerce, les télécommunications, etc. Rendre intelligent l'environnement dans lequel une personne évolue permet de mieux caractériser et anticiper ses besoins. Par exemple, dans le cadre d'une session d'e-learning, lorsque le système comprend que l'apprenant perd son intérêt ou rencontre des difficultés de compréhension, il peut adapter la façon dont l'apprentissage se déroule en mettant en place des stratégies adéquates. Dans le domaine médical, la capacité d'évaluer l'état de l'individu en amont de la mise en relation avec un professionnel de santé peut améliorer sa prise en charge. Dans un musée, des vitrines intelligentes capables d'interpréter et d'anticiper les attentes des visiteurs permettraient à ces derniers de mieux s'immerger dans la visite. Cette multitude d'usages potentiels met la technologie face à des situations d'interaction très variées, ce qui soulève de nombreux défis technologiques.

Dans ce qui suit, nous considérons essentiellement des situations d'interaction dans un environnement personnel. Nous analysons les informations fournies par le visage d'une personne afin de pouvoir caractériser au mieux la personne et son comportement et ainsi enrichir le contexte d'interaction.

De nombreux travaux ont été réalisés pour estimer l'orientation de la tête, la reconnaissance du genre ou la reconnaissance des expressions faciales. Les travaux existants enregistrent de bonnes performances sur des corpus de données enregistrés dans des situations contraintes de type laboratoire. Par exemple, les corpus dédiés à l'analyse des expressions faciales dans des situations où les expressions sont exagérées offrent un cadre idéalisé pour la reconnaissance des expressions. Cependant, plus les corpus de données reflètent des conditions naturelles (pose variable, intensité d'expression variable, etc.), moins les solutions existantes se montrent performantes. Souvent, les approches développées ne traitent que partiellement le problème en s'intéressant à certains défis, selon les corpus de données traités. Cela restreint la capacité de généralisation des solutions proposées.

Plusieurs défis peuvent être posés lorsque l'on souhaite analyser un flux vidéo sans contraintes particulières. Certains relèvent de la manière dont la captation de flux vidéo a été effectuée : les occultations du visage, les variations d'illumination et de pose de la tête. D'autres défis ne découlant pas directement du processus de captation peuvent être communs à différents domaines applicatifs. Par exemple, pour l'estimation de l'orientation de la tête pour la reconnaissance du genre ou des expressions, il faut tenir compte de la variabilité inter-personne. Il faut privilégier des outils qui s'abstraient de l'identité de la personne pour pouvoir maximiser la proximité intra-pose, intra-genre ou intra-expression.

D'autres défis sont spécifiques à un domaine applicatif et ils sont parfois contradictoires. Par exemple, lorsque l'on s'intéresse à la reconnaissance du genre, la reconnaissance doit être indépendante de l'orientation de la tête ou bien de l'expression faciale. Ainsi, on privilégie les descripteurs et les approches qui essaient de normaliser ou de s'abstraire le plus des variations en termes d'expression. Lorsque l'on s'intéresse à la reconnaissance des expressions, il faut que l'on privilégie des descripteurs et approches qui permettent de séparer le plus possible la représentation des différentes expressions tout en étant indépendants de l'orientation ou du genre de la personne.

Ainsi, au-delà des défis liés aux environnements de capture pour lesquels des solutions génériques existent, il faut envisager chaque domaine de l'analyse faciale de façon spécifique.

L'estimation de l'orientation de la tête (Chapitre 1) est l'un des processus centraux de l'analyse faciale. Il renseigne sur la manière dont le visage est orienté face à la caméra. La disposition du visage face à la caméra a des répercussions importantes sur les processus sous-jacents d'analyse. En effet, cette information est essentielle, car un visage de profil partage relativement peu d'information avec un visage de face si aucune procédure de normalisation ou d'alignement spatial n'intervient. Une fois l'orientation caractérisée, il est aisé d'envisager des traitements spécifiques. En revanche, l'estimation de l'orientation doit être effectuée sans aucun a priori sur la variabilité induite par des identités, des ethnies, des états affectifs, des occultations, etc.

Pour **la reconnaissance du genre** (Chapitre 2), la grande variabilité intra-classe (ethnie, couleur, longueur de cheveux) impose que l'extraction des caractéristiques soit faite avec soin afin d'optimiser les capacités de reconnaissance. Au contraire, dans le cadre de la reconnaissance des expressions, ces mêmes attributs n'interviennent pas dans le processus en tant qu'éléments de décision, mais plutôt comme des éléments perturbateurs occasionnant des occultations et induisant du bruit dans le processus d'analyse.

Lorsque l'on s'intéresse à la **reconnaissance des expressions** (Chapitre 3), en présence de mouvements de la tête, il faut dissocier la caractérisation du mouvement facial induit par l'expression et le mouvement induit par le changement d'orientation de la tête. Ce problème devient encore plus difficile lorsque l'on s'intéresse à la reconnaissance des expressions de faible intensité. En effet, dans une interaction naturelle, il peut y avoir une forte variabilité entre les micro- et les macro-expressions. Les micro-expressions, difficilement perceptibles même pour l'oeil humain, sont des expressions de très courte durée et de faible intensité. À l'opposé, les macro-expressions sont des expressions qui ont une durée et une intensité suffisantes pour pouvoir être perçues et décodées facilement.

1 Estimation de l'orientation de la tête dans des environnements peu contraints

L'estimation de la pose de la tête est un des problèmes fondamentaux de la vision par ordinateur. Le mouvement non-contraint de la tête génère des changements de perspective dans le plan (roulis) ou hors-plan (tangage, lacet) qui rendent l'analyse du visage plus difficile. Le roulis correspond à un mouvement autour de l'axe entre le sujet et l'observateur (caméra ou une autre personne), comme illustré dans la Figure 1.1 (rotation autour de l'axe $0z$). Le tangage correspond à un changement d'orientation dans le plan vertical (rotation autour de l'axe $0x$). Le lacet correspond à une rotation autour de l'axe $0y$ générant un mouvement de gauche à droite ou inversement.

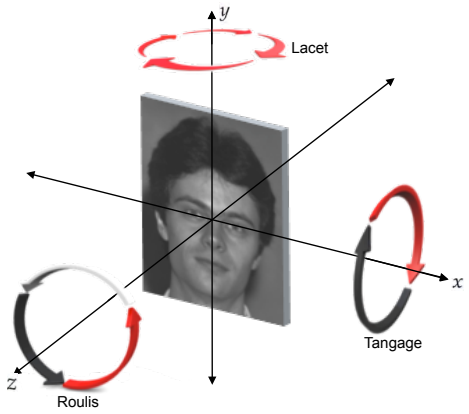


Figure 1.1 – *Illustration des angles de rotations (roulis - roll, lacet - yaw et tangage - pitch.)*

L'information concernant l'estimation de l'orientation de la tête est souvent exploitée dans les domaines applicatifs suivants : a) l'analyse du comportement de personnes, b) la contextualisation de processus sous-jacents d'analyse faciale.

En effet, premièrement, l'orientation de la tête d'une personne est une information exploitable dans différents processus, à des fins très diversifiées. L'estimation de l'orientation de la tête permet de connaître les éléments qui rentrent dans le champ visuel de la personne. Même si, l'estimation de l'orientation de la tête est moins précise que l'estimation du point de regard (Deravi et Guness 2011), lorsque l'utilisateur fait face à un écran, il est possible de répondre à des questions de type : *est-ce que la personne regarde l'écran ? est-ce que la personne observe le côté gauche de l'écran ?* En complément, en analysant la fréquence et l'amplitude des changements de l'orientation de la tête

dans le temps, il est également possible d'inférer la manière dont la personne perçoit l'interaction avec l'écran ou son niveau d'attention.

Deuxièmement, l'information sur la pose de la tête permet de caractériser de manière plus spécifique l'apparence et la géométrie du visage. Selon qu'il est de face ou qu'il subit une rotation importante par rapport à l'axe principal de la caméra, le visage se montre sous des angles très différents. Dans la littérature, la majorité des études concernant l'analyse faciale (par exemple, la reconnaissance des personnes, du genre, des expressions ou l'estimation de l'âge) suppose que le visage est de face. Cela limite les situations dans lesquelles l'analyse faciale peut se dérouler de manière optimale. En effet, dans des situations d'usage peu contraintes, souvent, les visages subissent des rotations dans le plan ou hors-plan mettant au défi le processus d'analyse faciale. Selon Demirkus et al. (2013), les rotations hors-plan déforment la projection 2D du visage dans l'image, ce qui rend l'analyse beaucoup plus compliquée qu'en présence de rotations dans le plan. Ainsi, il est important d'estimer la pose pour adapter le processus d'analyse.

Des outils performants existent pour estimer l'orientation de la tête, mais souvent ils requièrent une forte instrumentalisation (matériel spécifique, lumière infra-rouge contrôlée, etc.). Ces outils répondent bien à des situations d'interaction en présence de ces dispositifs, mais ne peuvent pas s'appliquer sur des données captées dans des situations non-instrumentalisées.

La difficulté de l'estimation de l'orientation de la tête à partir d'images 2D réside dans la perte de l'information de profondeur. L'estimation de l'orientation de la tête dans un contexte 2D reste un problème ouvert. Cela est surtout dû au fait que dans ce cas de figure, nous devons étudier une large palette de perspectives potentielles sur le visage. Cette grande variabilité pose de nombreux défis car, au cours du processus de caractérisation, il faut trouver des descripteurs qui s'abstraient autant que possible de l'identité de l'individu tout en renforçant la proximité entre visages présentant la même orientation.

Dans le but de disposer de solutions peu contraintes, nous visons le développement de méthodes qui, en partant uniquement des informations contenues dans une image standard RGB, permettent d'estimer le roulis et le lacet d'un visage. Nous nous intéressons, d'une part, à la symétrie bilatérale du visage et, d'autre part, à la manière dont les visages répondent à des détecteurs de visages frontaux. Nous étudions l'impact des changements d'orientation sur la préservation de ces informations (symétrie et réponse aux détecteurs frontaux).

Approche par symétrie Premièrement, dans la thèse d'Afifa Dahmane [1], nous nous concentrons sur l'étude de la symétrie et les apports de la symétrie dans la détection de l'orientation de la tête (Dahmane et al. 2015) [2]. L'hypothèse principale suppose qu'un visage frontal comporte plus de symétrie qu'un visage non-frontal. L'exploitation de cette hypothèse nous permet d'estimer de manière robuste le roulis et le lacet de la tête même en présence d'occultations partielles du visage.

1. Afifa Dahmane, doctorante de mai 2010 à février 2015, équipe FOX, lab. CRIStAL en co-tutelle avec l'Université de Sciences et Technologies Houari Boumediene, Algérie.

2. A. Dahmane ; S. Larabi ; I.M. Bilasco ; C. Djeraba - Head pose estimation based on face symmetry analysis - Signal, Image and Video Processing (SIViP), 2014, pp 1-10 (Facteur d'impact : 1,643 selon JCR 2018).

Approche par transformation inverse En parallèle, avec Taner Danisman [3] et José Mennesson [4], nous avons décidé de prendre avantage des contraintes de pose que doivent respecter les visages afin d'être localisés par les détecteurs de visages frontaux tels que le détecteur de Viola et Jones (2004). Par exemple, les détecteurs à base de filtres de Haar qui encodent la succession des zones lumineuses et sombres sur un visage permettent de localiser des visages quasi-frontaux, présentant des roulis et lacets limités. Ainsi, si un visage est localisé par le détecteur de Viola et Jones, cela implique que son roulis et son lacet sont compris, respectivement, dans les intervalles [-15°..+15°] et [-30°..+30°]. Au lieu de caractériser directement la pose d'un visage, nous étudions donc comment les changements d'orientation successifs induits par une rotation contrôlée font varier les réponses d'un détecteur frontal de visages. Notre hypothèse de travail est que les transformations maximisant les réponses d'un détecteur frontal renseignent sur l'orientation de la tête avant rotation, en matière de roulis dans (Danisman et al. 2015) [5], et en matière de lacet dans (Mennesson et al. 2016) [6].

Dans la suite de ce chapitre nous commençons par présenter certains travaux de l'état de l'art sur les méthodes d'estimation de l'orientation de la tête, qui ont servi d'étalon pour nos réalisations. Ensuite, nous présentons en détail la méthodologie et les protocoles expérimentaux qui ont permis de valider nos deux familles d'approches : par symétrie (Section 1.2) et par transformation inverse (Section 1.3).

1.1 État de l'art

Six catégories couvrant un large spectre d'approches pour l'estimation de l'orientation ont été introduites par Murphy-Chutorian et Trivedi (2009) : les approches à base de motifs, les approches à base de modèles, les approches par classification, les approches par régression, les approches de type *Manifold Embedding* et les approches à base de suivi. Les frontières entre ces différentes catégories sont perméables et il est difficile de proposer une taxonomie pour les approches d'estimation de l'orientation de la tête. La plupart du temps, les solutions existantes utilisent plusieurs niveaux d'analyse, constituant ainsi des méthodes hybrides.

Nous nous intéressons à une nouvelle structuration de travaux de l'état de l'art. Nous prenons en compte la nature des caractéristiques dominantes utilisées pour estimer l'orientation indépendamment des descripteurs et des méthodes d'estimation sous-jacentes :

— approches à base de détection et suivi de points caractéristiques ;
— approches globales - considérant le visage comme un tout ;
— approches à base de suivi d'un modèle de tête.

Approches à base de points caractéristiques L'orientation du visage peut être estimée également en analysant les zones spécifiques du visage, comme illustré par Danisman et al. (2010) et Zhao

3. Taner Danisman, post-doctorant septembre 2010 à juin 2011 et ingénieur de recherche de juin 2012 à octobre 2014, équipe FOX, lab. CRIStAL, actuellement enseignant-chercheur à Akdeniz Üniversitesi, Turquie.

4. José Mennesson, ingénieur de recherche de novembre 2014 - décembre 2015, équipe FOX, lab. CRIStAL, actuellement Maître Assistant à l'Institut Mines-Telecom Lille Douai, équipe FOX, lab. CRIStAL.

5. T. Danisman ; I.M. Bilasco - In-plane face orientation estimation in still images - Multimedia Tools and Applications, Springer Verlag, 2016, 75 (13), pp.7799-7829 (Facteur d'impact : 1,541 selon JCR 2018).

6. J. Mennesson ; A. Dahmane ; T. Danisman ; I.M. Bilasco - Head Yaw Estimation using Frontal Face Detector - Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2016, Portugal, vol. 4, pp. 517-524.

et al. (2013). Par exemple, l'angle θ (par rapport à l'axe horizontal) de la ligne traversant le centre des yeux est directement corrélé avec le roulis du visage. Toutefois, les deux yeux doivent être détectés de manière précise et robuste. Des solutions plus complexes basées sur des systèmes de suivi de points ont été également explorés dans la littérature. Asteriadis et al. (2010) s'intéressent au problème d'estimation de la pose de la tête en utilisant une caméra en environnement non-contrôlé. Les auteurs utilisent le descripteur *distance vector fields* (distance aux coins les plus proches) extraits de coordonnées des points caractéristiques du visage par rapport à la boîte englobante du visage. Suite à l'initialisation, les visages sont suivis en appliquant un filtre de peau. Oka et al. (2005) s'appuient sur un système de stéréovision qui extrait des informations sur le visage en six points caractéristiques : sourcils intérieurs et coins des yeux, du nez et de la bouche. Les travaux de Kazemi et Sullivan (2014) permettent de localiser et de suivre avec un très bon niveau de précision et de robustesse les points caractéristiques. Toutefois, des changements dûs au bruit local peuvent produire des variations qui influent sur l'estimation de la pose.

Approches globales Afin de faire face à l'instabilité inhérente à la détection de points caractéristiques, certains auteurs s'intéressent aux caractéristiques globales de visages. Ainsi, Zhou et al. (2002) ont étudié l'estimation de l'orientation de la tête en s'appuyant sur les histogrammes d'orientation de gradients, en appliquant un seuil par rapport à la magnitude du gradient observée en chaque pixel. Osadchy et al. (2007) proposent une méthode pour simultanément détecter des visages et estimer leur orientation dans le plan (roulis). Un réseau convolutionnel est utilisé pour projeter les images de visages sur des variétés (une par orientation) et les images ne comportant pas de visage à l'extérieur de ces variétés. Guo et al. (2011) proposent un modèle d'apprentissage multilinéaire supervisé (*supervised multilinear learning model*) pour l'estimation de la tête en utilisant la régression à base de vecteurs de support (*support tensor regression - STR*). Ils ont montré que STR fournit des caractéristiques permettant de mieux distinguer les diverses poses dans le processus d'estimation de l'orientation de la tête qu'une solution de régression à base de vecteurs supports (*support vector regression - SVR*). My et Zell (2013) proposent une méthode qui combine efficacement le détecteur de Viola et Jones avec un filtre à base de corrélations adaptatives pour le suivi non-contraint du visage. Afin d'améliorer la rapidité du suivi, ils estiment la taille du visage en utilisant une caméra de profondeur.

D'autres caractéristiques telles que la symétrie ont été explorées. Malgré le fait que le visage humain n'est pas parfaitement symétrique, la symétrie faciale d'une personne est significative et peut être employée pour distinguer différentes poses. Vinod Pathangay et Greiner (2008) estiment l'orientation de la tête (roulis et tangage) en exploitant la structure symétrique du visage. La symétrie est mesurée en considérant l'image miroir et en extrayant les points fiduciaires. Les positions relatives des points dans l'image originale et l'image miroir servent à inférer le niveau de symétrie et par la suite, la pose. Ma et al. (2013) introduisent des filtres de Gabor afin de faciliter le calcul de la zone de symétrie et d'estimer le tangage. Un modèle d'illumination basé sur la symétrie est proposé par Gruendig et Hellwich (2004). Il s'appuie sur trois caractéristiques principales du visage : les deux yeux et le bout du nez. Pour chaque combinaison des positions relatives de ces éléments, la pose de la tête est calculée en utilisant une projection géométrique et en considérant une calibration interne de la caméra. Dans le contexte de la reconnaissance de personnes, Harguess et al. (2008) utilisent la symétrie bilatérale du visage afin de déduire si la pose est frontale ou non.

Des travaux évaluant la symétrie et caractérisant la pose à partir de données 3D ont été proposés dans (Hattori et al. 1998) et (Gui et Zhang 2006).

Approches à base de suivi d'un modèle de tête Les approches à base de suivi de modèles exploitent l'information temporelle afin d'affiner l'estimation de la pose. Même si elles ne peuvent pas s'appliquer sur des images statiques, elles offrent un contexte très performant pour comparer les résultats obtenus par les méthodes n'exploitant pas d'information temporelle. Sung et al. (2008) analysent une séquence d'images afin de trouver les paramètres caractérisant le mouvement global de la tête. Ceci permet de les mettre en correspondance avec un modèle cylindrique de la tête. La Cascia et al. (2000) utilisent un suivi de modèle cylindrique et des techniques de mise en correspondance des images pour estimer l'orientation de la tête en présence de rotations dans le plan et hors-plan. Valenti et al. (2009) utilisent de manière conjointe le suivi d'un modèle cylindrique de la tête et un détecteur des centres des yeux. Murphy-Chutorian et Trivedi (2008) utilisent un filtre particulaire exploitant l'apparence de la tête, doublé par une estimation de la cohérence du suivi. Les particules générées encodent les paramètres de projections et de déplacements d'un modèle générique de la tête en considérant des petites variations en termes de roulis, tangage, lacet et déplacement entre deux trames.

Synthèse Les approches mettant l'accent sur l'exploitation directe de l'apparence des visages répondent mal aux défis posés par la variabilité en termes d'identité, d'occultation partielle ou d'intensité lumineuse. Dans le contexte des méthodes globales, les méthodes à base de motifs peuvent induire un sur-apprentissage en associant davantage l'identité des individus aux motifs qu'à la pose elle-même. Ce défaut est également présent dans les approches à base de réduction de dimensionnalité, où les sous-espaces sont attachés à la pose, mais également à l'identité des individus. Les approches exploitant les modèles géométriques ne sont pas assujettis au biais de l'identité car ils sont liés aux points caractéristiques de la morphologie d'une personne en général. En revanche, la construction de ces modèles pâtit des mêmes défauts que les approches exploitant l'apparence des visages. La localisation des points constituant le modèle se fait en partie sur la base de l'apparence locale du visage, qui peut subir des variations importantes en présence de variations d'illumination, d'occultations, etc. Ceci restreint le champ d'application de ces approches à des situations contrôlées, n'offrant pas de solution pour des environnements complexes tels que le poste de pilotage d'une voiture (assistant de conduite) ou une salle dont la luminosité est non-contrôlée (e-learning). En ce qui concerne les méthodes à base de suivi, la phase d'initialisation de diverses méthodes nécessite la détection d'un visage frontal. Ainsi, mise à part l'étape impliquant le suivi, qui permet de suivre l'orientation de la tête en dehors des poses reconnues par une détecteur frontal, cette famille d'approches présentent des défauts similaires à des méthodes nécessitant des poses identifiables par des détecteurs frontaux.

Au regard des constatations faites dans cet état de l'art, nous avons exploré dans nos travaux deux directions de recherche concernant l'estimation de l'orientation de la tête :

Symétrie bilatérale Afin de ne pas être tributaire de la détection de certains points et de maintenir la possibilité d'estimer l'orientation, nous nous sommes intéressés à une méthode globale portant sur l'étude de la symétrie bilatérale du visage. En effet, en partant du constat qu'un visage

frontal est fortement symétrique, nous estimons l'orientation de la tête en quantifiant la quantité de symétrie bilatérale du visage.

Transformation inverse Les approches d'estimation requierent souvent une initialisation du processus d'estimation s'appuyant sur un détecteur frontal de visages. Inspirés par ce fait, nous avons étudié les caractéristiques de détecteurs de visages frontaux. Nous nous en sommes servi afin d'inférer l'estimation de l'orientation de la tête. Nous étudions la plage des rotations qui appliquées à un visage permettent d'obtenir de réponses positives de la part d'un détecteur de visages frontaux. En identifiant les rotations ramenant le visage dans une configuration frontale, nous déduisons l'orientation de la tête.

1.2 Approche à base de symétrie bilatérale

Dans cette section nous abordons le problème de l'estimation de l'orientation de la tête en exploitant la symétrie bilatérale du visage. Selon Wilson et al. (2000), la perception humaine de la pose de la tête est basée sur deux indices : la déviation de la forme de la tête par rapport à la symétrie bilatérale et la déviation de l'orientation du nez par rapport à la verticale. En partant du principe que le problème d'estimation de pose est étroitement lié à la géométrie du visage, nous considérons que la symétrie du visage est un bon estimateur de la configuration géométrique et, par conséquent, de l'orientation de la tête. Dans la littérature, l'exploitation de la symétrie du visage dans la reconnaissance de poses reste un sujet abordé par une minorité de chercheurs. Toutefois, nous estimons que cette approche de caractérisation peut répondre à certains défis tels que les occultations locales.

Nous nous intéressons à l'estimation discrète de l'orientation de la tête en étudiant la symétrie bilatérale du visage et en considérant deux degrés de liberté du mouvement : le roulis et le lacet. Les changements induits par le tangage affectent à moindre mesure la symétrie bilatérale, ce qui rend l'usage de cette méthode moins efficace.

Notre approche à base de symétrie vise à être la moins intrusive possible, ne nécessitant pas une collaboration explicite de la part des utilisateurs. Elle se doit d'être indépendante de l'identité de l'utilisateur et doit fonctionner à la fois dans un contexte image ou vidéo. Notre système utilise un modèle géométrique qui ne nécessite pas de détection des points caractéristiques du visage, ni de calibration spécifique. Elle peut être déployée facilement en utilisant une caméra standard. La symétrie est mesurée sur la globalité du visage. En s'affranchissant d'une détection spécifique de points caractéristiques, qui peut souffrir des occultations locales, elle se montre plus robuste. La symétrie est caractérisée en analysant l'intensité des pixels.

Nous proposons une approche qui marie de manière intelligente un traitement local et un traitement global du visage. Nous sélectionnons les régions symétriques du visage en analysant l'intensité des pixels de peau. Nous utilisons les propriétés géométriques (par exemple, orientation et étendue) de la région de symétrie bilatérale afin d'estimer le roulis et le lacet. La structure générale de notre proposition est illustrée dans la Figure 1.2. Le processus démarre par une détection de visages, en employant par exemple l'algorithme de Viola et Jones (2004). Des prétraitements (tels que l'égalisation d'histogrammes) sont appliqués au visage détecté afin de réduire les variations d'illumination. Ensuite, l'axe de symétrie est recherché sur le visage. Dès que l'axe principal de

symétrie est identifié, nous extrayons les caractéristiques géométriques de la région contenant la symétrie. Nous déduisons l'angle de roulis à partir de l'orientation de l'axe de symétrie et nous estimons le lacet en utilisant la taille et l'orientation de la région de symétrie. Un arbre de décision est entraîné afin de classifier les visages selon les caractéristiques de symétrie mesurées. L'intérêt majeur de cette approche réside dans le fait qu'elle ne nécessite pas la détection des points caractéristiques et gère mieux les occultations locales, car la symétrie est calculée globalement sur le visage.

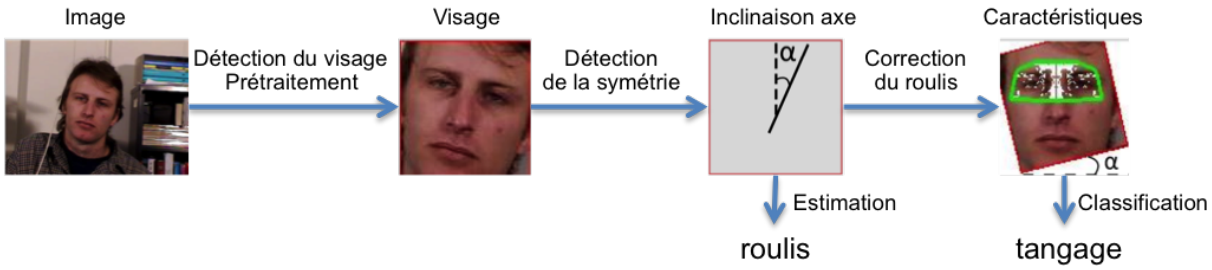


FIGURE 1.2 – *Schéma général de notre approche basée sur la symétrie.*

Autour de cette proposition nous pouvons mettre en exergue trois contributions principales :

— premièrement, nous construisons une méthode pour détecter la position et l'orientation de l'axe de symétrie du visage dans une image ;
— deuxièmement, l'angle de roulis est déduit en analysant l'inclinaison de l'axe de symétrie ;
— troisièmement, l'angle de lacet est calculé à partir des caractéristiques de la région de symétrie bilatérale (taille, position).

Dans la suite, nous montrons la corrélation existante entre les caractéristiques de la région de symétrie et l'orientation de la tête. Nous détaillons le processus de détection de l'axe de symétrie et la caractérisation des régions symétriques. À partir de ces caractéristiques, nous abordons l'entraînement d'un arbre de décision pour classifier les poses.

1.2.1 Symétrie bilatérale et orientation

Nous étudions l'impact des mouvements de la tête (lacet, roulis) sur les propriétés géométriques des régions de symétrie bilatérale. La Figure 1.3 montre les variations en termes de taille et inclinaison des régions de symétrie bilatérale en présence de variations de lacets et de roulis. Des propriétés basiques caractérisant la région de symétrie bilatérale, telles que la taille de la région, sont utilisées pour estimer le lacet et le roulis. Pour mettre cela en évidence nous analysons la différence de symétrie avant et après la rotation de la tête en présence d'une variation de type lacet.

Lacet et symétrie Lorsque le visage est présenté de face à la caméra, la symétrie entre les deux parties du visage apparaît clairement et la ligne qui passe entre les deux yeux et par le bout du nez définit l'axe de symétrie. Lorsque la tête tourne, par exemple en présence d'un changement de lacet, les régions symétriques diminuent en taille. La Figure 1.4 illustre les changements de ratio des pixels symétriques en présence des variations de lacet.

Dans la Figure 1.5, nous illustrons et analysons la manière dont les caractéristiques de la région de symétrie évoluent. Soit a et b deux points symétriques sur le visage. m est le milieu du segment

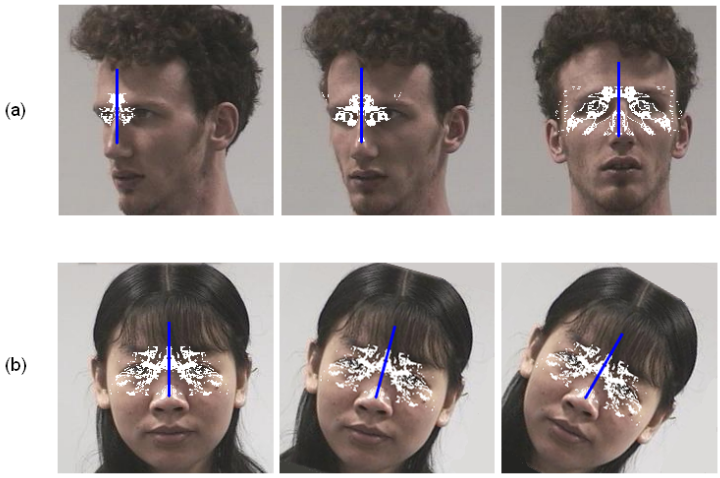


FIGURE 1.3 – *(a) Variation de la taille de la région symétrique en présence d'un mouvement de lacet ; (b) Variation de l'inclinaison de l'axe de symétrie en présence d'un mouvement de roulis.*

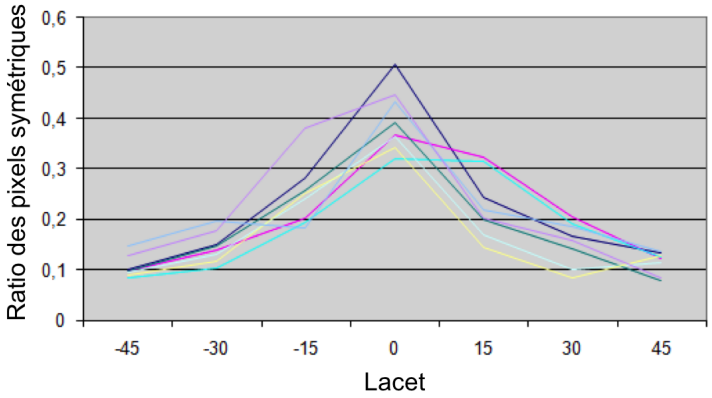


FIGURE 1.4 – *Le changement du ratio des pixels symétriques en présence de variations de lacet.*

$[ab]$. Les projections de ces points dans le plan image sont a_i, b_i, m_i. Lorsque le visage est face à la caméra, les segments $[a_i m_i]$ et $[m_i b_i]$ sont symétriques par rapport à m_i comme illustré dans la Figure 1.5 (a). Lorsque la tête effectue un mouvement de lacet, les points caractéristiques (a, b, m) sont projetés dans le plan image aux points (a'_i, b'_i, m'_i) (voir Figure 1.5 (b)).

Ainsi, suite à un mouvement de lacet, la symétrie dans le plan image n'est que partielle. La partie symétrique du segment reliant deux points symétriques sur le visage est plus petite que la partie symétrique observée dans l'instant précédant le mouvement. Ainsi, nous avons montré qu'une variation dans la taille de la région de symétrie peut résulter d'un mouvement de lacet.

Roulis et symétrie En ce qui concerne le roulis, les changements induits sur les régions de symétrie bilatérale n'affectent pas la taille de la région de symétrie mais l'orientation de l'axe de symétrie. Ainsi, nous observons que le roulis correspond à l'angle de l'axe de symétrie (voir Figure 1.3 (b)). Nous estimons l'angle de rotation de la tête à partir de l'inclinaison de l'axe de symétrie calculé dans une situation de pose frontale.

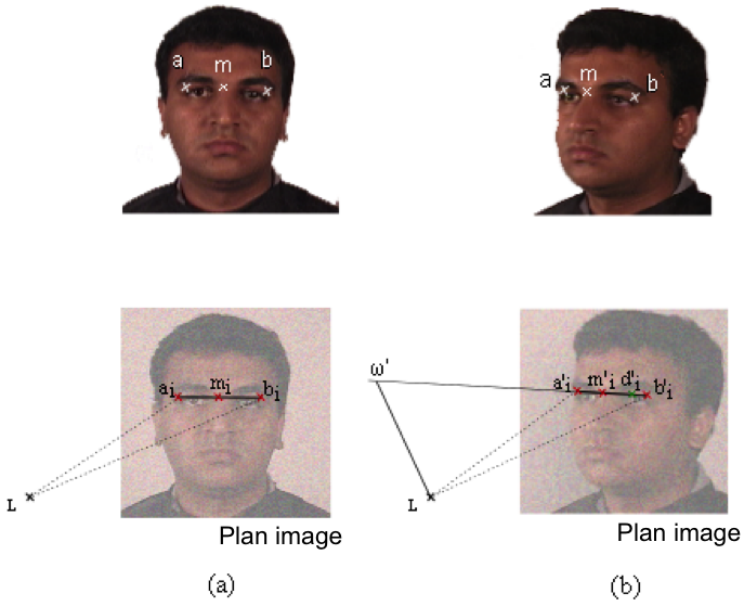


FIGURE 1.5 – *(a) Projection du segment [ab] lorsque le visage est face à la caméra ; (b) Projection du segment après une rotation hors-plan (lacet).*

Détection de la symétrie bilatérale Notre but est de caractériser la symétrie morphologique du visage sous différentes poses, tant que cette symétrie est observable et ne disparaît suite aux rotations hors-plan dépassant 45°. Stentiford (2005) a proposé un algorithme qui met en évidence la symétrie au niveau de l'image toute entière. Il est nécessaire d'adapter cet algorithme qui ne permet pas de cibler la détection de la symétrie au sein du visage. Suite à la détection du visage, nous considérons une ellipse inscrite dans la boîte englobante associée au visage. Nous divisons l'ellipse dans le plan horizontal et nous considérons sa partie supérieure pour caractériser la symétrie. La partie supérieure du visage est choisie car les effets de symétrie sur son apparence sont plus marqués en présence de rotations de la tête. En effet, la variation de symétrie après des rotations gauche/droite est plus importante dans la région des yeux que dans la région de la bouche.

Le calcul de la région de symétrie bilatérale commence par l'identification de l'axe de symétrie principal du visage. L'estimation de l'axe se fait en se servant de la mise en correspondance de régions similaires d'un point de vue de l'intensité des pixels. Nous utilisons l'intensité des pixels afin de détecter la symétrie faciale bilatérale dans l'image. Deux pixels sont considérés symétriques si sur l'ensemble des canaux définissant la couleur, la différence d'intensité est inférieure à un seuil fixé. Une fois l'axe identifié, il est possible de caractériser précisément l'étendue de la région de symétrie.

L'axe de symétrie est défini par un point P et une orientation α par rapport à l'axe $0y$. Afin de détecter l'inclinaison α de l'axe de symétrie du visage, nous faisons varier α de α_{min} à α_{max} avec un pas de α_{pas}. Pour chaque inclinaison considérée, nous cherchons les positions des axes de symétrie potentiels. La région d'intérêt de l'image est divisée en plusieurs blocs superposés de taille $s * s$. Dans nos expérimentations nous avons choisi une taille de bloc correspondant à 5% de la taille du visage. Cela nous garantit une granularité suffisante pour estimer l'inclinaison de l'axe de symétrie. Nous cherchons la symétrie locale par rapport aux blocs en recherchant pour chaque bloc non-homogène d'une région le bloc symétrique au sein de la région d'intérêt.

Nous testons la similarité entre le bloc original et les blocs miroirs candidats par rapport à l'inclinaison α (voir Figure 1.6) jusqu'à ce qu'une correspondance soit trouvée. La position du bloc miroir est calculée en utilisant l'Équation 1.1. Les coordonnées du pixel $P(x,y)$ dans la position miroir sont (x_i,y_i). Nous déplaçons les valeurs x_i le long de la largeur de la région d'intérêt et nous obtenons les valeurs y_i.

$$y_i = y + ((\tan\alpha) \times (x_i - x)) \tag{1.1}$$

Les blocs miroirs se situent sur la bande qui part du bloc initial et qui suit une orientation correspondant à un angle $\alpha + \pi/2$. Deux blocs sont mis en correspondance si chaque pixel de la diagonale du bloc d'origine est similaire au pixel correspondant dans le bloc miroir considéré. Un pixel est mis en correspondance avec un autre si la différence d'intensité sur les trois canaux est située en dessous d'un certain seuil ε.

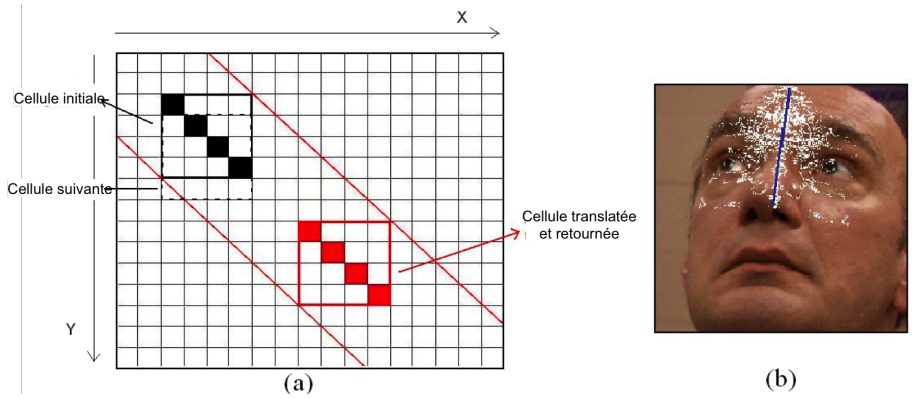


FIGURE 1.6 – *Détection de l'axe de symétrie.*

Lorsque nous trouvons deux blocs symétriques, nous déterminons la position de l'axe de symétrie de ces deux blocs. Cet axe de symétrie est perpendiculaire au milieu du segment reliant les deux blocs. Lorsque nous avons détecté l'ensemble des axes de symétrie $A_{i\{P_i,\alpha\}}$ avec i de 1 au nombre d'axes pour une inclinaison α, l'axe qui correspond au plus grand nombre de régions symétriques et qui est le plus proche du centre du visage est retenu. Le critère de proximité par rapport au centre du visage est introduit pour éviter, par exemple, de se focaliser sur la symétrie se trouvant au niveau du centre des yeux. Nous votons pour le meilleur axe $A_{\{P,\alpha\}}$ en utilisant le mécanisme suivant. Nous considérons les n premiers maxima de la distribution et nous votons pour l'axe de symétrie $A_{\{P,\alpha\}}$ qui minimise la distance par rapport au centre du visage C selon l'Équation :

$$d(C, A_{\{P,\alpha\}}) = \min_{i\in[1..n]} \{d(C, A_{i\{P_i,\alpha\}})\} \tag{1.2}$$

Nous répétons ce processus pour l'ensemble des inclinaisons entre α_{min} et α_{max}. Nous obtenons ainsi un ensemble de triplets (inclinaison, position et nombre de régions) répondant au critère de symétrie bilatérale pour l'inclinaison et la position correspondante. Un nouveau processus de vote et de sélection similaire à celui introduit dans l'Équation 1.2 permet de choisir l'inclinaison et la position de l'axe de symétrie optimal.

Caractérisation de la région de symétrie Lorsque l'axe de symétrie a été détecté, nous recherchons les mises en correspondance entre pixels par rapport à l'axe de symétrie retenu. Cette fois-ci,

contrairement à l'étape d'identification de l'axe, nous considérons l'intégralité des blocs (y compris les blocs homogènes). Au sein des blocs, les pixels sont comparés un à un afin de définir l'étendue de la région de symétrie (voir Figure 1.7). De cette manière, la détection de l'étendue de la région de symétrie est un processus qui exploite l'intégralité de la texture. Nous sélectionnons les paires de pixels symétriques (P_1, P_2) en utilisant l'Équation 1.1. Ensuite, nous construisons l'enveloppe convexe entourant l'ensemble des pixels symétriques et nous extrayons des caractéristiques géométriques telles que la taille de la région de symétrie (comme illustré dans la Figure 1.7).

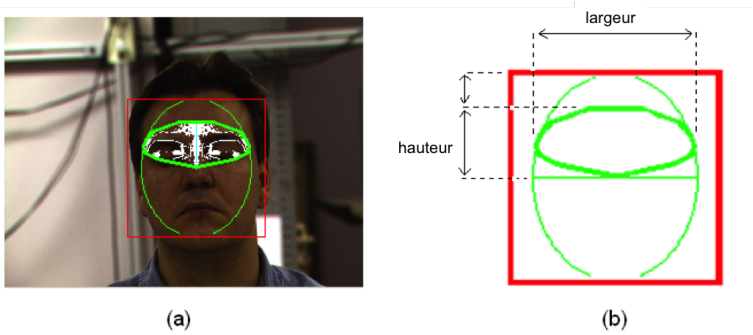


FIGURE 1.7 – *Exemples d'extraction de caractéristiques de symétrie : (a) Calcul de l'enveloppe convexe englobant les pixels symétriques ; (b) Mesures caractérisant la région de symétrie.*

Comme évoqué précédemment, l'estimation du roulis découle directement de l'inclinaison de l'axe de symétrie. Dans la suite, nous discutons de la manière d'estimer le lacet.

1.2.2 Estimation du lacet à l'aide de la symétrie bilatérale

Afin d'estimer le lacet, nous utilisons un classifieur à base d'arbres de décision entraîné avec les caractéristiques de la région de symétrie. Après une série d'expérimentations, nous avons décidé de ne pas conserver les mesures verticales de la région car elles ne sont que faiblement influencées par les mouvements de lacet. Ainsi, la caractérisation de la symétrie est effectuée en considérant la largeur de l'enveloppe convexe de la région ainsi que la distance moyenne entre les pixels symétriques et l'axe de symétrie. La largeur de l'enveloppe est définie comme la plus grande distance entre deux pixels de l'ensemble.

Nous commençons par extraire les caractéristiques (ratio des pixels symétriques, largeur de l'enveloppe convexe des pixels symétriques) de la région d'intérêt. Ensuite, nous construisons le modèle d'apprentissage à partir des vecteurs de caractéristiques extraits d'images correspondant à différentes poses et différentes personnes. Ici nous considérons une classification discrète de la pose. Ainsi, chaque classe de pose correspond à l'un des intervalles considérés. Les poses gauche et droite correspondant au même angle sont rassemblées dans une même classe, car les caractéristiques de la région de symétrie sont similaires. Ainsi, afin d'estimer $2 * n + 1$ poses discrètes (n poses vers la droite, n poses vers la gauche et 1 pose frontale), le modèle apprend $n + 1$ classes.

Afin de rendre la prédiction plus précise, nous suivons la méthodologie proposée par Holmes et al. (2001) et nous utilisons plusieurs arbres binaires de décision en cascade (*alternating decision tree*). L'arbre de décision présente une succession de noeuds de prédiction et de noeuds de décision. Le noeud racine est un noeud de prédiction et contient un vecteur dont les composantes reflètent la probabilité d'appartenance à chacune des classes de pose retenues. La racine de l'arbre contient

des valeurs nulles de prédiction pour les $n+1$ classes. Le premier niveau contient des noeuds de décision qui analysent les vecteurs de caractéristiques. S'ensuivent des noeuds de prédiction pour chaque classe et, ainsi, nous alternons des noeuds de décision et de prédiction jusqu'aux feuilles. La somme des valeurs, dans les noeuds de prédiction, le long des chemins dont les noeuds de décision sont vrais, est utilisée afin de classifier l'entrée. La classe avec la plus grande valeur de prédiction est considérée comme étant la classe associée à l'entrée.

Poses gauche et droite

Afin de différencier les poses droite et gauche, nous exploitons la différence d'intensité entre les pixels de la peau et ceux de l'arrière-plan. Notre hypothèse est qu'un pixel du visage est plus similaire à un autre pixel du visage qu'à un pixel de l'arrière-plan. Nous choisissons les pixels situés sur l'axe de symétrie afin d'être sûr qu'ils se situent au sein du visage. Nous calculons l'intensité moyenne des pixels dans le voisinage de l'axe et nous nous servons de cette valeur comme valeur de référence. Si l'axe de symétrie est plus proche du bord gauche du visage (respectivement, du bord droit), alors le visage est orienté à gauche (respectivement, à droite). Afin d'identifier le bord le plus proche, nous explorons les pixels situés sur la normale à l'axe de symétrie et nous conservons les pixels ayant des intensités proches de la valeur de référence. Si le pixel ayant une intensité cohérente avec la valeur de référence se situe le plus à gauche (respectivement, à droite), nous concluons que le visage est orienté vers la gauche (respectivement, la droite), comme illustré dans la Figure 1.8. En effet, la tête étant orientée vers la gauche, le côté droit du visage remplit davantage la partie gauche de l'image. Le pixel ayant une intensité cohérente avec la valeur de référence se situe à l'extremité gauche de l'image du visage.

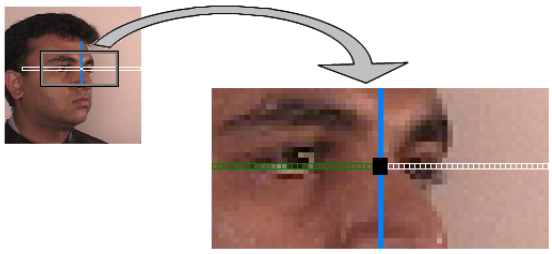


FIGURE 1.8 – *Critère pour différencier entre les poses gauche et droite.*

Avec cette méthode nous déterminons de quel côté le visage est orienté et cette information est combinée avec l'estimation de l'orientation fournie par l'arbre de décision afin d'obtenir le lacet.

1.2.3 Évaluation de l'estimation de l'orientation par symétrie bilatérale

Nous réalisons plusieurs expérimentations pour valider notre approche sur différents corpus tels que FacePix (Black et al. 2002), CMU-PIE (Sim et al. 2003) et Boston University Facial Tracker (La Cascia et al. 2000).

Estimation à partir d'images

Premièrement, nous évaluons l'approche en utilisant le corpus FacePix qui est idéal pour la détection de l'orientation car ses conditions de capture sont contrôlées et une grande variété de

poses est disponible. Le corpus est constitué de poses dans l'intervalle $[-90°, 90°]$ avec un incrément de $1°$. Ceci nous permet de construire différentes configurations pour l'apprentissage afin de constituer différentes classes. Nous avons notamment utilisé trois partitions de ce corpus :

— une en 4 classes, soit 7 poses discrètes (avec un incrément de $15°$) ;
— une en 5 classes, soit 9 poses discrètes (avec un incrément de $10°$) ;
— une en 10 classes, soit 19 poses discrètes (avec un incrément de $5°$).

Nous avons également testé notre méthode sur le corpus CMU-PIE, qui fournit plus de variabilité en termes d'illumination et d'expressions (yeux fermés, sourires). Cela permet de prouver la robustesse de notre approche à ce type de défis. En plus de ces corpus d'images, nous avons également pris en considération des séquences vidéos du corpus BUFT. Dans ce corpus, les sujets sont libres de bouger leur tête selon les six degrés de liberté, incluant lacet et roulis. Dans le cadre de ce corpus, nous pouvons également évaluer la capacité de notre solution à détecter le roulis.

De par la nature des corpus, nous présentons principalement les résultats concernant le lacet pour FacePix et CMU-PIE, où les visages ne subissent que de très légers roulis. Toutefois, dans le corpus BUFT, les expérimentations concernent à la fois le roulis et le lacet.

Dans l'ensemble des expérimentations concernant les différents corpus, nous avons utilisé les mêmes paramètres pour la détection de l'axe de symétrie : $\varepsilon = 25$, $s = 20$ et $n = 3$. Les valeurs d'α sont choisies dans l'intervalle $[85°, 95°]$ pour FacePix et CMU-PIE, où les visages ne subissent qu'un très léger roulis, et $[45°, 135°]$ pour le corpus BUFT et nos propres vidéos avec un pas de $3°$ pour l'inclinaison de l'axe.

FacePix Le corpus FacePix (Black et al. 2002) contient 3 séries d'images pour chacune des 30 personnes ayant participé aux enregistrements. La première série d'images contient pour chaque personne 181 images de visages correspondant à des angles de lacet entre $+90°$ et $-90°$. La deuxième et la troisième séries contiennent uniquement des visages frontaux et servent à étudier l'impact des variations de lumière. Pour la construction de l'arbre de décision à partir du corpus FacePix, nous considérons uniquement un sous-ensemble d'images de la première série qui correspond à des angles de lacet compris dans l'intervalle $[-45°; 45°]$. Au-delà de ces angles, l'information concernant le visage est parcellaire et ne permet pas de retrouver la symétrie bilatérale.

Afin de réaliser l'apprentissage et de tester le modèle ainsi appris, nous divisons le corpus en 6 sous-ensembles de même taille et nous appliquons un processus de validation croisée à 6 échantillons. Pour chaque échantillon nous mesurons les performances, en nous servant des cinq autres échantillons pour réaliser l'apprentissage. Le découpage en 6 échantillons est fait de telle sorte à ce qu'une même personne ne se retrouve pas dans deux échantillons différents. Cela permet de garantir que pour chaque expérimentation réalisée dans le cadre du processus de validation croisée aucune personne ne se trouve à la fois dans l'ensemble d'apprentissage et dans l'ensemble de test.

Lors de l'apprentissage, nous considérons différentes configurations en termes de nombre de classes et poses. La Figure 1.9 fournit les matrices de confusion pour trois configurations :

— 19 poses discrètes associées à des lacets entre -45° et 45° avec un incrément de 5° (10 classes) ;
— 9 poses discrètes associées à des lacets entre -40° et 40° avec un incrément de 10° (5 classes) ;
— 7 poses discrètes associées à des lacets entre -45° à 45° avec un incrément de 15° (4 classes).

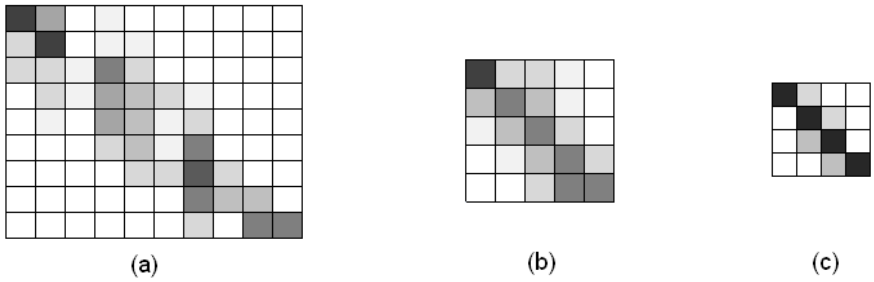


FIGURE 1.9 – *Matrices de confusion associées à différentes configurations d'apprentissage : (a) 19 poses avec un incrément de* 5° *; (b) 9 poses avec un incrément de* 10° *; (c) 7 poses avec un incrément à* 15°.

La diagonale des matrices renseigne sur la capacité de l'arbre à classifier correctement les poses. Le modèle à 7 poses offre des performances de classification plus importantes sur le corpus FacePix. Toutefois, le modèle à 19 poses présente l'intérêt de classifier plus finement la pose. Comme les erreurs de classification se font entre classes voisines, l'erreur globale reste maîtrisée.

Sur ce corpus, nous avons également testé la dépendance du processus complet à la précision de la détection de l'axe de symétrie. Ainsi, nous avons annoté manuellement la position et l'orientation de l'axe de symétrie. Cela nous permet de comparer le fonctionnement complètement automatique de l'estimation de l'orientation de la tête (détection d'axe, caractérisation, classification) et un fonctionnement semi-automatique où l'axe de symétrie est fourni par annotation manuelle. De par les résultats obtenus (voir les matrices de confusions précédentes), nous avons choisi de réaliser les tests dans une configuration à 7 classes. Une description détaillée des résultats est fournie dans la Table 1.1. En l'absence des imprécisions liées à la détection de la tête et de l'axe de symétrie, les résultats de classification sont meilleurs. Un gain d'environ 3% est enregistré. Toutefois, même dans le cas d'une automatisation complète du processus, la précision reste assez haute (79,6%) pour le modèle à 7 classes.

TABLE 1.1 – *Précision et erreur moyenne absolue (MAE) pour le corpus FacePix en mode automatique et en mode semi-automatique.*

Données	Précision (%)	MAE (°)
Position de la tête et axe de symétrie annotés	82,38	2,71
Position de la tête annotée manuellement et axe de symétrie détecté automatiquement	81,90	2,78
Position de la tête et axe de symétrie détectés automatiquement	79,63	3,14

CMU-PIE Le corpus CMU-PIE contient des prises de vues réalisées avec un réseau de caméras placées en demi-cercle autour du sujet. Des flashs lumineux modifient les conditions de capture. Les sujets doivent également reproduire certaines expressions faciales ou parler. Ces variations intervenant dans le processus de captation donnent naissance à trois version du corpus : CMU-PIE Expression, CMU-PIE Talking et CMU-PIE Lighting. Dans nos expérimentations, nous avons utilisé uniquement la partie du corpus correspondant aux caméras à hauteur de la tête et aux angles d'orientation compris dans l'intervalle $[-45° \,..\, 45°]$.

Nous construisons un arbre de décision pour chaque catégorie de données afin d'étudier de manière indépendante la robustesse de notre approche à des conditions d'illumination et à des expressions variables. Nous calculons le taux de performance en considérant une validation croisée à 6 échantillons. Nous effectuons ensuite une expérimentation en intégrant l'ensemble des images,

en garantissant qu'aucune personne se trouvant dans l'ensemble d'apprentissage ne se trouve dans l'ensemble de test.

La Table 1.2 présente les résultats pour chaque catégorie de données, incluant également la totalité des images. Le principal challenge introduit par ce corpus réside dans la variabilité de l'illumination. En effet, cela influe directement sur l'intensité des pixels et, par conséquent, sur l'extraction de la région de symétrie. Afin de limiter l'impact de l'illumination, l'égalisation d'histogramme RGB ne suffit pas. Nous appliquons une transformation en cosinus discrète globale (DCT) sur chaque image. La transformation DCT a été proposée par Chen et al. (2006). Un certain nombre de paramètres de la transformation sont tronqués afin de minimiser les variations de l'illumination. Les variations concernées se situent principalement dans les basses fréquences. En revanche, la transformation en cosinus discrète nuit au processus d'estimation dans le sous-ensemble CMU-PIE Expression. Les performances du système baissent de 72,57% (en RGB) à 49,81% (avec DCT). Au contraire, dans les sous-ensembles CMU-PIE Talking et CMU-PIE Lighting, la normalisation à base de DCT améliore les résultats. L'illumination affecte de manière plus forte la mise en correspondance lors du calcul de la symétrie que le bruit induit par la normalisation. D'un autre côté, les résultats moins bons obtenus par la DCT peuvent être également expliqués en partie par le nombre réduit d'images dans les sous-ensembles concernés. Le bruit introduit par la DCT semble mieux géré par l'apprentissage en présence d'un nombre plus important de données. Dans le sous-ensemble CMU-PIE Expression, pour chaque personne et pour chaque pose nous disposons de 3 à 4 images (neutre, clignement, sourire et, pour certains, le port de lunettes). Dans le sous-ensemble CMU-PIE Talking, nous disposons pour chaque pose de 60 images. Pour le sous-ensemble CMU-PIE Lighting, 23 images sont mises à disposition. Le nombre plus important d'images de l'ensemble CMU-PIE Talking permet de compenser la perte induite par l'élimination de certains paramètres de la DCT.

TABLE 1.2 – *Résultats obtenus sur différentes configurations d'apprentissage sur le corpus CMU-PIE.*

Data	Précision de classification (%)	
	Égalisation RGB	DCT
CMU-PIE Expression	72,57	49,81
CMU-PIE Talking	81,04	87,63
CMU-PIE Lighting	72,51	85,90
CMU-PIE	72,48	82,26

Après avoir étudié le comportement de notre approche face à différents challenges (illumination, expression, etc.), dans la section suivante nous étudions la manière dont notre approche peut répondre à des situations où l'on s'intéresse au suivi de l'orientation de la tête dans le cadre d'une séquence retraçant le mouvement de la tête.

Estimation à partir de vidéos

Nous avons réalisé également des expérimentations sur les vidéos car nous envisageons d'utiliser cette approche dans le cadre d'un système permettant de renseigner en continu sur la pose d'une personne face à une caméra. Pour réaliser nos tests nous avons utilisé le corpus BUFT. Dans cette expérimentation, nous considérons à la fois le lacet et le roulis.

Boston University Face Tracking (BUFT) Le corpus BUFT est un corpus vidéo enregistré en basse résolution (320×240 pixels) contenant 45 séquences enregistrées par 5 personnes effectuant une

série de 9 mouvements. Dans ces expérimentations, pour l'estimation du roulis nous nous servons de l'inclinaison de l'axe de symétrie et pour le lacet nous avons utilisé l'arbre de décision appris sur la base FacePix car il couvre mieux l'espace de poses comparativement à l'arbre appris sur la base CMU-PIE. Pour rappel, les poses discrètes sur CMU-PIE sont séparées d'approximativement $22,5°$, alors que sur FacePix nous avons pu construire un ensemble d'apprentissage plus précis. Nous réalisons un prétraitement en réalisant une mise à l'échelle similaire entre les visages des deux bases. Les meilleurs résultats pour le lacet ont été obtenus en utilisant le modèle appris sur 19 poses discrètes en considérant un pas de $5°$ entre poses. Nous obtenons $5,24°$ d'erreur moyenne absolue (*mean absolute error - MAE*), et $6,80°$ d'erreur moyenne quadratique (*root mean square error - RMSE*) avec une déviation standard (*standard deviation - STD*) de $4,33°$. Les résultats sont présentés dans la Table 1.3.

TABLE 1.3 – *Résultats d'estimation de pose sur la base BUFT en utilisant les arbres de décision appris sur le corpus FacePix.*

	RMSE (°)	MAE (°)	STD (°)
Roulis	4,39	2,57	3,56
Lacet (model FacePix - 5° intervalle)	7,60	5,12	5,62
Lacet (model FacePix - 15° intervalle)	6,80	5,24	4,33

Afin d'améliorer le temps de calcul et les performances du système, nous exploitons le fait que les images successives vont correspondre à des axes de symétrie relativement proches. Ainsi, nous commençons par estimer l'orientation de l'axe de symétrie en cherchant un axe ayant une inclinaison similaire à celui précédemment détecté. Afin d'éviter toute déviation vers une solution dénaturée, nous relançons le calcul initial une fois toutes les 10 images (environ une fois par seconde).

Positionnement par rapport à l'état de l'art

Nous comparons les résultats obtenus avec d'autres approches de l'état de l'art essentiellement sur les corpus FacePix et BUFT. Nous séparons le positionnement des résultats entre les corpus images et le corpus vidéo. Les Tables 1.4 et 1.5 montrent les performances obtenues respectivement sur les corpus FacePix et BUFT, mesurées en erreur moyenne absolue (*MAE*), erreur quadratique moyenne (*RMSE*), écart type (*STD*) et précision. Ces résultats publiés dans (Dahmane et al. 2015)[7] montrent que notre méthode fournit des résultats comparables avec d'autres propositions de l'état de l'art. Sur le corpus FacePix, les méthodes de type *manifold embedding* fournissent de bons résultats. Toutefois, ces méthodes ne présentent pas un haut niveau d'automatisation et de généralisation comme c'est le cas pour notre proposition. De nouvelles données peuvent être classifiées en s'appuyant sur les modèles créés à partir de données indépendantes. À ce titre, les résultats expérimentaux sur le corpus BUFT ont été obtenus en utilisant les modèles entraînés sur le corpus FacePix. Notons cependant que les résultats que nous avons obtenus pour l'estimation de roulis sur le corpus BUFT présentent un écart-type d'erreur (STD) supérieur à l'erreur moyenne (MAE). Ce fait est également souligné par l'écart important entre les valeurs MAE et RMSE obtenues dans le cadre de cette expérimentation. Cela montre que l'estimation du roulis à partir de l'axe de symétrie présente des erreurs plus importantes que la moyenne sur certaines vidéos. Cette faiblesse est en

7. A. Dahmane ; S. Larabi ; I.M. Bilasco ; C. Djeraba - Head pose estimation based on face symmetry analysis - Signal, Image and Video Processing (SIViP), 2014, pp 1-10 (Facteur d'impact : 1,643 selon JCR 2018).

partie partagée également par (Valenti et al. 2012) au regard de la faible différence entre les valeurs RMSE et STD rapportées. Cela est en partie dû à la nature de vidéos composant le corpus BUFT qui présente une grande variabilité en termes de roulis, lacet et tangage.

TABLE 1.4 – *Comparaison des résultats d'estimation du lacet sur le corpus FacePix.*

Méthode	Résolution	MAE (°)	Précision (%)
(Hao et al. 2011) (Regression)	60×60	6,1	-
(Liu et al. 2010) *(K-manifold clustering)	16×16	3,16	-
(Balasubramanian et al. 2007) *(Biased Isomap)	32×32	5,02	-
(Balasubramanian et al. 2007) *(Biased LLE)	32×32	2,11	-
(Balasubramanian et al. 2007) (Biased LE)	32×32	1,44	-
Notre méthode	80×80	3,14	79,63

* Le principal défaut des techniques d'apprentissage sur les variétés est l'absence de généricité du modèle appris, mis en difficulté en présence de nouvelles données.

TABLE 1.5 – *Comparaison des résultats d'estimation du lacet sur le corpus BUFT.*

		RMSE (°)	MAE (°)	STD (°)
(Valenti et al. 2012)	Lacet	6,10[a]	-	5,79[a]
	Roulis	3,00[a]	-	2,82[a]
(Morency et al. 2010)	Lacet	-	4,97	-
	Roulis	-	2,91	-
Notre méthode	Lacet	6,80	5,24	4,33
	Roulis	4,39[b]	2,57[b]	3,56[b]

[a] Les positions des yeux sont utilisées, ainsi la pose est estimée uniquement si les yeux sont détectés.

[b] Le roulis est estimé en présence de vues frontales.

Le principal avantage de notre méthode est sa capacité à calculer la pose dès la première image présentée, sans passer par une phase d'initialisation ou de calibrage. Ceci est rendu possible par le fait que l'axe de symétrie peut être automatiquement détecté pour les poses comprises entre -45° et +45°. Les caractéristiques de la région de symétrie ainsi que l'arbre de décision permettent de généraliser le problème. En effet, les tests réalisés dans un contexte inter-corpus (FacePix pour l'apprentissage et BUFT pour les tests) montrent la capacité du modèle à répondre de manière satisfaisante lorsque de nouvelles images lui sont présentées. Nous avons ainsi montré que le calcul de la symétrie se maintient à des niveaux satisfaisants en présence de différents défis tels que : les conditions de lumière, les expressions, l'identité de la personne, les occultations locales (yeux fermés, joue cachée par la main), les basses résolutions, etc.

1.3 Approche à base de transformation inverse

Les détecteurs de visages sont souvent utilisés en amont du processus d'estimation de l'orientation du visage. Selon qu'ils réalisent la détection sur une image (Viola et Jones 2004) ou sur une séquence d'images (Kapfer et Benois-Pineau 1997), les détecteurs présentent certaines limitations (rotations dans le plan, vitesse de déplacement, etc.). Ces limitations peuvent être exploitées afin de caractériser le visage. Par exemple, dans un contexte statique, en exploitant les limitations liées aux rotations dans le plan et hors-plan d'un algorithme de détection de visages frontaux tels que l'algorithme de Viola et Jones (2004), nous étudions des solutions pour estimer de manière robuste le roulis et le lacet de la tête. Selon Viola et Jones (2004), les performances de ce détecteur frontal (que nous dénommons VJ dans la suite de cette section) baissent lorsque la tête présente un roulis supérieur à $\pm 15\,^{\circ}$. Autrement dit, en présence de roulis, un visage est détectable approximativement sur

30 ° des 360 ° par le détecteur de visages frontaux VJ. Ce comportement est illustré dans la Figure 1.10 où toutes les fenêtres de détection de visages à différentes échelles sont représentées par les rectangles rouges. Le détecteur de visages VJ s'appuie sur ces fenêtres pour calculer les régions correspondant à de véritables visages. Dans cet exemple, une image contenant un visage frontal subit des rotations par rapport au centre de l'image. Le nombre de rectangles localisant potentiellement le visage croît lorsque la rotation appliquée au visage s'approche de la pose frontale. La rotation dans le plan a des effets négatifs sur le processus de détection de visages, car indépendamment du sens de rotation, le nombre de rectangles candidats décroît en présence de rotations importantes dans le plan. Le même comportement est décelable lorsque nous appliquons une rotation hors-plan de type lacet.

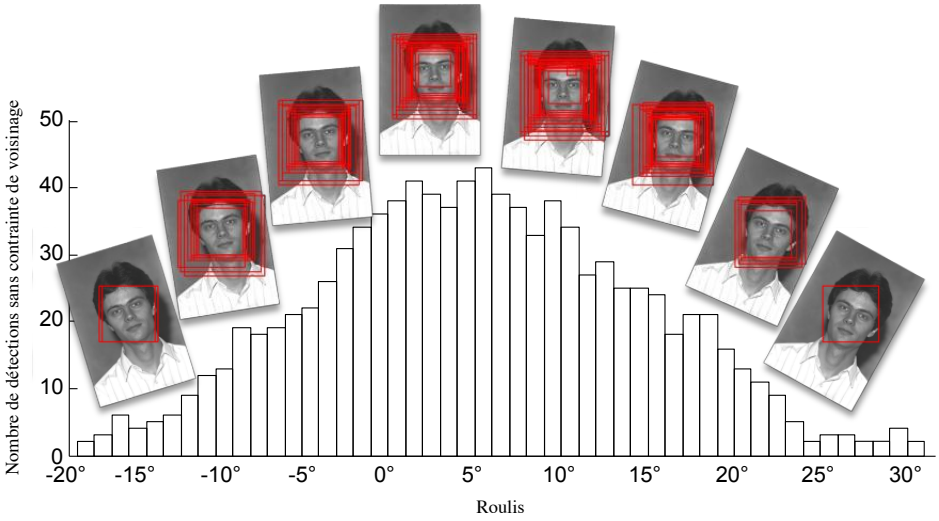


FIGURE 1.10 – *Variations du nombre de détections de visages par VJ en considérant différents angles de roulis sans contrainte de voisinage. Chaque boîte illustre le nombre de détections obtenu pour différentes échelles et pour différents degrés de rotation.*

En s'appuyant sur ces observations, nous appliquons une séquence de transformations (rotation dans le plan ou hors-plan, respectivement) et étudions le nombre de réponses positives. La transformation induisant un comportement optimal de détecteur s'apparente à l'opposé de la rotation subie initialement par la tête. Même si le point commun entre l'estimation du lacet et du roulis réside dans la dualité détection-transformation, nous appliquons des stratégies différentes afin de refléter les spécificités de rotations subies (hors-plan, dans le plan). Les rotations dans le plan préservent la géométrie du visage et, en appliquant des rotations successives dans le plan, nous conservons l'aspect général du visage. Lorsque nous considérons les rotations hors-plan, la difficulté consiste en l'application d'une rotation 3D à une image 2D tout en conservant au mieux la géométrie du visage. Dans ce but, nous extrayons la région de la tête que nous projetons sur un modèle 3D sphérique et effectuons la rotation dans l'espace 3D. Nous relançons le détecteur VJ sur les projections 2D des visages ainsi générées.

En d'autres mots, indépendamment du visage analysé, nous appliquons un ensemble de rotations dans le plan et hors-plan sur ce visage et nous étudions le comportement d'un détecteur frontal en présence de ces rotations. L'angle inverse de la rotation ayant généré une projection 2D conforme à un certain nombre de critères (par exemple, le plus grand nombre de visages détectés) constitue un bon candidat pour le roulis et le lacet du visage initial. L'idée principale de

cette approche n'est pas de caractériser l'orientation en soi. Par exemple, à aucun moment nous n'envisageons d'extraire des descripteurs caractérisant le visage. En revanche, nous recherchons la transformation inverse permettant de ramener le visage dans un état proche de l'état optimal attendu par le détecteur.

Dans la suite de cette section, nous détaillons les processus d'estimation du roulis. Compte tenu des nombreuses similarités entre le processus d'estimation du roulis et du lacet, nous présentons dans ce manuscrit uniquement l'estimation du roulis. Des détails concernant l'estimation du lacet peuvent être retrouvés dans Mennesson et al. (2016). Ensuite, nous présentons les expérimentations conduites pour valider l'approche d'estimation de l'orientation par transformation inverse.

1.3.1 Estimation du roulis par transformation inverse

Dans cette section, nous nous intéressons uniquement à l'estimation du roulis en utilisant une technique par transformation inverse. Dans cette première partie de l'étude nous démontrons que l'algorithme de détection de visages VJ peut être utilisé pour estimer de manière fine l'orientation du visage en termes de roulis sans nécessiter une initialisation précise ou des méthodes spécifiques de suivi.

Nous proposons un algorithme d'estimation structuré en deux niveaux. La Figure 1.11 montre le diagramme de processus de notre algorithme. L'algorithme démarre par une détection grossière des visages frontaux en appliquant de manière itérative le détecteur frontal en considérant différentes orientations. Ensuite, le deuxième niveau exploite les détections grossières provenant du premier niveau afin d'identifier les angles de rotation minimum et maximum qui, appliqués aux visages, permettent de maintenir les détections. La valeur moyenne des angles reflète la valeur de roulis du visage initial. Lors de cette opération, nous calculons également un indice de confiance en relation avec la continuité de ces multiples détections.

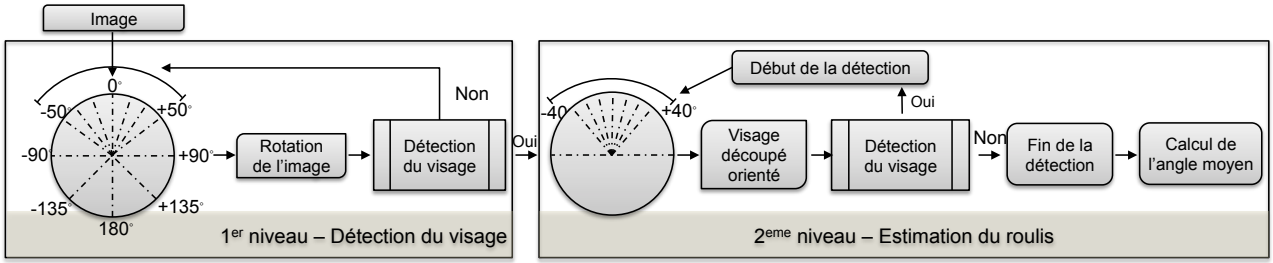
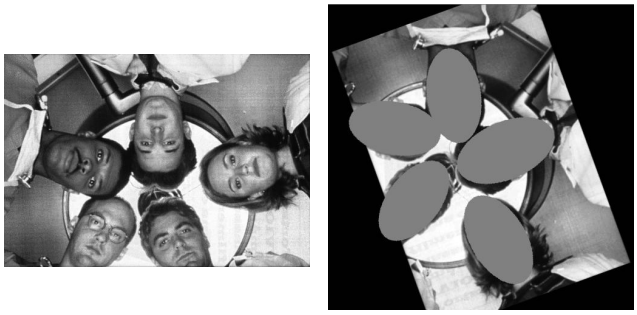


FIGURE 1.11 – *Diagramme de flux de processus de l'algorithme proposé.*

Premier niveau - Sélection des visages candidats Le premier niveau s'intéresse à la détection grossière des visages. Nous considérons différentes orientations reflétant des roulis importants (0 °, 90 °, 135 °, 180 °, −135 ° et −90 °) ainsi que des orientations modérées (de −50 ° à +50 ° avec un pas de 15 °). Sur les visages ainsi détectés, nous superposons des rectangles noirs et la détection reprend jusqu'à ce que l'ensemble des visages soit détecté (voir Figure 1.12).

1.3.2 Évaluation de l'estimation du roulis par transformation inverse

Dans cette section nous revenons sur la méthodologie d'évaluation, l'étude des paramètres et les résultats expérimentaux obtenus pour l'estimation du roulis en utilisant la transformation inverse.



a) b)

FIGURE 1.12 – *a) Image originale extraite depuis le corpus CMU Rotated ; b) Visages validées dans le premier niveau (état intermédiaire).*

Afin de démontrer l'efficacité et la généricité de nos approches, nous réalisons des expériences quantitatives sur plusieurs corpus captés en conditions contrôlées et non contrôlées. Les corpus retenus incluent : CMU Rotated (Rowley et al. 1998b), Boston University Face Tracking (BUFT) (La Cascia et al. 2000), Caltech (Weber 1999), FG-NET Aging (Lanitis 2000), BioID (Jesorsky et al. 2001), Manchester Talking-Face (Cootes 2004) et YouTube Faces (Wolf et al. 2011). Ces corpus ont été sélectionnés car ils contiennent les annotations concernant la position des yeux. En ce qui concerne le corpus YouTube Faces, les annotations fournies n'ont pas été réalisées manuellement et certaines images présentent des erreurs importantes. Nous avons annoté mannuellement les 4000 images contenant les plus importantes erreurs d'annotation. Nous avons ainsi constitué un nouveau sous-ensemble YouTube Faces C qui est utilisé dans nos expérimentations. Les résultats de ces expériences ont été publiés dans (Danisman et al. 2015)[8].

La position des yeux permet d'estimer de manière directe la vérité-terrain en ce qui concerne la rotation dans le plan de la tête. Nous avons choisi d'inclure des corpus qui n'ont pas été principalement conçus pour étudier la pose de la tête. Cela permet de montrer la généricité de notre approche et sa capacité d'estimer le roulis en conditions proches d'un usage non-contraint (où la pose comporte également des variations légères en termes de lacet et/ou tangage). Les informations concernant les rotations dans le plan présentes dans ce corpus sont synthétisées par la moyenne absolue des rotations (RMA), décrite dans l'Équation 1.3 :

$$RMA = \frac{1}{n}\sum_{k=1}^{n} |(\theta_{A_k})| \tag{1.3}$$

où n est le nombre de visages détectés dans le corpus et θ_A est la vérité-terrain concernant le roulis. Les corpus quasi-frontaux ont une RMA proche de zéro, alors que les corpus plus difficiles, ayant une plus large palette de rotations dans le plan, ont une RMA supérieure. La Table 1.6 synthétise les caractéristiques des corpus utilisés dans les expériences. La variété des caractéristiques de corpus sélectionnés garantit une validation poussée de notre méthode pour une large palette de configurations.

Pour chaque corpus, nous calculons le roulis en considérant l'angle θ_A par rapport à l'axe vertical en utilisant la pente m du segment reliant les yeux gauche et droit. La Figure 1.13 montre un exemple extrait du corpus CMU Rotated avec les angles θ_A annotés manuellement. La différence

8. T. Danisman ; I.M. Bilasco - In-plane face orientation estimation in still images - Multimedia Tools and Applications, Springer Verlag, 2016, 75 (13), pp.7799-7829 (Facteur d'impact : 1,541 selon JCR 2018).

TABLE 1.6 – *Synthèse des caractéristiques des corpus utilisés. C=Contrôlé, F=Frontal, VT=Vérité-terrain, VM=Visages multiples faces, R=Rotations, NC=Non-Contraint, RMA = Rotation moyenne absolue.*

Corpus	Type	Images	Visages	RMA	Roulis (étendue)
BioID (Jesorsky et al. 2001)	C, F,VT	1521	1521	$2,20\,^\circ$	$[-32,6\,^\circ + 14,8\,^\circ]$
BUFT (La Cascia et al. 2000)	C, R,VT	8955	8955	$7,03\,^\circ$	$[-38,7\,^\circ + 39,8\,^\circ]$
Caltech (Weber 1999)	C, F,VT	450	450	$2,03\,^\circ$	$[-11,6\,^\circ + 13,2\,^\circ]$
CMU (Rowley et al. 1998b)	U, R,VM,VT	50	223	$35,60\,^\circ$	$[-101,4\,^\circ + 180\,^\circ]$
FG-NET (Lanitis 2000)	C, F,VT	1002	1002	$5,07\,^\circ$	$[-32,0\,^\circ + 29,4\,^\circ]$
Manchester (Cootes 2004)	C, F,VT	5000	5000	$5,03\,^\circ$	$[-23,6\,^\circ + 11,5\,^\circ]$
YouTube Faces C (Wolf et al. 2011)	U, R,VM,VT	4000	> 4000	$8,86\,^\circ$	$[-32,4\,^\circ + 29,0\,^\circ]$

entre l'angle annoté θ_A et l'angle détecté θ_D correspond à l'erreur $\theta_{Err} = \theta_D - \theta_A$, par rapport à l'axe vertical.

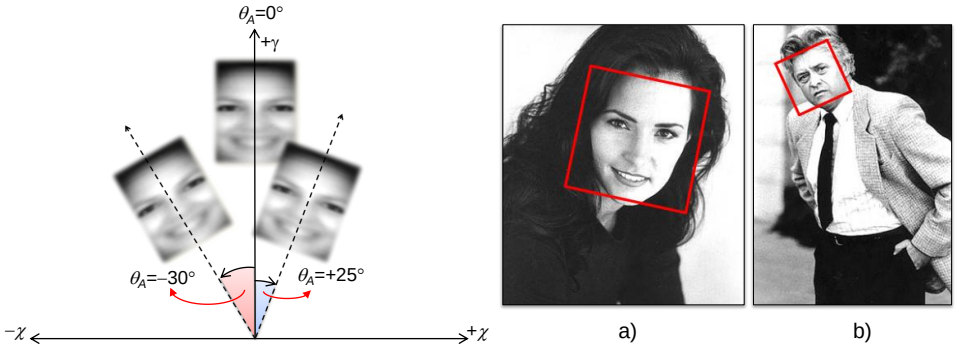


FIGURE 1.13 – *Exemple de vérité-terrain calculé sur le corpus CMU Rotated : a) $\theta_A = +12,63\,^\circ$ b) $\theta_A = -26,56\,^\circ$.*

Afin de fournir une évaluation quantitative de l'estimation du roulis, nous présentons nos résultats en termes d'erreur moyenne quadratique $RMSE$ (*root mean square error*), d'erreur moyenne absolue MAE (*mean absolute error*) et d'écart type standard STD (*standard deviation*).

Comme CMU Rotated et YouTube Faces C sont des corpus ayant des images contenant plusieurs visages, nous employons un appariement des visages en nous appuyant sur la distance euclidienne entre les visages annotés et les visages détectés. Premièrement, nous trouvons le centre du segment entre l'oeil gauche et l'oeil droit dans le visage annoté représenté par Ct_A. Ensuite, nous trouvons le centre de gravité du visage détecté Ct_D. ΔCt est la distance euclidienne entre Ct_A et Ct_D. Nous utilisons la distance inter-pupillaire (IPD) comme un seuil pour apparier les visages annotés et détectés. IPD est la distance euclidienne entre les deux yeux fournis par les annotations de la vérité-terrain. Finalement, un visage détecté est apparié à un visage de la vérité-terrain si $\Delta Ct \leq 2,0 \times IPD$, comme indiqué dans l'Équation 1.4 :

$$\text{Visage}_\text{i} = \begin{Bmatrix} \text{Apparié} & \Delta Ct_i \leq 2,0 \times IPD_i \\ \text{Non Apparié} & \text{sinon} \end{Bmatrix} \tag{1.4}$$

Nous présentons dans la Table 1.7 les résultats de l'estimation du roulis (en considérant les résultats trame par trame pour les corpus vidéo : BUFT et YouTube Faces). La valeur moyenne de MAE pour l'intégralité des sept corpus est de $2,16\,^\circ$. Les deux plus petites valeurs MAE ($1,13\,^\circ$ et $1,69\,^\circ$) et STD ($0,91\,^\circ$ et $1,18\,^\circ$) sont obtenues sur les corpus Caltech et Manchester, respectivement.

La valeur maximale pour MAE ($3,50\,^\circ$) est obtenue pour le corpus YouTube Faces C. Ici la difficulté réside dans la large palette de conditions d'illumination, la taille réduite des visages et la

TABLE 1.7 – *Résultats expérimentaux pour l'estimation du roulis en termes de RMSE, MAE et STD pour différents corpus.*

Corpus	RMSE	MAE	STD
BioID	$2,73\,^\circ$	$2,20\,^\circ$	$1,61\,^\circ$
BUFT	$2,68\,^\circ$	$1,93\,^\circ$	$1,83\,^\circ$
Caltech	$1,46\,^\circ$	$1,13\,^\circ$	$0,91\,^\circ$
CMU	$3,75\,^\circ$	$2,87\,^\circ$	$2,42\,^\circ$
FG-NET	$2,63\,^\circ$	$1,86\,^\circ$	$1,86\,^\circ$
Manchester	$2,07\,^\circ$	$1,69\,^\circ$	$1,18\,^\circ$
YouTube Faces C	$5,59\,^\circ$	$3,50\,^\circ$	$4,36\,^\circ$

présence de tangages et de lacets de grande ampleur (avoisinant $30\,^\circ$). Ces deux derniers facteurs (lacets, tangages) impactent de manière assez forte la stabilité de notre estimation. Certaines erreurs d'estimation obtenues sont importantes. Il en résulte un écart-type standard (STD) élevé. Cette constatation peut être faite également dans le cadre du corpus CMU présentant d'importantes variations dans l'orientation dans et hors-le plan des visages. Par ailleurs, la taille réduite des visages induit également un biais dans l'évaluation, car, une erreur d'un pixel dans l'annotation de la position verticale (valeur d'y) de l'oeil gauche ou de l'oeil droit dans un visage large de 10 pixels génère une erreur d'approximativement $5,71\,^\circ$. Ainsi, la largeur du visage par rapport à la résolution de l'image est un facteur important pour les corpus annotés manuellement. La Figure 1.14 montre quelques détections et des erreurs d'estimation du corpus YouTube Faces C où le segment blanc correspond à la vérité-terrain. Quelques exemples problématiques (par exemple, les cinq premières colonnes des deux dernières lignes, entourées par un rectangle jaune) correspondent à des incidents singuliers et sont corrigés dans les trames suivantes. Ce problème apparaît lorsque le premier niveau fournit une fausse détection incluant toutefois une partie du visage réel. Les erreurs mesurées dans ces exemples sont également inclues dans le calcul global des métriques *MAE*, *RMSE* et *STD*.

FIGURE 1.14 – *Bonnes (cinq premières lignes) et mauvaises (deux dernières lignes) estimations dans le corpus YouTube Faces C. La ligne blanche indique la véritable orientation annotée manuellement.*

Conformément aux résultats, notre méthode réduit la *MAE* à moins de $2,0\,^\circ$ pour les corpus frontaux (par exemple, BioId, Caltech, Manchester) ayant une *RMA* (rotation moyenne) comprise

entre $2,03\,^\circ$ et $5,03\,^\circ$ par rapport à l'axe vertical. Dans le cas de corpus comportant d'importantes rotations dans le plan, l'ensemble des valeurs de $RMSE$ est inférieures à $3,0\,^\circ$ à l'exception du CMU. Néanmoins, en comparaison d'autres corpus, CMU a la RMA la plus importante ($35,60\,^\circ$), tout en comportant un nombre réduit d'images et de visages (voir Table 1.6). La Figure 1.15 (a) montre que notre méthode fonctionne bien indépendamment du roulis. La MAE reste stable (parallèle à l'axe horizontal) sur l'ensemble des corpus. La Figure 1.15 (b) illustre la couverture θ_{Err} pour chaque corpus. En moyenne, notre méthode couvre à $91,69\%$ et $95,14\%$ les corpus de tests avec un $\theta_{Err} = \pm 5,0\,^\circ$ et $\theta_{Err} = \pm 6,0\,^\circ$ respectivement.

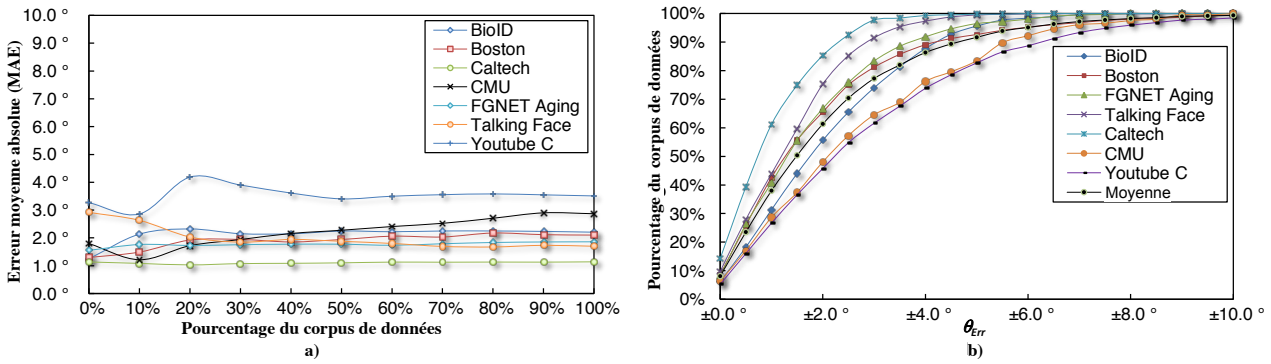


FIGURE 1.15 – *a) Variations en termes de $MAE(\hat{\theta})$ dans le corpus b) Couverture d'erreur angulaire θ_{Err}.*

La Table 1.8 compare l'estimation du roulis sur le corpus BUFT avec d'autres méthodes de l'état de l'art en termes de $RMSE$, MAE et STD. Les performances de notre méthode en termes de MAE sont comparables à celles communiquées dans (Xiao et al. 2002) et (Lefevre et Odobez 2010). Notons qu'à aucun moment nous n'avons utilisé l'information temporelle pour affiner nos estimations. Par ailleurs, nous remarquons également que malgré le fait que la méthode de Tran et al. (2013) présente un écart-type standard (STD) plus réduit que le notre, nous obtenons une erreur moyenne absolue plus réduite. Cela traduit le fait que globalement notre méthode est plus performante en moyenne, mais elle présente également une plus grande instabilité à certaines conditions d'estimation de roulis par rapport à la moyenne.

TABLE 1.8 – *Comparaison de l'estimation du roulis sur le corpus BUFT en termes de RMSE, MAE et STD.*

Méthode	RMSE	MAE	STD
AAM + cylinder head model (Sung et al. 2008)	-	$3,10\,^\circ$	-
3D ellipsoidal head model (An et Chung 2008)	-	$2,83\,^\circ$	-
3D cylinder head model (An et Chung 2008)	-	$3,22\,^\circ$	-
2D planar model (An et Chung 2008)	-	$2,99\,^\circ$	-
Fixed templates and eye cue (Valenti et al. 2009)	$3,00\,^\circ$	-	$2,82\,^\circ$
Updated templates and eye cue (Valenti et al. 2009)	$3,93\,^\circ$	-	$3,57\,^\circ$
Distance vector fields (Asteriadis et al. 2010)	$3,56\,^\circ$	-	-
Generalized Adaptive View-based Appearance Model (Morency et al. 2010)	-	$2,91\,^\circ$	-
Support Tensor Regression (Guo et al. 2011)	-	$5,60\,^\circ$	-
Support Vector Regression (Guo et al. 2011)	-	$5,80\,^\circ$	-
3D face model + Eigen-Decomposition Based Bayesian Approach (Tran et al. 2013)	-	$2,40\,^\circ$	$1,40\,^\circ$
3D deformable face tracking (Lefevre et Odobez 2010)	-	$1,90\,^\circ$	-
3D ellipsoidal head pose model + gait (Jung et Nixon 2012)	-	$2,10\,^\circ$	-
Adaptive correlation filter + Viola-Jones detector (My et Zell 2013)	-	$3,53\,^\circ$	-
Dynamic templates and re-registration (Xiao et al. 2002)	-	$1,40\,^\circ$	-
Notre méthode (avec un pas de rotation = 3)	**2,89** $^\circ$	**2,10** $^\circ$	**1,99** $^\circ$
Notre méthode (avec un pas de rotation = 1)	**2,68** $^\circ$	**1,93** $^\circ$	**1,83** $^\circ$

La Figure 1.16 montre les images ayant le θ_{Err} le plus grand (première ligne) et le plus petit (deuxième ligne) pour l'estimation du roulis. Les pires erreurs d'estimation se produisent lorsque

le visage présente un lacet de plus de 30° (voir Figure 1.16 (e)(f)(g)(h)). Davantage d'exemples extraits des corpus BUFT et CMU peuvent être trouvés dans les Figures 1.17 et 1.18.

FIGURE 1.16 – *Exemples des meilleurs (première ligne) et pires (deuxième ligne) résultats obtenus dans le corpus BUFT. Les rectangles orientés montrent la pose estimée et les lignes vertes montrent la vérité-terrain.*

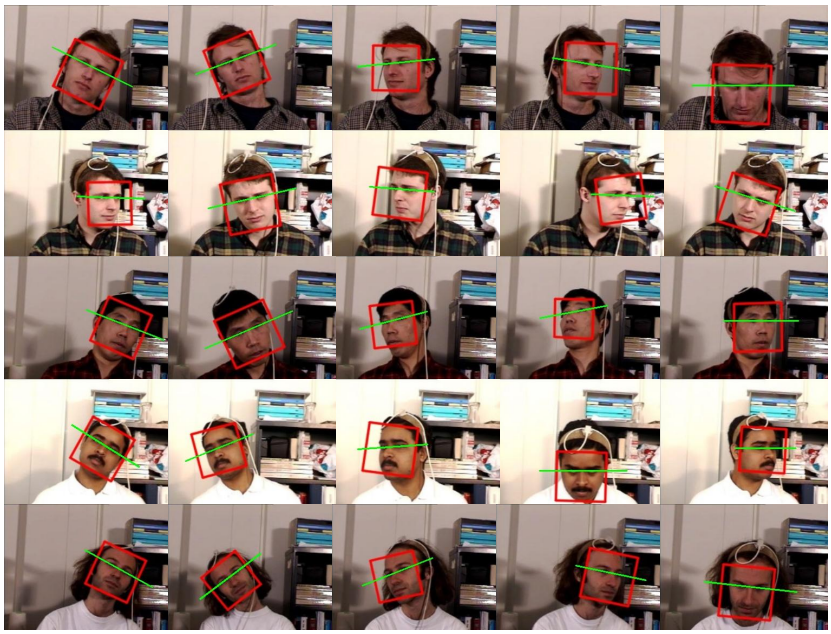


FIGURE 1.17 – *Exemples d'estimation du roulis sur le corpus BUFT.*

1.4 Résumé des contributions

Nous avons présenté deux approches exploitant la symétrie bilatérale et en détournant intelligemment l'utilisation des détecteurs frontaux de visages pour estimer la pose. Les approches proposées obtiennent des résultats comparables à l'état de l'art sur divers corpus.

Symétrie bilatérale Nous avons présenté une approche innovante pour étudier l'orientation de la tête en exploitant la symétrie bilatérale du visage pour l'estimation du lacet et du roulis. L'orientation de l'axe de symétrie bilatérale correspond au roulis, alors que les caractéristiques de l'enveloppe convexe englobant la symétrie servent à estimer le lacet. La symétrie bilatérale ainsi que ses caractéristiques peuvent être extraites sans passer par la détection de points caractéristiques sur le visage. Aucune calibration ou initialisation du système en situation de pose frontale n'est nécessaire. Les résultats obtenus sur les corpus publics communément utilisés pour ces tâches montrent des performances équivalentes aux autres méthodes de l'état de l'art. En plus, au regard des tests

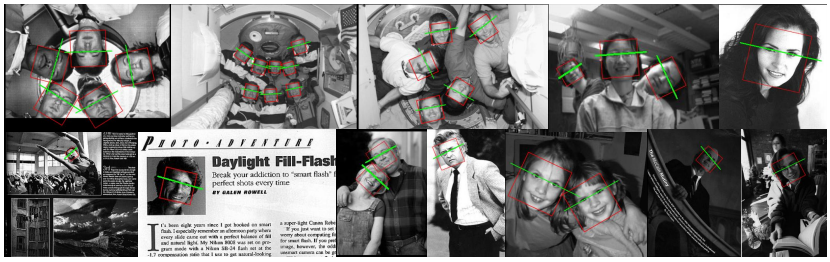


FIGURE 1.18 – *Exemples d'estimation du roulis sur le corpus CMU Rotated.*

inter-corpus FacePix vs BUFT, nous soulignons que notre système est moins contraignant et il a une capacité de généralisation plus forte.

Cette méthode a été validée sur différents corpus (FacePix, CMU-PIE, BUFT), répondant ainsi à des situations d'estimation en présence de variations d'illumination ou d'expressions faciales. Les résultats démontrent que les changements en taille et en axe de symétrie de la région contenant la symétrie bilatérale du visage reflètent les changements de la pose et permettent de l'inférer de manière assez précise.

Transformation inverse Nous avons proposé deux algorithmes pour l'estimation du roulis et du lacet en se servant essentiellement de l'algorithme standard de détection de visages frontaux VJ. Nous avons utilisé le même algorithme de détection pour extraire une évaluation grossière et, dans un deuxième temps, une estimation fine de la pose. Nous avons montré que le roulis et le lacet peuvent être estimés convenablement en utilisant un détecteur de visages frontaux.

Les nombreuses expérimentations conduites montrent que le roulis peut être estimé à l'aide du détecteur frontal VJ avec une précision comparable à d'autres méthodes d'estimation plus complexes de l'état de l'art. Encouragé par ces résultats concernant le roulis, nous avons poursuivi l'étude de l'estimation de la pose à l'aide d'un détecteur frontal en considérant les rotations hors-plan, et notamment le lacet. Nous avons ciblé le lacet car les transformations qu'il apporte en termes de projection 2D de visages répondent de manière différentiée à un détecteur frontal. Ce n'est pas le cas du tangage qui permet de conserver sur des plages de rotations assez étendues les éléments principaux du visage.

En comparaison avec d'autres études, la principale contribution de cette étude est d'utiliser le même détecteur frontal du visage pour la détection et pour l'estimation à grain fin de l'orientation. Ainsi, dans notre approche, l'estimation de la rotation dans le plan est un résultat immédiat du processus traditionnel de détection de visages sans aucune méthode complémentaire. Des modèles de suivi plus complexes (les modèles à base de suivi 3D de la tête ou des modèles cylindrique ou ellipsoïde) pourraient bénéficier de l'information temporelle. Toutefois, ils nécessitent une initialisation précise et des mécanismes de réenregistrement et suivi. Les occultations, la disparition de certains points suivis et l'accumulation des erreurs dans le domaine temporel sont tout autant de problèmes qui demanderaient de réinitialiser le processus d'estimation. Par ailleurs, ces méthodes ne peuvent être appliquées que sur des données ayant une dimension temporelle.

En se rapportant aux lignes directrices proposées par Murphy-Chutorian et Trivedi (2009), les deux méthodes d'estimation proposées peuvent être considérées comme étant précises, autonomes, monoculaires, multi-personnes, indépendantes de l'identité de la personne et de la résolution de l'image.

Reconnaissance du genre 2

Dans ce chapitre nous nous intéressons à la reconnaissance du genre qui est une branche de la biométrie douce [1]. La reconnaissance du genre est une composante nécessaire à la réalisation d'études démographiques portant sur le genre, l'âge et l'ethnie. Les résultats de ces études peuvent servir dans des domaines d'application divers (par exemple, marketing, sécurité ou santé). La reconnaissance du genre s'effectue à partir de modalités hétérogènes : le texte, la parole, l'information visuelle (image ou vidéo). Dans le cas de la modalité visuelle, malgré le fait que l'essentiel de la recherche se concentre sur le visage, des études concernant le corps tout entier, où la marche des personnes sont également disponibles dans la littérature. Toutefois, par la suite, nous nous concentrons sur la reconnaissance du genre à partir d'images 2D, facilement généralisable aux flux vidéo 2D.

La reconnaissance faciale du genre n'est pas une tâche triviale et hérite de certains défis de l'analyse du visage en général, tels que : les variations d'illumination, les changements d'orientation de tête, les occultations. En effet, il existe de nombreux facteurs qui affectent le processus de reconnaissance. Le premier groupe de facteurs correspond aux facteurs externes affectant le processus de capture : illumination, résolution de la caméra, perspective. Le deuxième groupe de facteurs découlent de l'humain : variations dans l'orientation de la tête, l'âge, le maquillage, l'ethnie, les accessoires, les occultations (cheveux, barbe) ou les expressions faciales. En effet, d'importantes et nombreuses variations intra-classe, parmi les sujets hommes et femmes, rendent la tâche de classification ardue.

Les modèles appris dans des contextes démographiques circonscrits ne peuvent pas répondre à la grande variabilité des visages, autres que ceux de ces corpus. L'ensemble d'apprentissage issu de ces corpus ne représente qu'une toute petite partie de l'espace du problème (7 milliards de personnes). Ainsi, une méthode de reconnaissance du genre robuste doit être apprise sur des corpus de taille conséquente, et disposer de méthodes de prétraitement permettant d'homogénéiser les visages captés en situation non-contrôlée. Dans l'ensemble des travaux que nous avons réalisés, nous avons eu à coeur de privilégier les méthodes capables de répondre à un large panel de données. Pour cela nous adoptons des protocoles de validation inter-corpus. Les expériences que nous proposons sont réalisées sur des corpus de taille conséquente et non-contrôlés tels que Groups (Gallagher et Chen 2009), GENKI-4K (MPLab 2011) et Labelled Faces in the Wild (LFW) (Huang et al. 2007). En plus de ces corpus non-contrôlés, nous avons également inclus dans certaines expériences le corpus FERET (Phillips et al. 1998) capté en conditions contrôlées.

1. La biométrie douce permet de différencier les individus sur la base de leurs traits caractéristiques (par exemple, la couleur des yeux, la forme du visage, le genre), sans toutefois les identifier précisément.

Dans ce qui suit, nous dressons un état de l'art des travaux concernant la reconnaissance du genre. Nous poursuivons par décrire nos deux contributions liées à la reconnaissance du genre en utilisant, dans un premier temps, l'information visuelle seule, ainsi que d'autres indices relatifs à la présence et aux caractéristiques de certains éléments distinctifs (moustache, cheveux, etc.).

Reconnaissance du genre à partir de visages normés En nous appuyant uniquement sur l'information visuelle, nous proposons une solution de reconnaissance dont la généricité est prouvée dans le cadre d'évaluations inter-corpus. Afin d'offrir un cadre de validation encore plus exigeant nous avons collecté un nouveau corpus d'apprentissage non-contrôlé WebDB à partir d'images disponibles sur le web. Notre contribution publiée dans (Danisman et al. 2014) [2] est double. Premièrement, nous montrons, dans le cadre d'une expérimentation inter-corpus, que l'alignement de visages et la normalisation des intensités des pixels offrent de meilleures performances que les méthodes par extraction des descripteurs complexes sur des visages non-alignés. Deuxièmement, nous montrons que la spécification d'histogrammes déjà largement étudiée et utilisée par la communauté dans d'autres domaines (Zhang 1992, Morovic et Sun 2003, Coltuc et al. 2006, Wan et Shi 2007) est une technique de normalisation plus adéquate que l'égalisation d'histogrammes pour la reconnaissance du genre. La technique de spécification suppose une transformation de l'image en respectant une distribution cible reflétant les caractéristiques globales des corpus analysés. Les nombreuses expérimentations inter-corpus réalisées sur les corpus non-contrôlés Groups, GENKI-4K et LFW montrent que notre approche offre des capacités de généralisation supérieures à d'autres méthodes de l'état de l'art. Les nombreuses expérimentations inter-corpus réalisées sur les corpus non-contrôlés Groups, GENKI-4K et LFW montrent que notre approche offre des capacités de généralisation supérieures à d'autres méthodes de l'état de l'art.

Système d'inférence floue à partir d'informations visuelles hétérogènes Nous étudions le gain obtenu en adoptant une approche d'inférence floue en présence de règles concernant d'autres caractéristiques faciales complémentaires telles que la chevelure ou la pilosité intra-faciale (moustache et barbe). Les résultats obtenus dans le cadre d'une validation inter-corpus publiés dans (Danisman et al. 2015) [3] attestent que le système d'inférence floue proposé, obtient de meilleures performances que lors de l'utilisation de l'information fournie par les visages normés uniquement.

2.1 État de l'art

L'enchaînement générique des processus contribuant traditionnellement à la reconnaissance du genre à partir d'images 2D est résumé dans la Figure Fig. 2.1. Dans la littérature de nombreuses techniques de prétraitement, de normalisation ou d'extraction de caractéristiques sont proposées pour faire face aux défis découlant de la grande variabilité intra-classe.

2. T. Danisman ; I.M. Bilasco ; C. Djeraba - Cross-database evaluation of normalized raw pixels for gender recognition under unconstrained settings - Proc. of International Conference on Pattern Recognition 2014, Stockholm, Sweden, pp. 3144-3149.

3. T. Danisman ; I.M. Bilasco ; J. Martinet - Boosting gender recognition performance with a fuzzy inference system - Expert Systems with Applications, Volume 42, Issue 5, 1 April 2015, pp. 2772-2784 (Facteur d'impact : 3,768 selon JCR 2018).

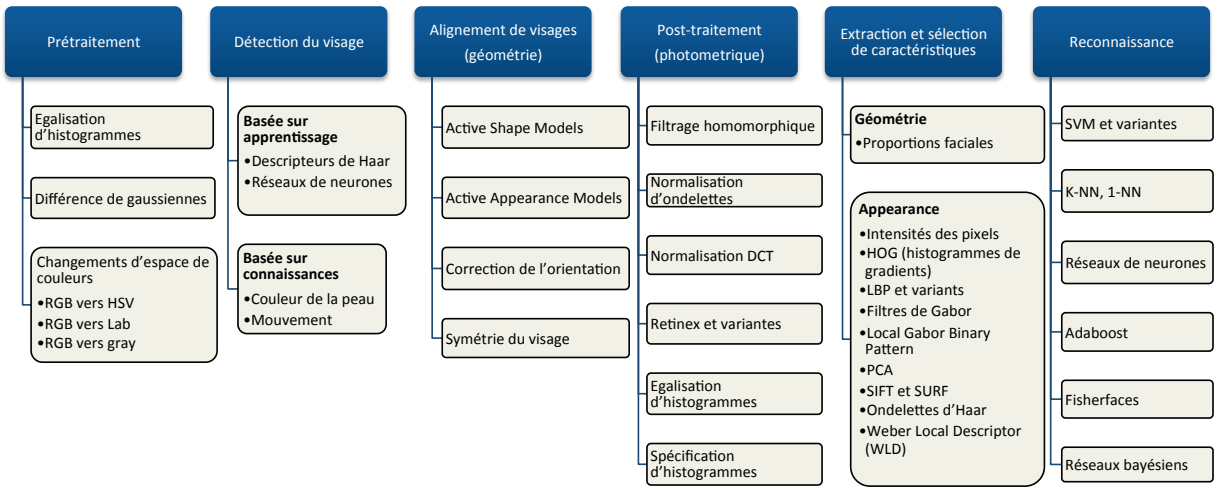


FIGURE 2.1 – *Organisation type du processus traditionnel de reconnaissance du genre à partir d'images 2D.*

Les premiers travaux menés pour la reconnaissance du genre utilisent des méthodes simples de caractérisation de l'apparence du visage en s'intéressant directement à l'intensité des pixels (Golomb et al. 1990, Gutta et al. 2000, Moghaddam et Yang 2002, Walawalkar et al. 2003). Les études de Shan (2012), Santana et al. (2013), Ramon-Balmaseda et al. (2012), Dago-Casas et al. (2011) menées au début des années 2010 s'intéressent à des méthodes employant des descripteurs évolués tels que les histogrammes orientés de gradients ou filtres de Gabor. L'opérateur LBP de Ojala et al. (2002) et ses déclinaisons constituent l'une des familles de descripteurs la plus utilisée. Ng et al. (2012) passent en revue l'ensemble des méthodes pour la reconnaissance du genre à partir d'information visuelle. L'une des conclusions de cette étude est que les classifieurs SVM avec un noyau RBF sont prédominants dans la reconnaissance du genre grâce à leur capacité de généralisation. D'autres méthodes s'appuient sur des classifieurs de type Adaboost, plus proche voisin ou réseau de neurones.

Makinen et Raisamo (2008a) étudient l'impact du prétraitement sur les performances en comparant l'alignement automatique, l'alignement manuel et l'absence d'alignement de visages. Ils mènent leurs études sur le corpus FERET en considérant différentes méthodes de classification de l'état de l'art. Ils concluent que l'alignement automatique n'augmente pas la précision du système et que l'utilisation des niveaux d'intensité des pixels offre de meilleurs taux de classification similaire à l'utilisation des descripteurs LBP. Toutefois, leurs expérimentations ne concernent qu'un petit sous-ensemble du corpus FERET (60 hommes et 47 femmes). Nous pensons que les corpus contrôlés, tels que FERET, et leurs sous-ensembles ne sont pas représentatifs pour répondre de manière fiable au problème de reconnaissance du genre pour les visages inconnus.

De nombreuses études se concentrent sur des techniques de validation croisée sur un même corpus de données. Cela peut produire des validations biaisées à cause des protocoles d'acquisition utilisés. De manière générale, comme indiqué par Bekios-Calfa et al. (2011), les performances retrouvées dans le cadre des expériences intra-corpus sont optimistes. Souvent, les images d'un même corpus partagent des attributs démographiques et des protocoles d'acquisition similaires. Selon Ng et al. (2012), la précision moyenne en termes de reconnaissance du genre sur le corpus contrôlé FERET est de $96 \pm 2,5\%$ notamment lorsqu'une validation de type croisée intra-corpus est employée. Toutefois, ces performances élevées décroissent rapidement lorsque l'on utilise une

validation inter-corpus avec des moyennes autour de $83 \pm 6,7\%$. Plusieurs chercheurs proposent des validations inter-corpus en utilisant les corpus non-contrôlés LFW et Groups. Dans ces corpus non-contrôlés, les performances moyennes baissent encore, atteignant $79,05 \pm 5,6\%$, comme illustré dans la Table 2.1. Ces résultats mettent en évidence la faible capacité de généralisation des solutions de classification proposées dans la littérature lorsque l'on considère des corpus différents. Ainsi, nous estimons que les recherches doivent s'orienter vers des propositions montrant des bonnes capacités de généralisation en présence de conditions d'acquisition non contrôlées et dans un contexte de validation inter-corpus.

TABLE 2.1 – *Synthèse des expériences menées sur les corpus Groups, LFW et Genki-4K.*

Travaux	Apprentissage/Test	Nb. images test	Méthode	Pertinence%
(Dago-Casas et al. 2011)	LFW/Groups[a]	14760	LBP+PCA+LDA	81,02
(Dago-Casas et al. 2011)	LFW/Groups[a]	14760	Pixels+PCA+SVM	72,09
(Dago-Casas et al. 2011)	Groups/LFW	13088	LBP+PCA+SVM	89,77
(Bekios-Calfa et al. 2014)	Groups/LFW	13233	PCA+LDA+KNN	79,11
(Bekios-Calfa et al. 2014)	Groups[b]/LFW	13233	PCA+LDA+KNN	79,53

[a] distance inter-pupillaire ≥ 20 pixels
[b] apprentissage sans visage d'enfant
[c] âge > 12
[d] âge ≥ 20

La plupart des méthodes se concentrent uniquement sur l'intérieur du visage. Les zones entourant le visage sont trop souvent assujetties au bruit. Lorsque ces zones sont considérées directement comme une composante primaire de l'information visuelle, elles nuisent au processus de classification. Ainsi, elles sont souvent ignorées. Cependant, certains travaux de l'état de l'art considèrent une caractérisation à base de descripteurs LBP, de filtre de Gabor ou d'intensité de pixels pour caractériser la pilosité faciale : cheveux, moustache et barbe. Les régions du visage et du cou sont étudiées par Ueki et Kobayashi (2008). Les cheveux et les vêtements du haut de corps sont étudiés par Li et al. (2012). La reconnaissance à base de la région tête-épaule est étudiée par Li et al. (2013a). Tome et al. (2014) s'intéressent à la longueur des bras et des cheveux. L'information contextuelle est exploitée par Satta et al. (2014) pour compléter la caractérisation de l'apparence du visage. Ces méthodes obtiennent de meilleurs résultats que les méthodes classiques de reconnaissance du genre. Ils utilisent des techniques automatiques basées sur des heuristiques concernant la localisation des éléments. Cependant, l'effet précis de l'information contextuelle n'est pas clairement mis en évidence. Les nombreuses expériences réalisées par Makinen et Raisamo (2008b) montrent que la normalisation et l'alignement des visages sont plus importantes que la prise en compte d'information contextuelle apportée par les cheveux.

Le corpus Part Labels (Kae et al. 2013) est le premier à proposer des annotations concernant la pilosité faciale, les cheveux et l'arrière plan en s'appuyant sur l'utilisation de superpixels. L'utilisation de descripteurs concernant l'apparence de l'intérieur du visage peut être combinée avec l'information de haut niveau concernant les caractéristiques de la pilosité faciale et des cheveux (emprise spatiale, localisation). Cela permet d'inclure des éléments de raisonnement de haut niveau dans la prise de décision. Par exemple, généralement, les cheveux d'une femme sont plus longs que les cheveux d'un homme comme illustré dans la Figure 2.2.

Cette connaissance extérieure au processus d'analyse d'images peut constituer une règle complémentaire permettant au système de décision d'améliorer sa prédiction. Même s'il reste difficile

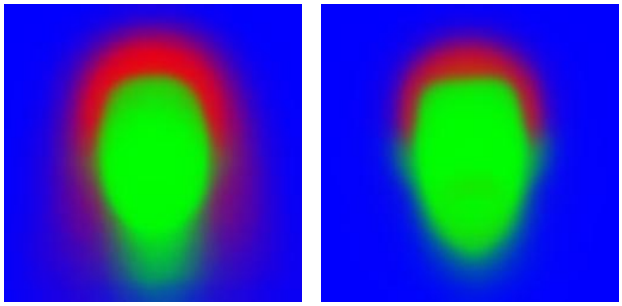


FIGURE 2.2 – *Visages moyens d'une femme et d'un homme obtenus à partir du corpus Part Labels. La couleur rouge indique les cheveux. Le vert correspond à l'intérieur du visage et le cou. Le bleu est associé à l'arrière plan.*

de définir une règle précise qui puisse convenir à l'ensemble des hommes et des femmes, un système d'inférence floue peut prendre avantage de ces connaissances formulées en tant que règles d'expert. Les systèmes d'inférence floue constituent l'une des applications phares de la logique floue. Ces derniers ont été adaptés pour les processus de reconnaissance par des auteurs tels que Melin et al. (2010), Polat et Yildirim (2008), Zadeh (2010). En ce qui concerne, la reconnaissance du genre, plusieurs études ont déjà été menés. Leng et Wang (2008) utilisent les SVM flous afin d'augmenter la capacité de généralisation. Leurs expériences réalisées sur différents corpus confirment l'hypothèse attendue : les classifieurs SVM flous montrent une robustesse plus importante aux variations démographiques. Moallem et Mousavi (2013) utilisent l'information de forme et de texture afin de concevoir un système de décision flou.

Le principal avantage des systèmes d'inférence floue réside dans leur habilité à prendre en compte l'information de haut-niveau et de réaliser des mises en correspondance non-linéaires entre les variables d'entrée et les variables de sortie. Les systèmes d'inférence floue sont construits en intégrant des connaissances provenant des experts et de l'information extraite depuis les données de bas-niveau. Nous pouvons ainsi formuler des règles prenant en compte ces deux éléments afin de résoudre le problème de reconnaissance du genre. L'utilisation de ces deux éléments, connaissances de l'expert et information extraite depuis les données de bas-niveau, est cruciale. En effet, l'utilisation seule de la connaissance de l'expert ne suffit pas pour tenir compte de la variabilité que l'on peut observer parmi les sujets. Cela mène à une solution moins robuste comme illustré dans (Guillaume 2001). La création d'un système d'inférence floue robuste dépend fortement de la structuration et l'optimisation du système pour inclure des données de bas-niveau, des ensembles et des règles floues.

Compte tenu des observations faites dans l'état de l'art, nous avons proposé une première approche qui traite de la reconnaissance du genre en exploitant de manière directe l'intensité des pixels à l'aide d'histogrammes spécifiés. Ensuite, afin de prendre avantage du pouvoir discriminant des règles d'experts, nous avons augmenté cette première solution en considérant des informations de haut niveau concernant la présence de certains attributs intra- et extra-faciaux. Nous utilisons l'information relative à la segmentation de visages en prenant en compte les caractéristiques géométriques, la pilosité faciale et les caractéristiques de la chevelure pour améliorer la reconnaissance du genre.

2.2 Reconnaissance du genre à partir de visages normalisés

Dans cette section nous présentons notre première contribution liée à la reconnaissance du genre. Nous étudions l'impact des processus de normalisation de la géométrie et de l'apparence sur le processus de reconnaissance en utilisant directement l'intensité de pixels comme descripteur principal.

2.2.1 Normalisation de visages

La première étape dans le processus de reconnaissance que nous proposons correspond à une détection de visages et à un prétraitement. Le prétraitement vise à normaliser les visages traités, tant d'un point de vue géométrique que photométrique. Les méthodes de normalisation géométrique et photométrique de visages ont une importance cruciale pour les processus d'analyse faciale. L'information normalisée est plus facilement exploitable par le processus d'apprentissage. En effet, l'analyse est moins assujettie au bruit de capture, ce qui permet d'avoir des données (d'apprentissage et de test) en meilleure adéquation avec les modèles mathématiques utilisés dans la vision par ordinateur.

Normalisation géométrique

Nous commençons le processus de normalisation géométrique (ou alignement) par la détection de visages dans l'image en utilisant un détecteur de visages frontaux tel que celui proposé par Viola et Jones (2004). Ensuite, la détection des yeux est réalisée afin d'être en mesure de corriger le roulis que le visage peut présenter (tête inclinée vers la droite ou la gauche). Afin de localiser les pupilles, nous adaptons le réseau de neurones proposé par Rowley et al. (1998a). Par la suite, pour normaliser les visages, nous considérons la distance inter-pupillaire DIP qui correspond à la distance euclidienne entre les centres des yeux ainsi que la différence entre les ordonnées des pupilles. La différence entre l'ordonnée de la pupille de l'oeil gauche et la pupille de l'oeil droit est utilisée pour remettre le visage en position droite, en supprimant une éventuelle rotation dans le plan de la tête. La région du visage obtenue en appliquant le détecteur de visage est redéfinie en fonction de la position des yeux et de la distance DIP. Nous suivons les équations ci-dessous où V_x, V_y, V_w et V_h représentent les coordonnées de la boîte englobante considérée pour caractériser le visage. Les valeurs $Oeil_{Gauche_x}$ et $Oeil_{Gauche_y}$ correspondent aux positions x et y de la pupille de l'oeil gauche dans le repère global de l'image.

$$V_x = Oeil_{Gauche_x} - DIP/4,0,\ V_y = Oeil_{Gauche_y} - DIP,\ V_w = DIP \times 1,5 \text{ et } V_h = DIP \times 2,5 \quad (2.1)$$

Les valeurs scalaires utilisées dans l'Équation 2.1 sont choisies en fonction de la morphologie moyenne constatée dans les corpus de données utilisés. Ces valeurs permettent de définir une région d'intérêt au sein du visage incluant un maximum d'informations pertinentes pour l'analyse (peau, pilosité faciale, etc.), tout en limitant le bruit amené par la présence d'occultations (avec les cheveux) ou par l'inclusion des éléments de l'arrière plan. Les visages normalisés sont redimensionnés à une taille de 20×24 pixels. La Figure 2.3 montre la région initiale du visage, ainsi que la région obtenue en appliquant les Équations 2.1.

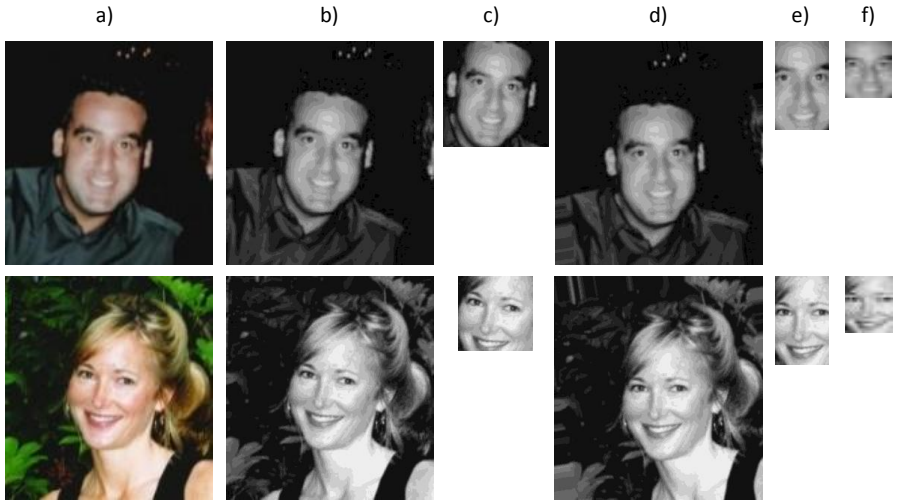


FIGURE 2.3 – *Normalisation géométrique de visages : a) image de départ, b) image en niveau de gris, c) détection de visage, d) détection des yeux et correction de l'orientation, e) définition de la région à analyser et f) mise à l'échelle* 20×24.

Normalisation photométrique

Une normalisation photométrique est appliquée en utilisant le procédé de spécification d'histogrammes afin de répondre aux variations d'illumination qui peuvent être observées dans un contexte non-contrôlé. La spécification d'histogrammes et l'égalisation d'histogrammes sont deux techniques communément utilisées pour améliorer et normaliser la qualité photométrique des images. Le procédé d'égalisation d'histogrammes attribue un nombre de pixels équivalent à l'ensemble des classes de l'histogramme. Toutefois, cette méthode ne prend pas en compte les caractéristiques de la distribution de niveaux de gris sur un visage. La spécification d'histogrammes est une généralisation de l'égalisation d'histogrammes. L'image n'est plus normalisée selon une distribution uniforme des poids entre les différents niveaux de gris, mais plutôt par rapport à une fonction de densité de probabilités (fdp) spécifique au visage. Tenant compte du fait qu'il est possible de calculer la répartition des niveaux de gris sur un visage moyen extrait depuis les corpus disponibles, nous pouvons ainsi proposer une fonction de densité de probabilités qui sert comme base pour normaliser la distribution d'histogramme d'une image. La Figure 2.4 montre l'effet de l'application d'un procédé de spécification d'histogramme sur une image, en considérant comme fonction de distribution de probabilité celle correspondant au visage moyen du corpus WebDB.

Comme illustré dans la Figure 2.4 (c), le nouveau histogramme obtenu est plus proche de l'histogramme du visage moyen et les visages normalisés préservent mieux les caractéristiques faciales. Cette étape de prétraitement permet de réaliser une correction plus adaptée en présence de conditions d'illumination variables.

2.2.2 Normalisation de l'espace de caractéristiques

Suite à ces deux étapes de normalisation, les intensités de pixels composant les visage normalisés de taille 20×24 constituent le vecteur de caractéristiques pour le processus de classification sous-jacent. Graf et Borer (2001) montrent que l'étape de prétraitement est assimilable à l'optimisation de la fonction noyau des classifieurs SVM. Autrement dit, un prétraitement efficace prépare le terrain pour l'identification des paramètres optimaux pour la fonction noyau du classifieur. En

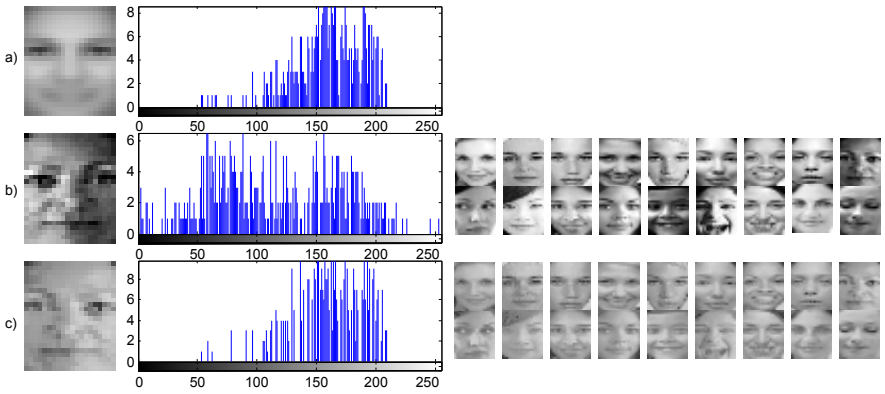


FIGURE 2.4 – *a) Image moyenne de visage et histogramme de niveaux de gris obtenus à partir de WebDB. b) Image test et histogramme correspondant. c) Normalisation photométrique de l'image suite à l'application de la spécification d'histogramme a) sur b).*

addition, ils concluent que la normalisation de l'espace de caractéristiques est plus efficace que l'optimisation des paramètres du noyau SVM en absence de normalisation des descripteurs en entrée.

Comme nous utilisons l'intensité des pixels sur les images normalisées nous obtenons 20 × 24 = 480 intensités de pixels caractérisant chaque visage. Ainsi, nous normalisons l'espace des caractéristiques issues des prétraitements géométrique et photométrique introduits dans la section précédente. Ce procédé de normalisation est similaire à celui proposé par Carminati et al. (2003) pour identifier les régions d'une image contenant des visages dans un contexte de suivi de personnes. La normalisation des caractéristiques est réalisée de manière individuelle pour chaque pixel caractérisant l'image. En considérant l'ensemble d'images du corpus d'apprentissage, nous avons extrait les valeurs minimum et maximum pour chaque pixel et de manière linéaire nous ramenons l'ensemble des valeurs entre -1 et 1.

2.2.3 Évaluation

Nous avons sélectionné plusieurs corpus représentatifs de l'état de l'art qui remplissent les critères exigés par une reconnaissance robuste du genre : nombre conséquent de visages, variabilité démographique, variabilité de poses et d'expressions faciales, variabilité de conditions de capture. Nous avons retenu les corpus LFW, GENKI-4K et Groups qui s'inscrivent parmi les corpus les moins contraints de la literature et qui présentent également les moins bonnes performances dans le cadre des tests inter-corpus. Malgré le fait que le corpus GENKI-4K est souvent utilisé dans le cadre des travaux liés à la reconnaissance d'expressions, nous avons décidé de l'inclure dans cette étude. GENKI-4K offre une répartition équitable entre les sujets hommes et les sujets femmes, ce qui est important pour constituer des modèles d'apprentissage non-biaisés par la sur-représentativité d'un genre. Nous incluons aussi le corpus WebDB, que nous avons collecté sur le Web (Flickr, Google Images) pendant les campagnes de validations que nous avons menées. Les résultats de ces expérimentations ont été publiés dans (Danisman et al. 2014) [4].

4. T. Danisman ; I.M. Bilasco ; C. Djeraba - Cross-database evaluation of normalized raw pixels for gender recognition under unconstrained settings - Proc. of International Conference on Pattern Recognition 2014, Stockholm, Sweden, pp. 3144-3149.

La Table 2.2 offre une synthèse de caractéristiques de la population initiale et de la population normalisée utilisée dans les différentes expériences menées sur ces corpus. La variabilité des caractéristiques des corpus sélectionnés rend le contexte de validation particulièrement difficile.

TABLE 2.2 – *Synthèse des caractéristiques des corpus. AG=différents groupes d'âge, E=différentes ethnies, NVE=nombre de visages exploitables (prétraitement et normalisation), EF= différentes expressions faciales, I=différentes conditions d'illumination, STDI=image couleur standard, NC=conditions de capture non-contraintes.*

Corpus	Type	Nombre de visages	Visages (NVE) normalisés	Visages homme	Visages femme	Taille normalisée (l×h)
FERET	STDI, E	2369	2337	908	1429	20×24
GENKI-4K	STDI, EF, E, I, U	4000	3045	1539	1506	20×24
Groups	STDI, EF, AG, E, I, U	28231	19835	10303	9532	20×24
LFW	STDI, EF, AG, E, I, U	13236	11106	8539	2567	20×24
LFW-P	STDI, EF, AG, E, I, U	2927[a]	1533[b]	399	1134	20×24
WebDB	STDI, EF, AG, E, I, U	2927[a]	1533[b]	399	1134	20×24

[a] Plusieurs visages de la même personne
[b] Un seul visage par personne

Les expériences ont été conçues pour valider la capacité de généralisation de notre approche sur des corpus peu contraints tels que : WebDB, LFW, GENKI-4K et Groups. Afin de mettre en place un protocole cohérent, nous utilisons les mêmes paramètres pour toutes les expériences de classification. Il y a un déséquilibre notable entre le nombre d'instances entre les classes hommes et femmes dans le corpus LFW : 2978 femmes vs 10258 hommes. Ainsi, nous privilégions comme base d'apprentissage le corpus Groups, GENKI-4K et WebDB.

Nous avons réalisé une série d'expériences pour comparer les effets de l'utilisation de la spécification d'histogrammes par rapport à l'égalisation d'histogrammes. Nous avons également étudié la manière dont les capacités de notre système de reconnaissance varient en fonction de l'âge des sujets en réalisant également des tests inter-corpus. Finalement, nous comparons notre résultats inter-corpus avec d'autres résultats reportés dans les autres méthodes de l'état de l'art.

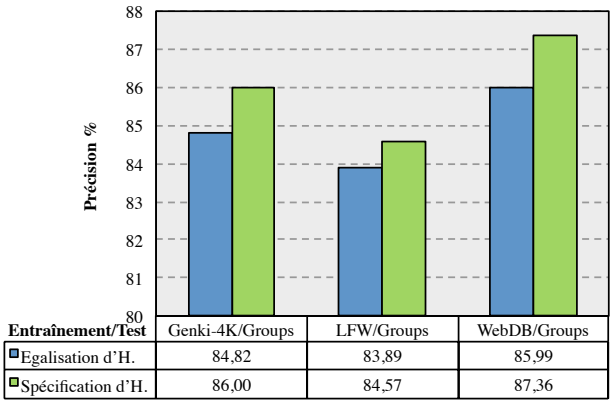
Spécification d'histogrammes versus égalisation d'histogrammes

Nous avons comparé les résultats obtenus sur le corpus Groups en utilisant deux techniques de normalisation photométriques proches : l'égalisation d'histogrammes et la spécification d'histogrammes. Pour l'apprentissage nous avons utilisé les corpus GENKI-4K, LFW et WebDB. La Figure 2.5 montre l'effet des techniques d'égalisation d'histogrammes et de spécification d'histogrammes sur le corpus Groups où l'âge des sujets est supérieur à 12 ans.

Conformément aux résultats illustrés dans la Figure 2.5 nous obtenons de meilleurs résultats de classification quand la technique de spécification d'histogrammes est utilisée. Nous pensons que la principale raison de l'amélioration notée correspond à une meilleure correction de l'illumination. En effet, la densité des probabilités utilisée pour normaliser l'illumination est construite sur la base d'un visage moyen.

Expériences sur la reconnaissance du genre en fonction de l'âge des sujets

Par le passé, Guo et al. (2009) ont montré que le taux de pertinence des solutions de reconnaissance du genre peut augmenter de plus de 10% lorsque l'on s'intéresse aux adultes par rapport aux

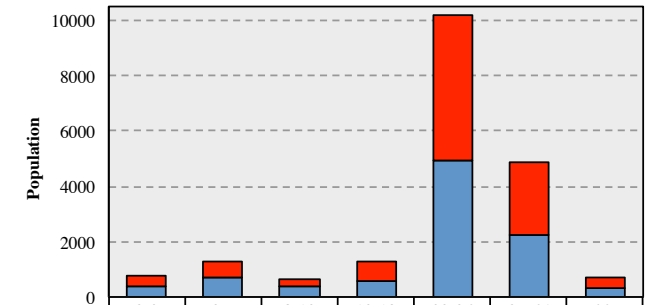


Entraînement/Test	Genki-4K/Groups	LFW/Groups	WebDB/Groups
Egalisation d'H.	84,82	83,89	85,99
Spécification d'H.	86,00	84,57	87,36

FIGURE 2.5 – *Comparaison des techniques d'égalisation d'histogrammes et de spécification d'histogrammes sur l'ensemble des visages des personnes ayant plus de 12 ans dans le corpus Groups en utilisant un protocole expérimental de validation inter-corpus.*

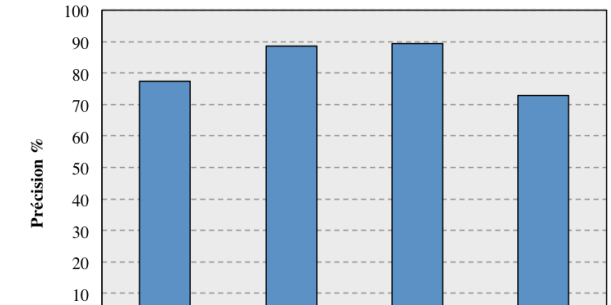
enfants ou seniors. Ils ont également montré que ce fait est indépendant du type de descripteur utilisé pour caractériser le visage : l'intensité normalisée des pixels ou d'autres descripteurs plus complexes tels que HOG, BIF ou LBP.

Les visages du corpus Groups sont annotés avec des informations concernant l'âge que nous structurons dans les catégories suivantes : 0-2, 3-7, 8-12, 13-19, 20-36, 37-65 et 65+. Les corpus GENKI-4K, LFW et WebDB utilisés pour l'apprentissage ne comportent pas de visages d'enfant en nombre significatif. La Figure 2.6-a) montre l'histogramme des âges du corpus Groups. Pour les expériences que nous exposons par la suite, nous avons retenu uniquement les visages associés aux catégories à partir de 13 ans. La Figure 2.6-b) montre la précision obtenue dans le cadre de la validation WebDB/Groups en considérant indépendamment chaque catégorie d'âge. Comme attendu, les meilleurs résultats sont obtenus pour les catégories 20-36 et 37-65 ans avec 88,61% et 89,51%, respectivement. Nous obtenons des meilleurs résultats pour la catégorie des adolescents (13-19) que pour la catégorie des seniors (66+) avec des taux de 77,52% et 72,91%, respectivement. La principale raison pour l'obtention des meilleurs résultats pour la catégorie des adultes est la proximité d'âge entre les visages du corpus d'apprentissage WebDB (essentiellement des adultes) et cette catégorie.



Groupe d'âge	0-2	3-7	8-12	13-19	20-36	37-65	66+
Femme	346	604	265	738	5278	2659	413
Homme	405	701	381	561	4935	2230	318

a)



Groupe d'âge	13-19	20-36	37-65	66+
Précision	77,52	88,61	89,51	72,91

b)

FIGURE 2.6 – *a) Histogramme d'âges des sujets du corpus Groups ; b) Taux de reconnaissance pour les différentes catégories d'âge. L'apprentissage est réalisé sur WebDB et les tests sur les catégories d'âge du corpus Groups.*

Validation croisée inter-corpus

Dans la Table 2.3, nous comparons les résultats obtenus dans un contexte inter-corpus, avec les résultats représentatifs de la littérature. Les entrées en gras correspondent aux meilleures performances de classification enregistrées par rapport aux corpus tests. Nous avons obtenu 88,16% de taux de bonne classification sur les personnes âgées de plus de 20 ans du corpus Groups et un taux de 87,36% pour les personnes âgées de plus de 12 ans. De manière similaire, nous avons obtenu 91,87% de taux de bonne classification sur la validation inter-corpus WebDB/LFW et 91,07% sur la validation inter-corpus Groups/GENKI-4K. Par ailleurs, il est à noter que nous n'avons pas restreint nos ensembles de tests pour une distance inter-pupillaire minimum (garantissant une taille significative de visage). Pour cette raison nous obtenons des ensembles de tests plus grands que ceux utilisé dans les études présentées dans l'état de l'art. En effet, nous utilisons 19834 image tests du corpus Groups sans aucune limitation d'âge ou de distance inter-pupillaire, 17132 visages tests pour des âges ≥ 12 et 15833 visages de tests pour les âges ≥ 20.

TABLE 2.3 – *Synthèse des expériences inter-corpus sur les corpus Groups, LFW et GENKI-4K.*

Travaux	Apprentissage/Test	Nb. images test	Méthode	Pertinence%
(Dago-Casas et al. 2011)	LFW/Groups[a]	14760	LBP+PCA+LDA	81,02
(Dago-Casas et al. 2011)	LFW/Groups[a]	14760	Pixels+PCA+SVM	72,09
(Dago-Casas et al. 2011)	Groups/LFW	13088	LBP+PCA+SVM	89,77
(Ramon-Balmaseda et al. 2012)	Morph/LFW[b]	1149	LBP et SVM+Linear	75,10
(Bekios-Calfa et al. 2014)	Groups/LFW	13233	PCA+LDA+KNN	79,11
(Bekios-Calfa et al. 2014)	Groups[b]/LFW	13233	PCA+LDA+KNN	79,53
Notre méthode	WebDB/Groups	**19834**	Pixels+SVM+RBF	**82,09**
Notre méthode	WebDB/Groups[c]	17132	Pixels+SVM+RBF	87,36
Notre méthode	WebDB/Groups[d]	15833	Pixels+SVM+RBF	**88,16**
Notre méthode	GENKI-4K/Groups[d]	15833	Pixels+SVM+RBF	86,78
Notre méthode	LFW/Groups[c]	17132	Pixels+SVM+RBF	84,57
Notre méthode	LFW/Groups[d]	15833	Pixels+SVM+RBF	85,00
Notre méthode	LFW/GENKI-4K	3045	Pixels+SVM+RBF	87,62
Notre méthode	Groups/GENKI-4K	3045	Pixels+SVM+RBF	**91,07**
Notre méthode	WebDB/LFW	11106	Pixels+SVM+RBF	**91,87**

[a] distance inter-pupillaire ≥ 20 pixels
[b] apprentissage sans visage d'enfant
[c] âge $\geq 12 ans$
[d] âge $\geq 20 ans$

Les études menées par Bekios-Calfa et al. (2014) s'intéressent à la reconnaissance du genre sur des visages non-alignés. Lorsqu'ils considèrent les différentes poses sous la forme d'ensembles représentatifs, ils obtiennent 79,53% et 78,33% avec et sans la prise en compte explicite de la pose. Toutefois, ces résultats sont inférieurs au taux de 91,62% obtenu dans le cadre de la validation inter-corpus Groups/LFW. Par rapport aux approches de Dago-Casas et al. (2011), Ramon-Balmaseda et al. (2012), qui emploient également une classification à base de SVM mais qui utilisent les descripteurs LBP, notre méthode offre de meilleurs résultats. Même si LBP est un descripteur de texture puissant, il n'est pas adapté à l'utilisation dans des contextes où la qualité de l'image est faible et la résolution petite. Généralement, les chercheurs utilisent LBP sur des visages d'une taille minimum de 100×100, car il est nécessaire d'appliquer une grille pour calculer les histogrammes relatifs. Comme les images du corpus LFW contiennent des visages en basse résolution, les performances obtenues en appliquant une caractérisation à base de LBP sont moins élevées. Au contraire, la ca-

ractérisation à base de l'intensité des pixels offre de meilleurs résultats pour les basses résolutions, lorsque les visages sont alignés et normalisés.

TABLE 2.4 – *Synthèse des expériences de validation croisée inter-corpus menées sur l'ensemble des corpus (pertinence %).*

Apprentissage \ Test	GENKI-4K	Groups	LFW	WebDB
GENKI-4K	×	86,78	88,18	×
Groups	91,07	×	91,62	88,70
LFW	87,62	85,00	×	92,05
WebDB	×	88,16	91,87	×

Dans la Table 2.4 nous synthétisons l'intégralité des validations inter-corpus que nous avons réalisées. Il est à noter que tous les résultats de validation inter-corpus s'étalent entre 85,00% et 92,05%, ce qui est un indicateur du bon niveau de généralisation de la méthode que nous proposons. Nous garantissons que tous les visages appartenant à une même personne n'apparaissent pas à la fois dans la partie du corpus consacrée à l'apprentissage et à la fois dans la partie du corpus consacrée aux tests. Nous appliquons cette séparation afin de ne pas biaiser la reconnaissance du genre en bénéficiant de la proximité entre les visages d'une même personne. Ainsi, comme le corpus GENKI-4K est un sous-ensemble de WebDB (cf. section 2.2), nous ne réalisons aucune expérience de validation croisée inter-corpus concernant GENKI-4K et WebDB.

Les résultats obtenus peuvent être encore améliorés en considérant les caractéristiques intra- et extra-faciales (pilosité, cheveux). En effet, certaines situations où l'intérieur du visage n'est pas discriminant peuvent être résolues en considérant les caractéristiques de la chevelure ou de la présence d'une pilosité faciale importante.

2.3 Reconnaissance du genre à partir de multiples critères dans un processus d'inférence floue

Dans cette section nous présentons le système d'inférence floue qui nous permet de désambiguïser la prédiction du genre qui s'appuie uniquement sur l'apparence visuelle des visages normés. À cette fin, nous exploitons les connaissances d'experts et les informations complémentaires qui peuvent être amenées par les caractéristiques des régions délimitant la pilosité intra- et extra-faciale.

Notre but est d'explorer les effets de la prise en compte des informations de haut niveau concernant la pilosité faciale et la chevelure dans leur globalité et non pas en tant qu'éléments de bas niveau caractérisés par l'intensité des pixels ou autres descripteurs visuels. Ainsi, nous nous concentrons sur un ensemble de connaissances provenant des experts humains (règles d'inférence, la caractérisation de la pilosité faciale et des cheveux) et une méthodologie de classification. Nous utilisons le volume des cheveux et le ratio de la pilosité faciale par rapport au visage tout entier afin d'alimenter le moteur de règles construit par un expert. L'information extraite depuis l'intérieur du visage en utilisant l'approche présentée précédemment est utilisée comme une variable d'entrée dans le cadre du système d'inférence floue, conjointement à l'information sur les cheveux et la pilosité faciale. Nous avons défini une base de connaissances contenant six règles qui réalisent un appariement non linéaire entre les variables d'entrée et la classification du genre.

Un système d'inférence floue (SIF) offre un moyen d'associer des variables d'entrée caractérisant des faits à un espace de solutions en utilisant la logique floue. Les SIF communément rencontrés dans la littérature s'inspirent des modèles proposées par Mamdani et Assilian (1975) et Takagi et Sugeno (1985). La principale différence entres les deux modèles de SIF repose sur la manière dont les conclusions sont inférées. Dans le modèle de Mamdani, la fonction d'appariement des sorties peut être évaluée de manière indépendante des variables en entrées, alors que dans le modèle de Takagi-Sugeno, la calcul des fonctions d'appartenance de sortie dépend des valeurs des variables en entrée. Nous avons adopté le modèle de Mamdani pour sa capacité à disposer de fonctions d'appartenance indépendantes, pour sa structure basée essentiellement sur les opérations de types min-max et pour le fait d'être largement reconnu pour ses capacités à encoder la connaissance d'experts.

Un SIF est composé des éléments suivants :

— **fuzzification :** modifie les variables en entrée en appliquant les fonctions d'appartenance d'entrée (FAE), afin de pouvoir être utilisées par le moteur de règles.

— **base de connaissances :** consiste en un ensemble de règles et en la définition des fonctions d'appartenance. Les règles correspondent à des structures logiques de type If-Then. Une règle est également appelée inférence floue ayant un antécédent et une conséquence. Les fonctions d'appartenance servent à la fois au processus de fuzzification, qu'à celui de défuzzification.

— **moteur d'inférence :** évalue les règles concernées en fonction des valeurs des variables d'entrée.

— **défuzzification :** convertit les résultats du moteur d'inférence dans une décision en utilisant une fonction spécifique. Le centre de gravité (COG), la moyenne des sommes (COS) et la moyenne des maximums (MOM) sont des fonctions de défuzzification largement répandues dans la littérature.

Le schéma global du système proposé est illustré dans la Figure 2.7. Les variables d'entrée impliquées dans le processus d'inférence floue incluent la prédiction obtenue à partir de l'intensité de pixels et les informations de haut niveau issues des annotations. Les variables ainsi obtenues illustrées dans la Figure 2.7(a) alimentent le système d'inférence floue comme illustré dans la Figure 2.7(b). Le processus de fuzzification évalue l'apport informationnel de chaque variable en considérant les fonctions d'appartenance pour obtenir les ensembles flous. Ensuite, le moteur d'inférence évalue les ensembles flous et génère les résultats évalués par le processus de défuzzification. À la fin, nous obtenons le résultat de reconnaissance du genre.

Dans la section suivante, nous définissons la manière dont les variables d'entrée sont calculées en amont du processus de fuzzification.

2.3.1 Extraction de descripteurs

Nous extrayons à partir d'un visage, trois variables d'entrée qui alimentent le système d'inférence floue : la pilosité faciale, les cheveux et la prédiction du genre à partir de l'intensité normée des pixels de l'intérieur du visage. Afin de transformer ces annotations en information servant à alimenter le système d'inférence, nous normalisons la taille du volume de la chevelure par la taille du visage. De manière similaire, nous normalisons l'aire de la zone couverte par la pilosité faciale. Nous utilisons pour cela la position du nez et de l'aire de la bouche, représentés par un carré si-

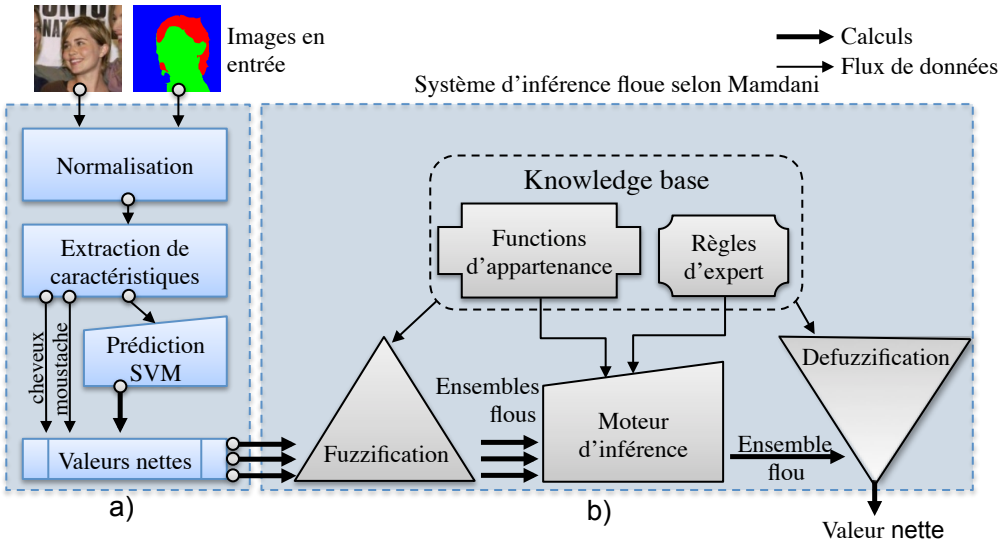


FIGURE 2.7 – *Schéma global de l'approche proposée. a) Extraction de variables d'entrée, alimentant le système d'inférence floue. b) Système d'inférence floue selon Mamdani et Assilian (1975).*

tué aux coordonnées suivantes : x=16, y=20 et de taille 8×8 dans un visage normalisé de 40×40 comme illustré dans la Figure 2.8.

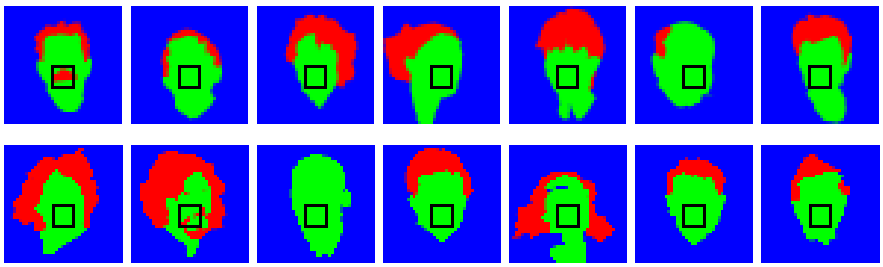


FIGURE 2.8 – *Localisation grossière de la bouche sur quelques images du corpus Part Labels.*

La variable reflétant la prédiction du genre à partir de l'intensité des pixels est obtenue par le classifieur SVM. Ses valeurs se situent entre [-1, +1]. Les réponses positives indiquent la reconnaissance d'un visage féminin et les réponses négatives celle d'un visage masculin.

Avant de détailler le fonctionnement du système d'inférence floue, nous illustrons les contributions individuelles de chacune des variables considérées pour la reconnaissance du genre dans le cadre du corpus LFW - Part Labels. Afin de faciliter la visualisation des réponses de prédictions obtenues en analysant l'intensité des pixels avec le classifieur SVM, nous appliquons une transformation linéaire sur l'intervalle [-10 .. +10] comme illustré dans la Figure 2.9. Les résultats correspondent à une validation inter-corpus Groups/LFW - Part Labels. L'axe vertical correspond aux résultats obtenus par le senseur visuel, pour chaque image du corpus. Afin de mieux mettre en exergue la corrélation entre la prédiction et la vérité du terrain, les premières données sur l'axe des abscisses correspondent aux images de femmes et le reste aux images d'homme. Ainsi, il est aisé de visualiser les fausses détections : pour les femmes, elles se situent en dessous de 0 et pour les hommes elles situent au dessus de 0.

En utilisant cette fois-ci une échelle logarithmique, nous illustrons la corrélation entre le genre et les caractéristiques géométriques de la chevelure et de la pilosité faciale dans les Figures 2.10 et 2.11 respectivement. On peut noter notamment qu'à partir d'un certain seuil la quantité des cheveux

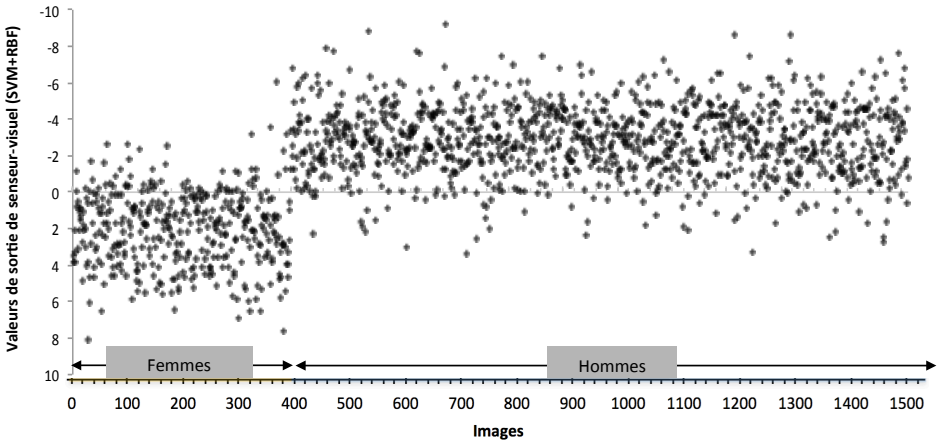


FIGURE 2.9 – *Résultats de tests de validation croisée inter-corpus Groups/LFW-P en utilisant un classifieur SVM+RBF exploitant uniquement l'information visuelle - intensité des pixels (Taux de bonne classification=91,25%).*

permettent d'orienter clairement la reconnaissance vers le genre féminin. Peu d'images d'hommes se retrouvent au dessus du seuil retenu. Une observation similaire peut être tirée également à propos de la pilosité faciale, même si, ici, l'identification d'un seuil précis est plus difficile.

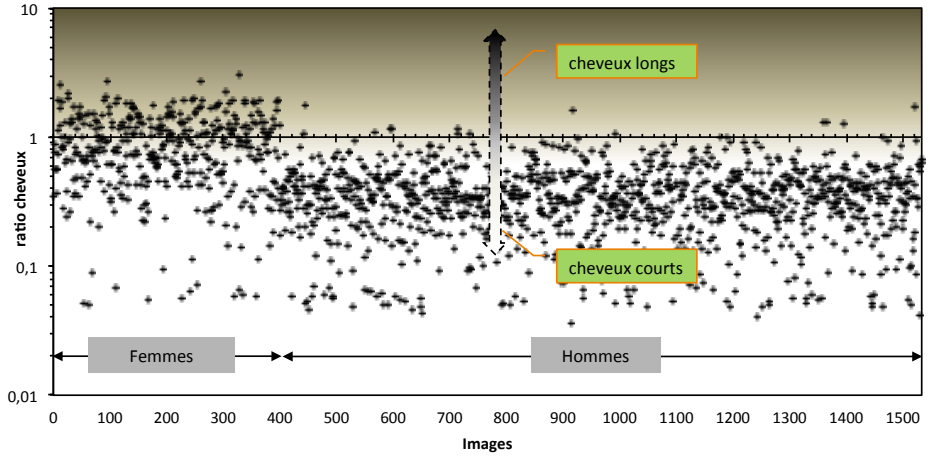


FIGURE 2.10 – *Représentation sur une échelle logarithmique du ratio des cheveux dans le corpus Part Labels.*

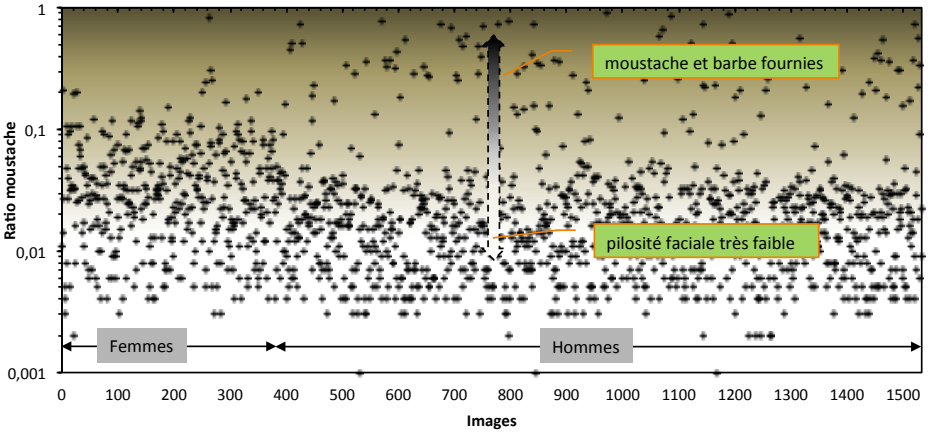


FIGURE 2.11 – *Représentation sur une échelle logarithmique du ratio de la moustache dans le corpus Part Labels.*

Lorsque l'on considèrent chacune de ces variables de manière indépendante des incertitudes

persistent et il est évident qu'une seule variable à la fois ne peut pas prédire précisément le genre. Dans la section suivante nous présentons en détail le système d'inférence floue qui permettra de tirer parti de l'analyse conjointe de ces éléments.

2.3.2 Système d'inférence floue

Nous construisons un SIF en suivant une configuration plusieurs entrées - une sortie (*Multiple Inputs Single Output* - MISO). Nous utilisation un mélange de gaussiennes pour définir les fonction d'appartenance qui aideront à construire les ensembles flous. La fonction d'appartenance à base d'un mélange de gaussiennes est une fonction d'appartenance lisse qui dépend de quatre paramètres pour deux gaussiennes : $\sigma_1, c_1, \sigma_2, c_2$. La fonction est définie comme suit :

$$\mu(x; \sigma_1, c_1, \sigma_2, c_2) = \begin{cases} \exp\left[\frac{-(x-c_1)^2}{2\sigma_1^2}\right] & : x < c_1 \\ 1 & : c_1 \leq x \leq c_2 \\ \exp\left[\frac{-(x-c_2)^2}{2\sigma_2^2}\right] & : c_2 < x \end{cases} \tag{2.2}$$

La Figure 2.12 illustre les fonctions d'appartenance de variables d'entrée et de sortie. Pour chaque graphique, la courbe la plus à gauche correspond à c_1, σ_1 et la courbe la plus à droite correspond à c_2, σ_2.

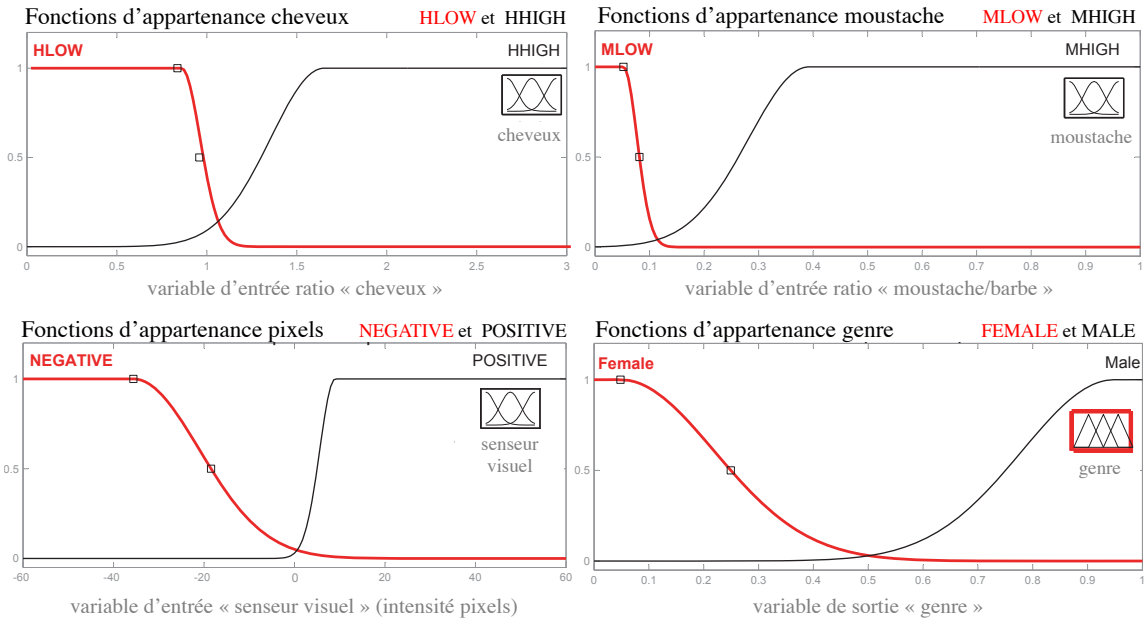


FIGURE 2.12 – *Graphiques pour les fonctions d'appartenance des variables d'entrée et de sortie couvrant la prise en compte des cheveux, de la pilosité faciale, de la caractérisation de l'intérieur du visage et, finalement, de la sortie du système, le genre.*

Un ensemble flou A définit sur le domaine X correspond à une suite de paires ordonnée :

$$A = \{(x, \mu_A(x)) \mid x \in X\} \tag{2.3}$$

où μ_A est une fonction d'appartenance, $\mu_A : X \to M$, M est l'espace d'appartenance où chaque élément de X est projeté. Ainsi, $\mu_A(x)$ représente le niveau d'appartenance de x à A, ce qui permet

de faire correspondre l'ensemble des X à l'espace d'appartenance. Considérant les Équations 2.2 et 2.3, la Table 2.5 présente le détail de chaque mélange de gaussienne correspondant aux fonctions d'appartenance utilisées dans le système.

Nous avons créé six règles de mise en correspondance logique entre les variables d'entrée et des sorties du système comme illustré dans la Figure 2.13. Afin de prendre en charge la logique floue dans un système à base de règles, l'opérateur "And" est interprété comme l'intersection des fonctions d'appartenance correspondantes. Ainsi pour deux ensembles flous A et B nous définisons :

$$A \cap B, \mu_{A \cap B}(x) = min(\mu_A(x), \mu_B(x)) \tag{2.4}$$

TABLE 2.5 – *Détails de la base de fonctions d'appartenance. VL=Variables d'appartenance, FA=Fonction d'Appartenance.*

VL	Type	FA	Intervalle	σ_1	c_1	σ_2	c_2
cheveux	entrée	bcpCheveux	[0,3]	0,251	-0,824	0,103	0,836
cheveux	entrée	peuCheveux	[0,3]	0,298	1,652	0,067	3,800
pilosité	entrée	bcpPilosité	[0,1]	0,217	-0,223	0,024	0,052
pilosité	entrée	peuPilosité	[0,1]	0,110	0,392	0,227	1,157
senseur visuel	entrée	négatif	[-60,60]	13,690	-76,730	14,580	-35,610
senseur visuel	entrée	positif	[-60,60]	3,397	9,000	8,901	73,020
genre	sortie	femme	[0,1]	0,338	-0,101	0,170	0,0481
genre	sortie	homme	[0,1]	0,168	0,949	0,391	1,114

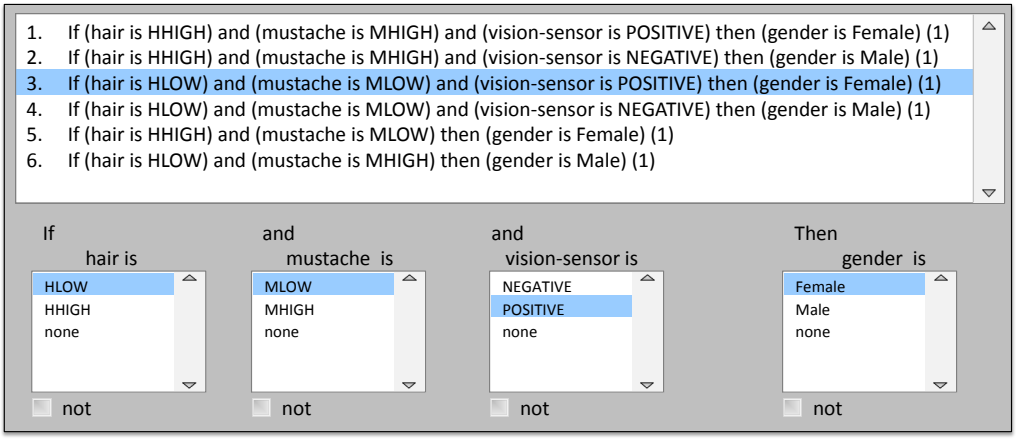


FIGURE 2.13 – *Détails de la base de règles.*

Compte tenu des variables et des fonctions d'appartenance correspondantes, la Figure 2.14 illustre l'aire de décision des variables d'entrée et de sortie.

La relation entre les cheveux, la pilosité et le genre est présentée dans la Figure 2.14(a). Nous mesurons la pilosité faciale uniquement autour de la bouche. La quantité des pixels associés à la pilosité faciale autour de la bouche peut être biaisée par des mouvements de la tête ou bien des occultations. Il se peut que ces bruits génèrent des situations où la quantité de pilosité faciale autour de la bouche et le volume des cheveux sont, en même temps, conséquents. Dans ce cas de figure, c'est le senseur visuel analysant l'intérieur du visage qui influencera majoritairement la décision finale.

La Figure 2.14(b) montre la relation entre le cheveux, le senseur visuel et le genre. Des valeurs hautes fournies par l'estimation du volume de cheveux et de l'analyse visuelle de l'intérieur du

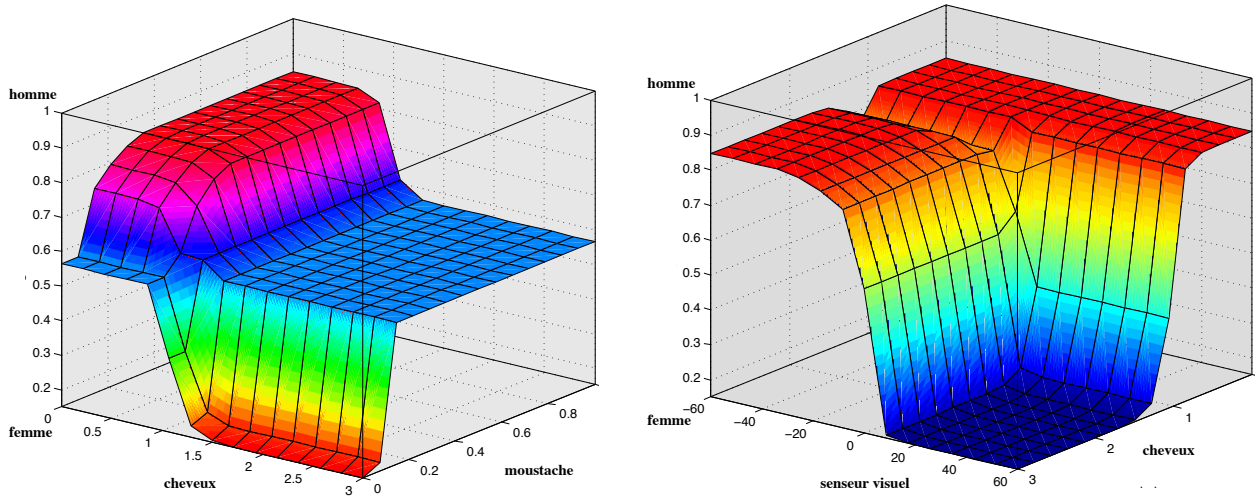


FIGURE 2.14 – *Aire de décision pour les variables d'entrée et de sortie. a) Interdépendance entre les cheveux, la moustache et le genre. b) Interdépendance entre le senseur visuel, les cheveux et le genre.*

visage traduisent une haute probabilité que le sujet soit une femme. Ainsi, dans la Figure 2.14(b) la partie bassse de l'aire est activée.

Dans l'étape de défuzzification, nous choisissons la méthode COG (centre de gravité) afin de trouver le point où les lignes verticales séparent l'ensemble d'agrégats dans deux masses égales comme indiqué dans l'Équation 2.5. La méthode COG trouve le point représentant le centre de gravité de l'ensemble flou A sur l'intervalle $[a, b]$.

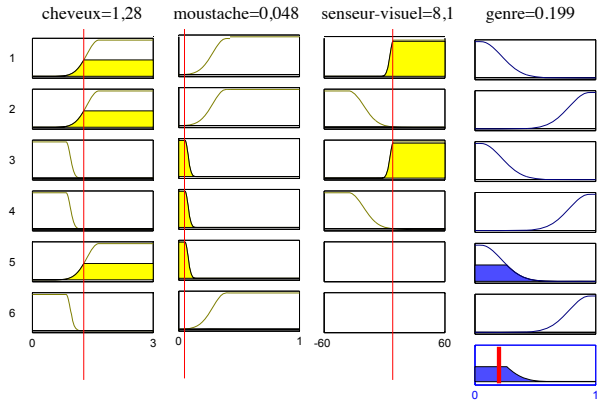
$$COG = \frac{\int_a^b \mu_A(x) x \mathrm{d}x}{\int_a^b \mu_A(x) \mathrm{d}x} \tag{2.5}$$

La Figure 2.15 montre quelques exemples d'entrées caractérisant des hommes et des femmes, ainsi que les résultats en termes de COG (lignes verticales rouges dans la colonne du genre) obtenus par le SIF. Chaque ligne dans la Figure 2.15 illustre l'évaluation d'une seule règle de la base de règles en considérant les valeurs spécifiques des variables d'entrée. Compte tenu de l'évaluation finale réalisée par la méthode de défuzification, la décision concernant le genre est prise. Une réponse femme est retenue, lorsque la valeur COG obtenue est inférieure à 0,5. Une réponse homme est retenue lorsque la valeur COG obtenue est supérieure à 0,5.

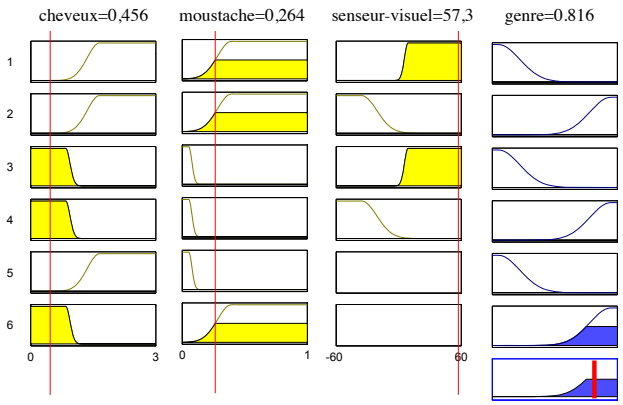
2.3.3 Évaluation

Des validations croisées inter-corpus sont menées en utilisant différents corpus publics : FERET, GENKI-4K, Groups et LFW. Des classifieurs SVM sont entraînés de manière individuelle sur chaque corpus afin d'obtenir une première prédiction basée uniquement sur l'intensité de pixels. Des modèles optimisés sont ensuite testés sur le corpus Part Labels, sous-ensemble de LFW, en utilisant le système d'inférence floue. Le corpus Part Labels est utilisé pour les tests, car c'est le seul disposant des informations caractérisant la pilosité faciale et la chevelure.

Nous utilisons les informations concernant la pilosité faciale et les cheveux afin d'enrichir le processus de reconnaissance. Des méthodes existent dans la littérature pour obtenir ces informations



a) Valeurs des variables [1,28; 0,048; 8,1]

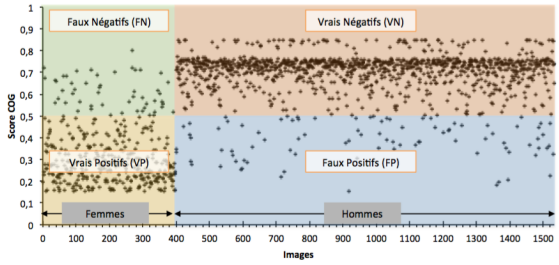


b) Valeurs des variables [0,465; 0,264; 57,3]

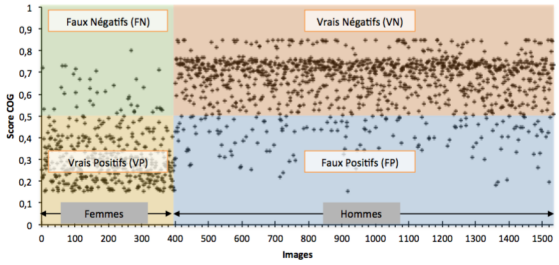
FIGURE 2.15 – *Quelques exemples de valeurs en entrées et en sortie pour différents genres.*

de manière automatique comme dans (Kae et al. 2013). Toutefois, ici, nous utilisons directement les annotations concernant la segmentation du visage disponible dans le corpus Part Labels constitué par Kae et al. (2013) à partir du corpus LFW. Comme, notre but est de montrer qu'il est possible d'inclure avec succès l'information de haut niveau concernant la reconnaissance du genre dans un système d'inférence floue, nous avons préféré utiliser des annotations manuelles. Cela évite d'introduire du bruit lié à une mauvaise segmentation de la pilosité intra- et extra-faciale.

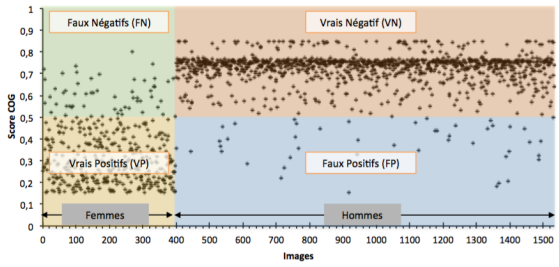
Les Figures 2.16(a–c) illustrent les résultats du SIF sur le corpus LFW-P en utilisant les corpus Groups, GENKI-4K et FERET respectivement, comme corpus d'apprentissage pour la modalité visuelle. En comparaison avec les performances illustrées dans la Figure 2.16(d) obtenues en utilisant uniquement le senseur visuel, nous observons une plus nette séparation entre les hommes et les femmes lorsque le SIF est employé.



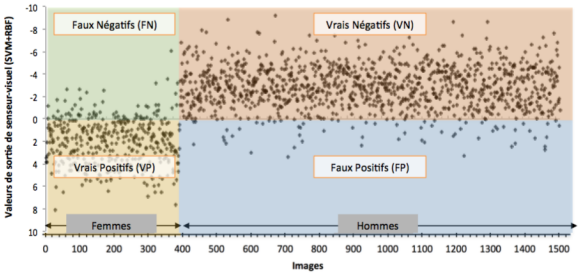
(a) *GENKI-4K/LFW-P en utilisant SIF.*



(b) *GENKI-4K/LFW-P en utilisant SIF.*



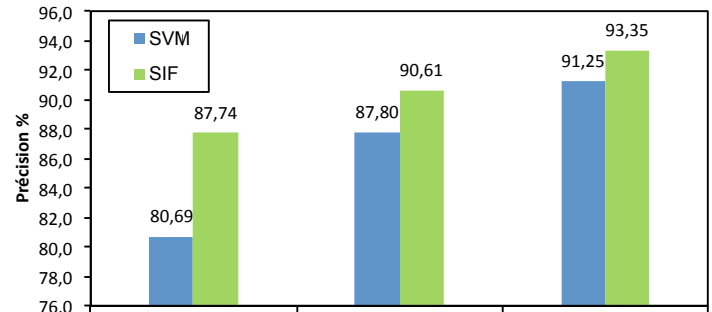
(c) *Groups/LFW-P en utilisant SIF.*



(d) *Groups/LFW-P en utilisant l'intensité de pixels.*

FIGURE 2.16 – *Séparation entre les exemples de femmes et d'hommes obtenue suite à l'application d'un protocole de validation croisée.*

La Figure 2.17 illustre les résultats obtenus par le SIF englobant les informations sur l'intérieur du visage, la pilosité faciale et les cheveux. Ces résultat sont comparés à ceux obtenus par le classifieur SVM en utilisant uniquement les intensités des pixels à l'intérieur du visage. Pour chaque expérience, SIF produit des résultats meilleurs que la méthode à base de SVM. En utilisant SIF, nous obtenons un taux de bonne classification de 93,35% dans le cadre de la validation croisée inter-corpus Groups/LFW-P. Ce taux est supérieur au taux de classification de 91,25% obtenu lorsque nous utilisons uniquement l'intensité de pixels dans la cadre de la validation croisée inter-corpus Groups/LFW-P. Du fait que les corpus Groups et GENKI-4K ont été constitués en agrégeant des données issues de sources variées, ils constituent des bases d'apprentissage plus génériques que le corpus FERET. La capacité de généralisation plus importante de l'apprentissage sur les corpus Groups et GENKI-4K est mise en évidence par les tests de validations croisées inter-corpus. Le corpus FERET est une base contrainte (sujets qui posent) et captée dans un environnement contrôlé ce qui limite la généralisation du processus d'apprentissage. Par ailleurs, nos expériences ont également montré qu'au delà de la variabilité en termes de conditions de capture, le nombre d'exemples disponibles pour l'apprentissage a également un impact sur les performances.



Entr./Test	Feret/LFW-P	Genki/LFW-P	Groups/LFW-P
SVM	80,69	87,80	91,25
SIF	87,74	90,61	93,35

FIGURE 2.17 – *Résultats de validations inter-corpus ayant comme cible le corpus LFW-P.*

Les détails concernant le positionnement de notre solution par rapport à un ensemble représentatif des validations inter-corpus réalisées dans la littérature sont présentés dans la Table 2.6. Ces résultats ont été publiés dans (Danisman et al. 2015) [5].

Les résultats des expériences menées montrent que le principal avantage du SIF par rapport à des méthodes traditionnelles réside dans sa capacité à réaliser des appariements non-linéaires entre les entrées et les sorties. La Table 2.7 offre une analyse comparative plus détaillée des résultats obtenus en utilisant SVM et SIF respectivement. Lorsque le SIF est utilisé en aval d'un classifieur (de type SVM, par exemple) il élimine davantage les faux positifs et les faux négatifs en se référant à une base de connaissances. La Table 2.7 contient également les valeurs prédictives positives (VPP) et les valeurs prédictives négatives (VPN) pour illustrer cette tendance. Ces valeurs mesurent, respectivement, la possibilité qu'un résultat positif/négatif reflète une bonne classification au sein de la catégorie. Elles sont calculées selon les formules suivantes :

$$VPP = \frac{VP}{VP + FP} \; et \; VPN = \frac{VN}{VN + FN} \tag{2.6}$$

5. T. Danisman ; I.M. Bilasco ; J. Martinet - Boosting gender recognition performance with a fuzzy inference system - Expert Systems with Applications, Volume 42, Issue 5, 1 April 2015, pp. 2772-2784 (Facteur d'impact : 3,768 selon JCR 2018).

TABLE 2.6 – *Résumé des expériences de validations croisées inter-corpus en considérant comme corpus de test LFW et LFW-P. Ch=cheveux, PF=pilosité faciale, IP=intensité de pixels, ICP=inter-corpus, VC=validation croisée en 5 parties intra-corpus.*

Méthode vs. Test	Corpus Apprentissage Éval.	Modèle	Taille test	Taux.
(Dago-Casas et al. 2011)	Groups/LFW[a]	LBP+PCA et SVM	13088	89,77%
(Ramon-Balmaseda et al. 2012)	Morph/LFW[b]	LBP et SVM+Linear	1149	75,10%
(Bekios-Calfa et al. 2014)	Groups/LFW[a]	PCA+LDA et KNN	13233	79,11%
(Bekios-Calfa et al. 2014)	Groups[c]/LFW[a]	PCA+LDA et KNN	13233	79,53%
Notre methode SVM+RBF	WebDB/LFW[a]	IP et SVM+RBF	11106	91,87%
Notre methode SVM+RBF	Groups/LFW[a]	IP et SVM+RBF	11106	91,62%
Notre methode SVM+RBF	Groups/LFW-P[b]	IP et SVM+RBF	1533	91,25%
Notre methode SVM+RBF	GENKI-4K/LFW-P[b]	IP et SVM+RBF	1533	87,80%
Notre methode SVM+RBF	FERET/LFW-P[b]	IP et SVM+RBF	1533	80,69%
Notre méthode SIF	Groups/LFW-P[b]	Ch+PF+IP et SIF	1533	**93,35**%
Notre méthode SIF	GENKI-4K/LFW-P[b]	Ch+PF+IP et SIF	1533	**90,61**%
Notre méthode SIF	FERET/LFW-P[b]	Ch+PF+IP et SIF	1533	**87,74**%

[a] plusieurs visages par personne
[b] un seul visage par personne
[c] apprentissage réalisé sans les visages d'enfant

TABLE 2.7 – *Analyse comparative détaillée des expériences de validation croisée inter-corpus. VP=Vrais Positifs, FP=Faux Positifs, VN=Vrais Négatifs, FN=Faux Négatifs, VPP=Valeur prédictive positive, VPN=Valeur prédictive négative.*

Apprentissage/Test	Modèle	VP	FP	VN	FN	VPP	VPN	Accuracy
FERET/LFW-P	SIF	361	149	984	39	70,78%	96,19%	**87,74**%
FERET/LFW-P	SVM	361	257	876	39	58,41%	95,74%	80,69%
GENKI-4K/LFW-P	SIF	346	90	1043	54	79,36%	95,08%	**90,61**%
GENKI-4K/LFW-P	SVM	340	127	1006	60	72,80%	94,37%	87,80%
Groups/LFW-P	SIF	343	45	1088	57	88,40%	95,02%	**93,35**%
Groups/LFW-P	SVM	336	70	1063	64	82,75%	94,32%	91,25%

2.4 Résumé des contributions

Dans notre première contribution nous avons présenté une solution de reconnaissance du genre qui s'appuie sur l'alignement des visages et la normalisation des intensités des pixels en utilisant la spécification d'histogrammes accompagnée d'une classification à base d'un SVM avec un noyau RBF. Pour montrer les capacités de généralisation de l'approche, nous avons suivi un protocole de validation croisée inter-corpus. Nous montrons que la spécification d'histogrammes offre de meilleurs résultats que l'égalisation d'histogrammes sur des corpus dont les conditions de capture sont non-contrôlées. En utilisant des visages alignés et normalisés de taille 20×24, l'intensité des pixels seule, permet d'obtenir des résultats compétitifs, généralisables et meilleurs que d'autres méthodes de l'état de l'art lorsque la validation se déroule dans un contexte inter-corpus. Par ailleurs, nous avons étudié l'impact de l'âge sur la reconnaissance du genre et nous avons montré que la reconnaissance du genre chez l'adulte est plus précise que chez l'enfant ou les seniors.

Le système d'inférence floue pour la reconnaissance du genre utilise l'intensité de pixels du visage en incluant également des informations sur certains éléments du visage (la moustache, la barbe) et certains éléments entourant le visage (les cheveux). Nous avons présenté des expériences fiables montrant la robustesse du système proposé en adoptant une méthodologie de validation croisée inter-corpus. Nous avons montré que la prise en compte des indices autres que l'intérieur

du visage, traité dans sa globalité, améliore les performances de classification. Afin d'évaluer l'influence d'autres éléments tels que les cheveux et la moustache, nous avons réalisé des tests sur le corpus LFW et, notamment sur le sous-ensemble Part Labels, en utilisant les annotations concernant la pilosité faciale et les cheveux des personnes.

Notre étude est la première à se servir des informations concernant la pilosité faciale et les cheveux disponibles dans le corpus Part Labels pour la reconnaissance du genre. Nos résultats corroborent les résultats de l'état de l'art qui montrent les effets positifs de la prise en compte des informations sur les cheveux. Par ailleurs, les corpus dont le contexte de capture n'est pas contrôlé, disposant d'un nombre important d'images, permettent d'obtenir des classifieurs plus génériques car les variations visuelles couvrent mieux l'espace des solutions.

En comparaison avec d'autres études, les principaux apports de cette deuxième contribution sont : la prise en compte explicite du volume des cheveux et de la surface couverte par la pilosité faciale et le système d'inférence floue qui surpasse le classifieur SVM en présence de règles et de fonctions d'appariement formulées par un expert humain.

Pour conclure, le principal avantage des approches proposées est de disposer des capacités de généralisation enrichies. La généralisation est l'une des plus importantes caractéristiques requises pour un système de reconnaissance du genre. L'utilisation de l'information de haut niveau (caractérisation explicite de la pilosité faciale et des cheveux) dans un cadre d'inférence floue améliore les résultats obtenus en s'intéressant uniquement aux caractéristiques visuelles, augmentant ainsi la capacité de généralisation.

Reconnaissance des expressions 3

La reconnaissance automatique des expressions faciales est un thème de recherche actif dans la communauté depuis le début des années 90. D'importantes retombées économiques et sociétales sont attendues suite à la mise à disposition de solutions fiables et robustes, capables de fonctionner dans un contexte non-contraint et peu instrumentalisé tel que dans une salle de réunion, un magasin, une maison ou un centre de soins. Par exemple, lors d'une visioconférence entre plusieurs participants distants, le feedback offert par un outil d'analyse des expressions faciales renforce le dialogue et l'interaction sociale entre les participants (par exemple, maintenir l'attention des participants car ils se sentent davantage impliqués dans l'interaction). Dans un contexte magasin, la détection de visages souriants peut être considérée comme un signe d'intérêt, alors que les grimaces de dégoût ou d'énervement peuvent être considérées comme un signe de répulsion.

Ces dernières années, des avancées importantes ont été marquées en matière de reconnaissance des expressions faciales. Les méthodes récentes obtiennent de très bonnes performances sur des corpus contrôlés contenant des expressions de forte intensité, dans de bonnes conditions d'illumination et en absence de mouvements de la tête. Cependant, construire des systèmes qui analysent les expressions spontanées à intensité variable dans un environnement non-contraint constitue encore un grand challenge.

L'apparition d'une expression sur le visage se traduit par des changements de texture et de géométrie. Un nombre important de travaux ont été menés afin de détecter les points caractéristiques du visage (la position des yeux, les sourcils, le nez, la bouche). Toutefois, il reste encore de nombreux défis à relever incluant l'illumination variable, les variations statiques et dynamiques des composants de visages, dues aux occlusions (par exemple, avec les mains ou avec les lunettes) ou aux changements d'orientation de la tête, ainsi que les intensités variables des expressions.

Les expressions faciales se traduisent par des changements apparaissant au cours du temps sur le visage. Cependant, la reconnaissance des expressions s'applique à la fois sur des images statiques que sur des séquences d'images. En présence d'une séquence, on étudie l'activation d'une expression dans son intégralité. En présence d'une image, on étudie un seul instant de l'activation. Souvent celui-ci correspond au moment où l'intensité de l'expression est la plus prononcée (l'*apex* de l'expression).

Les problématiques et les attentes vis-à-vis de ces deux cas d'usage sont très différents. La reconnaissance des expressions à partir d'images traite plutôt de l'identification des expressions plutôt marquées, de forte intensité. Dans un cadre dynamique, différents niveaux d'intensité peuvent être étudiés en partant de micro-expressions jusqu'aux macro-expressions.

Les macro-expressions sont des expressions faciales produites de manière volontaire par la personne et impactent souvent l'intégralité du visage. Au contraire, les micro-expressions ne

concernent que certaines parties du visage et pour des durées très courtes. Porter et Ten Brinke (2008) montrent que les micro-expressions ne peuvent pas impacter le haut et le bas du visage en même temps. Elles sont souvent invisibles à l'oeil nu et trahissent des émotions que la personne ne souhaite pas forcement exprimer. Les macro-expressions ont une durée comprise entre 0,5 et 4 s et s'étendent de manière notable sur le visage (Ekman et Rosenberg 1997). Selon certains auteurs, les micro-expressions ont une durée comprise entre 170 et 500 ms. Plus spécifiquement, Yan et al. (2013) montrent que la phase d'enclenchement constitue un bon indicateur pour déceler une micro-expression et sa durée est comprise entre 65 et 260 ms.

Les micro-expressions sont caractérisées non seulement par des courtes durées, mais également par de faibles intensités (Porter et Ten Brinke 2008, Yan et al. 2014b). Ces faibles intensités se traduisent par une quasi absence de mouvements faciaux et de changements d'apparence. Afin de pouvoir toutefois observer ces changements, de nouveaux protocoles de captation spécifiques ont été conçus. Par exemple, Li et al. (2013b) utilisent des caméras à haute fréquence (100-200 fps). À ces fréquences d'enregistrement il est possible d'observer les phases d'enclenchement d'une micro-expression, même s'il est toujours difficile de distinguer les véritables mouvements des bruits de capture.

Dans la suite de ce chapitre nous décrivons nos contributions relatives à la reconnaissance des expressions dans un contexte statique et dans un contexte dynamique.

Contexte statique Dans le cadre d'images statiques, nous nous intéressons à la construction de descripteurs globaux permettant d'optimiser les caractéristiques visuelles en vue de la reconnaissance des expressions actées. La sélection de caractéristiques permet d'obtenir des modèles plus robustes. Dans ce contexte, la majorité des travaux existants dans le domaine de la reconnaissance des expressions faciales utilisent une combinaison de régions rectangulaires (Viola et Jones 2004, Hadid et Pietikainen 2009). Ces régions sont utilisées pour localiser et extraire des caractéristiques importantes pour le processus de classification. Même si l'implémentation est relativement simple, ces régions couvrent également des zones non-pertinentes et peuvent contenir du bruit. Des techniques de sélection peuvent s'appliquer ensuite au niveau de ces régions pour identifier des sous-ensembles d'éléments ou de transformer l'espace de représentation pour optimiser la réponse de systèmes de reconnaissance dans une situation spécifique (Ververidis et Kotropoulos 2008). Dans cette lignée, nous proposons une méthode innovante pour estimer l'apport informationnel de l'intensité de chaque pixel en identifiant ainsi des pixels d'intérêts sur le visage. Les regroupements de pixels d'intérêt constituent des masques non-rectangulaires qui peuvent améliorer le processus de classification. Cette technique décrite dans (Danisman et al. 2013) [1] a été mise en œuvre pour la détection des expressions de joie et sera détaillée en section 3.2

Contexte dynamique Dans ce cadre nous nous intéressons à la reconnaissance des expressions exhibant des niveaux d'intensité variable. Nous traitons principalement de macro- et de micro-expressions qui constituent selon Ekman et Rosenberg (1997) les intensités extrêmes d'une expression. Disposer de descripteurs et d'un modèle facial capable d'encoder de manière unifiée les

1. T. Danisman ; I.M. Bilasco ; J. Martinet ; C. Djeraba - Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron - Signal Processing, Elsevier, 2013, Special issue on Machine Learning in Intelligent Image Processing, 93 (6), pp. 1547-1556 (Facteur d'impact : 3,470 selon JCR 2018).

changements induits par une micro- ou une macro-expression laisse présager la capacité de traiter différents niveaux d'intensités.

Nous proposons une solution à base de flux optique qui caractérise les mouvements au sein du visage de manière dense. Afin de pouvoir disposer de modèles de flux optique qui caractérisent de manière optimale une large palette d'intensité des expressions, nous mettons en œuvre un filtrage qui s'appuie sur la dynamique naturelle du mouvement au sein du visage. Un mouvement dans une région du visage se diffuse de manière homogène dans les régions voisines en perdant graduellement en intensité. Cette hypothèse nous permettent de séparer les véritables mouvements des bruits de mesure, même en présence d'expressions de faible intensité. Associé à un modèle facial construit en s'appuyant sur la points caractéristiques, notre approche publiée dans (Allaert et al. 2017) [2] est capable de reconnaître en présence de légers mouvements de la tête et sous différentes conditions d'illumination, à la fois des expressions de faible, de moyenne ou de forte intensité comme illustré dans la section 3.3. Nos travaux se détachent ainsi des autres méthodes de l'état de l'art qui restent souvent focalisées sur un seul niveau d'intensité.

3.1 État de l'art

Les approches de reconnaissance des expressions s'appuient généralement sur des descripteurs et des modèles faciaux. Les descripteurs encodent l'information concernant certaines caractéristiques faciales. L'information encodée peut être de nature différente : texture, mouvement ou géométrie. Le modèle facial correspond à la manière dont les régions étudiées sont disposées sur le visage. Le modèle facial est construit de sorte à ce qu'un maximum d'information discriminante puisse être pris en compte.

Selon l'étendue temporelle de l'analyse de visages pour caractériser l'expression, on identifie deux grandes familles : les approches statiques et les approches dynamiques. Les approches statiques caractérisent les déformations induites par les expressions à un instant donné en s'appuyant sur la caractérisation de la texture, de la géométrie faciale ou des deux, comme dans les travaux de Zhong et al. (2012a), Wan et Aggarwal (2014), Gonzalez et al. (2011). Généralement les auteurs appliquent ces descripteurs sur les images présentant des niveaux d'intensité forts, souvent assimilable à l'*apex* (instant culminant) de l'expression. Néanmoins, les expériences menées par Bassili (1979) montrent toutefois que les expressions faciales peuvent être reconnues plus précisément en considérant la séquence d'activation dans son intégralité. Ainsi, les approches dynamiques caractérisent les déformations induites par l'expression pendant un laps de temps. Cela peut se matérialiser soit par l'évolution dans le temps de caractéristiques de texture, de géométrie, ou directement par le mouvement.

La caractérisation des déformations faciales peut être faite à différents niveaux d'échelle. Les premiers travaux s'inscrivent dans la famille d'approches holistiques, qui traitent le visage comme un tout pour reconnaître l'expression faciale. Ensuite, certaines solutions mesurent les changements intervenus localement à proximité de certaines régions. Par exemple, on retrouve des approches qui étudient les déformations (apparition de rides, accentuation de fossettes) ou les changements géo-

2. B. Allaert ; I.M. Bilasco ; C. Djeraba - Consistent Optical Flow Maps for full and micro facial expression recognition - Proc. of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2017, Porto, Portugal, vol. 5, pp.235-242.

métriques (étirement ou levée) dans la proximité immédiate des coins de la bouche, des sourcils, etc. D'autres approches locales visent une caractérisation dense des régions qui s'étalent au-delà de la proximité immédiate des points caractéristiques. Cela permet de caractériser les déformations apparaissant dans les régions lisses au repos telles que les joues ou le front. L'ensemble de ces approches se décline pour un usage statique ou dynamique. Dans les situations statiques, les changements sont souvent mesurés par rapport à une situation de référence (visage neutre). Dans un cadre temporel les changements intervenant entre les trames successives d'une séquence sont corroborés afin de construire une signature temporelle de l'expression.

Dans la suite, nous commençons par passer en revue les descripteurs communément utilisés pour caractériser les régions du visage. Nous discutons ensuite les modèles faciaux agrégeant les descripteurs obtenus dans les différentes régions du visage considérées. Les descripteurs et les modèles faciaux restent les principaux objets d'études. Toutefois, d'autres problématiques pertinentes dans le contexte de la reconnaissance des expressions sont abordées dans cet état de l'art. Nous nous intéressons également à l'optimisation de la construction de descripteurs globaux issus de processus d'extraction de caractéristiques et à la capacité des systèmes de reconnaissance à traiter une large palette d'intensités, et notamment les micro- et les macro-expressions.

Descripteurs

Les algorithmes caractérisant les expressions faciales d'un point de vue statique (à l'instant t) peuvent être regroupés en trois catégories :

— les approches qui exploitent la géométrie des éléments du visage ;
— les approches qui s'intéressent essentiellement à l'apparence du visage ;
— les approches hybrides qui combinent à la fois des informations de texture et de géométrie.

Les premiers travaux dans le domaine se concentrer sur la géométrie. En employant des modèles actifs d'apparence (Active Appearance Model - AAM) ou des déclinaisons telles que les modèles actifs de forme (Active Shape Model - ASM), on étudie les correlations entre la disposition faciale des points caractéristiques et les expressions sous-jacentes. Par ailleurs, la position de ces points caractéristiques sert à la fois pour étudier les variations et déduire des mouvements (écartement des lèvres, levée des sourcils, etc.) ou bien pour définir les régions qui sont caractérisées ensuite par les descripteurs de forme ou de mouvement comme dans les travaux de Majumder et al. (2014).

Les approches s'intéressant à l'apparence exploitent les caractéristiques de texture soit localement (plusieurs descripteurs pour des régions spécifiques), soit globalement (un descripteur pour tout le visage). Local Binary Pattern (LBP) proposé par Ojala et al. (1996) est l'un des descripteurs de texture le plus largement utilisé dans la communauté. Il caractérise les variations relatives de l'intensité lumineuse autour d'un point de référence dans une région du visage. LBP est invariant aux changements d'orientation et d'illumination offrant ainsi une certaine robustesse aux processus de classification sous-jacents. Des descripteurs LBP calculés sur des régions distribuées de manière homogène sur le visage, comme suggéré par Tang et al. (2012), peuvent conserver de manière robuste les structures locales du visage. Shan et al. (2005) utilisent LBP afin d'obtenir une description plus complète du visage en utilisant des sous-fenêtres de différentes tailles recouvrant la quasi-totalité du visage.

A l'opposé de ces descripteurs conçus pour extraire des caractéristiques précises, d'autres descripteurs sont générés à partir de données elles-mêmes. Par exemple, inspirés des approches à base de mots visuels introduits par Sivic et Zisserman (2003), Chanti et Caplier (2018) mettent en place des mécanismes spécifiques qui visent à construire des mots visuels ayant un fort pouvoir discriminant dans le cadre de la reconnaissance des expressions. En parallèle, ces dernières années avec l'avènement des outils issus de la communauté d'apprentissage profond, de nombreux descripteurs appris en employant des architectures complexes de réseaux de neurones émergent. Gourmands en données, ces systèmes construisent des descripteurs performants qui encodent au mieux les caractéristiques implicites des données. Les expériences conduites par Lopes et al. (2017) montrent clairement la forte dépendance entre les données disponibles pour l'apprentissage et les performances obtenues. Ainsi, en présence d'un nombre limité de données, les descripteurs auto-encodés obtiennent des performances d'environ 86,67% sur les données initiales du corpus CK+ proposé par Lucey et al. (2010) et 96,76% sur les données augmentées générées à partir du même corpus.

Chaque catégorie évoquée ci-dessus se décline dans une version dynamique caractérisant l'évolution dans le temps de la géométrie ou de l'apparence du visage. Les approches de reconnaissances à base de descripteurs de type LBP ont été étendues afin de tenir compte de l'évolution des changements en termes de déformation sur le visage. Parmi ses extension notons Local Binary Pattern on Three Orthogonal Plans (LBP-TOP) proposé par Zhao et Pietikainen (2007) qui encode les variations d'intensités au sein d'une région dans l'espace et dans le temps. Ce type de descripteur nécessite un alignement parfait du visage sur les différentes trames considérées. Récemment, les approches à base de flux gagnent de l'intérêt auprès de la communauté s'inscrivant parmi les solutions les plus utilisées dans le domaine de la reconnaissance des expressions comme illustré par les travaux de Liao et al. (2013), Fortun et al. (2015). L'utilisation du flux optique est populaire car celui-ci reflette naturellement la dynamique locale de l'apparence dans le temps.

Les approches hybrides combinent les descripteurs géométriques et d'apparence. La combinaison de descripteurs de natures différentes offre généralement des résultats plus robustes. Kotsia et al. (2008) soulignent que les lacunes d'un type de descripteur peuvent être comblées par un autre. Han et al. (2014) utilisent un AAM pour adapter au mieux un modèle facial de caractérisation à base d'une grille afin de reconnaître les unités d'actions (AUs) composant une expression. Dans le domaine de l'apprentissage profond, Jaiswal et Valstar (2016) utilisent une combinaison de réseaux de neurones mélangeant une architecture convolutionnelle avec une architecture à base de mémoire à court et long terme (CNN-BLSTM). On apprend ainsi conjointement la forme, l'apparence et la dynamique du visage. Ce dernier travail montre que la combinaison des caractéristiques dynamiques extraites par le CNN et le BLSTM reflète de manière synthétique l'information temporelle et offre une solution performante pour la reconnaissance.

Les descripteurs introduits ci-dessus ont été évalués de manière extensive sur des corpus où les expressions sont produites par des acteurs produisant ainsi des déformations faciales exagérées. Peu de travaux ont tenté de valider les mêmes descripteurs dans des situations où l'intensité et le temps d'activation sont très faibles, comme c'est le cas des micro-expressions. Liu et al. (2016b) transposent directement les approches utilisées dans le domaine de la macro-expression à la micro-expression. Les auteurs s'accordent à dire que les changements subtils résultant de l'apparition d'une micro-expression sur le visage sont difficilement quantifiables par les descripteurs utilisés

dans le domaine de la macro-expression. En effet, la faible intensité et la nature locale des changements induits requiert une attention particulière, notamment en termes de séparation entre les véritables mouvements faciaux et le bruit induit par les légers mouvements de la tête ou la discontinuité du mouvement.

Selon Li et al. (2015), les micro-expressions sont difficilement détectables sans l'utilisation explicite de la dimension temporelle. Ainsi, afin de mieux caractériser et reconnaître les micro-expressions, les chercheurs utilisent des LBPs spatio-temporaux qui sont mieux adaptés aux micro-expressions. Liong et al. (2014) étendent LBP-TOP en utilisant des contraintes optiques comme des fonctions de pondération afin d'identifier les faibles mouvements. Wang et al. (2014c) proposent une extension de LBP-TOP basée sur l'intersection de trois lignes passant dans le point central de trois histogrammes. Les auteurs proposent ainsi une représentation plus légère et plus compact en réduisant la redondance au sein du descripteur LBP-TOP. Huang et al. (2016a) proposent un LBP spatio-temporel fortement discriminant en améliorant la projection intégrale et en mélangeant les caractéristiques de texture et de forme.

Même si la plupart des méthodes de reconnaissance de micro-expressions utilisent des descripteurs inspirés de LBP-TOP pour caractériser les changements en termes d'apparence du visage, certains auteurs s'intéressent à d'autres descripteurs dynamiques. Huang et al. (2016b) proposent des motifs spatio-temporels quantifiés localement (Spatio-Temporal Complete Local Quantification Pattern - STCLQP) et ils obtiennent de très bonnes performances. La justification de cette amélioration réside, probablement, dans le fait que le STCLQP encode de manière conjointe l'orientation et l'intensité du mouvement. Li et al. (2015) utilisent l'interpolation temporelle et la magnification du mouvement pour compenser la faible intensité des expressions. Les auteurs montrent que lors de l'interpolation, si les séquences obtenues sont trop longues, le mouvement perçu se dilue et cela ne permet pas d'améliorer les performances. Récemment, Liu et al. (2016b) ont construit un nouveau descripteur Main Directional Mean Optical-flow (MDMO) adapté à la reconnaissance de micro-expressions. En partant d'un flux optique dense, les auteurs extraient le flux moyen principal. Les résultats obtenus montrent que la magnitude est plus discriminante que l'orientation pour les micro-expressions. MDMO obtient de meilleurs résultats que les approches spatio-temporelles dérivées de LBP.

Dans le contexte des micro-expressions, les approches à base de caractérisation dynamique de l'apparence sont les plus performantes. En effet, la caractérisation dynamique de l'apparence permet de détecter les changements subtils perçus sur le visage. Ces approches ne requièrent pas d'amples variations ou de changements dans la texture ou la géométrie. Dans le paragraphe suivant, nous discutons la manière dont les visages sont segmentés sous la forme d'un modèle facial afin d'extraire à travers les descripteurs locaux ou globaux un maximum d'information discriminante.

Modèles faciaux

Les modèles faciaux décrivent la manière dont le visage est segmenté pour assurer une interpretation efficace, tout en maîtrisant la taille du vecteur de caractéristiques.

Les modèles faciaux, basés sur des informations géométriques, définissent une segmentation du visage qui privilégie l'extraction des informations pertinentes. En supposant que les régions du visage sont bien alignées, la construction d'histogrammes dans chaque cellule d'une grille ancrée

sur le visage est une pratique toujours d'actualité dans la caractérisation du visage comme illustré par les récents travaux de Fan et Tjahjadi (2017). Toutefois, des décalages dans l'alignement apparaissent souvent. Ils sont principalement causés par les déformations faciales induites par les expressions elles-mêmes. Dans la plupart des situations, les caractéristiques géométriques servent à aligner les régions significatives du visage (yeux, coins des lèvres, sourcils) au sein de modèles faciaux spécifiques. C'est à proximité de ces régions actives que certains auteurs tels que Happy et Routray (2015) extraient des descripteurs d'apparence augmentant de manière significative les taux de reconnaissance.

Des travaux récents utilisent les points caractéristiques afin de définir des régions faciales dont l'alignement est plus précis, même en présence de déformations dues aux expressions. Jiang et al. (2014) définissent un maillage construit sur la base d'un ASM, recouvrant l'intégralité du visage. Dans chacune des régions délimitées par le maillage, des descripteurs d'apparence sont calculés. Sadeghi et al. (2013) utilisent un modèle géométrique fixe pour normaliser les visages. Le visage est décomposé en régions de petite taille et ensuite, des LBPs sont extraits dans chaque région.

La disposition spatiale des régions d'intérêt permet d'optimiser la manière dont les caractéristiques locales sont extraites. Les modèles faciaux basés sur des régions, blocs, maillages ou masques de pondération ont été explorés dans la littérature. Toutefois, nous pouvons identifier deux problèmes toujours d'actualité que nous étudions plus en détails dans les paragraphes suivants. Premièrement, nous nous intéressons aux travaux relatifs à la réduction de la taille des vecteurs de caractéristiques résultant de la multiplication de régions locales. Deuxièmement, nous discutons de l'identification d'un cadre commun pour la reconnaissance de micro- et macro-expressions.

Optimisation de vecteurs globaux de caractéristiques

Indépendamment de la manière dont les descripteurs sont extraits, de manière locale ou globale, de manière statique ou dynamique, il est nécessaire d'éliminer les données à faible pouvoir discriminant. Prenons l'exemple des approches statiques locales qui sont omniprésentes dans le domaine de l'analyse du visage. On peut constater qu'elles s'appuient principalement sur la détection des éléments du visage tels que les yeux, les sourcils, le nez, la bouche comme résumé par Lam et Yan (1998). Ces éléments sont caractérisés soit par les relations géométriques, soit par des descripteurs renseignant les caractéristiques visuelles des régions sous-jacentes. Comme discuté dans les paragraphes précédents, la disposition spatiale des régions d'intérêt peut améliorer la caractérisation du visage. L'obtention d'une représentation optimale passe également par la sélection ou le remodelage des caractéristiques extraites dans le but d'obtenir une représentation plus compacte et plus discriminante.

Afin de disposer de représentations compactes et performantes, deux types d'approches se dégagent. L'une vise à sélectionner les caractéristiques les plus discriminantes, l'autre vise la transformation de l'espace de descripteurs dans un nouvel espace capable d'en proposer de nouveaux plus discriminants.

La factorisation matricielle NMF popularisée par Lee et Seung (2001) est une méthode reconnue de réduction de dimensionnalité. Le processus de factorisation consiste dans l'approximation de la matrice initiale par le produit de deux matrices non-négatives ayant des rangs inférieurs. NMF se distingue d'autres approches en tirant avantage de la nature non-négative des matrices. Utilisée souvent par les approches holistiques, NMF est capable de mettre en évidence les parties du visage

pertinentes selon l'analyse menée. Toutefois, l'utilisation de NMF peut poser certains problèmes : la factorisation ne peut se dérouler à la volée - l'ensemble des corpus doit être caractérisé en amont et la convergence est lente. Guan et al. (2012b) proposent une variante de NMF nommée OR-NMF qui peut s'appliquer de manière incrémentale et efficace sur de gros volumes de données. Les problèmes de convergence lente ont été adressés dans la variante NENMF décrite par Guan et al. (2012a) et la variante MD-NMF proposée par Guan et al. (2011).

Lorsque l'on s'intéresse à des approches holistiques, le visage dans sa totalité peut être encodé comme un point dans un espace de dimension très importante. Dans cette optique, le visage est considéré comme un tout. Des méthodes telles que Principal Component Analysis (PCA) (Turk et Pentland 1991), Linear Discriminant Analysis (LDA) (Etemad et Chellappa 1997) ou Independent Component Analysis (ICA) (Hyvärinen et Oja 2000) sont utilisées afin d'identifier les dimensions de l'espace de représentation qui présentent une variabilité limitée et qui ne permettent pas de contribuer de manière importante à l'analyse discriminante. Par exemple, PCA réduit la dimensionnalité en supposant que la variance d'une caractéristique marque son importance dans le processus de classification sous-jacent. Cependant, cette reduction de dimensionnalité ne convient pas à toutes les approches, notamment lorsque différents sous-ensembles du visage sont pertinents pour différentes tâches de classification.

Des algorithmes explicites de sélection séquentielle qui exploitent le pouvoir discriminant de chaque descripteur ont été proposés par Whitney (1971) et popularisés par Pudil et al. (1994). Des approches de type en aval ou en amont sont mises en place pour faire une sélection incrémentale de caractéristiques. Dans une approche de type aval on sélectionne en premier le descripteur qui individuellement atteint le meilleur score. Ensuite, on sélectionne le descripteur qui avec le précedemment sélectionné offre les meilleurs résultats. Et ainsi de suite, jusqu'à couvrir l'ensemble des descripteurs. Une solution sous-optimale peut être ainsi identifiée en s'affranchissant de la combinatoire inhérente à l'identification d'une solution optimale. De manière similaire, une approche de type amont part de l'ensemble global en enlevant un à un les descripteurs en minimisant à chaque fois l'impact sur les résultats obtenus. Des variantes combinant de manière intelligente les sélections en amont et en aval ont vu le jour (Somol et al. 1999). Toutefois, dans certaines situations qui considèrent de nombreux descripteurs (comme par exemple, l'intensité de pixels d'un visage) le coût exploratoire reste conséquent. La communauté continue à s'intéresser à ces approches comme illustré par Nakariyakul (2014). En effet, la sélection permet de conserver les données dans leur état initial. Cela permet d'avoir une maîtrise plus forte des descripteurs dérivés en cas de problèmes avec les données initiales (par exemple, bruit ou occultations). C'est pour ces raisons que nous explorons dans la section 3.2 une solution capable de trouver de manière intelligente les pixels contribuant le plus à la reconnaissance d'une expression tout en limitant le coût exploratoire.

Dans la suite, nous nous intéressons à la capacité des méthodes de reconnaissance des expressions à partir de vidéos de supporter un large spectre d'intensités en partant de micro-expressions jusqu'au macro-expressions.

Vers un traitement commun de micro- et macro-expressions

Les micro-expressions sont assez différentes des macro-expressions en termes d'amplitude de mouvement ou changement de texture, ce qui les rend difficilement reconnaissables. Dans le Table

3.1 nous présentons certaines approches représentatives issues de la reconnaissance de micro- et macro-expressions. Nous pouvons observer qu'il subsiste encore une forte différence en termes de taux de reconnaissance entre les deux contextes applicatifs.

La Table 3.1 passe en revue les performances relatives à la micro- et macro-expression obtenues par des approches issues des catégories discutées précédement : les approches caractérisant la texture statiquement ou dynamiquement, les approches s'intéressant à la géométrie du visage et les approches exploitant le mouvement dense. Les performances ont été obtenues sur deux corpus representatifs : CK+ (Lucey et al. 2010) pour les macro-expressions et CASME2 (Yan et al. 2014a) pour les micro-expressions. La Table 3.1 souligne la grande différence en termes de taux de reconnaissance entre les deux contextes (micro vs. macro) lorsque le même modèle facial et les mêmes descripteurs sont utilisés. Il est évident que les approches relatives aux deux contextes ne sont pas directement comparables, car les données elles-mêmes sont distinctes. Néanmoins, nous les présentons ensemble afin de mettre en évidence que les méthodes qui obtiennent de très bonnes performances dans une situation, ne permettent pas d'atteindre des performances comparables dans l'autre. Afin d'établir un classement entre les méthodes au sein d'une même catégorie, toutes les méthodes citées dans la Table 3.1 utilisent un classifieur SVM avec un protocole de validation croisée découpant le corpus en dix échantillons. Les méthodes visant les micro-expressions utilisent communément le protocole *leave-one-subject-out* (LOSO) dans le cadre d'une validation croisée.

TABLE 3.1 – *Méthodes représentatives et récentes pour la reconnaissance de micro- et de macro-expressions (* magnification).*

Basée sur	Macro-expression (CK+)		Micro-expression (CASME2)	
LBP	LBP Shan et al. (2009) Blocs	90,05%	LBP Li et al. (2015) Blocs	55,87%
HOG	PHOG Khan et al. (2012) Region Saillante	95,30%	HIGO Li et al. (2015) Blocs	57,09% *67,21%
LBP-TOP	LBP-TOP Zhao et Pietikainen (2007) Bloc	**96,26%**	DiSTLBP-IIP Huang et al. (2016a) Bloc	64,78%
Descr. Géom.	Gabor Jet Ghimire et Lee (2013) Points caractéristiques	95,17%	/	
Flux optique	Flux optique Allaert et al. (2017) Maillage facial	93,17%	MDMO Liu et al. (2016b) Maillage facial	**67,37%**

Comme illustré dans la Table 3.1, les descripteurs statiques reconnus (par exemple, LBP) ne présentent qu'un potentiel limité pour la reconnaissance de micro-expressions. L'absence d'adéquation est attribuable au fait que les descripteurs statiques ne sont pas capables de discriminer les mouvements de faible intensité comme suggéré par Li et al. (2015).

Les approches géométriques fournissent de très bons résultats pour les macro-expressions. En revanche, elles échouent clairement en présence de micro-expressions. Les mouvements subtils n'affectent la géométrie que de manière infime et négligeable. En pratique, la détection ou le suivi des points caractéristiques sur le visage ne sont suffisamment précis pour permettre, par exemple, de distinguer le bruit de détection et un véritable mouvement infime d'un coin de lèvre.

En ce qui concerne les approches à base d'apparence, Huang et al. (2016a) montrent que la caractérisation dynamique de texture semble la plus appropriée pour encoder de manière significative les mouvements faciaux de faible amplitude. Plus particulièrement, les méthodes à base de flux optique semblent présenter un grand potentiel pour la reconnaissance de micro-expressions comme attesté par les travaux de Liu et al. (2016b). Cependant, les approches à base de flux optique sont souvent critiquées, car le calcul du flux est fortement perturbé en présence de changements

de luminosité ou de discontinuités de mouvement. LBP-TOP affiche des performances encourageantes à la fois pour les micro- et les macro-expressions. Des nombreux travaux visant les micro-expressions se sont concentrés sur des descripteurs inspirés de LBP-TOP pour la reconnaissance de micro-expressions. En parallèle, les travaux récents autour de l'estimation de flux optique ont évolués afin de mieux supporter le bruit. La majorité des algorithmes de calcul de flux optique, comme celui de Revaud et al. (2015), est basée sur des processus de filtrage et de lissage du mouvement afin de réduire l'impact des discontinuités. Ces nouveaux algorithmes améliorent la qualité perçue visuellement d'un flux optique. En revanche, en présence de mouvements de forte et faible intensité, le lissage à tendance à induire des faux mouvements. Une deuxième approche permet de palier les difficultés de caractérisation de faibles mouvements en amplifiant artificiellement les déformations induites par le mouvement. Cette technique appliquée avec succès par Li et al. (2015) en présence de micro-expressions permet d'augmenter de significativement les performances. Naturellement, la magnification n'est pas adaptée pour les macro-expressions, car en présence de macro-expressions elle déforme de manière trop importante la morphologie du visage.

Au vu des résultats présentés, il est toujours difficile d'utiliser efficacement les mêmes descripteurs et les mêmes modèles faciaux dans la reconnaissance de micro- et macro-expressions. Les travaux représentatifs utilisent les mêmes descripteurs pour la reconnaissance de micro- et macro-expressions, en revanche, il est nécessaire de changer le modèle facial et le processus de traitement sous-jacent pour maximiser les résultats dans les deux situations.

Synthèse

Dans cette section nous avons discuté de différents problèmes encore d'actualité pour la reconnaissance des expressions dans un contexte d'analyse statique ou dynamique. Dans un contexte statique nous avons mis en exergue qu'en présence d'occultations et autres bruits de mesure, les méthodes de réduction de dimensionnalité, changeant d'espace de représentation, ne semblent pas adaptées. Comme les expressions impliquent un certain dynamisme de la part de la personne pouvant générer du bruit (par exemple, des occultations ou mouvements de la tête), il est intéressant de privilégier des approches de sélection de caractéristiques qui n'affectent pas la nature des données interprétées.

Dans un contexte dynamique, malgré le fait que des tendances similaires se dégagent pour les micro- et macro-expressions, il est toujours difficile de trouver une méthodologie commune pour analyser conjointement et de manière précise les macro- et micro-expressions. Toutefois, dans les deux situations, les approches dynamiques (LBP-TOP et flux optique) obtiennent les meilleurs résultats. En partant de ce constat, nous proposons une nouvelle méthode qui s'appuie sur la caractérisation cohérente du mouvement dense afin de reconnaître tant les micro- que les macro-expressions.

Dans la suite de ce chapitre, nous présentons (par ordre chronologique) nos contributions. Dans la section 3.2, nous exposons une approche automatique pour identifier, sur un visage, les pixels qui contribuent de manière importante à la reconnaissance de la *joie*. Au-delà d'une méthode pour identifier la *joie*, la principale contribution de cette étude est de proposer une solution pratique pour améliorer les performances de classification tout en conservant les caractéristiques de l'image initiale. Dans la section 3.3, nous proposons un nouveau descripteur de mouvement nommé Local

Motion Patterns (LMP) capable de caractériser fidèlement le mouvement facial indépendamment de l'intensité de l'expression. En partant d'un simple flux optique dense, nous obtenons un descripteur capable de réduire le bruit de caractérisation en présence de discontinuités de mouvements ou de légers mouvements de la tête. Nous explorons également la construction d'un modèle facial pour optimiser la caractérisation globale du visage en présence de macro- et de micro-expressions.

3.2 Construction de masques intelligents de pixels pour la reconnaissance de l'expression de joie

Dans ce travail, nous utilisons une approche analytique qui réalise une sélection de pixels en analysant les régions du visage pour retrouver les pixels qui contribuent de manière prépondérante au processus de reconnaissance des expressions. Par expression, nous pouvons déduire ainsi un masque spécifique qui permet d'améliorer le processus de classification. Nous explorons de manière intelligente et raisonnée les sous-ensembles de pixels permettant d'obtenir à eux seuls de bonnes performances de reconnaissance. À partir de sous-ensembles rectangulaires contenant les pixels d'intérêt qui alimentent le processus de classification, nous séparons les meilleures et les pires régions de détection. Des opérations morphologiques et un filtrage au niveau des pixels, qui tiennent compte de performances obtenues, nous permettent de délimiter de manière fine les contours des régions contenant les pixels contribuant le plus à la prise de décision. Pour évaluer les performances de chaque configuration, nous employons comme outil de classification un réseau de neurones multicouches (MLP). Toutefois, la démarche proposée est indépendante du processus de classification sous-jacent. La seule information exploitée dans le cadre de l'extraction de pixels d'intérêt réside dans les taux de classification obtenus par les régions couvrant chaque pixel individuellement par rapport à l'intégralité des régions considérées. Les opérations ensemblistes sur les meilleures et les pires régions (union, intersection, différence) produisent des régions à forme et étendue variables. Ces cartes permettent d'exclure les données non-pertinentes, améliorant ainsi le processus de classification, sans avoir recours à des filtres spécifiques transformant l'image initiale.

Dans les travaux de Chen et Chen (2010) il est montré que le problème de reconnaissance des expression faciales doit être envisagé plutôt comme un problème local et moins comme un problème concernant l'intégralité du visage. En effet, les auteurs utilisent une ellipse centrée sur le visage dont le but principal est d'enlever les points dans l'arrière plan, afin de disposer des données préenregistrées dans le processus de classification. Ce type de masque élimine les cheveux et le cou comme illustré dans la Figure 3.1.

FIGURE 3.1 – *Masques elliptiques de Chen et Chen (2010) appliqués sur des visages du corpus JAFFE (Shinohara et Otsu 2004).*

Toutefois, il subsiste d'autres pixels de peau qui n'enrichissent pas la qualité des descripteurs extraits et qui par ailleurs peuvent être assimilés à du bruit. Ainsi, des masques supplémentaires

et de nouvelles méthodes sont nécessaires pour éliminer les pixels ne contribuant pas de manière directe au processus de classification.

L'approche que nous avons retenue s'apparente aux travaux de Song et al. (2010) qui définissent la notion de motif de contexte visuel (*visual context patterns* - VCP) pour faciliter la détection des yeux sur un visage. De manière similaire, les auteurs explorent les plus petites régions de référence qui maximisent une fonction qui satisfait un critère lié à l'apparence stable tout en préservant au minimum l'apparence variable. Dans notre approche, nous recherchons d'un côté, des sous-régions qui maximisent et de l'autre, des régions qui minimisent les performances de classification. Toutefois, au lieu de sélectionner les plus petites régions, nous utilisons un critère d'élimination des régions par rapport à l'écart-type des performances de classifications observées dans le cadre du processus d'exploration. En termes de méthodologie d'apprentissage, Song et al. (2010) utilisent une méthode semi-supervisée où les faux positifs servent à alimenter une nouvelles fois le classifieur lors d'une deuxième étape d'apprentissage. Dans notre cas, nous créons le masque en fin de processus d'exploration, en considérant les performances obtenues sur chacune des régions considérées. Notre proposition est ainsi fortement parallélisable et indépendante de l'ordre d'exploration des régions.

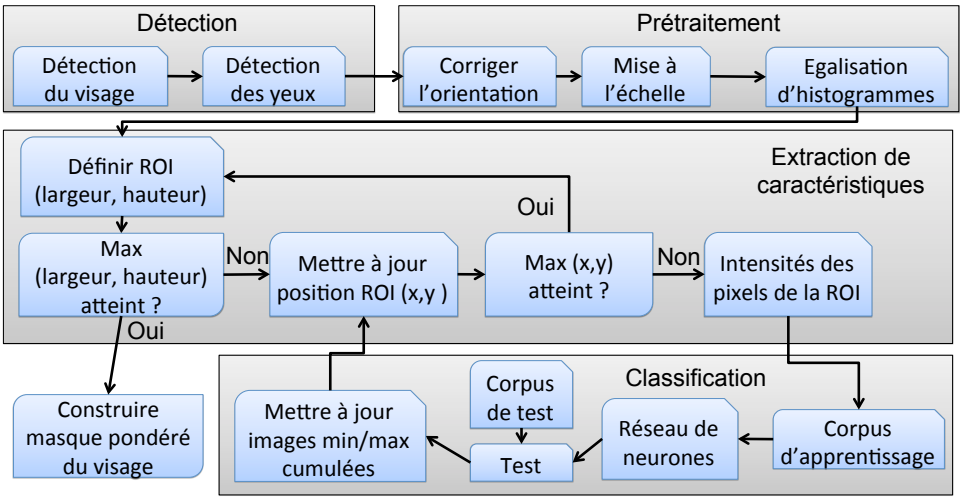


FIGURE 3.2 – *Diagramme de flux de notre méthode.*

Dans ce travail, nous extrayons des masques non-rectangulaires pour améliorer les résultats de classification dans le problème de reconnaissance des expressions faciales. La Figure 3.2 montre le diagramme général de notre solution. Les régions testées sont de forme rectangulaire, de taille variable et positionnées à divers endroits sur un ensemble d'images statiques utilisées pour l'apprentissage et le test. Les meilleures et pires régions rectangulaires sont ensuite analysées afin de produire des masques non-rectangulaires. Dans la suite, nous décrivons les étapes du processus en commençant par la normalisation du visage.

3.2.1 Normalisation de visages

Le but de cette étape est de préparer le visage contenu dans l'image pour le processus de caractérisation. En effet, les visages doivent être présentés dans des conditions similaires. Ainsi, nous avons adopté un prétraitement similaire à celui introduit dans le cadre de la reconnaissance du genre (voir section 2.2). Suite à la détection du visage, nous procédons à la détection des yeux

afin de pouvoir corriger l'orientation du visage à partir de la position verticale des pupilles gauche et droite.

Après la correction de l'orientation du visage, le visage est transformé dans une représentation en niveaux de gris comportant 50×50 pixels. La portion de l'image initiale ainsi transformée est calculée en fonction de la distance inter-pupillaire DIP. Les coordonnées du visage dans l'image suite à l'application de la correction d'orientation sont calculées ainsi :

$$V_x = Oeil_{Gauche x} - DIP/3,\ V_y = Oeil_{Gauche y} - DIP/2{,}5,\ V_l = DIP \times 1{,}6,\ V_h = DIP \times 1{,}9 \quad (3.1)$$

où V_x, V_y, V_l et V_h représentent l'origine (x, y), la *largeur* et la *hauteur* du visage dans l'image normalisée. $Oeil_{Gauche x}$ et $Oeil_{Gauche y}$ sont les coordonnées x et y de l'oeil gauche par rapport à l'origine de l'image. Les coordonnées du visage initialement obtenues en appliquant l'algorithme de Viola et Jones (2004) sont remplacées par les valeurs V_x, V_y, V_l et V_h calculées en fonction de la DIP comme ci-dessus. Les facteurs $1,6$ et $1,9$ appliqués à la DIP, pour déterminer la largeur et l'hauteur du visage normalisé, ont été choisis en relation avec les propriétés de visages afin de concentrer l'analyse sur le centre du visage en enlevant autant que possible des éléments de l'arrière-plan. Ce type de découpage limite également les occultations avec les autres éléments du visage et notamment les cheveux ou la barbe. Après normalisation, nous appliquons un processus d'égalisation d'histogrammes afin d'augmenter le contraste et accentuer les traits des images principales. La Figure 3.3 montre les visages initiaux et les visages normés.

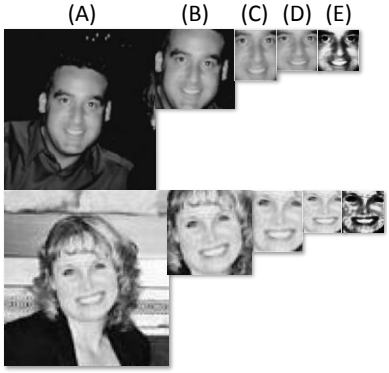


FIGURE 3.3 – *Étapes du prétraitement appliquées sur des images extraites depuis la source de données GENKI-4K. (A) Image originale. (B) Détection initiale du visage. (C) Rotation et recadrage du visage en fonction de la distance inter-pupillaire (IPD). (D) Normalisation du visage à une taille de* 50×50 *pixels. (E) Égalisation d'histogrammes.*

3.2.2 Méthode de construction de masques faciaux

Nous avons réalisé des expériences de classifications de visage joyeux/non-joyeux sur différents corpus de données. Dans cette étude, tous les corpus de données employés sont restructurés de sorte à obtenir deux classes : *joie* et *non-joie*. La classe *non-joie* regroupe l'ensemble des expressions à part la *joie*.

Nous considérons un ensemble de m régions $R_k = \{x, y, l, h\}$, $k = \{1, .., m\}$ où x, y, l et h représentent respectivement le coin haut gauche (x, y), la largeur (l) et la hauteur (h) de chaque région rectangulaire. $A(R_k)$ représente la précision de classification du MLP en considérant uniquement le contenu de la région R_k. Afin de constituer les descripteurs caractérisant les visages, les vecteurs D_k

sont extraits de chaque région R_k. Ces vecteurs D_k alimentent la couche d'entrée d'un réseau neuronal ayant 2 couches cachées comme illustré dans la Figure 3.4. Ainsi, afin de trouver les meilleures $A(R_k)_m$ et les pires $A(R_k)_p$ régions, les processus de classification se déroulent en considérant les descripteurs D_k. Pour chaque R_k, l'ensemble d'apprentissage est construit en sélectionnant pour chaque expression 50% de données. Les 50 autres pourcents constituent l'ensemble de test.

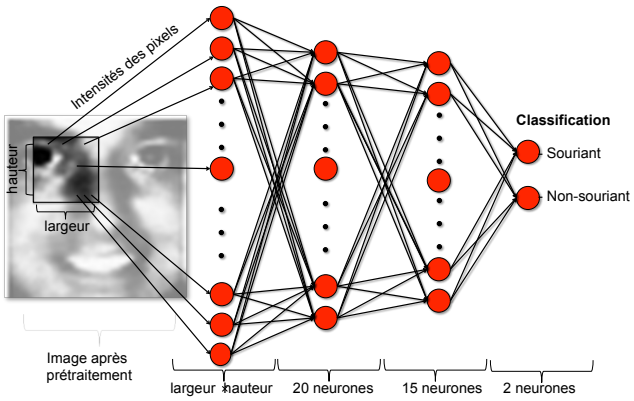


FIGURE 3.4 – *Architecture du réseau de neurones ayant deux couches cachées. Le nombre de neurones sur la couche d'entrée est déterminée de manière dynamique en fonction de la taille courante (largeur, hauteur) de la région de recherche.*

Soit $\overline{M(R_k)}_m$ et $\overline{M(R_k)}_p$ les moyennes obtenues en considérant les meilleures et les pires régions d'intérêt. Elles sont caractérisées respectivement par une distribution moyenne μ_m, μ_p et un écart-type $V(\mu_m)$, $V(\mu_p)$. La puissance de prédiction d'une région R_k est mesurée en termes de taux d'erreur de classification obtenu en considérant l'ensemble d'apprentissage construit à partir de la région R_k. Le masque C est créé en réalisant des opérations topologiques sur les régions sélectionnées comme indiqué dans l'Équation 3.2. Nous éliminons les meilleures régions parmi les pires et les pires régions parmi les meilleures en nous servant d'un seuil calculé en fonction de la moyenne et l'écart-type de la distribution des pires, respectivement meilleures régions.

$$C = (R_k | A(R_k)_m > \overline{A(R_k)}_m + V(\mu_m)) \wedge (Visage - R_k | (A(R_k)_p < \overline{A(R_k)}_p - V(\mu_p))) \tag{3.2}$$

3.2.3 Évaluation

Compte tenu du temps important requis pour réaliser l'apprentissage et les tests pour chaque région, nous avons optimisé et parallélisé l'implémentation réalisée afin de réduire le temps requis par notre approche. Sans perte importante de précision, en ce qui concerne le masque construit, nous déplaçons les régions de recherche à l'horizontale et à la verticale d'un pas de trois pixels pour chaque nouvelle région.

Nous avons utilisé la moitié du corpus GENKI-4K (MPLab 2011) comme ensemble d'apprentissage. Les corpus de données JAFFE (Lyons et al. 1999) et FERET (Phillips et al. 2000) ont été utilisés pour les tests et l'évaluation globale de l'approche. Dans chaque expérience les ensembles de test et d'apprentissage sont strictement séparés.

Pour une taille de visage de 50×50 pixels, nous avons considéré $m = 224$ types de régions de recherche différentes R_k disposées à différents endroits sur l'image. Cela a généré un total de 14 490 réseaux de neurones différents. Quand toutes les tailles ont servi à alimenter les réseaux

correspondants, nous gardons, par type de région, les positions où les meilleurs et les pires taux de classification ont été obtenus comme indiqué dans la Figure 3.5. La moyenne sur l'ensemble des régions est de 71,7% avec un écart-type de 8,71%.

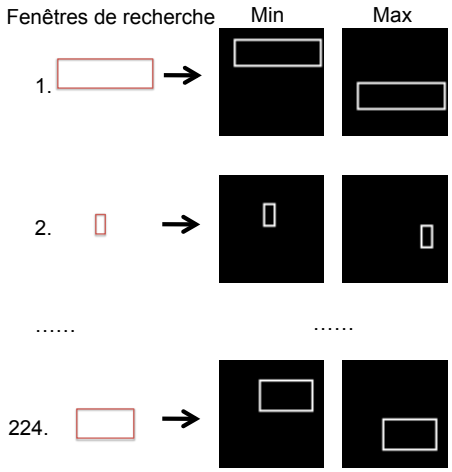


FIGURE 3.5 – *Meilleure et pire positions des régions de tailles différentes par rapport aux résultats de classification.*

La Figure 3.6 (A) et (B) montre les 224 régions et illustre les pires et les meilleures positions de ces régions, lorsqu'elles sont appliquées sur les visages de taille 50×50. Les régions placées en haut de visage donnent plutôt de mauvais résultats, et lorsque les régions sont placées autour de la bouche et le centre du visage les résultats de classification sont plutôt élevés. Cela conforte les théories qui défendent l'idée que les expressions positives se manifestent généralement dans la partie inférieure du visage, alors que la partie supérieure de visage reflète plus l'apparition des expressions négatives. Dans la Figure 3.6(A), de gauche à droite et du haut vers le bas, le taux de classification augmente de 33,1% à 85,7%. Alors que dans la Figure 3.6 (B), le taux de classification varie entre 74,9% et 87,8%.

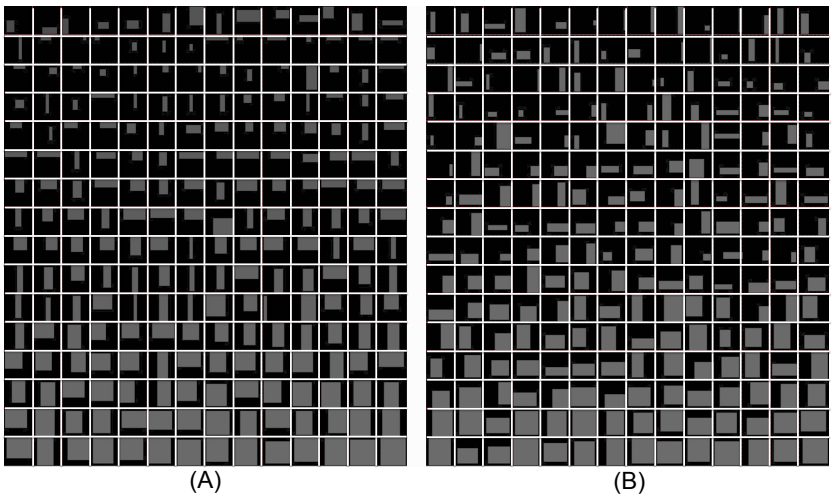


FIGURE 3.6 – *Pires (A) et Meilleures (B) régions d'intérêt pour le processus de classification.*

En se référant à la Figure 3.7, il est évident que plus la région de recherche est grande, plus les résultats de classification augmentent. Afin de quantifier la force du lien entre la taille des régions et les meilleurs taux de bonne classification, nous avons calculé le carré du coefficient de corrélation entre la taille des régions et les résultats de classification. Approximativement 64% des variations de taux de classification peut être expliqué par la taille de la région comme illustré dans

la Figure 3.7. On peut également noter que quelques petites régions produisent des taux de classification meilleurs que des régions plus grandes, ce qui justifie en partie la possibilité de résoudre le problème de reconnaissance en s'appuyant sur les techniques de sélection de caractéristiques.

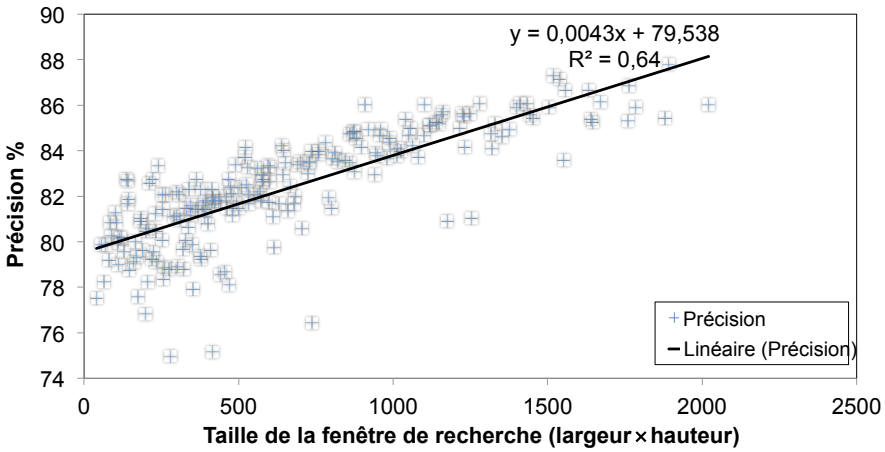


FIGURE 3.7 – *L'effet de la taille des régions sur les taux de bonne classification en considérant les meilleurs résultats obtenus sur la base de données GENKI.*

La Figure 3.8 illustre la construction du masque selon l'Équation 3.2 sur le corpus GENKI-4K. Par convention, nous estimons que les hautes valeurs de bonne classification sont corrélées avec l'importance de la région correspondante dans le processus de reconnaissance. Les valeurs basses traduisent une faible importance de la région dans le cadre de l'apprentissage. Les résultats obtenus en superposant le masque sur quelques images de la base de données sont visibles dans la Figure 3.9.

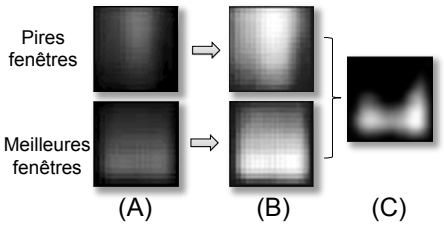


FIGURE 3.8 – *Construction du masque optimal. (A) Images moyennes pour les meilleures et pires régions (B) Étalement du contraste appliqué sur la version (A) pour une meilleure visualisation. (C) Soustraction à partir de meilleures régions des pires. L'effet d'échiquier dû au saut de pixels (3 pixels à la fois) est visible.*

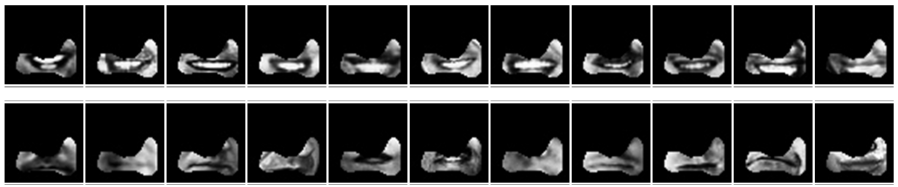
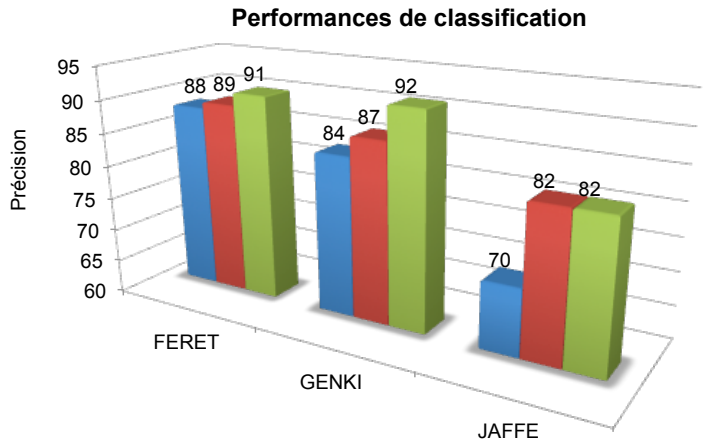


FIGURE 3.9 – *Un sous-ensemble des images de la base de données GENKI suite à l'application du masque pour classer les images de joie et non-joie.*

Nous avons comparé les précisions obtenues en utilisant, respectivement : notre masque, la région englobant tout le visage et la meilleure région rectangulaire de l'ensemble des régions R_k. D'après les résultats des expériences illustrées dans la Figure 3.10, les meilleurs résultats sont obtenus sur la base de données GENKI-4K. Les résultats expérimentaux montrent que, malgré la variété

de situations rencontrées, le masque construit par notre méthode offre les meilleurs résultats. Sur chaque base de données, notre masque a obtenu de meilleurs résultats que les masques rectangulaires classiques ou les masques couvrant le visage tout entier. Les résultats de ces expériences ont été publiés dans (Danisman et al. 2013) [3].



	FERET	GENKI	JAFFE
Visage entier	88	84	70
Meilleure fenêtre	89	87	82
Notre masque	91	92	82

FIGURE 3.10 – *Taux de précision (%) obtenus sur différentes bases de données en considérant le masque construit sur la base du corpus GENKI-4K.*

Dans la section suivante, nous analysons la dynamique des expressions faciales ainsi que les défis amenés par les variations d'intensités en considérant un spectre très large couvrant les micro- et les macro-expressions.

3.3 Variations d'intensité : de la micro- à la macro-expression

De nombreuses approches s'intéressent à la reconnaissance de macro-expressions faciales. Les résultats obtenus dans le domaine de macro-expressions ne sont pas directement transposables au domaine des micro-expressions à cause de l'intensité très faible du mouvement caractérisant ces dernières. En effet, les méthodes proposées dans la littérature traitent souvent soit des micro-expressions, soit des macro-expression et très peu, traitent les deux en même temps. Ceci est dû au fait que la caractérisation des changements induits par les expressions en termes de géométrie et texture faciale sont très différentes. Par exemple, le mouvement perçu lors d'une macro-expression caractérise clairement la déformation faciale induite. Alors que le mouvement perçu dans le cas d'une micro-expression se confond le plus souvent avec le bruit d'observation. Ainsi une attention particulière doit être portée à l'encodage des faibles variations. Ces éléments justifient en partie pourquoi les travaux obtenant de bons résultats pour la reconnaissance de macro-expressions ne peuvent pas être directement transposés aux micro-expressions comme montré par Li et al. (2015).

3. T. Danisman ; I.M. Bilasco ; J. Martinet ; C. Djeraba - Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron - Signal Processing, Elsevier, 2013, Special issue on Machine Learning in Intelligent Image Processing, 93 (6), pp. 1547-1556 (Facteur d'impact : 3,470 selon JCR 2018).

Trouver une méthodologie commune pour analyser à la fois les macro- et les micro-expressions de manière efficace reste un problème ouvert. Il convient de concevoir de nouveaux descripteurs spatio-temporels capables d'encoder de manière robuste l'intensité des mouvements faciaux, ainsi que leur propagation sur le visage. Malgré la grande différence entre les micro- et macro-expressions nous nous efforçons de proposer un modèle facial et des descripteurs capables de traiter de manière unifiée les micro- et les macro-expressions. En nous intéressant aux extrémités du spectre des intensités, nous visons également la mise à disposition d'une solution capable de traiter les niveaux d'intensités intermédiaires.

Dans ce qui suit, nous proposons une méthode qui s'intéresse à l'analyse conjointe de micro- et macro-expressions en proposant un nouveau descripteur de mouvement innovateur qui caractérise les motifs locaux de mouvement (en anglais, *local motion patterns* - LMP). LMP combine l'orientation et la magnitude pour obtenir des descripteurs robustes au bruit. Il exploite les hypothèses de propagation et cohérence locale (en termes d'orientation et de magnitude) du mouvement facial. Nous montrons que ce descripteur encode les mouvements faciaux de manière efficace indépendamment de leur intensité.

Dans le cadre de ce travail, nous partons de l'hypothèse que le seul mouvement présent sur le visage est celui de l'expression. Cela est rarement le cas, car les expressions faciales sont souvent accompagnées par des mouvements de la tête ou du haut du corps. Ces mouvements viennent perturber les mouvements relatifs aux expressions. Deux grandes familles d'approches peuvent répondre à ces défis. La première consiste à aligner les visages avant traitement. Nous avons étudié l'impact de ces techniques sur la reconnaissance de macro-expressions dans Allaert et al. (2018b). Des nombreux artefacts induits par les divers processus de normalisation viennent perturber le processus de reconnaissance. La deuxième famille d'approches consiste en la mise en place des méthodes de compensation de mouvement. Des nombreuses méthodes de compensation dans le plan existent et permettent de caractériser et éventuellement soustraire le mouvement induit par les mouvements de la caméra (Odobez et Bouthemy 1995, Kraemer et Benois-Pineau 2005, Szolgay et al. 2011, Jain et al. 2016). Les expérimentations qualitatives que nous avons réalisées montrent que transposer cela aux mouvements induits par la tête n'offre pas des bons résultats en présence de mouvements hors plan. Proposer des méthodes de séparation de mouvement adaptées aux visages fait partie de nos objectifs futurs tels que décrits dans la partie III de ce manuscrit. Toutefois, notons que dans la section 3.3.3 présentant l'évaluation de notre approche, nous étudions les performances obtenues sur le corpus MMI contenant divers mouvements de la tête.

Dans la section 3.3.1, nous introduisons le descripteur LMP qui exploite la cohérence du mouvement local sur le visage. En partant d'un flux optique brut, nous détaillons le processus de filtrage local qui enlève le bruit tout en préservant les mouvements faciaux. Dans la section 3.3.2, nous discutons plusieurs stratégies pour encoder l'intégralité du mouvement sur le visage en vue de la reconnaissance de macro- et micro-expressions. Les résultats expérimentaux sont détaillés dans la section 3.3.3.

3.3.1 Étude locale du mouvement - Local Motion Pattern

Les caractéristiques de la texture faciale (peau lisse, réflectance de la peau et élasticité) rendent difficile l'extraction précise du flux optique caractérisant le mouvement facial. Afin de répondre

à cette problématique nous mettons en place un filtrage du flux optique qui tient compte de la manière dont les déformations faciales se propagent physiquement sur le visage.

Compte tenu du fait que chaque partie morphologique du visage exhibe une propagation spécifique, nous abordons le problème de filtrage de manière locale, au niveau des régions définies en relation avec le système de codage d'unité d'actions FACS de Ekman et Friesen (1978). Cela nous permet de nous focaliser sur le mouvement concernant les régions intéressantes du visage et d'y appliquer un traitement spécifique. Ainsi, le mouvement extrait reflète uniquement les magnitudes et les orientations correspondantes aux véritables mouvements faciaux. Sur l'ensemble des régions, nous partons de l'hypothèse que le mouvement est uniforme et se propage de manière continue vers les régions voisines. Par exemple, le mouvement généré par le déplacement du coin droit de la bouche se propage dans la joue droite en conservant une orientation similaire, avec un affaiblissement régulier de l'intensité.

Afin de prendre en compte ces hypothèses caractérisant le mouvement et, ainsi, être en mesure de conserver uniquement les mouvements pertinents nous proposons un nouveau descripteur nommé motifs locaux de mouvements (*local motion pattern* - LMP). Le LMP décrit le mouvement cohérent observé autour de son épicentre $\epsilon(x, y)$ tout en analysant la propagation des mouvements dans les zones adjacentes. Chaque zone constitue une région du mouvement local (*local motion region* - LMR). Le LMR est caractérisé par un histogramme $H_{LMR_{x,y}}$. Les LMRs peuvent jouer un double rôle : celui de l'épicentre du mouvement appelé région centrale de mouvement (*central motion region* - CMR), ou celui d'une région voisine (*neighboring motion region* - NMR) impactée par la propagation nommée .

La Figure 3.11 illustre l'organisation d'un LMP. Le LMP est associé à un CMR. Huit NMRs sont générés autour d'un CMR. Le centre d'un NMR se trouve à une distance Δ du centre du CMR associé. Plus cette distance est petite, plus la cohérence entre les régions voisines est supposée forte. λ correspond à la taille de la zone à caractériser autour de l'épicentre. β indique le nombre de propagations à réaliser depuis l'épicentre afin de caractériser convenablement le LMP.

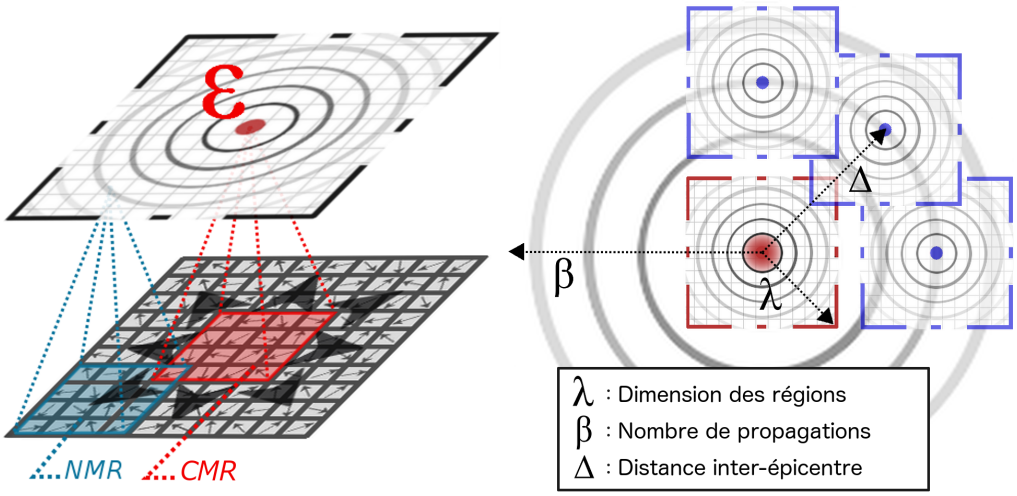


FIGURE 3.11 – *Illustration des éléments intervenant dans la caractérisation d'un motif local de mouvement (LMP).*

Dans la suite, nous présentons en détail la méthodologie utilisée pour extraire le mouvement cohérent à partir d'un flux optique dense. Nous filtrons le bruit en étudiant les hypothèses de mouvement facial caractérisé en évaluant la cohérence locale et la cohérence de la propagation.

Cohérence locale d'un CMR

Afin de séparer le bruit des vrais mouvements, nous analysons la distribution du mouvement au sein de chaque CMR. En présence d'un véritable mouvement, le flux optique a tendance à suivre les lignes principales des déformations induites par le mouvement. Cela se fait selon une ou plusieurs orientations bien identifiées et avec des magnitudes légèrement variées. Le bruit en revanche se retrouve généralement distribué de manière uniforme. Ainsi, dans un premier temps nous analysons le flux optique localement et nous conservons uniquement les directions principales qui se retrouvent de manière récurrente dans plusieurs intervalles de magnitude.

Comme illustré dans la Figure 3.12, en présence de mouvements ayant une forte intensité, la direction principale peut être déduite uniquement sur la base de l'orientation (courbe bleu). De plus, elle est similaire à la direction déduite lorsque la magnitude sert à pondérer l'orientation (courbe rouge). La différence entre les deux distributions (courbe grise) n'est pas significative dans le voisinage de la direction principale (bin 19).

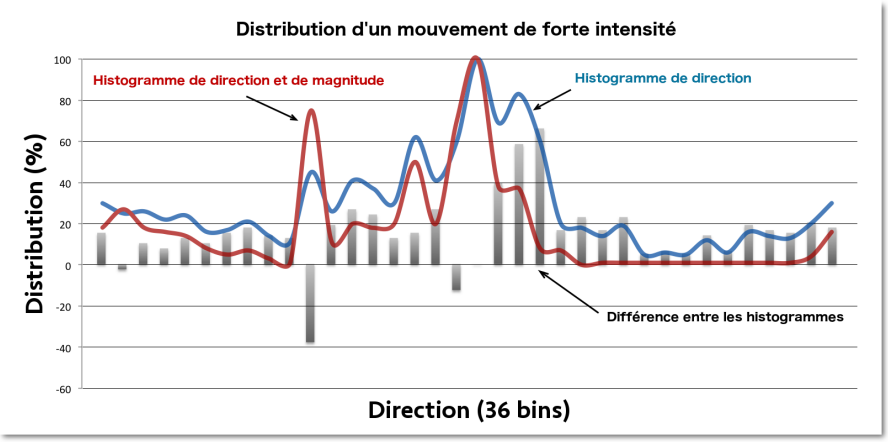


FIGURE 3.12 – *Distribution des orientations pour un macro mouvement et l'apport de la magnitude.*

Toutefois, en présence de faibles mouvements, la magnitude et l'orientation analysées conjointement permettent d'identifier de manière plus précise la direction principale (comme illustré dans la Figure 3.13). En effet, l'histogramme des orientations est difficilement exploitable, car aucune direction prédominante ne se dégage clairement. Dans ce cas, la combinaison de l'orientation et de la magnitude permet de mieux cibler la direction principale recentrant la détection vers le bin 16.

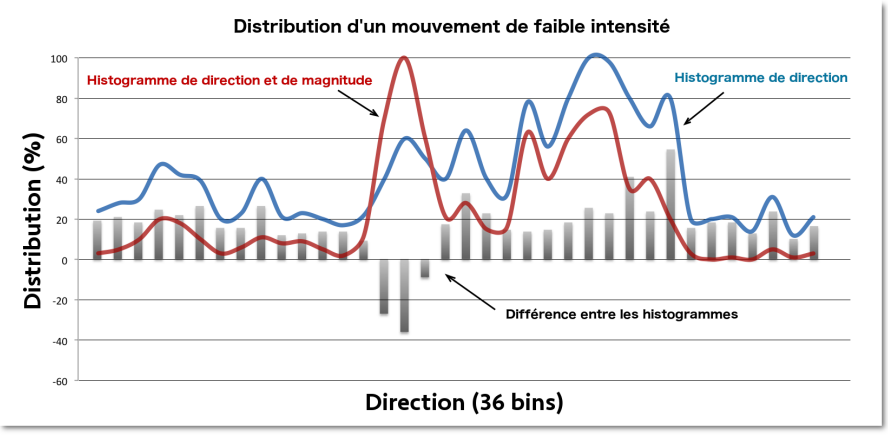


FIGURE 3.13 – *Distribution des orientations pour un micro mouvement et l'apport de la magnitude.*

Afin de mesurer la consistance en termes d'intensité et d'orientation au sein d'un LMP, nous analysons la distribution des orientations dans le CMR en considérant différentes couches de magnitude. Le mouvement facial se propage progressivement grâce à l'élasticité de la peau. Ainsi, nous supposons que pour des intervalles de magnitude voisins, nous devrons observer une superposition des directions principales. En considérant la caractérisation du mouvement selon plusieurs intervalles de magnitude, les directions principales tendent à se démarquer des mouvements parasites pour l'ensemble des intervalles de magnitude (voir Figure 3.14 A–B), alors qu'aucune démarcation claire apparaît en absence de véritables mouvements (voir Figure 3.14 C–D). En partant de cette observation et en analysant le mouvement par intervalle de magnitude, nous proposons une méthode qui permet d'extraire le véritable mouvement facial et de supprimer le bruit.

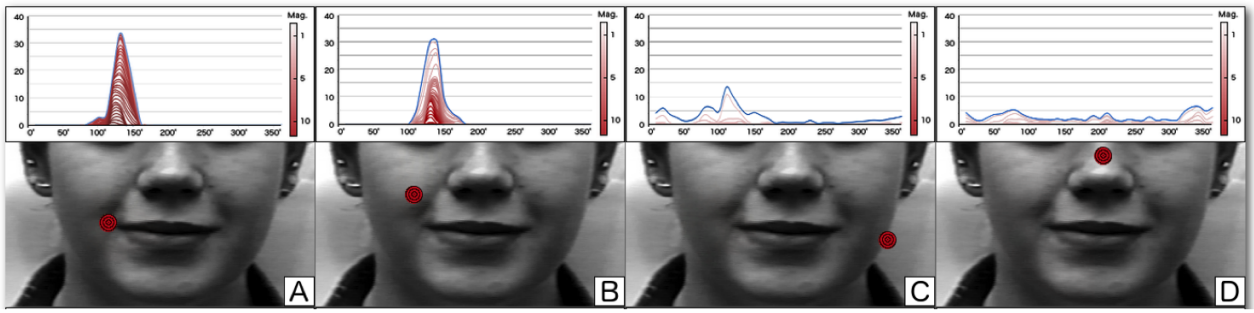


FIGURE 3.14 – *Distribution des orientations en présence d'une expression de joie, en considérant différents intervalles de magnitude et différents régions du visage.*

Le CMR correspond au LMR situé au centre du LMP situé aux coordonnées (x,y). La caractérisation du LMR résulte dans un histogramme des orientations $H_{LMR_{x,y}}$ composé de B classes. La distribution des orientations du LMR est analysée plus en détail en considérant q histogrammes correspondant à différents intervalles de magnitude. Les intervalles de magnitude sont définis en fonction de la nature des mouvements analysées. En présence de mouvements ayant une large palette d'intensités, il faut considérer des intervalles plus nombreux que lorsque l'on considère uniquement des mouvements de forte intensité. En effet, comme illustré dans les Figures 3.12 et 3.13 les mouvements de forte intensité sont plus facilement identifiables que les mouvements ayant des plus faibles intensités. Toutefois, même si les mouvements sont de faibles intensités, on peut observer une cohérence en termes de magnitudes pour certaines orientations. La distribution pour chaque intervalle de magnitude de LMR est obtenue ainsi :

$$MH_{LMR_{x,y}}(n,m) = \{(cls_i, mag_i) \in H_{LMR_{x,y}} \mid mag_i \in [n,m]\}. \tag{3.3}$$

Pour chaque orientation considérée, nous analysons la manière dont cette orientation est représentée dans les distributions construites pour chaque niveau de magnitude $MH_{LMR_{x,y}}(n,m)$. Les co-occurences des orientations à différents niveaux d'intensité servent à faciliter la distinction entre les directions principales même en présence de mouvements de faible intensité (qui ne se retrouvent pas dans l'intégralité des niveaux de magnitudes considérés). Ainsi nous construissons un histogramme $ML_{LMR_{x,y}}(m1,m2)$ qui caractérise la distribution des co-occurences des orientations pour les niveaux d'intensités compris entre $m1$ et $m2$.

$$\begin{aligned} ML_{LMR_{x,y}}(m1,m2) = \\ \{(cls_i, E\{(n,m) \mid \exists (cls_i, mag_i) \in \{MH_{LMR_{x,y}}(n,m) \mid m1 <= mag_i <= m2\}\}. \end{aligned} \tag{3.4}$$

À partir de ces histogrammes des co-occurences nous construisons un histogramme des directions magnifié $DMH_{LMR_{x,y}}$ en appliquant différents coefficients ω_1, ω_2 et ω_3 à chaque histogramme $ML_{LMR_{x,y}}(m1, m2)$ calculé précédemment.

$$\begin{aligned} DMH_{LMR_{x,y}} = ML_{LMR_{x,y}}(m1, m2) * \omega_1 + ML_{LMR_{x,y}}(m2, m3) * \omega_2 \\ + ML_{LMR_{x,y}}(m3, m4) * \omega_3. \end{aligned} \tag{3.5}$$

La distribution $DMH_{LMR_{x,y}}$ estime la cohérence du mouvement et identifie les directions principales. Afin de distinguer plus facilement entre les différentes directions principales, nous avons décidé d'appliquer un facteur d'échelle de 10 entre les différents poids associés aux parties constituant la distribution magnifiée ($\omega 1 = 1$, $\omega 2 = 10$ et $\omega 3 = 100$). Nous estimons que plus il y a une cohérence en matière d'orientation à travers différents intervalles de magnitude, plus le mouvement s'apparente à un mouvement facial réel.

L'intégralité du processus de caractérisation et de filtrage au niveau d'un LMR est illustré dans la Figure 3.15. Nous y montrons les mouvements perçus lors d'une macro- ou d'une micro-expression. La Figure 3.15-A illustre le calcul d'histogrammes par intervalles de magnitude considérés $MH_{LMR_{x,y}}(n, m)$. Nous choisissons de faire varier le paramètre n entre 0 et 10 avec un pas de 0,2. Le paramètre m est fixé à 10, afin de renforcer la prise en compte d'orientations communes à différentes magnitudes. En présence du mouvement caractérisant une macro-expression, l'empilement des distributions caractérisant les différents intervalles de magnitude permettent d'identifier clairement la direction principale. Ensuite, la segmentation des distributions en trois niveaux selon la co-occurence des orientations est illustrée dans la Figure 3.15-B. Chaque $ML_{LMR_{x,y}}$ correspond à une ligne, dont la valeur dans chaque colonne correspond au nombre de co-occurences en termes d'orientations identifiées pour l'ensemble des magnitudes et pour chaque classe d'orientations. Finalement, l'histogramme des orientations magnifié $DMH_{LMR_{x,y}}$ obtenu est visible dans la Figure 3.15-C.

Avant d'analyser le voisinage et rechercher des cohérences dans les LMR voisins, il faut qu'au moins une direction principale soit identifiée au sein du LMR courant. Pour cela, nous appliquons un seuil α aux valeurs de la distribution de l'histogramme des orientations magnifié $(DMH_{LMR_{x,y}})$. La valeur α précise l'importance de la co-occurence des orientations au sein de différents intervalles de magnitude.

Si une ou plusieurs directions principales ont été identifiées à travers les différents niveaux de magnitude, nous étudions la distribution locale des directions principales. Chaque direction principale doit couvrir un nombre limité de classes d'orientations (une à s classes). En effet, un mouvement facial cohérent s'étale rarement au-delà de 60°. Si l'une des directions principales s'étale au-delà, nous considérons qu'il s'agit d'un mouvement bruité qui risque de nuire à la qualité de la propagation du mouvement dans le voisinage. Ainsi, ces directions sont écartées de la distribution finale caractérisant le LMR. Les équations suivantes régissent cette dernière opération de filtrage concernant le LMR. La première équation détermine l'étendue des directions principales en associant à une orientation l'ensemble des classes adjacentes dont le cumul est similaire. La seconde équation impose le critère de filtrage pour les orientations s'étalant sur plus de s classes consécutives.

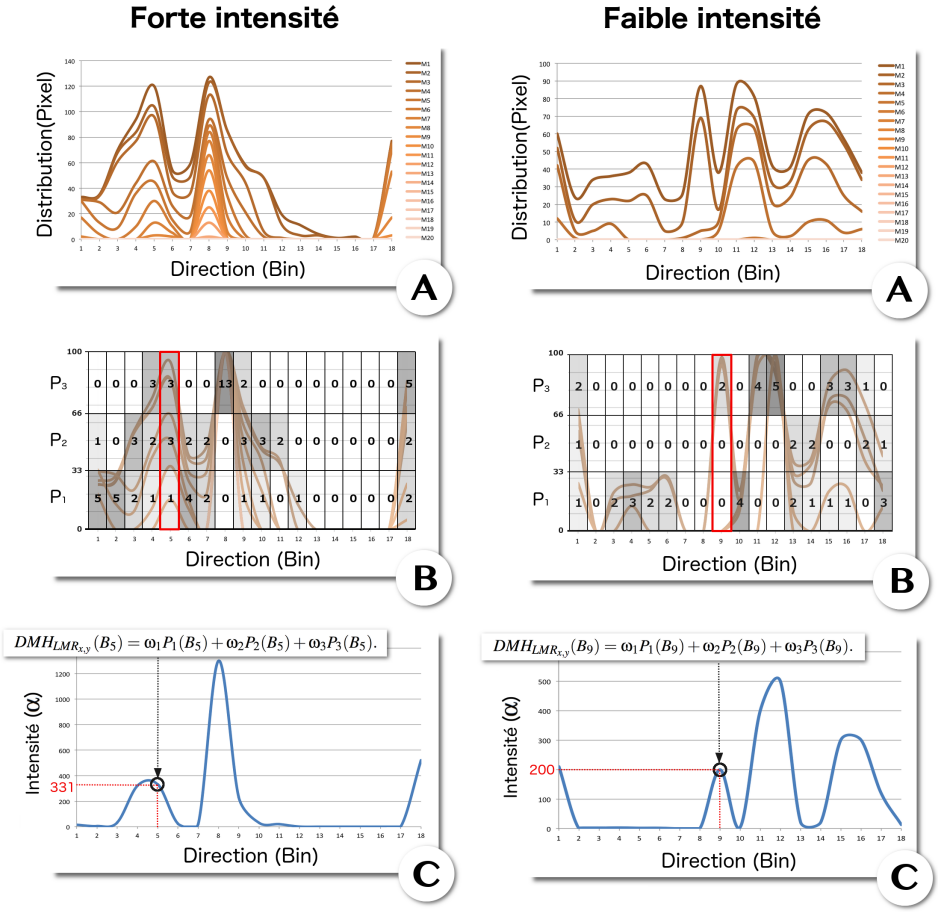


FIGURE 3.15 – *Le processus de caractérisation du mouvement cohérent localement au sein d'un LMR. (A) Histogrammes d'orientation pour différents intervalles de magnitude, (B) Histogramme de cumul d'orientations superposées pour différents magnitudes, (C) Identification des orientations cohérentes localement.*

$$C(DMH_{LMR_{x,y}}) = \{E = [a..b] \mid \forall_i \in [a..b] \mid DMH_{LMR_{x,y}}(i) > \alpha \\ \wedge \nexists j \in \{a-1, b+1\} \mid DMH_{LMR_{x,y}}(j) > \alpha\}. \tag{3.6}$$

$$C^{'}(DMH_{LMR_{x,y}}, s) = \{E \in C(DMH_{LMR_{x,y}}) \mid \|E\| < s\}. \tag{3.7}$$

où $[a..b]$ représentent l'intervalle des classes successives validant le seuil α de sélection en tant que direction principale. Ensuite, chaque direction principale est analysée afin de ne garder que celles qui s'étendent sur moins de s orientations successives.

Comme les directions principales résultent d'un mouvement cohérent, il est attendu que dans le voisinage de la direction principale un changement graduel des cumuls soit observé. Ainsi, nous imposons qu'entre deux classes d'orientations successives un différentiel de plus de Φ soit observé comme indiqué dans l'Équation 3.8. Suite à ces filtrages nous obtenons un histogramme des orientations magnifié et filtré $FDMH_{LMR_{x,y}}$ (cf. Equation 3.8) qui contient les k directions principales de $DMH_{LMR_{x,y}}$ et qui sont cohérentes d'un point de vue de leur étendue.

$$C^{''}(DMH_{LMR_{x,y}}) = \{E = [a..b] \in C^{'}(DMH_{LMR_{x,y}}, s) \mid \\ \forall_{i,j} \in E, \|i-j\| \leq 1 \mid \|DMH_{LMR_{x,y}}(i) - DMH_{LMR_{x,y}}(j)\| < \Phi\}. \tag{3.8}$$

$$FDMH_{LMR_{x,y}} = \{(b_i, m_i) \in DMH_{LMR_{x,y}} \mid \exists E = [a..b] \in C^{''}(DMH_{LMR_{x,y}}) \wedge b_i \in E\}. \quad (3.9)$$

Si aucune direction principale n'est conservée suite à l'ensemble de la chaîne de filtrage, le LMR est déclaré incohérent. Si le CMR associé à un LMP est incohérent, le LMP est déclaré incohérent. Cela signifie qu'aucun mouvement facial véritable n'est perçu dans la région. En revanche, si le CMR est considéré comme étant cohérent, la validation et la construction du LMP se poursuit en analysant les régions voisines (Neighboorhood Motion Regions - NMR). En effet, en présence de mouvements faciaux, nous assistons à une propagation du mouvement dans les régions voisines selon des directions et intensités proches.

La propagation du mouvement

Lorsqu'un mouvement est cohérent localement, nous essayons de vérifier que la propagation de ce mouvement s'effectue de manière cohérente vers les régions voisines. L'élasticité de la peau garantit que tout mouvement facial s'étend dans les régions voisines en perdant progressivement en intensité jusqu'à extinction complète. Toutefois, entre régions voisines, certaines caractéristiques du mouvement restent relativement similaires, même si certaines modifications peuvent apparaître (changement léger de l'orientation, diminution de l'intensité). À partir de ces observations, nous supposons qu'en cas de mouvement cohérent, la propagation du mouvement d'une région centrale permet de retrouver une caractérisation similaire en matière de direction principale dans au moins une région voisine.

Afin de mesurer la cohérence de la propagation, nous analysons, comme indiqué dans la section précédente, la cohérence locale du mouvement. Cette analyse résulte dans l'extraction d'un histogramme des directions magnifié et filtré $FDMH^{'}_{LMR_{x,y}}$ qui caractérise les directions principales dans cette nouvelle région. La cohérence en termes de propagation entre la région centrale et la région voisine se mesure en estimant le recouvrement entre les histogrammes caractérisant les deux régions : $FDMH_{LMR_{x,y}}$ et $FDMH^{'}_{LMR_{x_v,y_v}}$. Le coefficient de Bhattacharyya (1946) est utilisé afin d'évaluer ce recouvrement. Ce coefficient caractérise la similarité entre deux distributions et il est défini ainsi :

$$C^{'''}(FDMH_{LMR_{x,y}}, FDMH^{'}_{LMR_{x,y}}) = \sum_{i=1}^{B} \sqrt{FDMH_{LMR_{x,y}}(i) FDMH^{'}_{LMR_{x,y}}(i)}. \quad (3.10)$$

où, $FDMH_{LMR_{x,y}}$ et $FDMH^{'}_{LMR_{x,y}}$ sont les distributions locales du mouvement et B le nombre de classes. Deux régions sont considérées comme présentant une propagation cohérente du mouvement si le coefficient est supérieur au seuil ρ.

Le processus de propagation du mouvement dans les régions voisines en vue d'extraire le LMP après une itération est illustré dans la Figure 3.16. Les régions voisines (NMRs) sont grisées si le mouvement y est incohérent. Trois cas de figure peuvent rendre compte d'une incohérence en ce qui concerne le mouvement au sein d'une NMR :

— Le mouvement au sein du NMR est localement incohérent en termes de directions principales ;
— Les directions principales identifiées au sein d'un NMR couvrent de larges intervalles d'orientations.

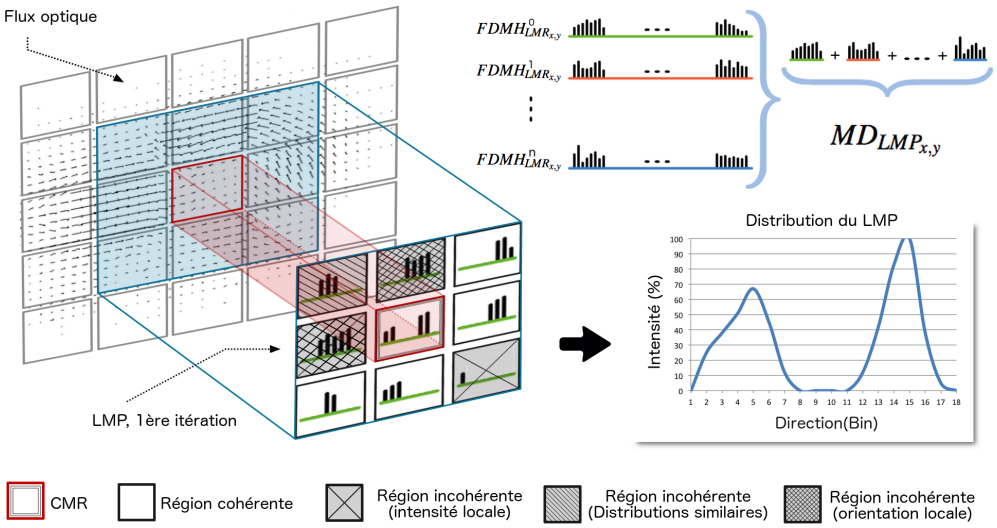


FIGURE 3.16 – *Illustration du processus de propagation dans les régions voisines dans le cadre de l'extraction de motifs locaux de mouvement (LMP).*

— Les distributions caractérisant les régions voisines ne sont pas similaires.

Si les trois critères énoncés précédemment sont validés, la région est marquée comme étant cohérente et elle sert comme nouveau point de départ pour les futures propagations. Ainsi, de manière récursive nous explorons l'étendue de la propagation du mouvement cohérent en considérant un maximum de β propagations. À cause de discontinuités du mouvement induites sur le visage en présence de rides, une région initialement considérée incohérente, peut changer d'état lors de la propagation du mouvement à partir d'une autre région voisine.

L'ensemble des régions cohérentes contribue à la construction du descripteur LMP. En effet, chaque distribution ($FDMH_{LMR_{x,y}}$) correspondant aux NMRs ayant des connexions directes ou indirectes avec le CMR est prise en compte dans le calcul du LMP en question. La distribution du mouvement $MD_{LMP_{x,y}}$ est caractérisée par un histogramme ayant B classes d'orientations. Chaque classe de cet histogramme cumule les intensités de chaque NMR et du CMR et est calculée de la manière suivante :

$$MD_{LMP_{x,y}} = \{\sum_{i=0}^{n} FDMH_{LMR_{x,y}} \mid FDMH_{LMR_{x,y}} \in LMP_{x,y}\}. \tag{3.11}$$

où n est le nombre de régions directement ou indirectement cohérentes avec le CMR (y inclus).

Pour résumer, le LMP reflète les directions principales du mouvement en renforçant la cohérence de ces orientations pour différentes intensités, tout en filtrant les mouvements incohérents. Le filtrage s'appuie sur trois hypothèses caractérisant le mouvement facial : la convergence de différents niveaux d'intensité dans la même direction, la cohérence locale des directions principales et la manière dont le mouvement se propage. Chaque critère est rendu opérationnel par une analyse statistique des distributions d'orientation à plusieurs niveaux d'intensité. Afin de prouver l'efficacité du filtrage et de l'identification de véritables mouvements, dans la section suivante, nous appliquons le LMP à l'analyse de micro- et macro-expressions faciales.

3.3.2 Reconnaissance des expressions en présence de variations d'intensité

La manière de segmenter et caractériser le visage joue en rôle très important dans le processus de reconnaissance. La disposition des épicentres servant à extraire les motifs de mouvement au sein du visage doit être choisie avec soin. Ainsi, nous étudions l'impact de la position des épicentres des LMPs afin de maximiser les informations extraites caractérisant au mieux les expressions.

Dans la suite nous nous intéressons aux spécificités des micro- et des macro-expressions et nous montrons que le choix des épicentres a un fort impact sur le mouvement extrait, notamment dans le cas des micro-expressions. Ensuite, à la lumière de ces observations, nous proposons un modèle de segmentation du visage permettant d'optimiser la caractérisation des mouvements indépendamment de leur intensité. Des motifs locaux de mouvement sont extraits autour des épicentres disposés selon la segmentation choisie. Les motifs de mouvement sont agrégés afin de caractériser le mouvement global tout le long d'une séquence vidéo. Ce descripteur global du mouvement est combiné avec un descripteur caractérisant la géométrie du visage afin de reconnaître efficacement les micro- et les macro-expressions.

Influence de la disposition des LMPs sur le visage

Afin d'illustrer l'influence de la position des LMPs sur un visage en présence de faibles et de forts mouvements, nous considérons deux visages exprimant une micro- et une macro-expression. Nous y disposons les LMPs comme illustré dans la Figure 3.17. La première ligne présente une macro-expression de *joie* et la deuxième ligne correspond à une micro-expression de *joie*. Dans les trois premières colonnes, nous considérons un épicentre différent (les points bleu, rouge et vert) pour caractériser simultanément à l'aide de trois LMP le mouvement facial à droite de la bouche. Avant caractérisation et filtrage, le mouvement est obtenu en calculant le flux optique dense sur l'intégralité du visage. Pour chaque visage et chaque LMP, nous présentons, en haut de la vignette correspondante, la distribution du mouvement extraite par le LMP et, en bas de la vignette, l'étendue du mouvement cohérent identifié depuis l'épicentre choisi. La colonne 4 présente la superposition des distributions obtenues à partir des trois positions précédentes.

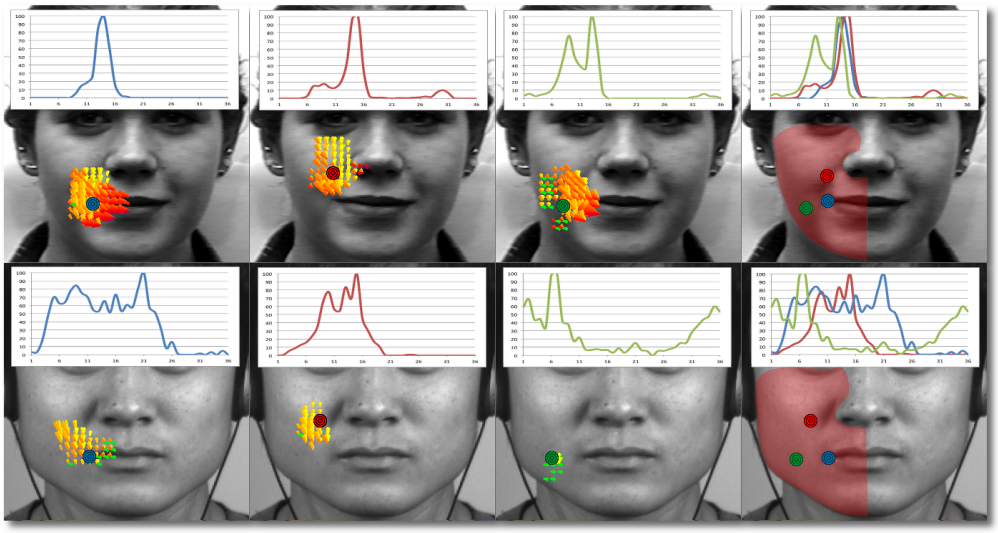


FIGURE 3.17 – *Le mouvement cohérent extrait à différents emplacements dans une même région lors d'une expression de joie.*

Dans le cas des macro-expressions, le mouvement est intense et sa propagation couvre de larges régions du visage. Si l'épicentre d'un LMP est placé dans une de ces régions, l'intégralité du mouvement et sa zone de propagation sont fidèlement identifiés. En effet, pour la macro-expression (première ligne), malgré le fait que les épicentres des LMPs sont différents, les distributions extraites présentent de larges recouvrements (colonne 4). Lorsque l'on traite des micro-expressions, le mouvement est de très faible intensité, et il se produit sur un très court intervalle de temps. La propagation se limite au voisinage immédiat du véritable épicentre. Les distributions qui correspondent aux trois épicentres sont très différentes. Cette expérimentation peut être reproduite pour mesurer l'influence de la disposition des LMP dans d'autres régions du visage et pour d'autres classes de micro- et macro-expressions.

Ainsi, afin de pouvoir traiter de manière unifiée les micro- et les macro-expressions, il est important de choisir avec soin les épicentres. Le modèle de segmentation du visage résultant doit extraire et caractériser efficacement le mouvement cohérent indépendamment de la classe et de l'intensité de l'expression. La section suivante est dédiée à l'étude de la disposition spatiale des épicentres qui vise à augmenter la fidélité de la caractérisation du mouvement facial cohérent et par conséquent, reconnaître de manière optimale les macro- et les micro-expressions.

Modèle unifié de segmentation pour la caractérisation de micro- et macro-expressions

Comme évoqué déjà à plusieurs reprises, les mouvements générés par les macro- et micro-expressions sont très différents en termes d'intensité et de propagation. Dans cette section nous cherchons à identifier un modèle facial optimal pour l'extraction des motifs locaux de mouvement. Nous étudions en parallèle six macro-expressions discrètes (*dgot*, *nervement*, *joie*, *peur*, *surprise* et *tristesse*) et cinq micro-expressions discrètes (*dgot*, *joie*, *rpression*, *surprise* et la classe *autres*).

Comme illustré dans le contexte de l'analyse statique d'une expression, il est important d'identifier les régions du visage portant l'information la plus discriminante. Dans le cas présent, il est intéressant d'identifier les régions où l'activation est la plus intense lors de la production d'une expression. Afin d'identifier ces régions pour les expressions retenues, nous construisons une carte de mouvement par expression et par intensité. Nous alignons l'ensemble des visages par rapport à la position des yeux, et nous calculons le flux optique sur l'ensemble des séquences correspondant à l'expression et à l'intensité considérées. Cette étape d'alignement du visage permet à la fois d'éliminer les effets de rotations dans le plan et de normaliser la taille des visages analysés. Ensuite chaque trame de la séquence est segmentée dans plusieurs blocs de taille 20×30 pixels. Les épicentres des LMPs sont placés au centre de chaque bloc. Chaque mouvement local cohérent filtré et extrait par les LMPs est intégré dans une carte de mouvements. Ces cartes obtenues pour chaque séquence sont ensuite regroupées par expressions et par intensité et agrégées afin de définir les motifs globaux de mouvement observés lors de l'apparition d'une micro- ou d'une macro-expression.

Motifs locaux de mouvement pour les macro-expressions Six cartes de motifs globaux de mouvement ont été construites pour les macro-expressions. Elles apparaissent sur la première ligne de la Figure 3.18. Les cartes ont été construites en analysant les séquences disponibles dans le corpus CK+ (Lucey et al. 2010). Les motifs de mouvement extraits indiquent que les mouvements pertinents pour l'analyse de macro-expressions se trouvent en dessous des yeux, sur le front, au-

tour du nez et de la bouche. Certaines régions apparaissent dans les motifs associés à différentes expressions. Néanmoins, le mouvement contenu est spécifique en termes d'intensité, orientation et densité. Par exemple, les expressions de *tristesse* et d'*nervement* enclenchent des mouvements spécifiques dans les régions autour de la bouche et des sourcils. Toutefois, lorsque la personne est énervée, le mouvement facial converge vers l'intérieur du visage. Les lèvres génèrent un mouvement vers le haut, et les sourcils un mouvement vers le bas. En cas de *tristesse*, le mouvement facial dans la partie supérieure du visage diverge du mouvement situé dans la partie inférieure du visage.

Motifs de mouvements pour les micro-expressions En appliquant la même stratégie que précédemment, il est aisé de construire des cartes de mouvements permettant de caractériser les motifs principaux de mouvement pour les micro-expressions. Nous utilisons les micros-expressions de *joie*, *dgot*, *surprise*, *rpression* et *autres* contenues dans le corpus CASME2 (Yan et al. 2014a). Comme illustré sur la seconde ligne de la Figure 3.18, les mouvements pertinents pour l'analyse des micro-expressions se situent dans le voisinage des sourcils et des coins de la bouche. Lorsque nous comparons les motifs obtenus pour les micro- et les macro-expressions, on observe que la propagation du mouvement est très différente. Par ailleurs, il est important de noter que le motif correspondant à la classe *autres* est très similaire aux autres classes ce qui rend le processus de reconnaissance plus difficile dans le cadre du corpus CASME2.

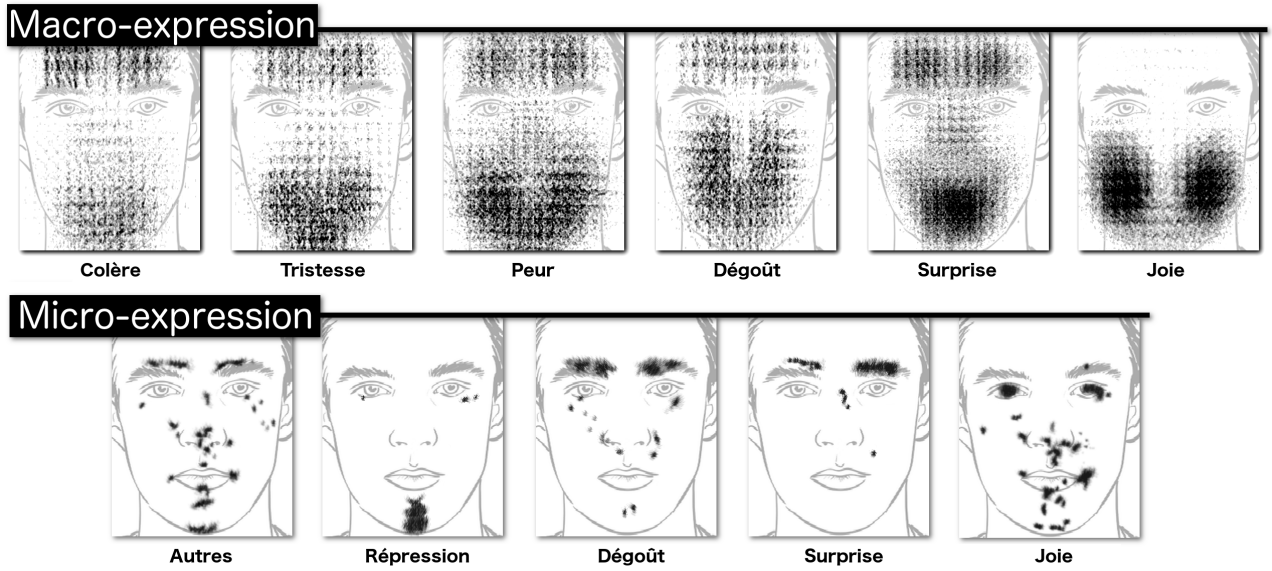


FIGURE 3.18 – *Motifs de mouvement rencontrés dans les six macro-expressions du corpus CK+ (ligne 1) et les micro-expressions du corpus CASME2 (ligne 2).*

L'expérience précédente nous a permis d'identifier les régions du visage participant de manière importante à la construction d'une expression faciale. De manière similaire à Jiang et al. (2014), nous nous appuyons sur les points caractéristiques du visage pour définir une segmentation qui suit au mieux la disposition des muscles faciaux sous-jacents et les déformations induites par leur activation. Les points caractéristiques et la géométrie du visage sont utilisés pour définir une segmentation qui couvre intégralement les régions mises en avant par l'analyse des motifs de mouvement : front, sourcils, yeux, nez, joues, lèvres.

La segmentation en régions d'intérêt pour la caractérisation du mouvement est illustrée dans la Figure 3.19. La Figure 3.19 présente à gauche les régions retenues pour la caractérisation de mouvements et à droite les points d'ancrage des régions. Au-delà des points caractéristiques, de nouveaux points d'ancrage dans les régions monotones du visage (par exemple, les joues, le front) sont calculés afin d'étendre de manière adéquate le modèle de segmentation et de recouvrir les parties porteuses d'information selon les cartes de la Figure 3.18. Par exemple, le point d'ancrage Q se situe au milieu du segment reliant les points 10 et 55 (en bleu) du modèle de Kazemi et Sullivan (2014). La distance entre les sourcils et les points d'ancrage recouvrant le front (A, B, ..., F) correspond à un quart de la distance entre les points 27 et 33 (en bleu). Ceci maintient en adéquation la segmentation avec les variations de taille au sein d'une séquence ou entre différentes personnes. Nous avons superposé partiellement les régions 19 et 22 avec les régions 18 et 23, respectivement, afin de caractériser précisément les mouvements subtiles relatifs aux coins des lèvres.

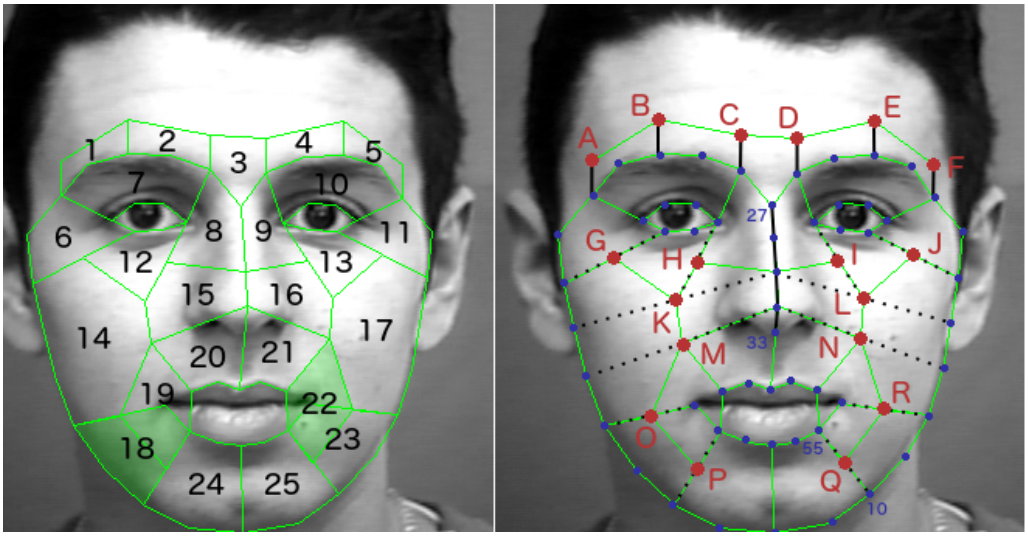


FIGURE 3.19 – *Modèle de segmentation faciale unifié (à gauche - régions, à droite - points d'ancrage).*

Dans la section suivante, nous discutons de la manière dont les LMPs disposés selon le modèle facial ci-dessus sont utilisés afin de procéder à la reconnaissance de micro- et macro-expressions.

Caractérisation globale du mouvement facial

Le mouvement facial est collecté dans les 25 régions d'intérêt découlant de la segmentation présentée précédemment. Dans chaque trame f_t, nous extrayons la distribution en termes de directions principales dans l'ensemble des régions R_t^k, où t est l'indice de la trame et $k = 1, 2, ..., 25$ est l'indice de la région concernée. Au sein de chaque R_t^k, la distribution du mouvement $MD_{LMP_{x,y}}$ est calculée. Au fur et à mesure que la séquence se déroule, les distributions de mouvements R_t^k sont sommées par région (η^k). En fin de séquence, dans chaque région, nous obtenons donc un cumul de directions principales dont le poids correspond à la fréquence d'apparition de la direction principale à travers les différentes trames. Cette somme permet de donner plus d'importance aux orientations qui accompagnent naturellement le mouvement et pénalise les bruits. Par exemple, lorsqu'on sourit, on peut observer un mouvement d'étirement au niveau des coins de la bouche sur plusieurs trames successives. L'orientation de l'étirement aura plus de poids que les éventuels

bruits qui apparaîtront de manière ponctuelle dans certaines trames.

$$\eta^k = \sum_{t=1}^{time} R_t^k. \tag{3.12}$$

Les distributions agrégées dans chaque région η^k sont concaténées dans un vecteur global nommé $GMD = (\eta^1, \eta^2, ..., \eta^{25})$. Le GMD caractérise l'évolution globale du mouvement pendant la séquence caractérisée. Le GMD est employé dans le processus de reconnaissance de macro- et micro-expressions. La taille du vecteur GMD est égale au nombre de régions multiplié par le nombre de classes d'orientations. Une illustration du processus de construction du GMD est présente dans la Figure 3.20. Dans cet exemple, les distributions de mouvements $MD_{LMP_{x,y}}$ sont illustrées par les régions R_t^1, R_t^2 et R_i^{25} à plusieurs moments de la séquence. Pour chaque région nous agrégeons les distributions sur l'intégralité de la séquence et nous obtenons les nouvelles distributions η^1, η^2 et η^{25}. Ces distributions sont juxtaposées pour caractériser de manière globale les changements intervenant sur le visage au sein du GMD.

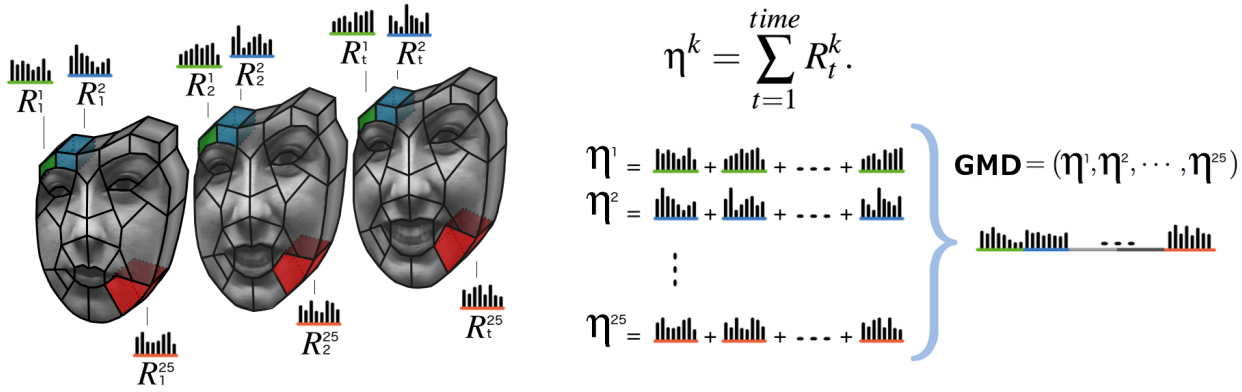


FIGURE 3.20 – *Construction de GMD : descripteur global retraçant l'évolution des mouvements cohérents sur le visage.*

Le processus de reconnaissance des expressions faciales

La Figure 3.21 met en évidence l'enchaînement des étapes du processus de reconnaissances : le prétraitement, l'extraction de descripteurs et la classification.

Prétraitement Le prétraitement inclut la détection des points caractéristiques du visage, l'alignement de visages par rapport à la position des yeux et la définition des régions d'intérêt sur le visage aligné. Malgré le fait que nous supposons que les visages sont presque de face, des petites rotations dans le plan sont présentes. L'alignement permet d'augmenter la cohérence entre la segmentation du visage tout le long de la vidéo. Le flux optique est calculé en utilisant l'algorithme de Farnebäck (2003). Le choix de cet algorithme a été fait par rapport à la vitesse de calcul et, par l'absence de processus de lissage du mouvement, que d'autres algorithmes tels que DeepFlow de Weinzaepfel et al. (2013) ou EpicFlow de Revaud et al. (2015) mettent en place. Un lissage qui ne prend pas en compte les spécificités du mouvement, peut détériorer le mouvement pertinent qui se confond avec le bruit pour les faibles intensités.

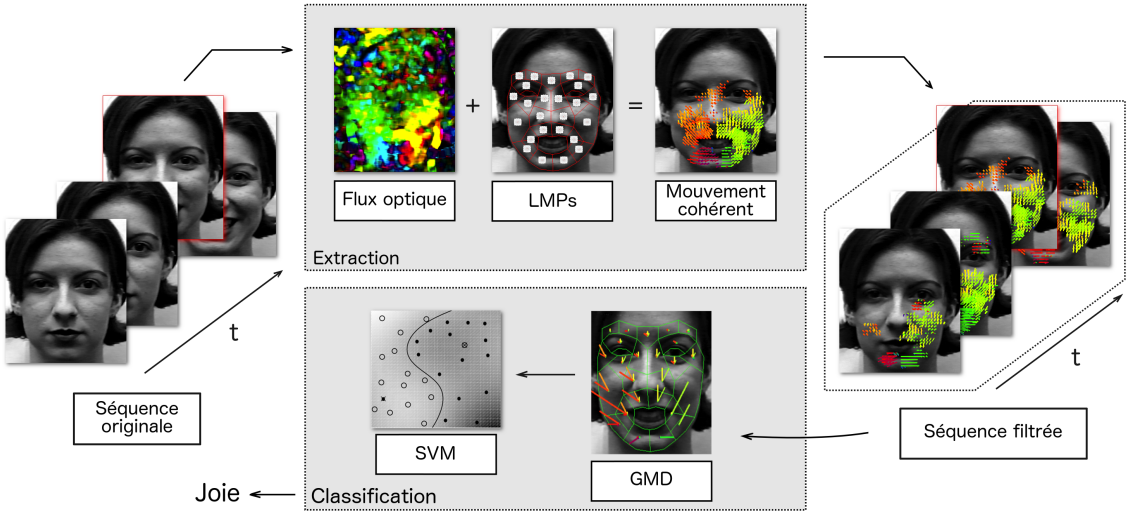


FIGURE 3.21 – *Aperçu du processus unifié de classification de micro- ou de macro-expressions.*

Caractérisation du mouvement Après la segmentation du visage en 25 régions comme illustrée dans la Figure 3.19, des LMPs sont ancrés dans le centre de chaque région faciale. Le mouvement cohérent collecté dans chaque LMP est synthétisé par un ensemble de direction principales. Finalement, nous cumulons les directions principales dans chaque région sur l'ensemble de la séquence. Ainsi, dans chaque région nous obtenons une distribution retraçant les évolutions en termes de directions principales. La concaténation des distributions locales forme le descripteur global qui sert dans l'étape de classification.

Classification L'étape de classification s'appuie, comme pour la plupart des travaux de l'état de l'art, sur un SVM avec noyau RBF. La séquence entière, représentant le passage de l'enclenchement de l'expressions (*onset*) jusqu'à la plus forte intensité de l'expression (*apex*), est caractérisée par le descripteur *GMD* qui sert dans le processus de classification.

Dans la section suivante, nous présentons en détail les expérimentations conduites dans le cadre de la validation de ce nouveau descripteur de mouvements et du nouveau modèle de segmentation faciale pour la reconnaissance des micro- et des macro-expressions.

3.3.3 Évaluation

Dans cette section, nous mesurons les performances de notre approche sur trois corpus de macro-expressions et deux corpus de micro-expressions :

- (Macro) Cohn-Kanade (CK+) de Lucey et al. (2010) - variabilité en termes de séquence d'activation ;
- (Macro) Oulu CASIA de Zhao et al. (2011) - diverses conditions d'illumination ;
- (Macro) MMI de Pantic et al. (2005) - mouvements de la tête ;
- (Micro) CASME2 de Yan et al. (2014a) - variabilité en termes de séquence d'activation ;
- (Micro) SMIC de Li et al. (2013b) - diverses conditions d'illumination et vitesses d'enregistrement.

Dans la suite de cette section, les résultats présentés sont obtenus en suivant le processus décrit en 3.3.2. En suivant les tendances dans la litérature, pour les corpus relevant de macro-expressions

nous appliquons une validation croisée à 10 échantillons. Nous adoptons une validation croisée de type *Leave-One-Subject-Out* (LOSO) pour les corpus relevant de micro-expressions.

Avant de présenter les résultats obtenus sur l'intégralité des corpus retenus, nous illustrons l'impact des différents paramètres régissant l'extraction des LMPs sur les performances de classification en considérant les corpus CK+ et CASME2.

Etude de l'impact de paramètres

Nous repassons en revue les paramètres régissant le filtrage et la caractérisation des directions principales lors de l'extraction du LMP. Nous proposons une méthodologies pour retrouver les paramètres optimaux de LMP selon le corpus analysé.

Dans la suite, les paramètres sont regroupés selon les hypothèses du processus de filtrage et caractérisation du LMP :

- cohérence locale de la distribution du mouvement en termes de magnitude et de direction
  - la dimension d'une région locale au sein du LMP (λ)
  - la granularité de l'histogramme d'orientation (B)
- cohérence locale des directions principales
  - le seuil de co-occurences accepté pour caractériser une direction principale (E)
  - le seuil de l'étendue d'une direction principale accepté (M)
  - le seuil de la variation des cumuls entre deux orientations successives acceptées pour caractériser une direction principale (V)
- la propagation du mouvement en termes de magnitude et d'orientation au sein du visage
  - la distance entre les épicentres de deux régions voisines (Δ)
  - le seuil de similarité entre les distributions des mouvements de deux regions connexes (ρ)
  - le nombre d'itération de propagations pour l'étude du mouvement (β)

i - Cohérence locale du mouvement en termes de magnitude et d'orientation

i.1 - La dimension d'une région locale au sein de LMP (λ) La taille d'une région épicentre (CMR) d'un LMP doit prendre en compte les variations dans la taille des visages (proximité de la caméra, différentes morphologies). Un facteur de mise à l'échelle est employé afin de maintenir une cohérence dans la taille des régions et garantir ainsi une propagation locale cohérente. Comme illustré dans la Figure 3.22-a, la valeur idéale pour λ est de 3% de la taille du visage, pour les micro- et les macro-expressions. En considérant un facteur de 3%, les régions composant le LMP sont suffisamment petites pour permettre une validation du critère de propagation locale du mouvement. Une plus large région serait caractérisée par une distribution trop éparse. Ceci empêcherait de distinguer une direction principale et de renforcer sa caractérisation par l'étude de propagations successives. Ceci se traduit par une réduction de la cohérence du mouvement extrait du visage.

i.2 - La granularité de l'histogramme des orientations (B) - Un autre paramètre qui doit être pris en compte pour étudier la cohérence locale du mouvement est le nombre de classes d'orientations considérées B. Lorsque λ est petit, il est également intéressant de considérer un nombre réduit de classes, car cela permet d'identifier de manière plus efficace la direction principale. Comme

illustré dans la Figure 3.22-b, la meilleure performance est obtenue lorsque nous considérons des histogrammes ayant 9 ou 12 classes d'orientation. Un grand nombre de classes d'orientation tend à réduire les performances pour deux raisons : cela augmente la taille du descripteur, tout en rendant plus difficile l'identification des directions principales car leur cumul se voit distribuer entre la classe principale et les classes voisines.

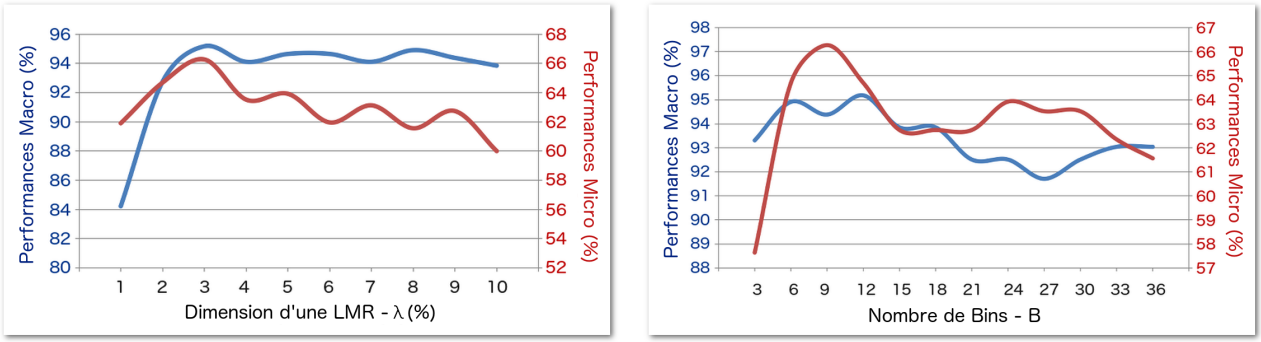


FIGURE 3.22 – *Paramètres impactant la cohérence locale du mouvement en termes de magnitudes et d'orientation : dimension d'une région (Δ) et nombre de classes d'orientation (B).*

ii - Cohérence locale dans la distribution de directions principales

ii.1 - Cumul de co-occurences pour accepter une direction principale (*E*) Concernant le seuil de cumul de co-occurences des orientations *E*, la Figure 3.23-A montre que plus ce seuil est élevé, plus les performances sont limitées. En effet, en étant très sélectifs sur la co-occurence d'une orientation à travers différents niveaux de magnitude, plus l'information restante suite au filtrage est réduite jusqu'à ce que tout le mouvement soit filtré. Ceci est d'autant plus vrai pour les micro-expressions où les distributions de mouvement ne présentent pas de directions franches.

ii.2 - Etendue d'une direction principale (*M*) La Figure 3.23-B, montre que jusqu'à un certain point, la prise en compte d'une contrainte concernant l'étendue d'une direction principale améliore les taux de reconnaissance à la fois pour les micro- et les macro-expressions.

ii.3 - Variation en termes de cumul entre orientations voisines (*V*) En ce qui concerne la variation en termes de cumul entre deux classe d'orientations influe de manière modérée les performances de reconnaissance comme illustré dans la Figure 3.23-C.

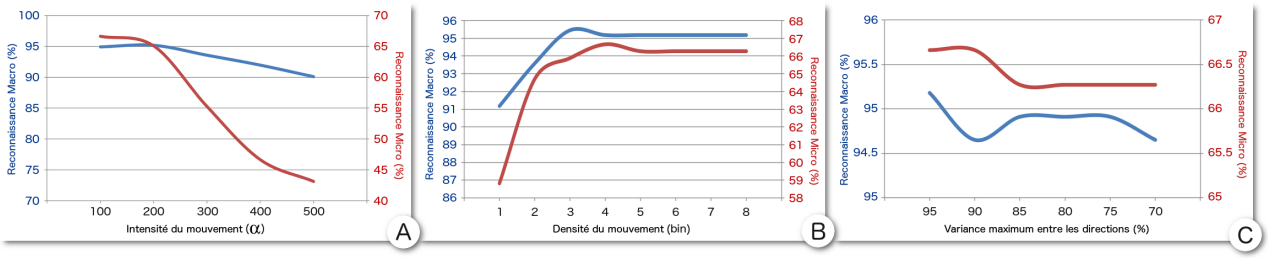


FIGURE 3.23 – *Taux de reconnaissance en présence de différentes intensités de mouvement (A), densités de mouvement (B) et variations de mouvement(C).*

iii - La propagation du mouvement en termes de magnitude et de direction au sein du visage

iii.1 - Nombre de propagations (β) En ce qui concerne la nombre de propagations, la Figure 3.24 montre qu'une augmentation du nombre d'itérations se traduit jusqu'à un certain point, par une augmentation des performances de reconnaissance. En effet, les LMPs construits en considérant un nombre plus importants de régions voisines renforcent d'avantage la cohérence en termes de propagations. Ainsi, le filtrage et la caractérisation sont plus précis et permettent de mieux reconnaître les expressions faciales. Un nombre de propagations trop grand (β) rendra le LMP bruité notamment en présence de macro-expressions. Dans ce cas, la propagation du mouvement risquera de s'effectuer sur l'intégralité du visage. De part l'hétérogénéité des directions principales au niveau du visage tout entier, le LMP deviendra trop générique, perdant son pouvoir discriminant en termes de directions principales.

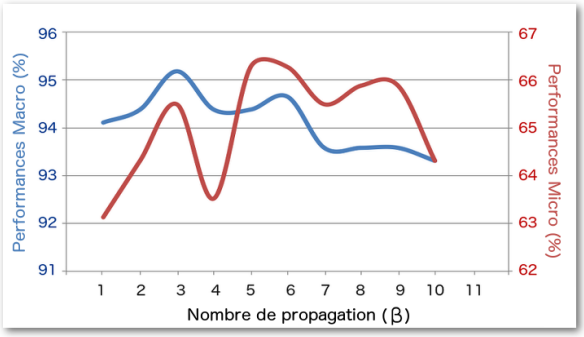


FIGURE 3.24 – *Taux de reconnaissances en fonction de différentes valeurs de nombre d'itération.*

iii.2 - Recouvrement entre régions voisines Le second paramètre intervenant dans le processus de vérification de la propagation est la distance entre les épicentres des régions voisines (Δ). En effet, afin de vérifier la propagation cohérente de mouvement en termes de direction et intensité, il faut s'assurer que les régions voisines se recouvrent partiellement.

iii.3 - Cohérence en termes de distribution entre les régions voisines Le taux de recouvrement spatial influe sur la cohérence en termes de distribution entre les régions voisines. Nous mesurons la cohérence de distribution en calculant le coefficient de Bhattacharyya et en la positionnant par rapport à un seuil de cohérence (ρ).

Il est aisé de comprendre qu'il existe une forte corrélation entre les valeurs de ces deux derniers paramètres Δ et ρ. Plus la valeur de Δ est grande, plus il est attendu que les distributions soient similaires car le recouvrement est plus important, et donc la valeur de seuil pour le coefficient de Battacharya élevé. En effet, comme illustré, dans les matrices de corrélations présentées dans la Figure 3.25, un recouvrement (Δ) de 50% et un seuil (ρ) de 50% donnent les meilleures résultats à la fois pour les micro- et les macro-expressions. Un seuil peu sélectif ne filtrera pas suffisamment les perturbations induites par le bruit.

Les valeurs des paramètres identifiées dans cette section assurent un comportement idéal en termes de filtrage de mouvement pour la reconnaissance des micro- et des macro-expressions telles

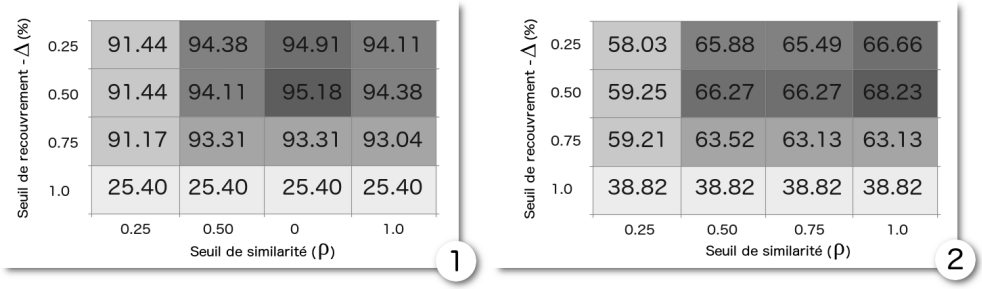


FIGURE 3.25 – *Matrice de corrélation entre la superposition des régions Δ et du coefficient de Bhattacharyya ρ pour les macro- (1) et les micro-expressions (2).*

que présentées dans les corpus CASME2 et CK+. Malgré le fait que ces valeurs sont spécifiques, la méthodologie mise en oeuvre est générique et, le protocole expérimental permettant d'identifier ces valeurs est transposable à d'autres corpus de données.

Dans les sections suivantes, nous présentons les résultats obtenus pour la reconnaissance de micro- et macro-expressions en utilisant des valeurs de paramètres proches de ceux identifiés dans cette section.

Micro-expressions

Pour reconnaître les micro-expressions nous utilisons le modèle de segmentation faciale unifié (voir section 3.3.2). Le flux optique est extrait à partir de deux trames successives sans aucune interpolation ou magnification. Avant d'extraire le flux optique nous procédons à une normalisation géométrique et photométrique du visage comme indiqué dans la section 2.2. Nous n'utilisons que la partie correspondante à l'activation (*onset -> apex*) de l'expression dans les séquences du corpus CASME2 (haute vitesse d'enregistrement) et les différentes versions de SMIC : SMIC-VIS (lumière et vitesse normales), SMIC-HS (haute vitesse), SMIC-NIR (infrarouge). La Table 3.2 montre une comparaison de nos résultats avec quelques-unes des approches majeures de l'état de l'art dans le domaine de la reconnaissance de micro-expressions. Ces résultats ont été initialement publiés dans (Allaert et al. 2017) [4]. Une version plus aboutie de ces expérimentations est disponible dans (Allaert et al. 2018a) [5]

Par rapport aux résultats obtenus dans la Table 3.2, notre approche offre de meilleures performances en comparaison avec des approches de l'état de l'art. En analysant plus attentivement les travaux de l'état de l'art, certains auteurs tels que Wang et al. (2014a), Li et al. (2015), Wang et al. (2014b), Huang et al. (2016b) essayent d'exagérer les mouvements en réduisant le nombre de trames de la séquence. Notre méthode de caractérisation du mouvement est capable d'exploiter l'infime mouvement entre deux trames successives de CASME2. Pour rappel, le corpus CASME2 présente des expressions de très faible intensité du mouvement et il est enregistré avec une caméra ayant une vitesse de capture très rapide (environ 200 trames par seconde).

En plus d'interpoler les séquences pour en réduire la longueur et d'accentuer temporairement

4. B. Allaert ; I.M. Bilasco ; C. Djeraba - Consistent Optical Flow Maps for full and micro facial expression recognition - Proc. of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2017, Porto, Portugal, vol. 5, pp.235-242.

5. B. Allaert ; I.M. Bilasco ; C. Djeraba - Advanced local motion patterns for macro and micro facial expression recognition - Computing Research Repository (CoRR), https://arxiv.org/abs/1805.01951.

TABLE 3.2 – *Comparaison des taux de reconnaissance sur le corpus de micro-expressions avec des approches majeures de l'état de l'art. (^{i} - interpolation, ^{m} - magnification)*

Method	CASME2(%)	SMIC		
		VIS	HS	NIR
LBP-TOP Li et al. (2013b)	48,78%	52,11%	38,03%	
LBP-TOP Yan et al. (2014a)	63,41%	-	-	-
LBP-SIP Wang et al. (2014c)	67,21%	-	-	-
LSDF^i Wang et al. (2014a)	65,44%	-	-	-
TICS^i Wang et al. (2014b)	61,76%	-	-	-
STLBP-IIP Huang et al. (2016a)	62,75%	-	-	-
DiSTLBP-IPP Huang et al. (2016a)	64,78	-	-	-%
HIGO^{i,m} Li et al. (2015)	67,21	68,29%	81,69%	67,61%
CNN Patel et al. (2016)	* 53,60%	* 56,30%	-	
CNN + LSTM Kim et al. (2017)	* 60,98%	-	-	-
CNN + AUs + LSTM Breuer et Kimmel (2017)	* 59,47	-	-	-%
Local Motion Pattern LMP	**70,20%**	**67,68%**	**86,11%**	**80,56%**

les mouvements, certains auteurs dont Li et al. (2015) utilisent un processus de magnification dans le domaine des fréquences pour accentuer l'intensité du mouvement facial. Ce type d'approche fonctionne bien uniquement en présence de mouvements de type micro-expressions et d'absence de mouvements de la tête. Appliquer la magnification en présence de légers mouvements ou de mouvements faciaux intenses, induit de fortes déformations qui nuisent au processus de reconnaissance.

Même si les approches d'apprentissage profond telles que celles de Kim et al. (2017), Breuer et Kimmel (2017) utilisent l'augmentation de données, leurs résultats ne dépassent pas ceux des méthodes classiques.

Les résultats obtenus sur le corpus CASME2 et les différentes variantes SMIC placent notre méthodes en tête des méthodes de l'état de l'art sur ce corpus indépendamment de la structuration du corpus et du type de méthode employée (classique vs apprentissage profond). Dans la suite, nous évaluons les performances de notre méthode pour la reconnaissance de macro-expressions.

Macro-expressions

Nous mesurons les performances de reconnaissance des macro-expressions en caractérisant le mouvement à l'aide de LMP sur les corpus suivants CK+ (avec des séquences d'activation variables), Oulu CASIA (avec ses deux versions : VL - lumière naturelle et NI - lumière infrarouge) et MMI (avec de légers mouvements de la tête).

Le corpus CK+ est l'un des plus utilisé pour évaluer les performances de reconnaissance de macro-expressions. Malgré cela, il est souvent difficile de positionner précisément les résultats obtenus par rapport à l'état de l'art car les protocoles expérimentaux présentent souvent des variations (par exemple, le nombre de séquences retenues et le nombre de classes). Ces variations sont la conséquence immédiate de la présences de changements d'illumination, de légers mouvements de la tête ou de l'absence d'annotations dans certaines séquences. Ainsi, certains auteurs exploitent différemment le corpus sans toutefois laisser une trace précise des séquences et des annotations utilisées. Nous avons décidé d'utiliser les deux versions les plus communément utilisées. La première *CK374* contient 6 classes et 374 séquences (dégoût - 45 séquences, énervement - 42 séquences, joie - 100 séquences, peur - 64 séquences, surprise - 80 séquences et tristesse - 82 séquences). La seconde

CK327 contient 7 classes et 327 séquences (dégoût - 59 séquences, énervement - 45 séquences, joie - 69 séquences, mépris - 18 séquences, peur - 25 séquences, surprise - 83 séquences et tristesse - 28 séquences).

La Table 3.3 montre les performances de notre méthode par rapport à des approches récentes et réputées dans l'état de l'art. Initialement publiés dans (Allaert et al. 2017) [6] ces résultats ont été améliorés. Les chiffres rapportés dans la Table 3.3 proviennent de (Allaert et al. 2018a) [7]

TABLE 3.3 – *Comparaison de performances pour la reconnaissance de macro-expressions (* apprentissage profond) .*

Méthode	Protocole	CK+		CASIA		MMI (%)
		327(7)	374(6)	VL	NI	(%)
LBP-TOP Zhao et Pietikainen (2007)	10-fold	-	96,26%	72,13%	71,59%	59,51%
AdaLBP Zhao et al. (2011)	10-fold	-	-	73,54%	72,09%	-
PHOG-TOP + Optical flow Fan et Tjahjadi (2015)	LOO	83,7	-	-	-	-%
DTAGN (joint) Jung et al. (2015)	10-fold	* 97,25%	-	81,46%	-	70,24%
Dis-ExpLet Liu et al. (2016a)	10-fold	95,10%	-	79,00 %	-	77,60%
RBM-based model Elaiwat et al. (2016)	10-fold	95,66	-	-	-	-%
CNN de Mollahosseini et al. (2016)	5-fold	-	-	-	-	* 77,60%
MHI-OF + QLZM-MCF Fan et Tjahjadi (2017)	LOO	88,3	-	-	-	-%
Spatio-temporal RBM model Elaiwat et al. (2016)	10-fold	95,66%	-	-	-	-
LBP-TOP + Gabor Zhao et al. (2017)	10-fold	-	-	74,37%	-	71,92%
CNN + AUs + LTSM Breuer et Kimmel (2017)	LOO	* 98,62%	-	-	-	-
CNN + Spatial features Lopes et al. (2017)	8-fold	* 96,76%	-	-	-	-
CNN + Spatial features Lopes et al. (2017)	8-fold	86,67%	-	-	-	-
PHRNN-MCSNN Zhang et al. (2017a)	10-fold	* 98,50%	-	86,25%	-	81,18%
FN2EN Ding et al. (2017)	10-fold	-	-	* 87,71%	-	-
LMP	**10-fold**	**96,94%**	**96,26%**	**75,13%**	**81,88%**	**74,40%**
LMP + Geom. feat.	**10-fold**	**97,25%**	**96,79%**	**84,58%**	**81,49%**	**78,26%**

Malgré le bruit inhérent à l'extraction du flux optique sur le visage (surface lisse, peu texturé, avec des zones de discontinuité - rides, etc.), l'analyse conjointe de la magnitude et de l'orientation permet de filtrer le mouvement et de caractériser les directions principales du mouvement. Motivé par les améliorations observées dans l'état de l'art lors de la mise en place de méthodes hybrides, nous enrichissons la caractérisation à base de LMP, par une caractérisation géométrique. La caractérisation géométrique renseigne sur les proportions et la forme des 25 régions composant le modèle facial. La caractérisation géométrique est réalisée sur la dernière image de la séquence pour encoder une déformation maximale. Lorsque nous utilisons à la fois les descripteurs géométriques et les descripteurs de mouvement LMP, nous obtenons en général des taux de classification plus élevés.

Sur CK+, nos performances sont comparables avec les approches de Jung et al. (2015), Lopes et al. (2017), Breuer et Kimmel (2017), Zhang et al. (2017a) issues de l'apprentissage profond. Notre approche à la mérite de ne pas nécessiter de grandes quantités de données pour l'apprentissage, ni beaucoup de temps de calcul. En l'absence d'augmentation de données, les approches à base d'apprentissage profond sont moins performantes, comme témoignent les résultats obtenus par Lopes et al. (2017). Cette dernière approche sans augmentation de données, n'obtient que 86,67%, alors que les performances grimpent à 96,76% en présence d'augmentation de données.

6. B. Allaert ; I.M. Bilasco ; C. Djeraba - Consistent Optical Flow Maps for full and micro facial expression recognition - Proc. of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2017, Porto, Portugal, vol. 5, pp.235-242.

7. B. Allaert ; I.M. Bilasco ; C. Djeraba - Advanced local motion patterns for macro and micro facial expression recognition - Computing Research Repository (CoRR), https://arxiv.org/abs/1805.01951.

En condition d'illuminations différentes, comme dans le corpus Oulu-CASIA, notre méthode surpasse les méthodes n'employant pas d'apprentissage profond telles que (Zhao et Pietikainen 2007, Liu et al. 2016a, Zhao et al. 2017). Elle obtient aussi des résultats compétitifs par rapport aux approches d'apprentissage profond (Jung et al. 2015, Zhang et al. 2017a, Ding et al. 2017). La performance obtenue par notre méthode en condition d'illumination en infrarouge (81,88%) dépasse les autres approches. Cela est principalement dû à la robustesse de la caractérisation en présence d'erreurs de localisation de points caractéristiques. Tout comme pour CK+, la mise en commun des informations relatives au mouvement et à la géométrie améliore clairement les performances dans un contexte d'illumination naturelle comme rencontré dans CASIA-VL.

Pour le corpus MMI où de nombreux mouvements de la tête viennent perturber l'analyse, le mouvement et la géométrie se complètent très bien permettant d'obtenir de bonnes performances (78,26%). Ainsi, nous dépassons d'autres méthodes classiques de la littérature comme celles de Zhao et Pietikainen (2007), Zhong et al. (2012b), Liu et al. (2016a), Zhao et al. (2017). Par rapport à d'autres méthodes issues de la mouvance de l'apprentissage profond, nous obtenons de meilleurs résultats que Jung et al. (2015) et Mollahosseini et al. (2016), tout en restant compétitifs par rapport aux résultats présentés par Zhang et al. (2017a).

3.3.4 Synthèse des évaluations des micro- et macro-expressions

La Table 3.4 offre une vue globale sur le positionnement de nos résultats par rapport à l'état de l'art dans les contextes ciblés : la micro- et la macro-expression.

TABLE 3.4 – *Synthèse des résultats obtenus sur les corpus de micro- et macro-expressions (* augmentation de données / apprentissage profond) .*

Méthode	Micro-expression				Macro-expression			
	CASME II	SMIC			CK+	CASIA		MMI
		HS	VIS	NIR	7 classes	VL	NI	
LBP-TOP Zhao et Pietikainen (2007)	-	-	52,11%	-	96,26%	68,13%	-	59,51%
LBP-TOP + GaborZhao et al. (2017)	-	-	-	-	95,80%	74,37%	-	71,92%
AdaLBP Zhao et al. (2011)	-	-	-	-	-	73,54%	72,09%	-
Dis-ExpLet Liu et al. (2016a)	-	-	-	-	95,10%	79,00%	-	77,60%
HIGO + magnification Li et al. (2015)	67,21%	68,29%	81,69%	67,61%	-	-	-	-
LMP	**70,20%**	**67,68%**	**86,11%**	**80,56%**	**97,25%**	**84,58%**	**81,46%**	**78,26%**
CNN + LSTM* Kim et al. (2017)	60,98%	-	-	-	-	-	-	-
CNN* Patel et al. (2016)	47,30%	53,60%	56,30%	-	-	-	-	-
CNN + LSTM* Breuer et Kimmel (2017)	59,47%	-	-	-	98,62%	-	-	-
PHRNN-MSCNN* Zhang et al. (2017a)	-	-	-	-	98,50%	86,25%	-	81,18%
FN2EN* Ding et al. (2017)	-	-	-	-	-	87,71%	-	-

Les résultats montrent que la méthode que nous proposons présente la singularité de répondre de manière unifiée aux défis posés pour les micro- et les macro-expressions. La méthode dépasse les performances obtenues par les méthodes de l'état de l'art pour les micro-expressions. En moyenne, nous obtenons des résultats meilleurs de 4,93% par rapport aux approches classiques et de 17,7% par rapport aux approches d'apprentissage profond. De plus, nous obtenons des résultats compétitifs pour la reconnaissance de macro-expressions dans différentes conditions d'acquisition (séquence d'activation, illumination, mouvements de la tête). En moyenne nous devançons de 4,15% les meilleures approches classiques pour tous les corpus considérés. Les approches issues de l'apprentissage profond s'appuyant également sur des processus d'augmentation de données nous dépassent de 2,25% en moyenne.

Certaines méthodes issues de l'apprentissage profond réalisent de meilleures performances que

les notres. Toutefois, il est important de mettre en lumière le fait que notre approche est la seule, à notre connaissance, qui traite de manière unifiée et compétitive les micro- et les macro-expressions.

TABLE 3.5 – *Les paramètres utilisés pour obtenir les meilleures performances.*

Corpus		λ	Δ	ρ	E	M	V	β	bin
Micro	CASME II	4	0,5	0,75	100	4	5	6	9
	SMIC-HS	3	0,5	0,75	100	3	5	6	9
	SMIC-VIS	5	0,5	0,75	100	4	5	3	9
	SMIC-NIR	4	0,5	0,75	100	3	5	3	12
Macro	CK+	3	0,5	1	100	4	5	3	12
	MMI	3	0,5	1	100	4	5	6	12
	CASIA-VL	4	0,5	1	100	5	5	3	6
	CASIA-NI	5	0,5	0,75	100	5	5	6	9

Les paramètres utilisés pour mesurer les performances de notre méthode pour les différents corpus sont présentés dans la Table 3.5. Selon les conditions de capture des corpus de données (distance par rapport à la caméra, résolution, cadence de trames), les paramètres varient faiblement soulignant la capacité de généralisation de notre approche unifiée.

Les résultats obtenus pour la reconnaissance de micro- et macro-expressions montre l'efficacité et la robustesse du descripteur et du modèle facial proposé. Notre contribution se positionne comme un bon candidat pour la reconnaissance des expressions dans des contextes difficiles, proches de situations d'interaction naturelle (mouvements de la tête, différents types d'illumination, séquence d'activation et intensités variables).

3.4 Résumé des contributions

Dans ce chapitre nous avons illustré deux de nos contributions dans le domaine de la reconnaissance des expressions. Nous nous sommes intéressés à la reconnaissance dans un cadre statique en optimisant la construction de masques reflétant l'importance de certaines zones du visage. Dans un cadre dynamique, nous avons montré l'intérêt de filtrer le mouvement sur le visage en s'intéressant aux spécificités des mouvements faciaux.

Ainsi, dans un premier temps, nous avons abordé le problème d'optimisation de caractérisation globale d'un visage en étudiant le choix et l'étendue des régions à caractériser. Nous avons proposé une méthodologie de construction de masques non-rectangulaires qui tiennent compte de la contribution de chaque pixel de l'image dans le processus de reconnaissance. Nous réalisons une recherche exhaustive de fenêtres de tailles et formes différentes, disposées à différents endroits sur un visage normé. Cette recherche offre des indications quant aux meilleures et pires configurations (type fenêtre et position) en termes de résultats de classification. Suite à l'analyse des régions sélectionnées comme étant les meilleures ou les pires, des opérations topologiques nous permettent de dégager un masque dont les points ont un pouvoir discriminant important. Les expérimentations entreprises ont montré que des régions de petites tailles peuvent apporter plus de précision que des régions de taille supérieure. Cela défend la prédominance du caractère local dans la caractérisation des changements reflétant l'apparition d'une expression. Ce constat est également renforcé par le fait que des régions de même taille produisent des résultats assez différents lorsqu'elles sont distantes de quelques pixels seulement. Malgré le fait que la construction du masque soit laborieuse, son utilisation est simple et efficace. Superposé sur un visage, le masque permet de prendre en

considération uniquement les points labélisés comme pertinents pour la reconnaissance. Des travaux similaires sont menés actuellement dans le cadre de la thèse de Delphine Poux [8]. Nous avons transposé cette méthodologie à l'identification des cartes de mouvement spécifiques en présence d'une large typologie d'occultations. Les premiers résultats ont déjà été publié dans (Poux et al. 2018) [9].

In a second step, we showed that the coherent movement extracted on the face using LMP allows to obtain high performances for both micro- and macro-expressions. The constraints relating to the orientation and the intensity of the movement make it possible to separate the real movement from the noise. The results illustrated on the CASME2 and SMIC corpora show that our approach surpasses the recent approaches of the state of the art (including deep learning approaches). The precise characterization of the movement allows us to free ourselves from pretreatments artificially intensifying the movement such as magnification or interpolation. Similarly, without using manual annotations or data augmentation techniques, we obtain competitive performances on the CK +, CASIA-VL, CASIA-NI and MMI corpora. Thus, the facial model and the LMP descriptor meet both the challenges posed by the recognition of micro- and macro-expressions under different illumination conditions and in the presence of slight movements of the head.

8. Delphine Poux, doctorante depuis octobre 2015, équipe FOX, lab. CRIStAL cofinancement par l'Ecole d'ingénieurs des Hautes Technologies et du Numérique - ISEN YNCREA.

9. D. Poux ; B. Allaert ; J. Mennesson ; N. Ihaddadene ; I.M. Bilasco ; C. Djeraba - Mastering Occlusions by Using Intelligent Facial Frameworks Based on the Propagation of Movement - Proc. of International Conference on Content-Based Multimedia Indexing (CBMI), Sept. 2018, La Rochelle, France.

Synthèse 4

Ces dernières années, avec mes collègues et collaborateurs, nous avons œuvré à faciliter l'analyse faciale dans des contextes non-contraints. Nous avons étudié et nous avons conduit de nombreuses expérimentations dans les domaines de l'estimation de l'orientation de la tête, de la reconnaissance du genre et de la reconnaissance des expressions.

Dans le domaine de l'orientation de la tête, nous avons proposé deux approches innovantes. La première, réalisée dans le cadre de la thèse d'Afifa Dahmane [1], est basée sur la caractérisation de la symétrie du visage. La seconde, construite avec Taner Dansiman [2] et José Menneson [3], exploite les spécificités de détecteurs frontaux et, par le biais d'une transformation inverse, offre une estimation du roulis et du lacet.

Avec Taner Danisman, nous avons exploré la caractérisation de personnes en nous intéressant à la reconnaissance du genre. Des résultats intéressants ont pu être obtenus dans un contexte de validation entre différents corpus de données en appliquant un processus de normalisation de l'intérieur du visage. Ainsi, nous avons pu prouver que la représentation normalisée du visage, malgré sa petite taille, permet de répondre de manière convenable aux défis rencontrés dans un large panel de bases de données. Toutefois, la caractérisation du genre en se basant uniquement sur les éléments du visage ne peut pas répondre à la variété des situations que nous pouvons rencontrer dans la vie courante. Certains individus ont des traits plus proches du sexe opposé que du leur. Ainsi, nous avons poursuivi ces travaux en les enrichissant avec des informations caractérisant à la fois des éléments précis à l'intérieur (moustache, barbe) et à l'extérieur du visage (cheveux).

Les travaux autour de la reconnaissance du genre ont souligné l'importance des processus de normalisation du visage avant caractérisation. Ainsi, nous avons pu généraliser l'approche précédente à la caractérisation de l'état émotionnel des personnes en proposant une méthode globale pour la détection de la joie dans des conditions difficilement abordables (taille d'images très petite, faible illumination, etc.) avec des méthodes classiques. Ces travaux ont permis également de construire des masques de pixels sur les visages capables d'améliorer la reconnaissance de l'expression de joie à partir des niveaux de gris sans passer par des descripteurs complexes.

1. Afifa Dahmane, doctorante de mai 2010 à février 2015, équipe FOX, lab. CRIStAL en co-tutelle avec l'Université de Sciences et Technologies Houari Boumediene (USTHB), Algérie, actuellement Maître de Conférences à l'USTHB.

2. Taner Danisman, post-doctorant septembre 2010 à juin 2011 et ingénieur de recherche de juin 2012 à octobre 2014, équipe FOX, lab. CRIStAL, actuellement enseignant-chercheur à Akdeniz Üniversitesi, Turquie.

3. José Mennesson, ingénieur de recherche de novembre 2014 à décembre 2015, équipe FOX, lab. CRIStAL, actuellement Maître Assistant à l'Institut Mines-Telecom Lille Douai, équipe FOX, lab. CRIStAL.

Les travaux conduits sur des données statiques (images ou trames d'une vidéo) ont été poursuivis par les recherches menées dans le cadre de la thèse de Benjamin Allaert [4] sous ma co-direction. Ces travaux visent à réduire l'écart entre la reconnaissance des expressions exagérées et la reconnaissance des expressions à intensités variables. Nous avons adopté une approche permettant de filtrer et caractériser le mouvement facial en conservant un maximum d'information cohérente indépendamment de l'intensité de l'expression sous-jacente. La méthode mise en œuvre obtient d'excellents résultats pour la reconnaissance de micro- et macro- expressions.

4. Benjamin Allaert, doctorant de octobre 2014 à juin 2018, équipe FOX, lab. CRIStAL, actuellement ingénieur de recherche équipe FOX, lab. CRIStAL.

Troisième partie

Projet de recherche

L'essentiel de mes récentes activités de recherche s'est concentré sur la reconnaissance d'expressions faciales (micro et macro) dans un contexte non-contraint. Ce type de contexte présente de nombreux défis : intensités variables, séquences d'activation spécifiques, mouvements de la tête, occultations. Les travaux réalisés dans le cadre de la thèse de Benjamin Allaert [1] ont montré que le mouvement permet de répondre à plusieurs défis, tels que la variation d'intensité des mouvements faciaux (micro et macro mouvements) et la segmentation temporelle de l'activation musculaire. Cela me pousse à creuser l'utilisation de l'analyse du mouvement facial afin de répondre à d'autres défis posés (par exemple, occultations ou mouvements de la tête) dans le cadre d'une situation d'interaction naturelle.

Même si le mouvement à lui seul semble prometteur, répondre aux défis posés par les mouvements de la tête et les occultations suppose également une meilleure maîtrise de la caractérisation du visage dans son ensemble. Cela commence par un approfondissement des problématiques liées à la localisation de points caractéristiques en présence de mouvements de la tête. Avec Pierre Tirilly (MCF, équipe FOX, lab. CRIStAL) et Nacim Ihaddadene (MCF, ISEN YNCREA), nous avons initié dans le cadre de la thèse de Romain Belmonte [2] des travaux autour de la localisation de points caractéristiques. Ces travaux visent la mise en place des réseaux de neurones profonds qui encodent conjointement les dimensions spatiale et temporelle d'une séquence d'images, tout en offrant un bon compromis entre complexité et efficacité. Je considère qu'il est intéressant de poursuivre ces efforts pour identifier des architectures qui traitent nativement de séquences d'images sans se limiter à une analyse temporelle tardive, telle que mise en place au sein des réseaux de neurones récurrents (RNN) ou dérivés.

Bien qu'une certaine expertise se dégage de l'étude analytique de ces problèmes, les solutions passent souvent par des processus d'apprentissage. L'apprentissage a un rôle fondamental dans l'ensemble des processus de reconnaissance évoqués dans ce manuscrit, même si, souvent, il accompagne une étape d'analyse qui reflète notre compréhension des défis et notre expertise pour y répondre. L'ampleur prise par les processus d'apprentissage profond nous invite naturellement à consacrer du temps à une meilleure maîtrise et compréhension de différentes architectures capables d'encoder efficacement la caractérisation dynamique du visage. En lien avec ces architectures, les processus d'augmentation de données sont intimement liés au succès des méthodes d'apprentissage profond. De nombreuses méthodes d'augmentation ont été explorées pour générer artificiellement de grands volumes de données à partir de corpus ayant une taille modérée. Elles sont transposées pour des séquences en appliquant des opérateurs d'augmentation d'images sur l'ensemble de trames de la séquence. Toutefois, à mon sens, cela limite les possibilités d'augmentation de données dynamiques. Je souhaite explorer de nouvelles pistes de recherche concernant la mise en place des processus d'augmentation exploitant la nature temporelle des séquences.

Les tendances actuelles s'appuient sur des processus d'apprentissage de plus en plus gourmands en termes d'énergie et de temps de calcul pour faire converger d'importantes architectures neuronales dont la complexité ne cesse de grandir. En rupture avec cette course à l'énergie et à la complexité des architectures, je souhaite poursuivre les travaux initiés autour des réseaux

1. Benjamin Allaert, doctorant de octobre 2014 à juin 2018, équipe FOX, lab. CRIStAL, actuellement ingénieur de recherche équipe FOX, lab. CRIStAL.

2. Romain Belmonte, doctorant depuis octobre 2015, équipe FOX, lab. CRIStAL cofinancement par Ecole d'ingénieurs des Hautes Technologies et du Numérique - ISEN YNCREA et la Métropole Européenne de Lille.

de neurones impulsionnels (ou à spikes) qui offrent un bon compromis d'un point de vue de la consommation énergétique.

Dans la suite, je détaille les pistes évoquées ci-dessus que je résume ainsi :

— la reconnaissance des expressions faciales en présence d'occultations partielles du visage ;
— la séparation des sources de mouvement sur le visage ;
— la construction de nouvelles architectures pour l'analyse de séquences d'images ;
— la conception de nouvelles techniques d'augmentation de données pour les données spatio-temporelles ;
— l'élaboration de mécanismes d'apprentissage à base de réseaux de neurones impulsionnels.

Reconnaissance des expressions faciales en présence d'occultations partielles du visage

Les premiers travaux autour des occultations ont été menés dans la thèse de Delphine Poux [3] débutée en octobre 2017 et co-encadrée par Chaabane Djeraba (PR, équipe FOX, lab. CRIStAL) et Nacim Ihaddadene (MCF, ISEN YNCREA). Cette thèse s'inscrit naturellement dans la suite des travaux de Benjamin Allaert, offrant un cadre idéal pour continuer l'exploration des problématiques liées au mouvement en présence d'occultations. Les premiers résultats obtenus montrent que l'étude du mouvement offre des perspectives intéressantes lorsque l'on s'intéresse aux défis soulevés par les occultations. En effet, la propagation naturelle du mouvement au-delà des zones occultées permet d'exploiter partiellement le mouvement. Bien que l'épicentre du mouvement soit occulté, le mouvement de la région occultée se répand dans les régions voisines visibles.

Dans les travaux récents de Delphine Poux, nous montrons que, par expression et par type d'occultation, il est possible de sélectionner un nombre réduit de régions afin de retrouver des performances similaires à des situations ne comportant aucune occultation. Ces premières expériences prometteuses seront approfondies afin de proposer des solutions plus génériques supportant de manière concomitante l'ensemble des expressions et une large palette d'occultations. Par ailleurs, des méthodes de reconstruction du mouvement dans les zones occultées du visage sont à l'étude.

Séparation des sources de mouvement sur le visage

Les méthodes qui étudient le mouvement, ou qui, de manière plus générale, caractérisent l'évolution temporelle de la texture faciale, fonctionnent bien lorsque le seul mouvement perçu sur le visage est celui induit par l'expression. Cependant, dans une situation d'interaction spontanée, les expressions faciales sont souvent accompagnées par des mouvements de la tête. Lorsque la tête se déplace pendant qu'une personne manifeste une certaine expression, le mouvement facial perçu reflète en même temps le mouvement induit par la tête et le mouvement induit par l'expression faciale. Par exemple, les expressions de surprise ou de peur sont accompagnées généralement d'un mouvement en arrière de la tête. Dans ce cas, il est nécessaire d'explorer des outils et des méthodes afin de pouvoir exploiter le mouvement facial en présence de mouvements de la tête.

Des techniques d'alignement du visage ont souvent été employées pour corriger les variations de pose de la tête afin de disposer d'une image du visage ne comportant aucun mouvement de tête. Afin de traiter efficacement les visages en mouvement, certains proposent de segmenter de manière très fidèle ces visages. Ainsi, l'évolution de la texture d'une part, ou le mouvement d'autre

3. Delphine Poux, doctorante depuis octobre 2015, équipe FOX, lab. CRIStAL cofinancement par l'Ecole d'ingénieurs des Hautes Technologies et du Numérique - ISEN YNCREA.

part, peuvent être extraits de manière relative par rapport aux périmètres identifiés lors de la segmentation. Sans cette segmentation précise, les bénéfices d'une caractérisation dynamique du visage sont perdus. Les approches d'alignement facial s'améliorent de manière constante pour répondre aux défis posés par les situations d'interaction peu contraintes. Néanmoins, il est toujours difficile de trouver des solutions qui préservent efficacement les expressions faciales, notamment en présence de faibles intensités et de larges variations de pose. Les pertes de performance observées, lorsque des processus d'alignement sont employés pour corriger les variations de pose de la tête, sont en partie dues aux artefacts générés par l'alignement, comme souligné dans les travaux récents de Benjamin Allaert.

Ainsi, nous pensons qu'il est nécessaire de traiter le problème de séparation du mouvement de la tête et du mouvement du visage dans le domaine du flux optique. Cela reste un problème difficile à résoudre lorsque les deux sources de mouvement se confondent, car leurs spécificités rendent la séparation difficile. Le mouvement de la tête est souvent important et il prend le dessus sur le mouvement facial propre à l'expression. Deux pistes sont actuellement envisagées pour différencier les effets des deux mouvements : une méthode analytique et une méthode à base d'apprentissage. En collaboration avec François Lemaire (MCF, équipe Calcul Formel et Haute Performance, lab. CRIStAL), nous essayons de formuler mathématiquement l'impact du mouvement de la tête sur le mouvement facial afin d'être en mesure de les séparer ensuite. En parallèle, nous explorons l'usage des auto-encodeurs afin de séparer les deux sources de mouvement. Nous fournissons des séquences vidéo contenant d'une part, uniquement le mouvement induit par l'expression, et d'autre part, le mouvement induit par la variation de la tête englobant celui des expressions. Les séquences disponibles dans le corpus SNaP-2DFE (Allaert et al. 2018b) ont été captées simultanément dans le but de pouvoir étudier de manière implicite, à l'aide de l'apprentissage, l'extraction du mouvement facial en présence du mouvement de la tête.

Construction de nouvelles architectures pour l'analyse de séquences d'images Lorsque l'on s'intéresse à l'étude de phénomènes dynamiques impliquant le mouvement, souvent, on part des architectures conçues pour l'analyse d'images statiques. Les descripteurs intermédiaires synthétisés sont ensuite interprétés à une plus large échelle temporelle afin d'étudier le phénomène dans son ensemble. Appuyé par les résultats obtenus en étudiant directement le mouvement, j'estime qu'il est important d'étudier localement les évolutions temporelles des caractéristiques visuelles au sein d'une séquence. Cette proposition, qui vise à mettre en place des convolutions spatio-temporelles très tôt dans le processus d'analyse, ne se veut pas une alternative aux solutions mettant en œuvre une introduction tardive de l'analyse temporelle, mais plutôt un outil complémentaire.

Les résultats obtenus dans ce sens dans la thèse de Romain Belmonte sont très encourageants. En mettant en place des convolutions spatio-temporelles et en utilisant un nombre relativement réduit de couches, nous obtenons des performances comparables à l'état de l'art pour la localisation de points d'intérêts sur un visage. D'autres travaux tels que Baccouche et al. (2011), Zhang et al. (2017b) menés dans le domaine de la reconnaissance d'actions, montrent également l'intérêt de construire des architectures spatio-temporelles.

Dans un premier temps, je souhaite explorer des approches similaires pour caractériser les séquences d'activation des expressions faciales. Dans ce contexte, il est important d'étudier avec soin l'impact de la taille de la fenêtre temporelle et spatiale pour les convolutions utilisées. Les ensei-

gnements tirés des expériences mises en place lors de l'étude des motifs locaux de mouvements peuvent offrir des indices sur la taille de la fenêtre spatiale de convolution. De même, les nombreux travaux concernant le calcul de flux optique en s'appuyant sur des architectures profondes constitueront des objets d'étude privilégiés.

Conception de nouvelles techniques d'augmentation de données pour les données spatio-temporelles L'apprentissage profond est caractérisé par des architectures dépendant d'un nombre colossal de paramètres. Faire converger ces réseaux et identifier les valeurs optimales de paramètres requiert un nombre important de données d'apprentissage. Lorsque des techniques d'apprentissage profond sont appliquées sur des collections contenant peu de données, il est nécessaire de passer par des processus d'augmentation artificielle de données pour disposer d'un plus grand ensemble d'apprentissage. Les techniques d'augmentation recensées dans la littérature se concentrent majoritairement autour de l'augmentation d'images et consistent en des opérations de mise à l'échelle, de rotations ou de constructions de l'image en miroir horizontal ou vertical. La mise à disposition de données variées renforce les capacités de généralisation du processus d'apprentissage en apprenant des situations non prévues dans le corpus initial.

Les nouvelles architectures traitant directement de l'analyse spatio-temporelle des séquences pourraient, à mon sens, prendre avantage de processus d'augmentation qui ne s'appliquent pas uniquement image par image, mais qui traitent les séquences dans leur ensemble. Par exemple, des motifs d'activation variés peuvent être obtenus en partant de la séquence d'activation d'une expression en effectuant différentes opérations d'interpolation temporelle. Afin de rendre l'analyse faciale robuste à l'apparition de mouvements de la tête, des augmentations spécifiques pourraient être mises en œuvre pour induire des rotations dans le plan et hors-plan simulant les mouvements de la tête. La mise en place de ces nouvelles formes d'augmentation requiert une attention spéciale. Il faut veiller à ce que les augmentations ainsi mises en place ne nuisent pas au processus d'apprentissage par l'introduction d'artefacts.

J'estime que la mise en place de ces techniques d'augmentation accompagnera naturellement les nouvelles architectures d'apprentissage qui s'intéressent à la nature temporelle de l'analyse de séquences d'images.

Élaboration de mécanismes d'apprentissage à base de réseaux de neurones impulsionnels En parallèle de l'étude des architectures classiques d'apprentissage profond, je souhaite poursuivre mes recherches autour des architectures à base de neurones impulsionnels. Ces dernières se positionnent comme une alternative moins gourmande énergétiquement que les architectures classiques. Les premiers résultats obtenus dans la thèse de Pierre Falez [4] ont démontré la capacité des architectures à base de réseaux de neurones impulsionnels à reconnaître et de caractériser les mouvements d'unités élémentaires : points et lignes dans une image. Cela laisse entrevoir la possibilité d'étudier des mouvements plus complexes (par exemple, gestes ou expressions). Dans ce contexte, je souhaite explorer des méthodes encodant de manière efficace les informations de mouvement, au sens large, dans les architectures SNN. De nouveaux systèmes de vision dynamique (DVS) permettent d'encoder naturellement sous forme d'une séquence d'activation le mouvement dans

4. Pierre Falez, doctorant depuis octobre 2016, équipe FOX, lab. CRIStAL en co-direction avec l'équipe Émeraude, lab. CRIStAL.

une scène. Pour cela, il est nécessaire d'étudier différentes architectures SNN qui permettent d'exploiter efficacement la temporalité du phénomène étudié. L'étude des architectures SNN traitant efficacement du mouvement peut se révéler bénéfique également en présence de mouvements extraits des caméras classiques. Dans ces conditions, le mouvement est encodé sous la forme de flux optique gardant trace des changements intervenus dans une scène. Ainsi, il est pertinent d'étudier les mécanismes d'encodage de ces flux optiques sous forme de trains d'impulsions.

L'objectif à long terme est d'utiliser ces architectures pour résoudre des problèmes de vision dans un contexte d'analyse du comportement humain. La mise en place de systèmes de vision bio-inspirés de bout en bout nécessiterait l'usage de capteurs bio-inspirés (ex : rétines artificielles type DVS) et l'ajout de supervision dans les règles d'apprentissage. Les travaux effectués jusqu'ici ont permis d'identifier un certain nombre de verrous à lever pour atteindre ces objectifs :

— le choix des prétraitements et de l'encodage des données en entrée (notamment, pour préserver l'information de couleur) ;
— l'adéquation des règles d'apprentissage et des architectures de réseaux aux tâches visées ;
— le passage à l'échelle (nombre de couches, quantité et complexité des données).

Quatrième partie

Curriculum Vitae

Etat civil

Bilasco Ioan Marius — marius.bilasco@univ-lille.fr
Né le 8 janvier 1980 — 15 rue Fémy
Nationalité : roumaine, française — 59800 Lille
Marié, 2 enfants — tél. : 06 63 61 91 97

Expérience professionnelle

Enseignant-chercheur (MCF)

Depuis 2009 — Département Informatique, Faculté de Sciences et Technologies de l'Université de Lille
- Service d'enseignement effectué en Licence et Master Informatique et Master Méthodes Informatiques Appliquées à la Gestion des Entreprises (MIAGE)
- Directeur d'études Master Informatique parcours E-Services formation initiale (FI) et formation en alternance (FA) – depuis 2012
- Responsable de suivi des alternants inscrits en Licence et Master – toutes filières confondues

Membre de l'équipe de recherche Fouille de dOnnées compleXes du Centre de Recherche en Informatique, Signal et Automatique de Lille (UMR 9189)
Prime d'Encadrement Doctoral et de Recherche - Depuis 2015

Post doctorant CNRS

2008 – 2009 — Laboratoire d'Informatique Fondamentale de Lille (LIFL) - Unité Mixte de Recherche (UMR 8022) – équipe Fouille de dOnnées compleXes (FOX)
Participation au projet européen ITEA2 (Information Technology for European Advancement) - Collaborative Aggregated Multimedia for Digital Home – responsabilité de la sous-tâche relative à la modélisation et l'exploitation de métadonnées – www.cam4home-itea.org
Intégration de métadonnées et extraction d'information de la vidéo
Vacataire à Télécom Lille 1

Attaché temporaire à l'enseignement et à la recherche

2006 – 2008 — Institut Universitaire de Technologie I – Université Joseph Fourier, Grenoble
Service effectué au sein du département Réseaux et Télécommunications

Moniteur de l'enseignement supérieur

2003 – 2006 — Institut National Polytechnique Grenoble
Service effectué à l'Ecole Supérieure d'Ingénieurs en Systèmes Avancés Rhône-Alpes à Valence

Formation

2003 – 2007 — Université Joseph Fourier, Grenoble
Doctorat en Informatique : Une approche sémantique pour la réutilisation et l'adaptation de données 3D
Défendu le 19 décembre 2007. Directeurs : Hervé MARTIN et Marlène VILLANOVA-OLIVER *e-thèse* : http://tel.archives-ouvertes.fr/tel-00206220/fr/

2002 – 2003 — Université Joseph Fourier, Grenoble
Diplôme d'Etudes Approfondies (DEA) en Informatique : Systèmes et Communication ISC *mention Bien*

Synthèse de mes activités de recherche, pédagogique et administratives 1

Dans cette partie, je présente une vue temporelle de mes co-encadrements et de mes implications dans les activités scientifiques connexes, telles que le montage de projets ou la contribution à la vie scientifique de la communauté. Je poursuis en listant l'intégralité de mes publications et autres activités de dissémination depuis 2009. Afin de dresser un portrait fidèle de l'ensemble de mes activités, avant de conclure, je mets également en lumière les importantes responsabilités pédagogiques que j'ai eu l'occasion de porter en parallèle de mes activités de recherche.

1.1 Encadrements

Dans cette section, je passe en revue les co-encadrements de doctorants (8) et la coordination des travaux de post-doctorants et assimilés (4). La Figure IV.1 en offre une vue synthétique structurée autour des principaux domaines de recherche : les métadonnées et le multimédia (MM), l'estimation de la pose (P), la reconnaissance du genre (G) et des expressions (E), l'alignement de visages (AV), la fusion de données de plusieurs capteurs (FC), l'apprentissage profond (AP) et les réseaux de neurones impulsionnels (SNN - Spiking Neural Networks).

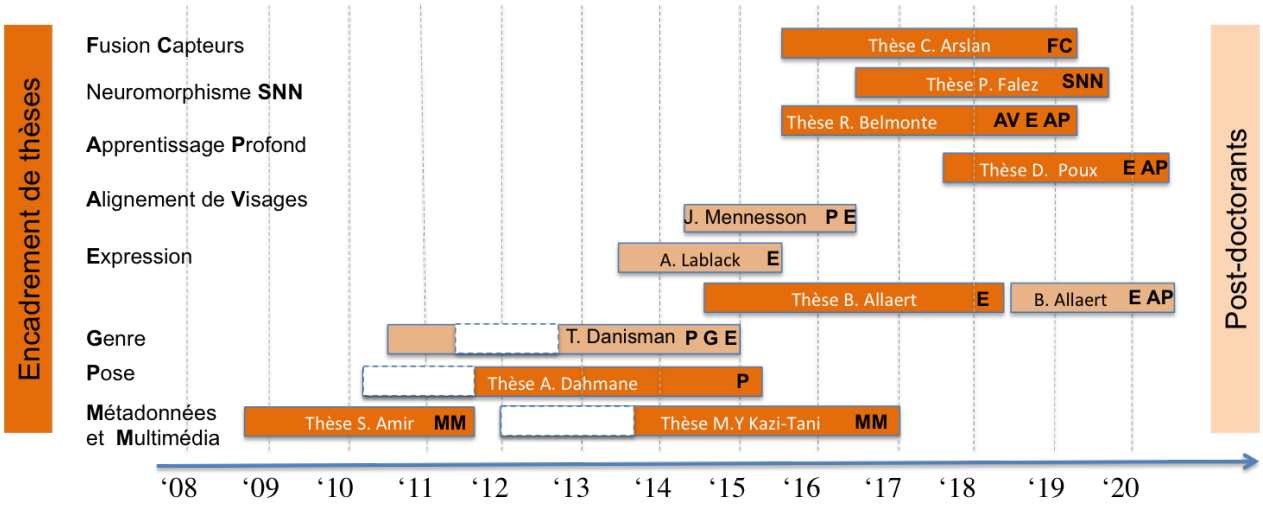


Figure IV.1 – *Synthèse des encadrements doctoraux et post-doctoraux.*

L'ensemble des encadrements (doctorant, post-doctorants) est décrit ci-dessous en précisant pour chacun le sujet principal, le taux d'encadrement, les publications réalisées en commun, ainsi que le devenir des doctorants et post-doctorants. Je structure la présentation des encadrements de

thèse en trois catégories : les thèses déjà soutenues, les thèses en cours au sein de l'équipe FOX et les thèses en cours en collaboration avec d'autres équipes et laboratoires de recherche. Dans chaque catégorie, les thèses sont présentées dans l'ordre chronologique.

Afin de renseigner de manière explicite la nature des publications produites dans le cadre de mes activités d'encadrement, les réferences citées dans la suite de cette section emploient le format suivant : A(rticle)/C(orpus de Données ou d'Annotations)/CH(apitre de livre) C(onférence)/J(ournal)/W(orkshop) N(ational)/I(nternational) ou P(re-)P(rint). La numérotation reflète l'ordre chronologique inverse de la date de publication (de la plus récente à la plus ancienne) dans chaque catégorie. Ces références sont détaillées dans Partie IV section 2.

1.1.1 Thèses soutenues (4)

Un système d'intégration de métadonnées dédiées au multimédia
Samir Amir, septembre 2008 - décembre 2011, inscrit à l'Université Lille 1
Soutenance le 6 décembre 2011 à Lille, France
Encadrement : moi-même à 50% et C. Djeraba (PR, équipe FOX, lab. CRIStAL)
Financement : projet ITEA2 CAM4HOME
Publications : [AJI7][AJI8][ACI16][ACI17][ACI21][ACI20][ACI24][ACI25][ACN10][ACN11][CH1][CH2]
Devenir : R&D Team Leader à Press'Innov (Lyon)

Estimation de l'orientation de la tête dans les images
Afifa Dahmane, mai 2010 - février 2015, inscrite à l'Université Lille 1 et à l'USTHB, Algérie
Soutenance le 1er février 2015 à Alger, Algérie
Financement : bourse du Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, Algérie
Encadrement : moi-même à 33% (à partir de nov. 2011), C. Djeraba (PR, équipe FOX, lab. CRIStAL) et S. Larabi (PR, Université de Sciences et Technologies Houari Boumediene, Algérie)
Publications : [AJI5][ACI8][ACI9][ACI15][ACN9][AWI4][CDA1]
Devenir : Maître de Conférences à l'USTHB, Algérie

Analyse d'expressions faciales dans un flux vidéo
Benjamin Allaert, octobre 2014 - juin 2018, inscrit à l'Université de Lille
Soutenance le 8 juin 2018 à Lille
Encadrement : moi-même à 50% et C. Djeraba (PR, équipe FOX, lab. CRIStAL)
Financement : contrat doctoral de l'ED SPI de Université Lille Nord-de-France
Publications : [AJI1][ACI7][ACI10][AWI1][ACN4][ACN5][ACN6][CDA3][PP1][PP3][PP4][PP5]
Devenir : Ingénieur de Recherche - projet ITEA3 PAPUD (2018-2020), équipe FOX, lab. CRIStAL

Intégration des règles d'inférence sémantiques dans les processus d'analyse vidéo
Mohamed Yassine Kazi Tani, janvier 2012 à mars 2018, inscrit à l'Université d'Oran Es Sénia
Soutenance le 24 Mars 2018 à Oran, Algérie
Financement : Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, Algérie
Encadrement : moi-même à 15% à l'occasion de séjours doctoraux (sept.-oct. 2013, avril-mai 2014, nov. 2014) et par A. Ghomari (Université Oran Es Sénia, Algérie)

Publications : [AJI2][ACI13][AWI3]
Devenir : Maître Assistant à l'École Supérieure en Informatique 8 mai 1945 - Sidi Bel Abbes, Algérie.

1.1.2 Thèses en cours au sein de l'équipe FOX (2)

Contextualisation de l'analyse du comportement centré individu en situation de captation non–contrainte
Romain Belmonte, octobre 2015 - juin 2019, inscrit à l'Université Lille
Encadrement : moi-même à 25%, P. Tirilly (MCF, équipe FOX, lab. CRIStAL), N. Ihaddadene (MCF, École d'ingénieurs des Hautes Technologies et du Numérique - ISEN YNCREA) et C. Djeraba (PR, équipe FOX, lab. CRIStAL)
Financement : ISEN YNCREA et la Métropole Européenne de Lille
Publications : [ACI1][ACI6][ACN2][PP4]

Reconnaissance des expressions faciales dans un environnement non-contraint
Delphine Poux, octobre 2017 - septembre 2020, inscrite à l'Université Lille
Encadrement : moi-même à 33%, N. Ihaddadene (MCF, École d'ingénieurs des Hautes Technologies et du Numérique - ISEN YNCREA) et C. Djeraba (PR, équipe FOX, CRIStAL)
Financement : contrat doctoral de l'ED SPI Université Lille Nord-de-France et ISEN YNCREA
Publication : [ACI2][ACN1]

1.1.3 Thèses en cours en collaboration avec d'autres équipes et laboratoires (2)

Data Fusion for Humain-Computer Interaction
Cagan Arslan, octobre 2015 - juin 2019, inscrit à l'Université de Lille (en cours)
Encadrement : moi-même à 20%, J. Martinet (MCF HDR, équipe FOX, lab. CRIStAL) et L. Grisoni (PR, équipe MINT, lab. CRIStAL)
Financement : contrat doctoral de l'ED SPI Université Lille Nord-de-France Publications : [ACI3][ACI5]

Exploration d'architecture d'un accélérateur neuromorphique pour la vision
Pierre Falez, octobre 2016 - septembre 2019, inscrit à l'Université de Lille
Encadrement : moi-même à 25%, P. Tirilly (MCF, équipe FOX, lab. CRIStAL), Ph. Devienne (CR, équipe Émeraude, lab. CRIStAL) et P. Boulet (PR, équipe Émeraude, lab. CRIStAL)
Financement : contrat doctoral de l'ED SPI Université Lille Nord-de-France
Publications : [ACI4][ACN3][PP2]

1.1.4 Encadrement Post-doc et assimilés (4)

Étude des mouvements de la tête dans les flux vidéo
José Mennesson, novembre 2014 - décembre 2015
Encadrement : moi-même à 100%
Financement : projet ITEA2 EMPATHIC
Publications : [AJI1][ACI8][ACI10][AWI1][ACN4][ACN5][CDA3]
Devenir : Maître Assistant à l'Institut Mines-Telecom Lille Douai

Étude de la valence des états affectifs
Adel Lablack, juin 2013 - août 2015
Encadrement : moi-même à 100%
Financement : projet ITEA2 EMPATHIC
Publications : [AJI2][ACI11][ACI13][ACN6][ACN7][AWI3][AWI4][AWI5][AWI6]
Devenir : Ingénieur R&D FLIR, Courtrai, Belgique

Reconnaissance du genre et des expressions faciales ; Estimation de l'orientation de la tête
Taner Danisman, septembre 2010 - juin 2011 et juin 2012 - octobre 2014
Encadrement : moi-même à 75%, J. Martinet (MCF HDR, équipe FOX, lab. CRIStAL) et C. Djeraba (PR, équipe FOX, lab. CRIStAL)
Financement : projet ITEA2 MIDAS, puis projet ITEA2 TWIRL
Publications : [AJI3][AJI4][AJI6][ACI8][ACI11][ACI12][ACI19][ACI22][AWI4][AWI6][ACN8][CDA2] [CDA4] [CDA5] [CDA6]
Devenir : enseignant-chercheur à l'Université d'Akendiz, Turquie

Fusion de données issues de bio-capteurs et de l'analyse faciale pour inférer l'état affectif
Benjamin Allaert, depuis septembre 2018
Encadrement : moi-même at 75%, J. Mennesson (*FOX team, CRIStAL lab.*)
Financement : projet ITEA3 PAPUD
Publications : [ACI2][ACN1]

La Figure IV.2 reprend par domaine de recherche et par ordre chronologique inversé les publications réalisées en collaboration avec les doctorants et post-doctorants.

Au-delà de ces encadrements de doctorants et post-doctorants, je me suis investi également dans l'initiation des étudiants de Master Informatique à la recherche. Ainsi, j'ai pu encadrer des stagiaires de niveau Master 2 (15 dont 5 stages de fin d'études et 10 sur des projets universitaires en lien avec mes recherches) et des étudiants de niveau Master 1 (21 dont 20 sur des projets universitaires en lien avec mes recherches et un stage). Ces stages sont principalement, en relation avec des projets dans le domaine de l'expression et de l'estimation de la pose.

1.2 Responsabilités scientifiques

Dans ce qui suit, l'ensemble de mes responsabilités scientifiques est présenté, tant en matière de montages et participations aux projets collaboratifs nationaux et européens, qu'en matière de participation aux jurys de thèses, aux comités de programme et aux comités de lecture. Le schéma de la Figure IV.3 offre une vue globale des projets scientifiques auxquels j'ai participé. Ceux pour lesquels je me suis fortement impliqué en termes de montage et pilotage local sont entourés en gras.

1.2.1 Montage et pilotage local de projets de recherche

Depuis septembre 2010 je me suis activement impliqué dans le montage et le pilotage local de 3 projets européens dans le cadre du Information Technology for European Advancement (ITEA)

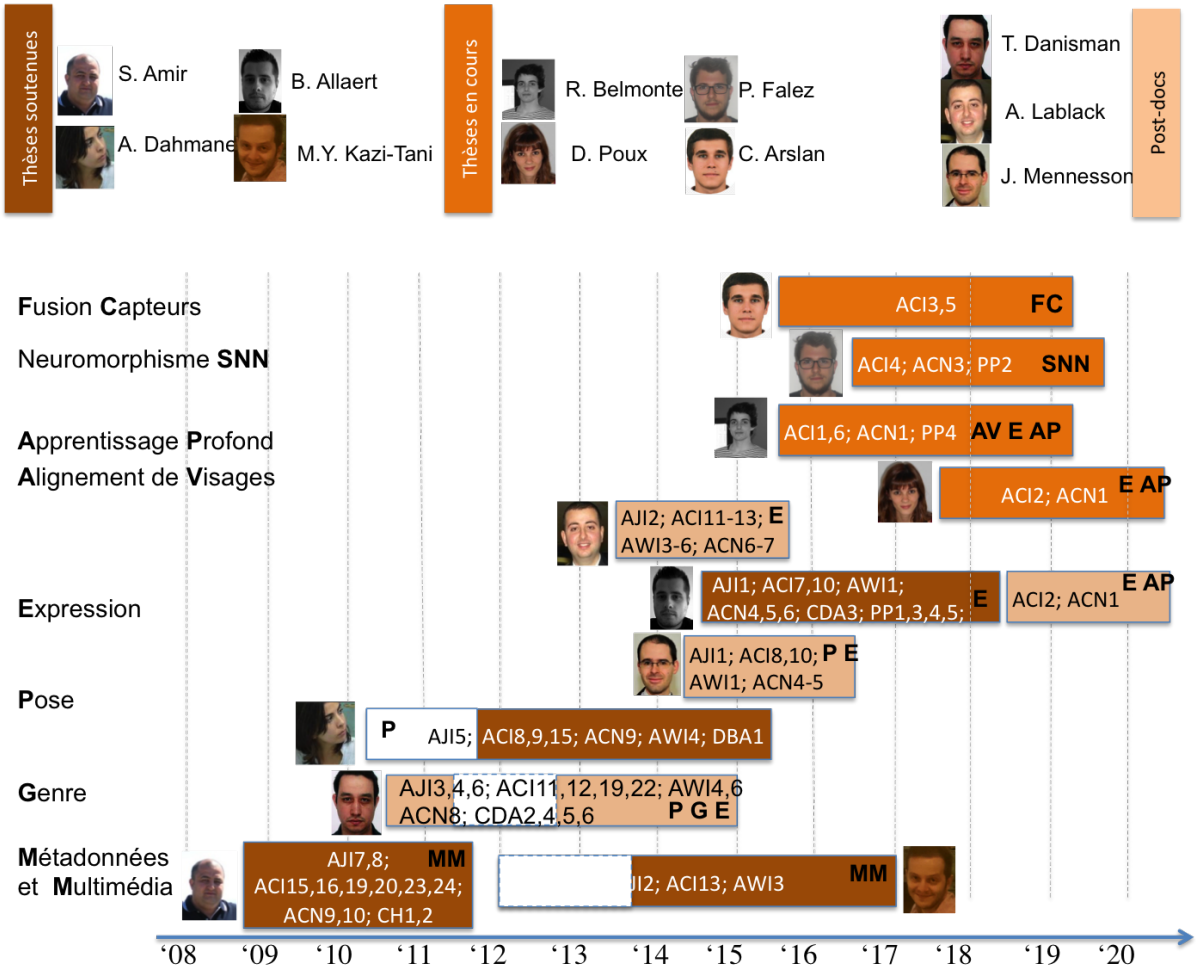


FIGURE IV.2 – *Répartition des publications par encadrement.*

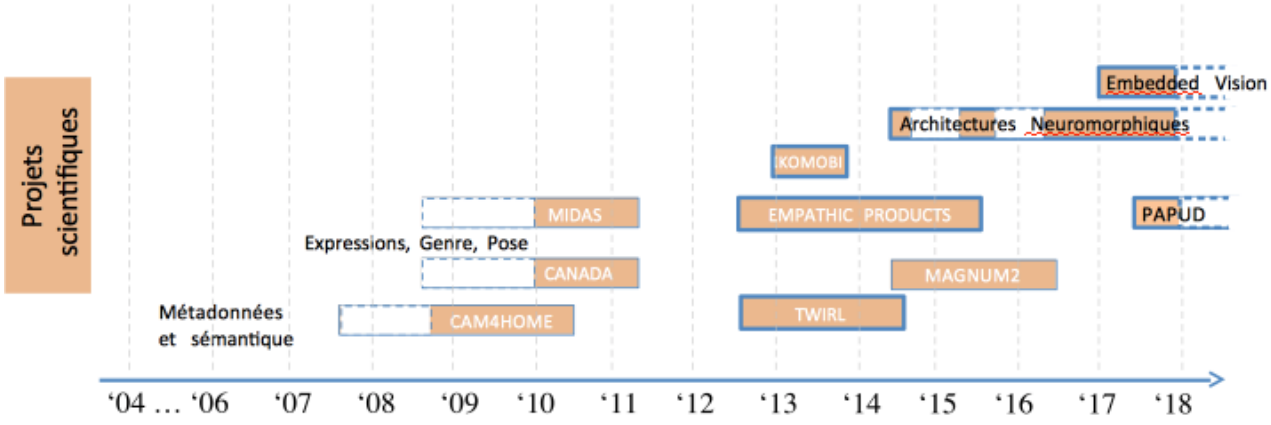


FIGURE IV.3 – *Vue d'ensemble des projets collaboratifs.*

d'EUREKA. Les projets en question sont : *Twinning virtual World (on-line) Information with Real world (off-Line) data sources* (TWIRL, 2012-2014, labellisé en 2010 et financé en 2012), *Empathic Products : enabling intentions and emotion aware products* (EMPATHIC, 2012-2015 - labellisé en 2011 et financé en 2012) et *Profiling and Analysis Platform Using Deep Learning* (PAPUD, 2017-2020 - labellisé et financé en 2017).

En plus de ces projets collaboratifs à l'échelle européenne, j'ai participé aux activités menées au sein de l'Institut de Recherche sur les Composants logiciels et matériels pour l'Information et

la Communication Avancée (IRCICA, USR CNRS 3380) qui accueille notre équipe depuis 2010. J'ai ainsi réalisé une étude scientifique pour une entreprise privée (IKOMOBI, 2013) et participé à deux projets inter-équipes. Le premier projet, *Architectures neuromorphiques pour la vision* traite des approches à base de neurones impulsionnels pour la vision. Le second, *Embedded Vision* étudie les compromis entre coûts énergétiques et précision de l'analyse dans le traitement vidéo embarqué.

TWIRL : Twinning virtual World (on-line) Information with Real world (off-Line) data sources
ITEA2 - 2012-2014, labellisé en 2010 - https://itea3.org/project/twirl.html
11 partenaires, 3 pays, budget global de 5 762 k€, coordonné par Cassidian Cybersecurity France, financement pour l'équipe FOX à hauteur de 246 k€
Ce projet européen a pour objectif de créer une plateforme capable de traiter, fouiller, établir des liens et fusionner des données en provenance d'applications en relation avec le monde réel (par exemple, estimations du trafic routier, prévisions météorologiques) avec des sources de données en ligne (par exemple, réseaux sociaux, forums, blogs, wikis, RSS). Notre ambition est de résoudre des problématiques liées à la quantité de données disponibles sur Internet, l'hétérogénéité des formats et types de données (par exemple, textes, commentaires, vidéos, images), et la difficulté d'agréger des données provenant de sources multiples. J'ai porté la participation de l'équipe dans ce projet dès ses prémices en février 2010. J'ai contribué aux échanges qui ont eu lieu lors des réunions de montage au niveau européen à Amsterdam (juin 2010) et au niveau français à Paris (août 2010). En octobre 2010, j'ai eu l'occasion de montrer l'impact positif de notre participation au projet auprès de la Direction Générale de la Compétitivité, de l'Industrie et des Services (DGCIS). Le projet a été labellisé par ITEA2 en décembre 2010. J'ai été sollicité pour monter une demande de financement et ai contribué à la définition de l'annexe technique en octobre 2011. J'ai eu la charge de définir le budget et d'organiser l'implication de l'équipe dans le projet. Dès le financement du projet, intervenu en février 2012, j'ai coordonné localement les activités de recherche et les recrutements jusqu'en février 2013. Pendant cette période, j'ai mis en place des outils de travail collaboratif (wiki) servant à faciliter les échanges et de capitaliser les avancées au sein du consortium. La tâche du pilotage a ensuite été transférée à mon collègue Jean Martinet (MCF HDR, équipe FOX, lab. CRIStAL). J'ai continué à participer au projet en travaillant dans le cadre du sous-lot 3,4 (Reconnaissance, Extraction, Catégorisation) sur l'extraction d'information portant sur les personnes présentes dans les collections d'images et de vidéos.
Mes publications en lien avec le projet TWIRL sont les suivantes : [AJI3][AJI4][ACI12][AWI2].

Empathic Products : Enabling intention- and emotion-aware products
ITEA2 - 2012-2015, labellisé en 2011 - https://itea3.org/project/empathic.html
29 partenaires, 8 pays, budget global de 15 285 k€, coordonné par VTT Technical Research Center, Finland, financement équipe FOX à hauteur de 276 k€
L'objectif de ce projet est d'améliorer l'expérience utilisateur en appliquant des "technologies empathiques" capables de comprendre et de répondre aux intentions et aux émotions des utilisateurs lors de l'utilisation des applications ou des systèmes déployés dans le monde réel. Nous nous efforçons à ce que les "applications empathiques" possèdent ce supplément de compréhension de leurs utilisateurs. Elles peuvent ainsi percevoir les émotions et les intentions de ces derniers.

Ces capacités leur permettent d'adapter en temps réel leur fonctionnement pour optimiser l'expérience utilisateur. J'ai coordonné le montage du projet pour le compte de l'équipe dès ses prémices en participant à l'événement « Project Outline Days » organisé à Paris en mars 2011. J'ai pu défendre les contributions de l'équipe dans le cadre du projet, auprès de la DGCIS au mois de novembre 2011. Le projet a été labellisé en décembre 2011. Le montage du dossier de demande d'aide et la répartition des coûts pour le compte de notre équipe ont été sous ma responsabilité. Depuis son financement, intervenu au mois de septembre 2012, j'ai coordonné localement le projet. Dans le cadre du projet, j'ai eu la responsabilité du sous-lot 3,3 (Modélisation des interactions/comportements utilisateur) et j'ai coordonné les contributions de l'équipe FOX au sein de plusieurs cycles : Empathic Video-Conferencing Systems, Empathic TV, Empathic Bilboard, EmoShop. En avril 2013, j'ai organisé une réunion de travail sur deux jours dans nos locaux à laquelle nous avons accueilli une trentaine de personnes. Dans ce projet, ma contribution porte sur l'extraction à partir de flux vidéo d'indices sur la manière dont les personnes perçoivent leur environnement, leur comportement individuel ou leur interaction avec d'autres personnes. Ce projet, organisé en cycles successifs, m'a permis de m'intéresser de manière graduelle aux verrous scientifiques sous différents angles et dans différents contextes applicatifs. Ainsi, tous les 3 mois divers contextes applicatifs tels que les vidéo-conférences, les systèmes d'e-learning ou les expériences multimédia interactives ont été abordés.
Mes publications en lien avec le projet EMPATHIC sont les suivantes : [AJI5][ACI10][ACI11][ACI14] [AWI1][AWI4][AWI5][AWI6][ACN4][ACN5][ACN6][ACN7]

PAPUD : Profiling and Analysis Platform Using Deep Learning
ITEA3 - 2018-2020, labellisé en 2017 - https://itea3.org/project/papud.html
23 partenaires, 6 pays, bugdet global 9 530 k€, coordonné par Atos Bull France, financement pour l'équipe FOX à hauteur de 216 k€
Le projet PAPUD a pour but de créer de nouveaux modèles et algorithmes dédiés à l'analyse de grandes masses de données textuelles et hétérogènes issues de différentes applications. Étant donné que les méthodes classiques d'apprentissage ont déjà montré leurs limitations, dans le cadre de ce projet, l'apprentissage profond est préconisé comme approche principale. L'apprentissage profond implique la prise en compte des problématiques liées au passage à l'échelle. Dans le cadre du projet, j'ai notamment participé à la définition d'un scénario qui concerne l'exploitation des traces d'usage des utilisateurs. Lors des interactions avec un système, une multitude de traces d'exécutions et de logs sont collectés. Souvent ces données sont exploitées dans la maintenance applicative ou dans la construction des profils utilisateurs basée sur l'historique de navigation. Les informations recueillies restent objectives et ne permettent pas de renseigner, par exemple, sur la perception subjective de l'expérience utilisateur, sur un produit ou un e-service. Je m'intéresse plus particulièrement à un processus d'inférence de mesures subjectives (état cognitif et affectif) à partir d'un ensemble de mesures objectives fournis par des caméras vidéo ou bio-capteurs. Pour cela, nous exploitons des corrélations entre mesures subjectives et objectives obtenues dans le cadre des expérimentations conduites dans un living lab instrumentalisé. Les traces objectives produites par les utilisateurs dans leur interactions, ainsi que les traces subjectives sont analysées conjointement afin d'identifier des corrélations qui sont exploitées pour inférer des traces subjectives lorsque,

dans un contexte réel, uniquement les traces objectives sont disponibles. Ces résultats permettent d'envisager de construire des profils cognitifs et affectifs des utilisateurs de manière transparente et sans aucune instrumentalisation spécifique. Le projet a été labellisé par ITEA en février 2017. Le consortium français a obtenu un financement sous la forme d'un projet FUI homonyme.
Mes publications en lien avec le projet PAPUD sont les suivantes : [ACI1][ACI2][ACN1].

Étude scientifique sur le suivi d'objets mobiles pour l'entreprise IKOMOBI
janv. 2013 - janv. 2014
en collaboration avec l'entreprise IKOMOBI, financement à hauteur de 36 k€
J'ai participé à la définition et la réalisation d'une étude scientifique autour du suivi d'objets mobiles dans un environnement indoor à partir de l'analyse de flux vidéo issus des caméras hémisphériques placées en hauteur. Suite à des réunions de concertation menées au mois de décembre 2012, le projet a démarré en janvier 2013. J'ai eu la charge d'assurer la communication avec IKOMOBI en organisant diverses réunions et en m'assurant de la bonne livraison de documents et prototypes. J'ai organisé et dirigé les activités de l'équipe dans ce projet en collaboration avec mes collègues Jean Martinet (MCF HDR, équipe FOX, lab. CRIStAL) et Pierre Tirilly (MCF, équipe FOX, lab. CRIStAL). Ma contribution scientifique s'est portée sur la mise en place d'un suivi de personnes en exploitant les mouvements enregistrés au sein de la scène, tout en tenant compte des situations d'occultation partielle ou totale et des changements de luminosité.

Architectures neuromorphiques pour la vision
IRCICA 2014-2018
en collaboration avec l'équipe Émeraude de CRIStAL et l'équipe CSAM de l'IEMN
L'objectif général de ce projet est de développer de nouveaux paradigmes de vision en adoptant un fonctionnement proche de celui du cerveau humain. Plus précisément, l'objectif de ce projet est d'étudier de nouvelles modalités de traitement de l'information visuelle, en privilégiant les paradigmes bio-inspirés en général et neuro-inspirés en particulier. Cela devrait offrir une efficacité énergétique proche de celle du cerveau et de permettre de gagner ainsi plusieurs ordres de grandeur. Les progrès récents en neurosciences et dans le domaine des dispositifs mémoire laissent entrevoir la possibilité de créer des architectures radicalement nouvelles qui bénéficient à des domaines comme la vision par ordinateur. L'utilisation des *spiking neural networks* (SNNs - réseaux de neurones impulsionnels) dotés de mécanismes d'apprentissage non-supervisés de type *spike timing dependent plasticity* (STDP) semble être une piste sérieuse. Dans le cadre de ce projet, j'ai participé à des travaux concernant l'apprentissage non-supervisé de motifs visuels à partir de collections d'images variées. Ces travaux ont été appliqués à la reconnaissance des chiffres manuscrits, ainsi qu'à la détection du visage.

Embedded Vision
IRCICA 2016-2017
en collaboration avec l'équipe Émeraude de CRIStAL et la plateforme Telecom de l'IEMN
Les algorithmes de reconnaissance d'images sont aujourd'hui utilisés dans les systèmes embarqués

temps-réel, comme les robots, les voitures autonomes, etc. Les traitements vidéo sont énergivores de par la nature des données traitées et de la chaîne complète de traitements (par exemple, captation, encodage et analyse). Des travaux pour optimiser les temps de traitement en distribuant l'analyse ou en faisant recours à des calculs sur GPU ou bien FPGA ont été explorés par la communauté. De grandes capacités de calcul sont nécessaires pour explorer l'espace complet de solutions. Peu d'études ont été menées pour mesurer l'impact sur les taux de reconnaissance des systèmes en présence d'une baisse de capacités de calcul ou du temps alloué. Dans le cadre de ce projet, nous nous intéressons notamment à la partie de reconnaissance en étudiant l'impact sur les performances d'une certaine maîtrise de capacités des calculs traduites par des temps processeurs et par des coûts énergétiques données. J'ai participé à la mise en place des protocoles expérimentaux pour mesurer les pertes en termes de performances lorsque l'on utilise des ressources maîtrisées. Cela permet d'anticiper les capacités du système à traiter certaines données en fonction des ressources disponibles.

1.2.2 Participations à d'autres projets collaboratifs

J'ai intégré l'équipe FOX en tant que post-doctorant dans le cadre du projet ITEA2 *Collaborative Aggregated Multimedia for Digital Home* (CAM4HOME, 2007-2010) où j'ai coordonné avec Chaabane Djeraba (PR, équipe FOX, lab. CRIStAL) les activités du groupe de travail autour de la modélisation sémantique des métadonnées. Ce projet a reçu la médaille d'argent lors du symposium ITEA-ARTEMIS organisé à Ghent, Belgique en 2010. J'ai participé également au projet ITEA2 *Multimodal Interfaces for Disabled and Ageing Society* (MIDAS, 2008-2011). Je suis intervenu avec Chaabane Djeraba (PR, équipe FOX, lab. CRIStAL) à la fois sur des aspects recherche autour de la reconnaissance des expressions [ACI22] et de l'endormissement [ACI19] ainsi que dans les développements en vue de préparer les démonstrateurs associés.

Je me suis fortement impliqué en tant que collaborateur, sous la coordination locale de l'équipe *Modeling and Analysis of Static and Dynamic Shapes* (3D-SAM) de CRIStAL, dans le montage d'un projet financé sur Fonds Unique Interministeriel (FUI) : *Mesure Analyse Gestion de flux Nativement Unifiée dans des Malls* (MAGNUM2, 2014-2016). Le projet vise la mise en place des interactions innovantes dans les magasins. Dans ce cadre, nous avons exploré l'étude d'expressions faciales à faible et forte intensité [ACI7].

En parallèle de cette riche activité dans un cadre collaboratif, je me suis également impliqué fortement dans la participation aux comités de programme et aux comités de lecture.

1.2.3 Participations aux jurys de thèse, comités de programme et aux comités de lecture

Jurys de thèse (2)

- Samin Mohammadi - Analysis of User Popularity Pattern and Engagement Prediction on Online Social Networks, sous la direction de Noël Crespi, le 4 décembre 2018 (matin) à Telecom Sud Paris, Evry - en tant qu'examinateur.
- Amir Mohammadinejad - Consensus Opinion Model in Online Social Networks based on

impact of Influential Users, sous la direction de Noël Crespi, le 4 décembre 2018 (après-midi) à Telecom Sud Paris, Evry - en tant qu'examinateur.

Comités d'organisation (2)

— Web chair et membre du comité d'organisation de la conférence internationale *Face and Gesture Recognition (IEEE)*, Lille, France, mai 2019.

— Membre du comité d'organisation de la conférence nationale *COmpression et REprésentation des Signaux Audiovisuels*, Lille, France, mai 2012.

Comités de programmes (5)

— Membre du comité de programme des conférences internationales *ACM Symposium on Applied Computing SWA*, 2010-2019.

— Membre du comité de programme de la conférence nationale *COmpression et REprésentation des Signaux Audiovisuels*, 2012.

— Membre du comité de programme de la conférence internationale *Grid and Pervasive Computing (GPC)*, 2011-2016.

— Membre du comité de programme de la *PECCS Special Session on Simulation and Interaction in Intelligent Environments (SIMIE)*, 2011.

— Membre du comité de programme de la conférence internationale *Geospatial Semantics*, 2011.

Comités de lecture (8)

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Transactions on Knowledge and Data Engineering* (IEEE Computer Society, depuis 2018).

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Journal of Electronic Imaging* (SPIE Digital Library, depuis 2018).

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Applied Computing and Informatics* (Elsevier, depuis 2016).

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Pattern Recognition* (Elsevier, depuis 2015).

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Neural Computing and Applications* (Springer, depuis 2015).

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Signal Image and Video Processing* (Springer, depuis 2014).

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Journal of Supercomputing* (Springer, depuis 2010).

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Multimedia Tools and Applications* (Springer, depuis 2010).

— Relecteur (reviewer) externe pour le comité de lecture de la revue *Knowledge and Information Systems* (Springer, depuis 2008).

1.3 Diffusion scientifique

1.3.1 Publications scientifiques

Ma production scientifique en termes de publications sur la période de septembre 2008 à août 2018 est synthétisée dans la Figure IV.4 et elle comporte :

— 8 articles de journaux internationaux avec comité de lecture ;
— 4 communications invitées ;
— 26 articles de conférences internationales avec comité de lecture et actes (dont une concernant mes activités de thèse) ;
— 6 workshops internationaux avec comité de lecture (dont 5 avec actes) ;
— 13 articles de conférences nationales avec actes ;
— 3 chapitres de livre (dont un concernant mes activités de thèse).

FIGURE IV.4 – *Synthèse de mes publications*

Parmi ces publications, je mets en avant ci-dessous, les 5 publications les plus représentatives liées à mes activités de recherche :

— T. Danisman ; I.M. Bilasco ; J. Martinet - Boosting gender recognition performance with a fuzzy inference system - Expert Systems with Applications, Volume 42, Issue 5, 1 April 2015, pp. 2772-2784. (Facteur d'impact : 3,768 selon Journal Citations Reports 2018) ;
— T. Danisman ; I.M. Bilasco ; J. Martinet ; C. Djeraba - Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron - Signal Processing, Elsevier, 2013, Special issue on Machine Learning in Intelligent Image Processing, 93 (6), pp. 1547-1556. (Facteur d'impact : 3,470 selon JCR 2017) ;
— B. Allaert ; J. Mennesson ; I.M. Bilasco ; C. Djeraba - Impact of the face registration techniques on facial expressions recognition - Signal Processing : Image Communication, EURASIP, Elsevier, 2017, 61, pp. 44-53. (Facteur d'impact : 2,073 selon JCR 2018) ;
— A. Dahmane ; S. Larabi ; I.M. Bilasco ; C. Djeraba - Head pose estimation based on face symmetry analysis - Signal, Image and Video Processing (SIViP), 2014, pp 1-10. (Facteur d'impact : 1,643 selon JCR 2018) ;
— T. Danisman ; I.M. Bilasco - In-plane face orientation estimation in still images - Multimedia Tools and Applications, Springer Verlag, 2016, 75 (13), pp.7799-7829. (Facteur d'impact : 1,541 selon JCR 2018).

Ces publications relatent les activités menées dans le domaine de la caractérisation de l'individu (genre, orientation de la tête) et de son comportement (expression faciale). Ces théma-

tiques se retrouvent également dans les projets européens ITEA2 TWIRL (2012-14), ITEA2 EMPATHIC (2012-15) et ITEA3 PAPUD (2017-2020) auxquels j'ai contribué activement. La liste intégrale des publications est disponible en Partie IV, Chapitre 2 et à l'adresse suivante : http://www.cristal.univ-lille.fr/~bilasco/single_publications.html

Par ailleurs, dans le cadre de mes activités de recherche, j'ai également participé à la conception de 3 corpus de données pour valider les différentes approches d'analyse faciale (estimation de pose et expression : *FoxFaces* et *SNaP-2DFE*, genre : *Web Gender Dataset*). De plus, j'ai contribué à la mise à disposition de nouvelles annotations concernant 4 corpus (*GENKI-4K*, *LFW Pro*, *Caltech Faces*, *Youtube Faces*) afin de pouvoir valider de manière exhaustive nos réalisations. Ces corpus complètent les corpus existants dans la communauté scientifique internationale. Une description plus détaillée de ces corpus est disponible dans la Partie IV, Chapitre 2.

1.3.2 Autres communications

Depuis mon arrivée à CRIStAL (anciennement LIFL), j'ai pu contribuer à de nombreuses activités de vulgarisation et dissémination. Dans ce qui suit, j'introduis ces activités par ordre chronologique. Mes premières activités de vulgarisation ont été réalisées, dès 2009, dans le cadre de la Plateforme Interactions, Réalité Virtuelle et Images (PIRVI, http://www.lifl.fr/PIRVI) de l'IRCICA. Mes contributions se sont organisées autour de la présentation de la salle «laboratoire magasin» hébergée par la plateforme et autour de la réalisation de démonstrations portant sur l'analyse des comportements humains (individu et groupe). Je suis notamment intervenu lors des événements organisés localement ou lors de visites d'extérieurs.

En octobre 2010 à Ghent (Belgique), j'ai participé aux activités de dissémination liées au projet ITEA2 CAM4HOME dans le cadre du *Joint ITEA2 - Artemis Symposium*, où le projet a été récompensé avec la médaille d'argent de l'*Achievement Award*. Dans le cadre de ce projet, j'ai coordonné la modélisation et la mise à disposition d'une plateforme générique de gestion et d'exploitation de métadonnées en vue d'adapter et d'enrichir le déploiement des contenus multimédia.

En mai 2011, j'ai été invité par Leif Hanlen (NICTA, Australia) pour présenter les travaux récents que nous menions sur la thématique Extraction of human behavior from video streams au workshop bilatéral EU-Australia sur la thématique e-Health organisé à Budapest, Hongrie.

Entre décembre 2011 et juin 2012, je me suis fortement impliqué dans la réalisation de l'œuvre interactive Tempo Scaduto [1] imaginée par Vincent Ciciliato, docteur en arts plastiques. Cette réalisation a été faite en collaboration avec l'équipe Méthodes et outils pour l'interaction à gestes (MINT) de CRIStAL et l'entreprise INSID Inc. Nous avons mis en place un dispositif visuel interactif immergeant un « joueur » dans la réalité sur fond de guerre des mafias à Palerme en Sicile dans les années 80. Le « joueur », en immersion, peut, grâce à une analyse fine des gestes de la main, suivre et tirer sur l'auteur « pressenti » d'un meurtre imminent. L'œuvre a été exposée en France et à l'étranger dans les manifestations suivantes :

— Panorama 14, Studio d'Art Le Fresnoy, Tourcoing (2012) ;
— Journée Recherche : Art, Recherche et Technologie à Polytech Lille (2013) ;
— Futur en Seine organisé par pôle de compétitivité Cap Digital à Paris (2013) ;

1. http://www.panorama14.net/142/tempo-scaduto-panorama-14 et http://vimeo.com/50522111

— Realidad Elastica, Laboral Centro de Arte y Creacion Industrial, Gijon, Asturias, Spain (2013).

En décembre 2012, j'ai participé à la présentation du projet ITEA2 TWIRL lors du *Joint ITEA2 - Artemis Symposium*. En décembre 2013, j'ai réitéré ma participation au *JointITEA2 - Artemis Symposium* (Stockholm, Suède) en disséminant les résultats du projet ITEA2 Empathic. En marge des démonstrations, j'ai pu échanger avec les visiteurs sur les solutions empathiques à base de vision que nous avons intégrées dans le projet (disponibles sur le portail du projet [2]). J'ai renouvelé cette expérience en 2015 lors du *Joint ITEA2 - Artemis Symposium* (Berlin, Allemagne) et du congrès *Innovate, Connect, Transform* - ICT 2015 (Lisbon, Portugal). J'ai également organisé une journée de démonstration fin 2014, à l'IRCICA, présentant les résultats du projet ITEA2 EMPATHIC à l'intention de Worldline (Atos). Je me suis également impliqué lors de l'inauguration du nouveau laboratoire CRIStAL (2 avril 2015 à Polytech Lille) dans la mise en place de démonstrations issues du projet européen ITEA2 EMPATHIC. Les démonstrations ont été vues par un large panel de personnes incluant également des personnes issues du monde politique et industriel de la Métropole Européenne de Lille (MEL). À l'occasion de l'inauguration de la Métropole Européenne de Lille (MEL) en 2015, les démonstrateurs issus du projet ont été également présentés aux collaborateurs de la MEL (le 16 janvier) et au grand public (le 17 janvier).

De 2012 à 2018, j'ai présenté et coordonné les divers démonstrateurs reflétant les travaux de l'équipe lors de la journée Recherche Innovation Création (RIC). Cette journée est organisée annuellement à l'attention des étudiants inscrits en L3, M1 et M2 dans les formations de l'Université de Lille, de l'Ecole Centrale de Lille et de l'Institut Mines Telecom Lille Douai. Elle vise à faciliter l'accès des étudiants aux équipes de recherche du laboratoire CRIStAL. Suite aux démonstrations faites, nous avons été sollicités pour plusieurs stages et projets de fin d'études (L3 et M2). Deux de ces stages se sont poursuivis par la thèse de Benjamin Allaert, qui a soutenu ses travaux, en juin 2018, et celle de Delphine Poux, qui a démarré en octobre 2017.

1.4 Responsabilités pédagogiques et administratives

Au-delà de mes activités de recherche, j'ai eu à coeur de m'impliquer fortement dans les activités pédagogiques et administratives inhérentes à la vie d'un enseignant-chercheur. Je détaille dans cette section les responsabilités pédagogiques prises dans le cadre du département Informatique de la Faculté de Sciences et Technologies de l'Université de Lille. Mes activités et responsabilités pédagogiques (détaillées par la suite) se structurent de la façon suivante :

— Direction d'études du Master Informatique parcours E-Services (Formation Initiale et Formation en Alternance) ;
— Coordination du suivi pédagogique des alternants ;
— Coordination du suivi de feuilles de présences.

Dans la suite, je présente ces activités dans l'ordre chronologique afin de mieux illustrer ma progression en termes de responsabilités au sein du département.

Coordination de suivi pédagogique des alternants Dès mon recrutement en tant que Maître de Conférences en septembre 2009, j'ai eu la charge de coordonner les suivis pédagogiques des alternants inscrits en contrat de professionnalisation en Licence MIAGE et en Master MIAGE (1ère

2. http://portal.empathic.eu/

et 2ème année) de l'Université de Lille. Au début, cette activité concernait une quarantaine d'étudiants chaque année. Dès 2011, j'ai eu en charge de coordonner le suivi de l'ensemble des alternants Master MIAGE et Master Informatique confondus (environ 90 étudiants par an). Les responsables d'études m'ont également confié la tâche d'organiser les soutenances des étudiants inscrits en Master 2 en alternance, tâche que j'ai assurée jusqu'en 2015 pour l'ensemble des parcours. Depuis, je continue à m'occuper des soutenances des étudiants inscrits en parcours E-Services en formation par alternance (entre 20 et 24 tous les ans).

Dans le département Informatique, le suivi d'un alternant se structure de la façon suivante :

— choisir un étudiant à suivre selon ses missions,
— prendre contact et faire connaissance avec l'étudiant,
— réaliser une première visite pédagogique pour s'assurer de la cohérence des missions annoncées avec le travail effectivement demandé à l'étudiant,
— définir ensemble la mission à présenter en soutenance,
— réaliser une seconde visite pédagogique,
— échanger de manière ponctuelle et informelle avec l'étudiant du déroulement de l'alternance.

Dès 2009, j'ai mis en place une application Web intitulée "Livret Électronique"[3] permettant de couvrir l'ensemble de ces besoins. Cette application est ouverte aux étudiants, qui peuvent en amont décrire l'essentiel de leurs missions aux tuteurs (universitaires) et aux responsables de formations. Les référents (en entreprise) reçoivent les comptes rendus consignés dans le livret électronique.

En 2011, j'ai étendu l'application afin qu'elle puisse intégrer également les stages de Licence, puis, les stages de Master 2 des étudiants en formation initiale. Actuellement, tous les ans, environ 350 stages et alternances sont complètement gérés par l'application. Au-delà de la mise à disposition de cette application, la coordination implique des tâches récurrentes pour informer et rappeler les consignes de suivi aux tuteurs, ainsi que des tâches de conseils et médiation en cas de conflits ou de difficultés dans le suivi. L'application évolue tous les ans pour intégrer de nouvelles fonctionnalités : historique du suivi sur plusieurs années, géolocalisation des bureaux entreprise accueillant les alternants, rappels automatiques, statistiques, génération d'ordres de missions prêts à signer, lien avec le site de candidatures.

Gestion de l'application des feuilles de présences Depuis 2011, l'application UniPres[4] assurant le suivi de présences des alternants a été conçue et déployée par mes soins. Les étudiants en alternance ont une obligation de présence auprès de leur employeur. Ils doivent être en mesure de fournir des feuilles de présence en fin de mois pour justifier de leur participation aux activités pédagogiques. Les responsables de formation, le directeur du département et moi-même avons eu de nombreux échanges avec le Service de la Formation Continue (entité de l'université de Lille gérant les relations avec les entreprises et les organismes de financement des formations) afin de faciliter la gestion et le suivi de ces feuilles. J'ai proposé une génération personnalisée de feuilles de présences pré-remplies en fonction du parcours spécifique de l'étudiant et des options choisies. Des discussions sont actuellement en cours pour étendre l'application au niveau de la Faculté de Sciences et Technologies dans son intégralité.

Direction d'études du Master Informatique parcours E-Services Le travail en étroite coopération avec Jean-Marie Lebbe et Yves Roos (directeurs d'études de la Licence et du Master MIAGE

3. http://stages.fil.univ-lille1.fr/suivi
4. http://stages.fil.univ-lille1.fr/unipres

en formation par alternance jusqu'en 2011) autour de la mise en place du livret électronique et l'organisation du suivi m'a permis d'acquérir les compétences et connaissances nécessaires pour assurer la direction d'une formation. Fort de mes expériences en termes de coordination du suivi pédagogique des stages et des alternances dans l'ensemble des parcours, j'ai été sollicité par Lionel Seinturier (le précédent directeur d'études du parcours E-Services) pour lui succéder. J'ai assuré la codirection du Master 2 pendant l'année 2012-2013 afin de préparer la transition et la définition des nouvelles maquettes. J'ai eu en charge de préparer le bilan de la maquette précédente et le montage de la maquette dans le cadre du parcours E-Services. Depuis septembre 2013, j'ai seul la responsabilité de la direction d'études tout en travaillant en étroite collaboration avec mes collègues responsables des principales UEs du parcours (Luigi Lancieri, Xavier Le Pallec, Lionel Seinturier et Jean-Claude Tarby).

Le parcours E-Services comporte deux groupes au fonctionnement différent. Les étudiants en formation initiale (FI) et les étudiants en formation par alternance (FA). Dans les faits, cela se traduit par une duplication des enseignements et une mise en adéquation des interventions pour le groupe FA. Depuis 2013, 247 étudiants ont obtenus leur diplôme de Master 2 Informatique parcours E-Services. Cela correspond à une moyenne de 42 étudiants par an.

1.5 Bilan

Ces dix dernières années, après l'obtention de ma thèse, ont été pour moi l'occasion de marier de manière équilibrée mes ambitions en termes de recherche et d'enseignement. La Figure IV.5 retrace les faits marquants de ma carrière en termes de parcours et activités pédagogiques, encadrement de thèses et projets scientifiques.

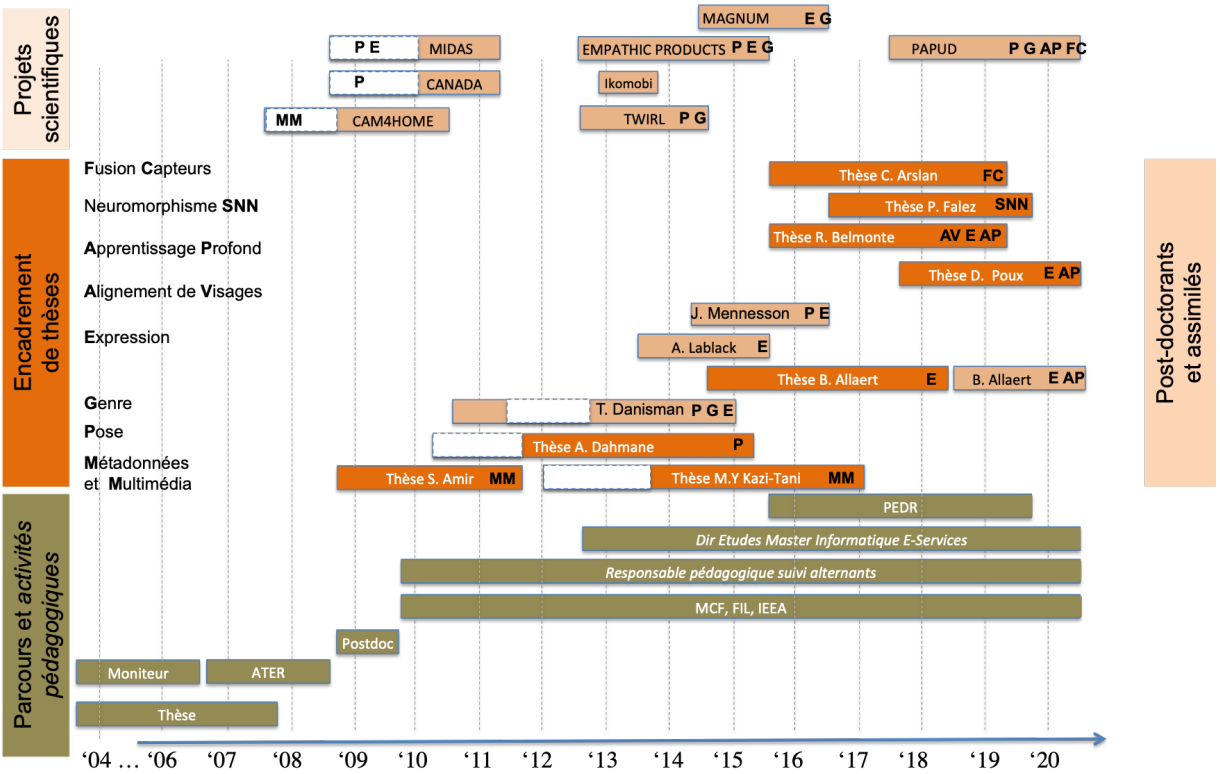


FIGURE IV.5 – *Vue d'ensemble de ma carrière professionnelle.*

Du côté de mes activités d'enseignement, j'ai pu monter en compétences graduellement en ce qui concerne les aspects administratifs et de gestion de l'enseignement. Ceci m'a permis de m'investir davantage dans les activités du département, tout en mettant en place de nouveaux enseignements pérennes dans le Master MIAGE et le parcours E-Services du Master Informatique.
Bien que je sois fortement impliqué dans les activités d'enseignement, j'ai pu continuer à m'investir dans mes activités de recherche en initiant et en portant plusieurs projets prioritaires pour l'équipe. Les nombreux encadrements valorisés avec succès dans des conférences et journaux reconnus par la communauté scientifique me font envisager de manière très positive la suite de ma carrière. L'obtention de mon HDR me permettra de donner une nouvelle dimension à ma carrière d'enseignant-chercheur.

Publications 2

Les publications qui retracent mes contributions sont regroupées en différentes catégories :
— articles de journaux internationaux avec comité de lecture (AJI)
— articles de journaux nationaux sans comité de lecture (AJN)
— articles ou communications invités (AI)
— articles de conférences internationales avec comité de lecture (ACI)
— articles de workshops internationaux avec comité de lecture (AWI)
— articles de conférences nationales avec actes et comité de lecture (ACN)
— chapitres de livres (CH)
— pre-prints (PP)
— corpus de données et annotations (CDA)

La numérotation reflète l'ordre chronologique inverse de la date de publication (de la plus récente à la plus ancienne dans chaque catégorie). L'ordre des auteurs reflète principalement le statut des personnes contribuant activement à chaque contribution (doctorant ou post-doctorant en premier, puis encadrants).

2.1 Articles de journaux internationaux avec comité de lecture - AJI (8)

[AJI1] B. Allaert ; J. Mennesson ; I.M. Bilasco ; C. Djeraba - Impact of the face registration techniques on facial expressions recognition - Signal Processing : Image Communication, EURASIP, Elsevier, 2017, 61, pp. 44-53 (Facteur d'impact : 2,073 selon Journal Citations Reports 2018).

[AJI2] M.Y. Kazi Tani ; A. Ghomari ; A. Lablack ; I.M. Bilasco - OVIS : Ontology video surveillance indexing and retrieval system - International Journal of Multimedia Information Retrieval, 2017, 6 (4), pp. 295 - 316 (SJR : 0,268 selon SCImago Journal Rank 2017).

[AJI3] T. Danisman ; I.M. Bilasco - In-plane face orientation estimation in still images - Multimedia Tools and Applications, Springer, 2016, 75 (13), pp.7799-7829. (Facteur d'impact : 1,541 selon JCR 2018, SJR : 0,287 selon SCImago Journal Rank 2017).

[AJI4] T. Danisman ; I.M. Bilasco ; J. Martinet - Boosting gender recognition performance with a fuzzy inference system - Expert Systems with Applications, Volume 42, Issue 5, 1 April 2015, pp. 2772-2784 (Facteur d'impact : 3,768 selon JCR 2018).

[AJI5] A. Dahmane ; S. Larabi ; I.M. Bilasco ; C. Djeraba - Head pose estimation based on face symmetry analysis - Signal, Image and Video Processing (SIViP), Springer, 2015, 9 (8), pp 1871-1880 (Facteur d'impact : 1,643 selon JCR 2018).

[AJI6] T. Danisman ; I.M. Bilasco ; J. Martinet ; C. Djeraba - Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron - Signal Processing, Elsevier, 2013, Special issue on Machine Learning in Intelligent Image Processing, 93 (6), pp. 1547-1556 (Facteur d'impact : 3,470 selon JCR 2018).

[AJI7] S. Amir ; I.M. Bilasco ; M. Rautiainen - CAM4Home : A generic ontology for a rich multimedia experience - International Journal of Computer Applications, 2013, 67 (12), pp. 19-25 (Facteur d'impact : 0,702 selon IJCA Citations Reports 2018).

[AJI8] S. Amir ; I.M. Bilasco ; C. Djeraba - MuMIe : Multi-level metadata mapping system - Journal of Multimedia, Academy Publisher, 2011, 6(3), pp. 225-235 (Facteur d'impact : 0,719 selon JCR 2013).

2.2 Articles de journaux nationaux sans comité de lecture - AJN (1)

[AJN1] I.M. Bilasco - La sémantique des scènes 3D : Une approche sémantique pour la recherche et la réutilisation de scènes 3D - Le monde des cartes - Revue du Comité Francais de Cartographie, 2008, 198, pp.31-35.

2.3 Articles ou communications invités - AI (4)

[AI1] I.M. Bilasco - Extracting human behaviour from video streams - EU-Australia Workshop on Bilateral Cooperation in e-Health, Mai 2011, Budapest, Hungary.

[AI2] I.M. Bilasco - Extracting human behaviour from video streams - Multitel Spring Workshop on Video Analysis, Juin 2010, Mons, Belgium.

[AI3] I.M. Bilasco - CAM4HOME metadata framework - FP7 NoTube 4th PCC Meeting, Décembre 2009, Munich, Allemagne.

[AI4] I.M. Bilasco - Metadata roles within CAM4HOME, ITEA2-CELTIC Joint Workshop, Juin 2009, Paris, France.

2.4 Articles de conférences internationales avec comité de lecture - ACI (26)

[ACI1] R. Belmonte ; P. Tirilly ; I.M. Bilasco ; C. Djeraba - Video-based face alignment with local motion modeling - Proc. of Winter Conference on Applications of Computer Vision (WACV), Jan. 2019, Hawaii.

[ACI2] D. Poux ; B. Allaert ; J. Mennesson ; N. Ihaddadene ; I.M. Bilasco ; C. Djeraba - Mastering occlusions by using intelligent facial frameworks based on the propagation of movement - Proc. of International Conference on Content-Based Multimedia Indexing (CBMI), Sept. 2018, La Rochelle, France (session meilleurs papiers).

[ACI3] C. Arslan ; I.M. Bilasco ; J. Martinet- Dynamic index finger gesture video dataset for mobile interaction - Proc. of International Conference on Content-Based Multimedia Indexing (CBMI), Sept. 2018, La Rochelle, France.

[ACI4] P. Falez ; P. Tirilly ; I. M. Bilasco ; Ph Devienne ; P. Boulet - Mastering the output frequency in spiking neural networks - Proc. of the International Joint Conference on Neural Networks (IJCNN), Jul. 2018, Rio de Janeiro, Brazil.

[ACI5] C. Arslan, F. Berthaut, J. Martinet, I.M. Bilasco, L. Grisoni - The Phone with the flow : combining touch + optical flow in mobile instruments - Proc. of New interfaces for musical expression (NIME), Juin 2018, Blacksburg, VA, United States.

[ACI6] R. Belmonte ; N. Ihaddadenne ; P. Tirrily ; I.M. Bilasco ; C. Djeraba - Towards spatio-temporal face alignment in unconstrained conditions - Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2018, Funchal, Portugal, pp. 433-438.

[ACI7] B. Allaert ; I.M. Bilasco ; C. Djeraba - Consistent optical flow maps for full and micro facial expression recognition - Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2017, Porto, Portugal, vol. 5, pp. 235-242.

[ACI8] J. Mennesson ; A. Dahmane ; T. Danisman ; I.M. Bilasco - Head yaw estimation using frontal face detector - Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2016, Portugal, vol. 4, pp. 517-524.

[ACI9] A. Aissaoui ; A. Dahmane ; J. Martinet ; I.M. Bilasco - Introducing FoxFaces : a 3-in-1 head dataset - Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2016, Porto, Portugal, vol. 4, pp. 533-537.

[ACI10] J. Mennesson ; B. Allaert ; I.M. Bilasco ; N. Van Der Aa ; A. Denis ; S. Cruz-Lara - Faces and thoughts : an empathic dairy - Proc. of IEEE International Conference on Automatic Face and Gesture Recognition 2015, Ljubljana, Slovenia, pp. 1-1 (présentation poster et démo).

[ACI11] A. Lablack ; T. Danisman ; I.M. Bilasco ; C. Djeraba - A local approach for negative emotion detection - Proc. of International Conference on Pattern Recognition 2014, Stockholm, Sweden, pp. 417-420 (présentation poster).

[ACI12] T. Danisman ; I.M. Bilasco ; C. Djeraba - Cross-database evaluation of normalized raw pixels for gender recognition under unconstrained settings - Proc. of International Conference on Pattern Recognition 2014, Stockholm, Sweden, pp. 3144-3149 (présentation poster).

[ACI13] M.Y. Kazi Tani ; A. Ghomari ; H. Belhadef ; A. Lablack ; I.M. Bilasco - An ontology-based approach for inferring multiple object events in surveillance domain - Proc. of Science and Information Conference 2014, London, United Kingdom, pp. 404-409.

[ACI14] O. Hadjerci ; A. Lablack ; I.M. Bilasco ; C. Djeraba - Affect recognition using magnitude models of motion - Proc. of MultiMedia Modelling 2014, Dublin, Ireland, LNCS 8362, pp. 339-344 (présentation poster).

[ACI15] A. Dahmane ; S. Larabi ; C. Djeraba ; I.M. Bilasco - Learning symmetrical model for head pose estimation - Proc. of 21st International Conference on Pattern Recognition, Nov 2012, Tsukuba, Japan. pp. 3614-3617 (présentation poster).

[ACI16] S. Amir ; Y. Benabbas ; I.M. Bilasco ; C. Djeraba - MuMIe : a new system for multimedia metadata interoperability - Proc. of 1st ACM International Conference on Multimedia Retrieval 2011, Trento, Italy (8 pages, http ://dl.acm.org/citation.cfm-id=1991997).

[ACI17] S. Amir ; I.M. Bilasco ; T. Danisman ; I. El Sayad ; C. Djeraba - Multimedia metadata mapping : towards helping developers in their integration task - Proc. of 8th International Conference on Advances in Mobile Computing and Multimedia (MoMM) 2010, Paris, Franc, pp. 205-212.

[ACI18] R. Auguste ; A. El Ghini ; I.M. Bilasco ; N. Ihaddadene ; C. Djeraba - Motion similarity measure between video sequences using multivariate time series modeling - Proc. Of International Conference on Machine and Web Intelligence (ICMWI), 2010, Algiers, Algeria, pp. 292-296.

[ACI19] T. Danisman ; I.M. Bilasco ; C. Djeraba ; N. Ihaddadene - Drowsy driver detection system using eye blink patterns - Proc. of International Conference on Machine and Web Intelligence (ICMWI) 2010, Algiers, Algeria, pp. 230-233.

[ACI20] S. Amir ; I.M. Bilasco ; T. Danisman ; T. Urruty ; I. El Sayad ; C. Djeraba - Schema matching for integrating multimedia metadata - Proc. of the International Conference on Machine and Web Intelligence (ICMWI) 2010, Algiers, Algeria, pp. 234-239.

[ACI21] S. Amir ; I.M. Bilasco ; T. Danisman ; T. Urruty ; C. Djeraba - Semi-automatic multimedia metadata integration - Proc. of EKAW 2010 Poster and Demo Track 2010, Lisboa, Portugal, ISSN 1613-0073, Vol 674 (présentation poster - http ://ceur-ws.org/Vol-674/Paper94.pdf).

[ACI22] T. Danisman ; I.M. Bilasco ; C. Djeraba ; N. Ihaddadene - Automatic facial feature detection for facial expression recognition - Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISAPP 2010, Angers, France, pp. 407-412.

[ACI23] H. Zhang ; H. Nguyen ; I.M. Bilasco : L.M. Gyu ; H. Wang - IPTV 2.0 from triple play to social TV - Proc. of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) 2010, Shanghai, China, pp.1-5.

[ACI24] I.M. Bilasco ; S. Amir ; P. Blandin ; C. Djeraba ; J. Laitakari ; J. Martinet ; E. Martinez Gracia ; D. Pakkala ; M. Rautiainen ; M. Ylianttila ; J. Zhou - Semantics for intelligent delivery of multimedia

content - Proc. of the International Symposium On Applied Computing 2010, Sierre, Suisse, pp. 1366-1372.

[ACI25] S. Amir ; I.M. Bilasco ; C. Djeraba - A semantic approach to metadata management in sensor systems - Proc. of Cognitive systems with interactive sensors (COGIS) 2009, Paris, France, pp. 112-119.

[ACI26] I.M. Bilasco ; R. Lozano Espinosa ; H. Martin - In-situ quantification of 3D scene complexity - Proc. of The International Conference on Advanced Geographic Information Systems and Web Services 2009, Cancun, Mexico, pp. 34-39.

2.5 Articles de workshops internationaux avec comité de lecture - AWI (6 dont 1 sans actes)

[AWI1] B. Allaert ; J. Mennesson ; I.M. Bilasco - EmoGame : Towards a self-rewarding methodology for capturing children Faces in an engaging context - Proc. of 7th International Workshop, HBU 2016 at ACMMM 2016 , Oct 2016, Amsterdam, Netherlands. 2016, pp 3-14.

[AWI2] T. Danisman ; J. Martinet ; I.M. Bilasco - Pruning near-duplicate images for mobile landmark identification : A graph theoretical approach - Proc. of 13th International Content-Based Multimedia Indexing, CBMI 2015, Prague, République Tchèque, pp 1-4.

[AWI3] M.Y. Kazi Tani ; A. Lablack ; G. Abdelghani ; I.M. Bilasco - Events detection using a video-surveillance Ontology and a rule-based approach - CONTACT Workshop in European Conference on Computer Vision 2014, Zurich, Suisse, pp. 299-308.

[AWI4] I.M. Bilasco ; A. Lablack ; A. Dahmane ; T. Danisman - Analysing user visual implicit feedback in enhanced TV scenarios - in Spontaneous Facial Behavior Analysis Workshop European Conference on Computer Vision 2014, Zurich, Suisse, pp. 315-324.

[AWI5] L. Ballihi ; A. Lablack ; B. Ben Amor ; I.M. Bilasco ; M. Daoudi - Positive/negative emotion detection from RGB-D upper body images - 1st International Workshop on Face and Facial Expression Recognition from Real World Videos (FFER) in International Conference on Pattern Recognition 2014, Stockholm, Sweden, pp. 109-120.

[AWI6] I.M. Bilasco ; A. Lablack ; T. Danisman - Data analysis of TV Shows viewers - 1st Workshop on Empathic Television Experiences (EmpaTeX 2014) in ACM International Conference on Interactive Experiences for Television and Online Video, June 2014, Newcastle-upon-Tyre, United Kingdom (présentation orale uniquement, sans actes).

2.6 Articles de conférences nationales avec actes et comité de lecture - ACN (13 dont 10 avec actes en ligne)

[ACN1] D. Poux ; B. Allaert ; N. Ihaddadene ; I.M. Bilasco ; C. Djeraba - Etude de l'apport de la reconstruction des régions occultées du visage pour la reconnaissance des expressions - COmpression et REprésentation des Signaux Audiovisuels (CORESA) 2018 (3 pages, actes en-ligne).

[ACN2] R. Belmonte ; N. Ihaddadene ; P. Tirilly ; I.M. Bilasco ; C. Djeraba - Vers un alignement spatio-temporel du visage en conditions non contrôlées - COmpression et REprésentation des Signaux Audiovisuels (CORESA) 2017 (6 pages, actes en-ligne).

[ACN3] P. Falez ; Ph. Devienne ; P. Tirilly ; I.M. Bilasco ; C. Loyez ; I. Sourikopoulos ; P. Boulet - Flexible simulation for neuromorphic circuit design : motion detection case study - Actes de Conférence d'informatique en Parallélisme, Architecture et Système (ComPAS), Juin 2017, Sophia Antipolis, France (6 pages, actes en-ligne).

[ACN4] J. Mennesson ; I.M. Bilasco ; B. Allaert - Fast head turns detection in low quality videos using optical flow - Actes de Reconnaissance des Formes et l'Intelligence Artificielle (RFIA), Jun 2016, Clermont-Ferrand, France (2 pages, actes en-ligne).

[ACN5] B. Allaert ; J. Mennesson ; I.M. Bilasco ; C. Djeraba - Etude de la dynamique du visage en situation d'interaction naturelle - Actes de COmpression et REprésentation des Signaux Audiovisuels (CORESA) 2016, May 2016, Nancy, France (6 pages, actes en-ligne) (meilleur papier - 2e place).

[ACN6] B. Allaert ; I.M. Bilasco ; A. Lablack - Vers une reconnaissance d'état affectif à base de mouvements du haut du corps et du visage - Colloque National Compression et Représentation des Signaux Audiovisuels (CORESA), Nov 2014, Reims, France (6 pages, actes en-ligne, présentation poster).

[ACN7] W. Adaidi ; A. Lablack ; I.M. Bilasco - Caractérisation locale des changements de texture pour la reconnaissance d'expressions faciales spontanées - Actes de Colloque National Compression et Représentation des Signaux Audiovisuels (CORESA), Nov 2014, Reims, France (6 pages, actes en-ligne, présentation poster).

[ACN8] T. Danisman ; I.M. Bilasco ; J. Martinet ; C. Djeraba - Construction de masques faciaux pour améliorer la reconnaissance d'expressions - Actes de COmpression et REpresentation des Signaux Audiovisuels, May 2012, Lille, France (6 pages, actes en-ligne, présentation poster).

[ACN9] A. Dahmane ; S. Larabi ; I.M. Bilasco ; C. Djeraba - Estimation discrète de l'angle pan de la tête - Actes de COmpression et REprésentation des Signaux Audiovisuels (CORESA), May 2012, Lille, France (6 pages, actes en-ligne, présentation poster).

[ACN10] S. Amir ; I.M. Bilasco ; T. Urruty ; C. Djeraba - Vers une interopérabilité multi-niveaux des métadonnées - Actes INFORSID 2011, May 2011, Lille, France (16 pages, actes en-ligne).

[ACN11] S. Amir ; I.M. Bilasco : T. Urruty ; C. Djeraba - MuMIE : Une approche automatique pour l'interopérabilité des métadonnées - Actes Extraction et Gestion des Connaissances (EGC) 2011, Jan 2011, Brest, France, pp. 347-352.

[ACN12] R. Auguste ; A. El Ghini ; I.M. Bilasco ; C. Djeraba - Prédiction de séries temporelles et applications à l'analyse de séquences vidéos - Actes Extraction et Gestion des Connaissances (EGC) 2011, Jan 2011, Brest, France, pp. 713-714.

[ACN13] Md.H. Sharif ; H. Alustwani ; I.M. Bilasco ; C. Djeraba - Détection des mouvements anormaux dans des vidéos - Actes Extraction et Gestion des Connaissances (EGC) 2011, Jan 2011, Brest, France, pp. 699-700.

2.7 Chapitres de livre - CH (3)

[CH1] S. Amir ; I.M. Bilasco ; Md.H. Sharif ; C. Djeraba - Towards a unified multimedia metadata management solution - Intelligent Multimedia Databases and Information Retrieval : Advancing Applications and Technologies (Ma, Zongmin), IGI Global, 2012, pp. 170-194.

[CH2] S. Amir ; I.M. Bilasco ; T. Urruty ; J. Martinet ; C. Djeraba - Designing intelligent content delivery frameworks using MPEG-21 - The Handbook of MPEG Applications : Standards in Practice (Agius, Harry and Angelides, Marios), John Wiley and Sons Ltd, 2010, ch. 19, pp. 455-476.

[CH3] I.M. Bilasco ; J. Gensel ; H. Martin ; M. Villanova-Oliver - Indexing three dimensional scenes. - Encyclopedia of Multimedia, Springer, 2008, pp.346-352.

2.8 Pre-prints (5)

[PP1] B. Allaert ; I.M. Bilasco ; C. Djeraba - Advanced local motion patterns for macro and micro facial expression recognition - Computing Research Repository (CoRR), https://arxiv.org/abs/1805.01951.

[PP2] P. Falez ; P. Tirilly ; I.M. Bilasco ; Ph. Devienne ; P. Boulet - Unsupervised visual feature learning with STDP : how far are we from traditional feature learning approaches ?

[PP3] B. Allaert ; I.M. Bilasco ; C. Djeraba ; Z. Zhang - Fully-connected neural networks and local motion patterns for micro and macro facial expression recognition.

[PP4] R. Belmonte ; B. Allaert ; P. Tirilly ; I.M. Bilasco - Study of the impact of face alignment on subsequent face analysis tasks in presence of head movements.

[PP5] B. Allaert ; I.M. Bilasco - Towards an adaptation of optical flow methods for facial expression analysis.

2.9 Corpus de données et annotations - CDA (6)

Corpus de données (3)

[CDA1] **Fox Faces Dataset** Cette collection contient des images de visages capturées en laboratoire avec différentes poses de la tête, conditions d'illumination, expressions faciales, modalités d'acquisition (caméra durée-de-vol ou time-of-flight, caméra stéréoscopique et Kinect). La collection contient 2624 images et 64 personnes distinctes. La mise à disposition d'une capture synchrone entre différentes modalités d'acquisition permet d'identifier la manière dont les modalités considérées permettent de répondre au mieux aux défis posés dans le cadre de la capture (pose, expressions, illumination). Ces corpus ont été présentés dans [ACI9].

[CDA2] **Web Gender Dataset** Cette collection contient des images de visages masculins et féminins. Les images ont été obtenues en interrogeant les moteurs de recherche Web avec des termes liés à chaque sexe dans différentes langues (Français, Anglais, Allemand, Chinois, Turc...), et filtrées par un détecteur de visages (Viola-Jones), puis manuellement. La collection contient 4700 images. Nous avons conçu ce corpus afin d'éprouver la capacité de généricité de classifieurs appris notamment dans un contexte de validation inter-corpus. La variété des conditions de captures et des sujets nous a permis de converger vers des classifieurs robustes. Ce corpus a été utilisé dans les expérimentations inter-corpus concernant le détection de genre rapportées dans [ACI12] et [AJI4].

[CDA3] **Synchronous Natural and Posed 2D Facial Expressions (SNaP-2DFE)** Cette collection contient des séquences d'images enregistrées simultanément par deux caméras reflétant des conditions d'usage distinctes. La première caméra est fixée sur un casque qui suit les mouvements naturels de la tête. La seconde caméra est fixée en face de la personne. La caméra attachée au casque enregistre de manière continuelle le visage en pose frontale, alors que la caméra qui est de face enregistre des images comportant des variations de la tête et de larges déplacements. La collection SNaP-2DFe permet de mesurer l'impact des mouvements de la tête sur la reconnaissance d'expressions faciales. La collection contient 1260 séquences enregistrées à l'aide de 15 volontaires effectuant simultanément des expressions et des mouvements de la tête. Chaque volontaire a été enregistré en suivant six types de mouvements (3 rotations, 2 translations et une pose fixe) en reproduisant pour chaque mouvement sept expressions (colère, dégoût, énervement, joie, neutre, peur, surprise, tristesse). Ce corpus a été utilisé dans [AJI1] afin d'étudier l'impact des méthodes de normalisation de visage sur la reconnaissance d'expression en comparant les performances mesurées sur la caméra frontale, par rapport à la caméra attachée au casque. Il a été également employé dans [ACI1] et [PP4] pour étudier l'impact des changements d'orientation de la tête sur les méthodes d'alignement de visages.

Corpus d'annotations (3)

[CDA4] **Eye Center Annotations** Ces données fournissent des annotations manuelles des positions du centre de l'oeil dans les collections d'images de visages Caltech Faces (Weber 1999) et

Youtube Faces (Wolf et al. 2011). Le nombre d'images annotées dans les deux collections est de 450, respectivement 5000. Ces annotations ont servi à estimer la pose de la tête, et notamment le roulis, dans le cadre des travaux publiés dans [AJ3]. Nous avons souhaité montrer la généricité de notre approche sur un nombre important de corpus d'images.

[CDA5] **Gender LFW Pro Annotations** Ces données fournissent des annotations manuelles pour catégoriser la collection d'images de visages LFW Pro (Huang et al. 2007) selon le genre des sujets. Les annotations concernent 7895 images. Ces annotations servent à disposer d'une large palette d'images qui valide les approches de caractérisation du genre dans un contexte inter-corpus proposées dans [ACI12] et [AJI4].

[CDA6] **GENKI4K Gender-Emotion Annotations** Ces données fournissent des annotations manuelles pour catégoriser la collection d'images de visages GENKI-4K (MPLab 2011) selon le genre et les expressions faciales des sujets. Les annotations concernent 4000 images du corpus. Comme dans le cas du corpus d'annotations Gender LFW Pro, nous avons souhaité mettre à disposition de la communauté de nouvelles données pour la détection de genre. Disposer des corpus riches et variés permet de mettre en œuvre des processus d'apprentissage ayant un degré important de généricité. Ces annotations ont été utilisées dans [ACI12] et [AJI4] pour valider la reconnaissance de genre dans un contexte inter-corpus.

Plus de détails sur ces corpus de données et corpus d'annotations sont disponibles à l'adresse suivante : http://www.cristal.univ-lille.fr/FOX/index.php?page=donnees

Bibliographie

Allaert, Benjamin, Bilasco, Ioan Marius, et Djeraba, Chaabane. Consistent optical flow maps for full and micro facial expression recognition. Dans *Proc. Int'l Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pages 235–242, 2017. (Cité pages 16, 73, 79, 105 et 107.)

Allaert, Benjamin, Bilasco, Ioan Marius, et Djeraba, Chaabane. Advanced local motion patterns for macro and micro facial expression recognition. *CoRR*, abs/1805.01951, 2018a. URL http://arxiv.org/abs/1805.01951. (Cité pages 105 et 107.)

Allaert, Benjamin, Mennesson, José, Bilasco, Ioan Marius, et Djeraba, Chaabane. Impact of the face registration techniques on facial expressions recognition. *Signal Processing : Image Communications*, 61, p. 44–53, 2018b. (Cité pages 16, 88 et 117.)

Amir, Samir, Bilasco, Ioan Marius, et Djeraba, Chaabane. MuMIe : multi-level metadata mapping system. *J. of Multimedia*, 6(3), p. 225–235, 2011. (Cité page 17.)

Amir, Samir, Bilasco, Ioan Marius, et Rautiainen, Mika. CAM4Home : A generic ontology for a rich multimedia experience. *Int'l J. of Computer Applications*, 67(12), p. 19–25, 2013. (Cité page 17.)

An, Kwang Ho et Chung, Myung Jin. 3D head tracking and pose-robust 2D texture map-based face recognition using a simple ellipsoid model. Dans *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, pages 307–312, 2008. (Cité page 45.)

Asteriadis, Stylianos, Karpouzis, Kostas, et Kollias, Stefanos. Head pose estimation with one camera, in uncalibrated environments. Dans *Proc. Workshop on Eye Gaze in Intelligent Human Machine Interaction (EGIHMI)*, pages 55–62, New York, NY, USA, 2010. ACM. (Cité pages 26 et 45.)

Baccouche, Moez, Mamalet, Franck, Wolf, Christian, Garcia, Christophe, et Baskurt, Atilla. Sequential deep learning for human action recognition. Dans *Proc. Int'l Workshop on Human Behavior Understanding*, Salah, Albert Ali et Lepri, Bruno (éditeurs), pages 29–39, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25446-8. (Cité page 117.)

Balasubramanian, Vineeth Nallure, Jiepinge, Ye, et Panchanathan, Sethuraman. Biased manifold embedding : A framework for person-independent head pose estimation. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007. (Cité page 39.)

Bassili, John N. Emotion recognition : the role of facial movement and the relative importance of upper and lower areas of the face. *J. of Personality and Social Psychology*, 37(11), p. 2049–2058, 1979. (Cité page 73.)

Bekios-Calfa, Juan, Buenaposada, José M, et Baumela, Luis. Revisiting linear discriminant techniques in gender recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 33(4), p. 858–864, 2011. (Cité page 51.)

Bekios-Calfa, Juan, Buenaposada, José M, et Baumela, Luis. Robust gender recognition by exploiting facial attributes dependencies. *Pattern Recognition Letters*, 36, p. 228–234, 2014. (Cité pages 52, 59 et 69.)

Belmonte, Romain, Ihaddadene, Nacim, Tirilly, Pierre, Djeraba, Chaabane, et Bilasco, Ioan Marius. Video-based face alignement with local motion modeling. Dans *Proc. Int'l Winter Conf. on Applications of Computer Vision*, 2019. (Cité page 17.)

Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhya : The Indian J. of Statistics (1933-1960)*, 7(4), p. 401–406, 1946. (Cité page 94.)

Black, A. John Jr., Gargesha, Madhusudhana, Kahol, Kanav, Kuchi, Prem, et Panchanathan, Sethuraman. A framework for performance evaluation of face recognition algorithms. Dans *ITCOM, Internet Multimedia Systems II*, Boston, 2002. (Cité pages 34 et 35.)

Breuer, Ran et Kimmel, Ron. A deep learning perspective on the origin of facial expressions. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. (Cité pages 106, 107 et 108.)

Carminati, Lionel, Benois-Pineau, Jenny, et Gelgon, Marc. Human detection and tracking for video surveillance applications in a low-density environment. Dans *Proc. Int'l. Conf. Visual Communications and Image Processing*, volume 5150, pages 51–56, 2003. (Cité page 56.)

Chanti, Dawood Al et Caplier, Alice. Improving bag-of-visual-words towards effective facial expressive image classification. Dans *Proc. Int'l Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 145–152, 2018. (Cité page 75.)

Chen, Mei-Yen et Chen, Chien-Chung. The contribution of the upper and lower face in happy and sad facial expression classification. *Vision Research*, 50(18), p. 1814–1823, 2010. (Cité page 81.)

Chen, Weilong, Er, Meng Joo, et Wu, Shiqian. Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *IEEE Trans. Systems, Man, and Cybernetics*, 36, p. 458–466, 2006. (Cité page 37.)

Coltuc, Dinu, Bolon, Philippe, et Chassery, Jean-Marc. Exact histogram specification. *IEEE Trans. Image Processing*, 15(5), p. 1143–1152, May 2006. ISSN 1057-7149. (Cité page 50.)

Cootes, Timothy F. Manchester talking face video dataset, 2004. (Cité pages 42 et 43.)

Dago-Casas, Pablo, Gonzalez-Jimenez, Daniel, Yu, Long Long, et Alba-Castro, José Luis. Single and cross-database benchmarks for gender classification under unconstrained settings. Dans *Proc. IEEE Int'l Conf. on Computer Vision Workshops (ICCVW)*, pages 2152–2159, 2011. (Cité pages 51, 52, 59 et 69.)

Dahmane, Afifa, Larabi, Slimane, Bilasco, Ioan Marius, et Djeraba, Chaabane. Head pose estimation based on face symmetry analysis. *Signal, Image and Video Processing (SIViP)*, 9(8), p. 1871–1880, 2015. (Cité pages 15, 24 et 38.)

Danisman, Taner et Bilasco, Ioan Marius. In-plane face orientation estimation in still images. *Multimedia Tools and Applications (MTAP)*, 75(13), p. 7799–7829, 2016. (Cité page 15.)

Danisman, Taner, Bilasco, Ioan Marius, et Djeraba, Chaabane. Cross-database evaluation of normalized raw pixels for gender recognition under unconstrained settings. Dans *Proc. Int'l Conf. on Pattern Recognition (ICPR)*, pages 3144–3149, Stockholm, Sweden, 2014. (Cité pages 15, 50 et 56.)

Danisman, Taner, Bilasco, Ioan Marius, Ihaddadene, Nacim, et Djeraba, Chaabane. Automatic facial feature detection for facial expression recognition. Dans *Proc. Int'l Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Richard, Paul et Braz, José (éditeurs), pages 407–412. INSTICC Press, 2010. (Cité page 25.)

Danisman, Taner, Bilasco, Ioan Marius, et Martinet, Jean. Boosting gender recognition performance with a fuzzy inference system. *Expert Systems with Applications*, 42(5), p. 2772–2784, 2015. (Cité pages 15, 25, 42, 50 et 68.)

Danisman, Taner, Bilasco, Ioan Marius, Martinet, Jean, et Djeraba, Chaabane. Intelligent pixels of interest selection with application to facial expression recognition using multilayer perceptron. *Signal Processing*, 93(6), p. 1547–1556, 2013. (Cité pages 15, 72 et 87.)

Demirkus, Meltem, Clark, James J., et Arbel, Tal. Robust semi-automatic head pose labeling for real-world face video sequences. *Multimedia Tools and Applications (MTAP)*, pages 1–29, 2013. (Cité page 24.)

Deravi, Farzin et Guness, Shivanand P. Gaze trajectory as a biometric modality. Dans *Proc. Int'l Conf. on Bio-Inspired Systems and Signal Processing (BIOSIGNALS)*, 2011. (Cité page 23.)

Ding, Hui, Zhou, Shaohua Kevin, et Chellappa, Rama. FaceNet2ExpNet : Regularizing a deep face recognition net for expression recognition. Dans *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 118–126, 2017. (Cité pages 107 et 108.)

Ekman, Paul et Friesen, Wallace. *Facial Action Coding System : Manual*. Numéro v. 1-2. Consulting Psychologists Press, 1978. (Cité page 89.)

Ekman, Paul et Rosenberg, Erika. *What the face reveals : Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997. (Cité page 72.)

Elaiwat, Said, Bennamoun, Mohammed, et Boussaid, Farid. A spatio-temporal RBM-based model for facial expression recognition. *Pattern Recognition*, 49, p. 152–161, 2016. (Cité page 107.)

Etemad, Kamran et Chellappa, Rama. Discriminant analysis for recognition of human face images. Dans *Proc. Int'l Conf. Audio- and Video-based Biometric Person Authentication*, Bigun, Josef, Chollet, Gerard, et Borgefors, Gunilla (éditeurs), volume 1206 de *Lecture Notes in Computer Science*, pages 125–142. Springer Berlin / Heidelberg, 1997. (Cité page 78.)

Falez, Pierre, Devienne, Philippe, Tirilly, Pierre, Bilasco, Marius, Loyez, Christophe, Sourikopoulos, Ilias, et Boulet, Pierre. Flexible simulation for neuromorphic circuit design : Motion detection case study. Dans *Conférence d'informatique en Parallélisme, Architecture et Système (ComPAS)*, Sophia Antipolis, France, 2017. (Cité page 17.)

Falez, Pierre, Tirilly, Pierre, Bilasco, Ioan Marius, Devienne, Philippe, et Boulet, Pierre. Mastering the output frequency in spiking neural networks. Dans *Proc. Int'l Joint Conf. on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, 2018. (Cité page 17.)

Fan, Xijian et Tjahjadi, Tardi. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognition*, 48(11), p. 3407–3416, 2015. (Cité page 107.)

Fan, Xijian et Tjahjadi, Tardi. A dynamic framework based on local Zernike moment and motion history image for facial expression recognition. *Pattern Recognition*, 64, p. 399–406, 2017. (Cité pages 77 et 107.)

Farnebäck, Gunnar. Two-frame motion estimation based on polynomial expansion. Dans *Image Analysis*, Bigun, Josef et Gustavsson, Tomas (éditeurs), volume 2749 de *Lecture Notes in Computer Science*, pages 363–370. Springer Berlin Heidelberg, 2003. (Cité page 100.)

Fortun, Denis, Bouthemy, Patrick, et Kervrann, Charles. Optical flow modeling and computation : a survey. *Computer Vision and Image Understanding*, 134, p. 1–21, 2015. (Cité page 75.)

Gallagher, Andrew et Chen, Tsuhan. Understanding images of groups of people. *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 256–263, 2009. (Cité page 49.)

Ghimire, Deepak et Lee, Joonwhoan. Geometric feature-based facial expression recognition in image sequences using multi-class Adaboost and support vector machines. *Sensors*, 13(6), p. 7714–7734, 2013. (Cité page 79.)

Golomb, Beatrice, Lawrence, David T., et Sejnowski, Terrence. SexNet : a neural network identifies sex from human faces. Dans *Proc. Int'l Conf. on Advances in Neural Information Processing Systems*, NIPS-3, pages 572–577, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. (Cité page 51.)

Gonzalez, Isabel, Sahli, Hichem, Enescu, Valentin, et Verhelst, Werner. Context-independent facial action unit recognition using shape and gabor phase information. Dans *Proc. Int'l Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2011. (Cité page 73.)

Graf, Arnulf B.A. et Borer, Silvio. Normalization in support vector machines. Dans *Proc. DAGM Symposium*, Radig, Bernd et Florczyk, Stefan (éditeurs), volume 2191 de *Lecture Notes in Computer Science*, pages 277–282. Springer, 2001. (Cité page 55.)

Gruendig, Martin et Hellwich, Olaf. 3D head pose estimation with symmetry based illumination model in low resolution video. Dans *Lecture Notes in Computer Science*, volume 3175, pages 45–53. Springer, 2004. (Cité page 26.)

Guan, Naiyang, Tao, Dacheng, Luo, Zhigang, et Yuan, Bo. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE Trans. Image Processing*, 20(7), p. 2030–2048, 2011. (Cité page 78.)

Guan, Naiyang, Tao, Dacheng, Luo, Zhigang, et Yuan, Bo. NeNMF : An optimal gradient method for nonnegative matrix factorization. *IEEE Trans. Signal Processing*, 60(6), p. 2882–2898, 2012a. (Cité page 78.)

Guan, Naiyang, Tao, Dacheng, Luo, Zhigang, et Yuan, Bo. Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Trans. Neural Networks and Learning Systems*, 23(7), p. 1087–1099, 2012b. (Cité page 78.)

Gui, Zhenghui et Zhang, Chao. 3D head pose estimation using non-rigid structure-from-motion and point correspondence. *Proc. IEEE Region Annual Int'l Conf. TENCON*, pages 1–3, 2006. (Cité page 27.)

Guillaume, Serge. Designing fuzzy inference systems from data : An interpretability-oriented review. *IEEE Trans. Fuzzy Systems*, 9(3), p. 426–443, 2001. (Cité page 53.)

Guo, Guodong, Dyer, Charles R., Fu, Yun, et Huang, Thomas S. Is gender recognition affected by age ? Dans *Proc. IEEE Int'l Conf. on Computer Vision Workshops (ICCVW)*, pages 2032–2039, 2009. (Cité page 57.)

Guo, Weiwei, Kotsia, Irene, et Patras, Ioannis. Higher order support tensor regression for head pose estimation. Dans *Proc. Int'l Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2011. (Cité pages 26 et 45.)

Gutta, Srinivas, Huang, Jeffrey R.J., Philips, Jonathon P., et Wechsler, Harry. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE Trans. Neural Networks*, 11 (4), p. 948–960, 2000. (Cité page 51.)

Hadid, Abdenour et Pietikainen, Matti. Combining appearance and motion for face and gender recognition from videos. *Pattern Recognition*, 42(11), p. 2818–2827, 2009. (Cité page 72.)

Han, Shizhong, Meng, Zibo, Liu, Ping, et Tong, Yan. Facial grid transformation : A novel face registration approach for improving facial action unit recognition. Dans *Proc. Int'l Conf. on Image Processing (ICIP)*, pages 1415–1419, 2014. (Cité page 75.)

Hao, Ji, Risheng, Liu, Fei, Su, Zhixun, Su, et Yan, Tian. Robust head pose estimation via convex regularizd sparse regression. *Proc. Int'l Conf. on Image Processing (ICIP)*, 2011. (Cité page 39.)

Happy, S.L. et Routray, Aurobinda. Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affective Computing (TAC)*, 6(1), p. 1–12, 2015. (Cité page 77.)

Harguess, Josh, Gupta, Shalini, et Aggarwal, J.K. 3D face recognition with the average-half-face. Dans *Proc. Int'l Conf. on Pattern Recognition (ICPR)*, pages 1–4, 2008. (Cité page 26.)

Hattori, Kazuyuki, Matsumori, Shinichi, et Sato, Yukio. Estimating pose of human face based on symmetry plane using range and intensity images. Dans *Proc. Int'l Conf. on Pattern Recognition (ICPR)*, volume 2, pages 1183–1187 vol.2, 1998. (Cité page 27.)

Holmes, Geoffrey, Pfahringer, Bernhard, Kirkby, Richard, Frank, Eibe, et Hall, Mark. Multiclass alternating decision trees. Dans *Proc. European Conf. on Machine Learning (ECML)*, pages 161–172. Springer, 2001. (Cité page 33.)

Huang, Gary B., Ramesh, Manu, Berg, Tamara, et Learned-Miller, Erik. Labeled Faces in the Wild : A database for studying face recognition in unconstrained environments. Rapport Technique 07-49, University of Massachusetts, Amherst, 2007. (Cité pages 49 et 149.)

Huang, Xiaohua, Wang, Sujing, Liu, Xin, Zhao, Guoying, Feng, Xiaoyi, et Pietikainen, Matti. Spontaneous facial micro-expression recognition using discriminative spatiotemporal local binary pattern with an improved integral projection. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016a. (Cité pages 76, 79 et 106.)

Huang, Xiaohua, Zhao, Guoying, Hong, Xiaopeng, Zheng, Wenming, et Pietikäinen, Matti. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing*, 175, p. 564–578, 2016b. (Cité pages 76 et 105.)

Hyvärinen, Aapo et Oja, Erkki. Independent component analysis : algorithms and applications. *Neural Networks*, 13(4-5), p. 411–430, 2000. (Cité page 78.)

Jain, Mihir, Jégou, Hervé, et Bouthemy, Patrick. Improved motion description for action classification. *Frontiers in ICT*, 2, p. 28, 2016. ISSN 2297-198X. (Cité page 88.)

Jaiswal, Shashank et Valstar, Michel. Deep learning the dynamic appearance and shape of facial action units. Dans *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1–8, 2016. (Cité page 75.)

Jesorsky, Oliver, Kirchberg, KlausJ., et Frischholz, Robert W. Robust face detection using the Hausdorff distance. Dans *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication*, Bigun, Josef et Smeraldi, Fabrizio (éditeurs), volume 2091 de *LNCS*, pages 90–95. Springer Berlin Heidelberg, 2001. (Cité pages 42 et 43.)

Jiang, Bihan, Martinez, Brais, Valstar, Michel F, et Pantic, Maja. Decision level fusion of domain specific regions for facial action recognition. Dans *Proc. Int'l Conf. on Pattern Recognition (ICPR)*, pages 1776–1781, 2014. (Cité pages 77 et 98.)

Jung, Heechul, Lee, Sihaeng, Yim, Junho, Park, Sunjeong, et Kim, Junmo. Joint fine-tuning in deep neural networks for facial expression recognition. Dans *Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, pages 2983–2991, 2015. (Cité pages 107 et 108.)

Jung, Sung-Uk et Nixon, Mark S. On using gait to enhance frontal face extraction. *IEEE Trans. Information Forensics and Security*, 7(6), p. 1802–1811, 2012. (Cité page 45.)

Kae, Andrew, Sohn, Kihyuk, Lee, Honglak, et Learned-Miller, Erik. Augmenting CRFs with Boltzmann machine shape priors for image labeling. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2026. University of Massachusetts Amherst and University of Michigan Ann Arbor, 2013. (Cité pages 52 et 67.)

Kapfer, M. et Benois-Pineau, Jenny. Detection of human faces in color image sequences with arbitrary motions for very low bit-rate videophone coding. *Pattern Recognition Letters*, 18(14), p. 1503 – 1518, 1997. (Cité page 39.)

Kazemi, Vahid et Sullivan, Josephine. One millisecond face alignment with an ensemble of regression trees. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, 2014. (Cité pages 26 et 99.)

Kazi-Tani, Mohammed Yassine, Ghomari, Abdelghani, Lablack, Adel, et Bilasco, Ioan Marius. OVIS : ontology video surveillance indexing and retrieval system. *Int'l J. of Multimedia Information Retrieval (IJMIR)*, 6(4), p. 295–316, 2017. (Cité page 16.)

Khan, Rizwan Ahmed, Meyer, Alexandre, Konik, Hubert, et Bouakaz, Saida. Human vision inspired framework for facial expressions recognition. Dans *Proc. Int'l Conf. on Image Processing (ICIP)*, pages 2593–2596, 2012. (Cité page 79.)

Kim, Dae Hoe, Baddar, Wissam, Jang, Jinhyeok, et Ro, Yong Man. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Trans. Affective Computing (TAC)*, 2017. (Cité pages 106 et 108.)

Kotsia, Irene, Zafeiriou, Stefanos, et Pitas, Ioannis. Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition*, 41(3), p. 833–851, 2008. (Cité page 75.)

Kraemer, P. et Benois-Pineau, Jenny. Camera Motion Detection in the Rough Indexing Paradigm. Dans *TREC Video Retrieval Evaluation (TRECVID05)*, Nov 2005. (Cité page 88.)

La Cascia, Marco, Sclaroff, Stan, et Athitsos, Vassilis. Fast, reliable head tracking under varying illumination : An approach based on registration of texture-mapped 3D models. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 22(4), p. 322–336, 2000. (Cité pages 27, 34, 42 et 43.)

Lam, Kin-Man et Yan, Hong. An analytic-to-holistic approach for face recognition based on a single frontal view. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 20(7), p. 673–686, 1998. (Cité page 77.)

Lanitis, Andreas. Fgnet aging dataset, 2000. (Cité pages 42 et 43.)

Lee, Daniel D. et Seung, Sebastian H. Algorithms for non-negative matrix factorization. Dans *Proc. Int'l Conf. on Advances in Neural Information Processing Systems*, Leen, T.K., Dietterich, T.G., et Tresp, V. (éditeurs), pages 556–562. MIT Press, 2001. (Cité page 77.)

Lefevre, Stéphanie et Odobez, Jean-Marc. View-based appearance model online learning for 3D deformable face tracking. Dans *Proc. Int'l Conf. on Computer Vision Theory and Applications*, Richard, Paul et Braz, José (éditeurs), pages 223–230. INSTICC Press, 2010. (Cité page 45.)

Leng, XueMing et Wang, Yiding. Improving generalization for gender classification. Dans *Proc. Int'l Conf. on Image Processing (ICIP)*, pages 1656–1659, 2008. (Cité page 53.)

Li, Bing, Lian, Xiao-Chen, et Lu, Bao-Liang. Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing*, 76(1), p. 18–27, 2012. (Cité page 52.)

Li, Min, Bao, Shenghua, Dong, Weishan, Wang, Yu, et Su, Zhong. Head-shoulder based gender recognition. Dans *Proc. Int'l Conf. on Image Processing (ICIP)*, pages 2753–2756. IEEE, 2013a. (Cité page 52.)

Li, Xiaobai, Hong, Xiaopeng, Moilanen, Antti, Huang, Xiaohua, Pfister, Tomas, Zhao, Guoying, et Pietikäinen, Matti. Reading hidden emotions : spontaneous micro-expression spotting and recognition. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 217–230, 2015. (Cité pages 76, 79, 80, 87, 105, 106 et 108.)

Li, Xiaobai, Pfister, Tomas, Huang, Xiaohua, Zhao, Guoying, et Pietikäinen, Matti. A spontaneous micro-expression database : Inducement, collection and baseline. Dans *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 1–6. IEEE, 2013b. (Cité pages 72, 101 et 106.)

Liao, Chia-Te, Chuang, Hui-Ju, Duan, Chih-Hsueh, et Lai, Shang-Hong. Learning spatial weighting for facial expression analysis via constrained quadratic programming. *Pattern Recognition*, 46(11), p. 3103–3116, 2013. (Cité page 75.)

Liong, Sze-Teng, See, John, Phan, Raphael C-W, Le Ngo, Anh Cat, Oh, Yee-Hui, et Wong, KokSheik. Subtle expression recognition using optical strain weighted features. Dans *Proc. Asian Conf. on Computer Vision (ACCV)*, pages 644–657. Springer, 2014. (Cité page 76.)

Liu, Mengyi, Shan, Shiguang, Wang, Ruiping, et Chen, Xilin. Learning expressionlets via universal manifold model for dynamic facial expression recognition. *IEEE Trans. Image Processing*, 25(12), p. 5920–5932, 2016a. (Cité pages 107 et 108.)

Liu, Xiangyang, Lu, Hongtao, et Li, Wenbin. Multi-manifold modeling for head pose estimation. *Proc. Int'l Conf. on Image Processing (ICIP)*, 2010. (Cité page 39.)

Liu, Yong-Jin, Zhang, Jin-Kai, Yan, Wen-Jing, Wang, Su-Jing, Zhao, Guoying, et Fu, Xiaolan. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Trans. Affective Computing (TAC)*, 7(4), p. 299–310, 2016b. (Cité pages 75, 76 et 79.)

Lopes, André Teixeira, de Aguiar, Edilson, De Souza, Alberto F, et Oliveira-Santos, Thiago. Facial expression recognition with convolutional neural networks : Coping with few data and the training sample order. *Pattern Recognition*, 61, p. 610–628, 2017. (Cité pages 75 et 107.)

Lucey, Patrick, Cohn, Jeffrey F, Kanade, Takeo, Saragih, Jason, Ambadar, Zara, et Matthews, Iain. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. Dans *Proc. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101. IEEE, 2010. (Cité pages 75, 79, 97 et 101.)

Lyons, Michael J., Budynek, Julien, et Akamatsu, Shigeru. Automatic classification of single facial images. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 21(12), p. 1357–1362, 1999. (Cité page 84.)

Ma, Bingpeng, Li, Annan, Chai, Xiujuan, et Shan, Shiguang. Head yaw estimation via symmetry of regions. Dans *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 1–6, 2013. (Cité page 26.)

Majumder, Anima, Behera, Laxmidhar, et Subramanian, Venkatesh K. Emotion recognition from geometric facial features using self-organizing map. *Pattern Recognition*, 47(3), p. 1282–1293, 2014. (Cité page 74.)

Makinen, Erno et Raisamo, Roope. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 30(3), p. 541–547, 2008a. (Cité page 51.)

Makinen, Erno et Raisamo, Roope. An experimental comparison of gender classification methods. *Pattern Recognition Letters*, 29(10), p. 1544–1556, 2008b. (Cité page 52.)

Mamdani, E.H. et Assilian, S. A fuzzy logic controller for a dynamic plant. *Man-Machine Studies*, 7 (7), p. 1–13, 1975. (Cité pages 61 et 62.)

Melin, Patricia, Mendoza, Olivia, et Castillo, Oscar. An improved method for edge detection based on interval type-2 fuzzy logic. *Expert Systems with Applications*, 37(12), p. 8527–8535, 2010. (Cité page 53.)

Mennesson, José, Dahmane, Afifa, Danisman, Taner, et Bilasco, Ioan Marius. Head yaw estimation using frontal face detector. Dans *Proc. Int'l Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Porto, Portugal, 2016. (Cité pages 15, 25 et 41.)

Moallem, Payman et Mousavi, B Somayeh. Gender classification by fuzzy inference system. *Int'l J. of Advanced Robotic Systems*, 10(89), 2013. (Cité page 53.)

Moghaddam, Baback et Yang, Ming-Hsuan. Learning gender with support faces. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 24(5), p. 707–711, 2002. (Cité page 51.)

Mollahosseini, Ali, Chan, David, et Mahoor, Mohammad H. Going deeper in facial expression recognition using deep neural networks. Dans *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016. (Cité pages 107 et 108.)

Morency, Louis-Philippe, Whitehill, Jacob, et Movellan, Javier. Monocular head pose estimation using generalized adaptive view-based appearance model. *Image Vision Computing*, 28(5), p. 754–761, 2010. (Cité pages 39 et 45.)

Morovic, J. et Sun, P.-L. Accurate 3d image colour histogram transformation. *Pattern Recognition Letters*, 24(11), p. 1725 – 1735, 2003. ISSN 0167-8655. (Cité page 50.)

MPLab. The mplab genki database, genki-4k subset, 2011. URL `http://mplab.ucsd.edu`. (Cité pages 49, 84 et 149.)

Murphy-Chutorian, Erik et Trivedi, Mohan Manubhai. HyHOPE : Hybrid head orientation and position estimation for vision-based driver head tracking. Dans *Proc. IEEE Intelligent Vehicles Symposium*, pages 512–517, 2008. (Cité page 27.)

Murphy-Chutorian, Erik et Trivedi, Mohan Manubhai. Head pose estimation in computer vision : A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 31(4), p. 607–626, 2009. (Cité pages 25 et 47.)

My, Vo Duc et Zell, Andreas. Real time face tracking and pose estimation using an adaptive correlation filter for human-robot interaction. Dans *Proc. European Conf. on Mobile Robots (ECMR)*, pages 119–124, 2013. (Cité pages 26 et 45.)

Nakariyakul, Songyot. Suboptimal branch and bound algorithms for feature subset selection : A comparative study. *Pattern Recognition Letters*, 45, p. 62–70, 2014. (Cité page 78.)

Ng, Choon-Boon, Tay, Yong-Haur, et Goi, Bok-Min. Recognizing human gender in computer vision : A survey. Dans *Proc. Pacific Rim Int'l Conf. on Artificial Intelligence (PRICAI)*, Anthony, Patricia, Ishizuka, Mitsuru, et Lukose, Dickson (éditeurs), volume 7458 de *Lecture Notes in Computer Science*, pages 335–346. Springer Berlin Heidelberg, 2012. (Cité page 51.)

Odobez, J.M. et Bouthemy, P. Robust multiresolution estimation of parametric motion models. *J. Visual Communication and Image Representation*, 6(4), p. 348 – 365, 1995. ISSN 1047-3203. (Cité page 88.)

Ojala, Timo, Pietikainen, Matti, et Harwood, David. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1), p. 51–59, 1996. (Cité page 74.)

Ojala, Timo, Pietikainen, Matti, et Maenpaa, Topi. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 24(7), p. 971–987, 2002. (Cité page 51.)

Oka, Kenji, Sato, Yoichi, Nakanishi, Yasuto, et Koike, Hideki. Head pose estimation system based on particle filtering with adaptive diffusion control. Dans *Proc. IAPR Conf. on Machine Vision Applications (IAPR MVA)*, pages 586–589, 2005. (Cité page 26.)

Osadchy, Margarita, Cun, Yann Le, et Miller, Matthew L. Synergistic face detection and pose estimation with energy-based models. *Machine Learning Research*, 8, p. 1197–1215, 2007. (Cité page 26.)

Pantic, Maja, Valstar, Michel, Rademaker, Ron, et Maat, Ludo. Web-based database for facial expression analysis. Dans *Proc. IEEE Int'l Conf. on Multimedia and Expo (ICME)*, 2005. (Cité page 101.)

Patel, Devangini, Hong, Xiaopeng, et Zhao, Guoying. Selective deep features for micro-expression recognition. Dans *Proc. IEEE Int'l Conf. on Pattern Recognition (ICPR)*, pages 2258–2263. IEEE, 2016. (Cité pages 106 et 108.)

Phillips, Jonathon P., Moon, Hyeonjoon, Rizvi, Syed A., et Rauss, Patrick J. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 22(10), p. 1090–1104, 2000. (Cité page 84.)

Phillips, Jonathon P., Wechsler, Harry, Huang, Jeffer, et Rauss, Patrick J. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5), p. 295–306, 1998. (Cité page 49.)

Polat, Ovunc et Yildirim, Tulay. Genetic optimization of GRNN for pattern recognition without feature extraction. *Expert Systems with Applications*, 34(4), p. 2444–2448, 2008. (Cité page 53.)

Porter, Stephen et Ten Brinke, Leanne. Reading between the lies : Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, 19(5), p. 508–514, 2008. (Cité page 72.)

Poux, Delphine, Allaert, Benjamin, Mennesson, José, Ihaddadene, Nacim, Bilasco, Ioan Marius, et Djeraba, Chaabane. Mastering occlusions by using intelligent facial frameworks based on the propagation of movement. Dans *Int'l Conf. on Content-Based Multimedia Indexing (CBMI)*, La Rochelle, France, 2018. (Cité pages 16 et 110.)

Pudil, Pavel, Novovičová, Jana, et Kittler, Josef. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11), p. 1119–1125, 1994. (Cité page 78.)

Ramon-Balmaseda, Enrique, Lorenzo-Navarro, Javier, et Castrillon-Santana, Modesto. Gender classification in large databases. Dans *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Alvarez, Luis, Mejail, Marta, Gomez, Luis, et Jacobo, Julio (éditeurs), volume 7441 de *Lecture Notes in Computer Science*, pages 74–81. Springer Berlin Heidelberg, 2012. (Cité pages 51, 59 et 69.)

Revaud, Jerome, Weinzaepfel, Philippe, Harchaoui, Zaid, et Schmid, Cordelia. EpicFlow : Edge-preserving interpolation of correspondences for optical flow. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1172, 2015. (Cité pages 80 et 100.)

Rowley, Henry A., Baluja, Shumeet, et Kanade, Takeo. Neural network-based face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 20(1), p. 23–38, 1998a. (Cité page 54.)

Rowley, Henry A., Baluja, Shumeet, et Kanade, Takeo. Rotation invariant neural network-based face detection. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 38–44. IEEE, 1998b. (Cité pages 42 et 43.)

Sadeghi, Hamid, Raie, Abolghasem-A, et Mohammadi, Mohammad-Reza. Facial expression recognition using geometric normalization and appearance representation. Dans *Machine Vision and Image Processing (MVIP)*, pages 159–163. IEEE, 2013. (Cité page 77.)

Santana, Modesto Castrillon, Lorenzo-Navarro, Javier, et Ramon-Balmaseda, Enrique. Improving gender classification accuracy in the wild. Dans *Proc. Iberoamerican Congress on Pattern Recognition (CIARP)*, Ruiz-Shulcloper, José et di Baja, Gabriella Sanniti (éditeurs), volume 8259 de *Lecture Notes in Computer Science*, pages 270–277. Springer, 2013. (Cité page 51.)

Satta, Riccardo, Galbally, Javier, et Beslay, Laurent. Children gender recognition under unconstrained conditions based on contextual information. Dans *Proc. Int'l Conf. on Pattern Recognition (ICPR)*, pages 357–362, Stockholm, Sweden, 2014. (Cité page 52.)

Shan, Caifeng. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters*, 33(4), p. 431–437, 2012. (Cité page 51.)

Shan, Caifeng, Gong, Shaogang, et McOwan, Peter W. Conditional mutual information based boosting for facial expression recognition. Dans *Proc. British Machine Vision Conf.*, volume 1, pages 399–408. Guide Share Europe, Berlin, 2005. (Cité page 74.)

Shan, Caifeng, Gong, Shaogang, et McOwan, Peter W. Facial expression recognition based on local binary patterns : A comprehensive study. *Image and Vision Computing*, 27(6), p. 803–816, 2009. (Cité page 79.)

Shinohara, Yusuke et Otsu, Nobuyuki. Facial expression recognition using fisher weight maps. Dans *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 499–504, 2004. (Cité page 81.)

Sim, Terence, Baker, Simon, et Bsat, Maan. The *cmupose, illumination, andexpression* database. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 25(12), p. 1615–1618, 2003. (Cité page 34.)

Sivic, Josef et Zisserman, Andrew. Video Google : A text retrieval approach to object matching in videos. Dans *Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, volume 2, pages 1470–1477, 2003. (Cité page 75.)

Somol, Petr, Pudil, Pavel, Novovičová, Jana, et Paclík, Pavel. Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20(11-13), p. 1157–1163, 1999. (Cité page 78.)

Song, Mingli, Tao, Dacheng, Sun, Zhuo, et Li, Xuelong. Visual-context boosting for eye detection. *IEEE Trans. Systems, Man, and Cybernetics, Part B : Cybernetics*, 40(6), p. 1460–1467, 2010. (Cité page 82.)

Stentiford, Fred. Attention based facial symmetry detection. Dans *Proc. Int'l Conf. on Advances in Pattern Recognition*, pages 112–119, 2005. (Cité page 31.)

Sung, Jaewon, Kanade, Takeo, et Kim, Daijin. Pose robust face tracking by combining active appearance models and cylinder head models. *Int'l J. Computer Vision*, 80(2), p. 260–274, 2008. (Cité pages 27 et 45.)

Szolgay, D., Benois-Pineau, J., Megret, R., Gaestel, Y., et Dartigues, J.-F. Detection of moving foreground objects in videos with strong camera motion. *Pattern Analysis and Applications*, 14(3), p. 311–328, Aug 2011. ISSN 1433-755X. (Cité page 88.)

Takagi, Tomohiro et Sugeno, Michio. Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Systems, Man and Cybernetics*, SMC-15(1), p. 116–132, 1985. (Cité page 61.)

Tang, Hengliang, Yin, Baocai, Sun, Yanfeng, et Hu, Yongli. 3D face recognition using local binary patterns. *Signal Processing*, 93(8), p. 2190–2198, 2012. (Cité page 74.)

Tome, Pedro, Fierrez, Julian, Vera-Rodriguez, Ruben, et Nixon, Mark S. Soft biometrics and their application in person recognition at a distance. *IEEE Trans. Information Forensics and Security*, 9 (3), p. 464–475, 2014. (Cité page 52.)

Tran, Ngoc-Trung, Ababsa, Fakhr-Eddine, Charbit, Maurice, Feldmar, Jacques, Petrovska-Delacretaz, Dijana, et Chollet, Gerard. 3D face pose and animation tracking via Eigen-decomposition based bayesian approach. Dans *Advances in Visual Computing*, volume 8033, pages 562–571, 2013. (Cité page 45.)

Turk, Matthew et Pentland, Alex. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1), p. 71–86, 1991. (Cité page 78.)

Ueki, Kazuya et Kobayashi, Tetsunori. Gender classification based on integration of multiple classifiers using various features of facial and neck images. *Information and Media Technologies*, 3(2), p. 479–485, 2008. (Cité page 52.)

Valenti, Roberto, Sebe, Nicu, et Gevers, Theo. Combining head pose and eye location information for gaze estimation. *IEEE Trans. Image Processing*, 21(2), p. 802–815, 2012. (Cité page 39.)

Valenti, Roberto, Yücel, Zeynep, et Gevers, Theo. Robustifying eye center localization by head pose cues. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 612–618. IEEE, 2009. (Cité pages 27 et 45.)

Ververidis, Dimitrios et Kotropoulos, Constantine. Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition. *Signal Processing*, 88(12), p. 2956–2970, 2008. (Cité page 72.)

Vinod Pathangay, Sukhendu Das et Greiner, Thomas. Symmetry-based face pose estimation from a single uncalibrated view. *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 1–8, 2008. (Cité page 26.)

Viola, Paul et Jones, Michael J. Robust real-time face detection. *Int'l J. Computer Vision*, 57(2), p. 137–154, 2004. (Cité pages 25, 28, 39, 54, 72 et 83.)

Walawalkar, L., Yeasin, Mohammad, Narasimhamurthy, Anand M., et Sharma, Rajeev. Support vector learning for gender classification using audio and visual cues. *Int'l J. Pattern Recognition and Artificial Intelligence*, 17(3), p. 417–439, 2003. (Cité page 51.)

Wan, Shaohua et Aggarwal, J.K. Spontaneous facial expression recognition : A robust metric learning approach. *Pattern Recognition (PR)*, 47(5), p. 1859–1868, 2014. (Cité page 73.)

Wan, Y. et Shi, D. Joint exact histogram specification and image enhancement through the wavelet transform. *IEEE Trans. Image Processing*, 16(9), p. 2245–2250, Sep. 2007. ISSN 1057-7149. (Cité page 50.)

Wang, Su-Jing, Yan, Wen-Jing, Zhao, Guoying, Fu, Xiaolan, et Zhou, Chun-Guang. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. Dans *Proc. European Conf. on Computer Vision Workshop (ECCVW)*, pages 325–338, 2014a. (Cité pages 105 et 106.)

Wang, Sujing, Yan, Wen-Jing, Li, Xiaobai, Zhao, Guoying, et Fu, Xiaolan. Micro-expression recognition using dynamic textures on tensor independent color space. Dans *Proc. Int'l Conf. on Pattern Recognition (ICPR)*, pages 4678–4683, 2014b. (Cité pages 105 et 106.)

Wang, Yandan, See, John, Phan, Raphael C-W, et Oh, Yee-Hui. LBP with six intersection points : Reducing redundant information in LBP-TOP for micro-expression recognition. Dans *Proc. Asian Conf. on Computer Vision (ACCV)*, pages 525–537, 2014c. (Cité pages 76 et 106.)

Weber, Markus. Caltech frontal face dataset, 1999. URL http://www.vision.caltech.edu/html-files/archive.html. (Cité pages 42, 43 et 148.)

Weinzaepfel, Philippe, Revaud, Jerome, Harchaoui, Zaid, et Schmid, Cordelia. DeepFlow : Large displacement optical flow with deep matching. Dans *Proc. IEEE Int'l Conf. on Computer Vision (ICCV)*, pages 1385–1392, 2013. (Cité page 100.)

Whitney, A. Wayne. A direct method of nonparametric measurement selection. *IEEE Trans. Computers*, C-20(9), p. 1100–1103, 1971. (Cité page 78.)

Wilson, Hugh R., Wilkinson, Frances, Lin, Li-Ming, et Castillo, Maja. Perception of head orientation. *Vision Research*, 40(5), p. 459–472, 2000. (Cité page 28.)

Wolf, Lior, Hassner, Tal, et Maoz, Itay. Face recognition in unconstrained videos with matched background similarity. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534, 2011. (Cité pages 42, 43 et 149.)

Xiao, Jing, Kanade, Takeo, et Cohn, Jeffrey F. Robust full-motion recovery of head by dynamic templates and re-registration techniques. Dans *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 163–169, Washington, DC, USA, 2002. (Cité page 45.)

Yan, Wen-Jing, Li, Xiaobai, Wang, Su-Jing, Zhao, Guoying, Liu, Yong-Jin, Chen, Yu-Hsin, et Fu, Xiaolan. CASME II : An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1), 2014a. (Cité pages 79, 98, 101 et 106.)

Yan, Wen-Jing, Wang, Su-Jing, Liu, Yong-Jin, Wu, Qi, et Fu, Xiaolan. For micro-expression recognition : Database and suggestions. *Neurocomputing*, 136, p. 82–87, 2014b. (Cité page 72.)

Yan, Wen-Jing, Wu, Qi, Liang, Jing, Chen, Yu-Hsin, et Fu, Xiaolan. How fast are the leaked facial expressions : The duration of micro-expressions. *J. of Nonverbal Behavior*, 37(4), p. 217–230, 2013. (Cité page 72.)

Zadeh, Ataollah Ebrahim. Automatic recognition of radio signals using a hybrid intelligent technique. *Expert Systems with Applications*, 37(8), p. 5803–5812, 2010. (Cité page 53.)

Zhang, Kaihao, Huang, Yongzhen, Du, Yong, et Wang, Liang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Trans. Image Processing*, 26(9), p. 4193–4203, 2017a. (Cité pages 107 et 108.)

Zhang, L., Zhu, G., Shen, P., et Song, J. Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition. Dans *Proc. IEEE Int'l Conf. on Computer Vision Workshops (ICCVW)*, pages 3120–3128, 2017b. (Cité page 117.)

Zhang, Y. J. Improving the accuracy of direct histogram specification. *J. Electronics Letters*, 28(3), p. 213–214, Jan 1992. ISSN 0013-5194. (Cité page 50.)

Zhao, Guoying, Huang, Xiaohua, Taini, Matti, Li, Stan Z, et Pietikäinen, Matti. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9), p. 607–619, 2011. (Cité pages 101, 107 et 108.)

Zhao, Guoying et Pietikainen, Matti. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 29(6), p. 915–928, 2007. (Cité pages 75, 79, 107 et 108.)

Zhao, Lei, Wang, Zengcai, et Zhang, Guoxin. Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and Gabor multiorientation fusion histogram. *Mathematical Problems in Engineering*, 2017. (Cité pages 107 et 108.)

Zhao, Sicheng, Yao, Hongxun, et Sun, Xiaoshuai. Video classification and recommendation based on affective analysis of viewers. *Neurocomputing*, 119(0), p. 101–110, 2013. (Cité page 25.)

Zhong, Lin, Liu, Qingshan, Yang, Peng, Liu, Bo, Huang, Junzhou, et Metaxas, Dimitris N. Learning active facial patches for expression analysis. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012a. (Cité page 73.)

Zhong, Lin, Liu, Qingshan, Yang, Peng, Liu, Bo, Huang, Junzhou, et Metaxas, Dimitris N. Learning active facial patches for expression analysis. Dans *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2562–2569. IEEE, 2012b. (Cité page 108.)

Zhou, Jie, Lu, Xiao Guang, Zhang, David, et Wu, Chen-Yu. Orientation analysis for rotated human face detection. *Image and Vision Computing*, 20(4), p. 257–264, 2002. (Cité page 26.)