

Apprentissage automatique pour l'analyse des expressions faciales

Kevin Bailly

▶ **To cite this version:**

Kevin Bailly. Apprentissage automatique pour l'analyse des expressions faciales. Intelligence artificielle [cs.AI]. Sorbonne Université, 2019. tel-02489704

HAL Id: tel-02489704
https://theses.hal.science/tel-02489704

Submitted on 26 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.







HABILITATION À DIRIGER DES RECHERCHES
Sorbonne Université – Faculté des Sciences et Ingénierie

Présentée par
Kévin BAILLY

Apprentissage automatique pour l'analyse des expressions faciales

---

Soutenue le 12 février 2019 devant le jury composé de :

Pr. Laurent Heutte	LITIS – Université de Rouen	Rapporteur
Dr. Jean-Marc Odobez	IDIAP – École Polytechnique Fédérale de Lausanne	Rapporteur
Pr. Björn Schuller	University of Augsburg – Imperial College London	Rapporteur
Pr. Mohamed Chétouani	ISIR/CNRS – Sorbonne Université	Examinateur
Pr. Matthieu Cord	LIP6/CNRS – Sorbonne Université	Examinateur
Pr. James Crowley	LIG/INRIA – Grenoble INP	Examinateur

Table des matières

1 Introduction **1**

1.1 Contexte et objectifs . . . 1

1.1.1 Caractérisation des expressions faciales . . . 1

1.1.2 Les défis de l'analyse automatique des expressions faciales . . . 3

1.2 Bases de données et métrique d'évaluation . . . 4

1.2.1 Bases de données disponibles . . . 4

1.2.2 Métrique d'évaluation . . . 5

1.3 Plan du mémoire . . . 7

2 Positionnement et contributions **9**

2.1 2000-2008 : les premiers grands succès de l'analyse faciale . . . 9

2.2 2008-2012 : l'essor de l'analyse des expressions faciales . . . 10

2.3 2012-2018 : vers l'analyse faciale "*in the wild*" . . . 12

2.3.1 Robustesse au faible nombre de données d'apprentissage . . . 12

2.3.2 Robustesse aux fortes variations d'apparence . . . 13

2.4 L'émergence de l'apprentissage profond (*Deep Learning*) . . . 15

3 SVM Multi-Noyaux pour la combinaison de descripteurs hétérogènes **19**

3.1 Choix du descripteur et de la fonction noyau . . . 19

3.1.1 Les descripteurs d'apparence . . . 19

3.1.2 Les fonctions noyaux . . . 21

3.1.3 Résultats expérimentaux . . . 21

3.2 Prise en compte de l'identité . . . 22

3.3 Combinaisons de descripteurs hétérogènes . . . 23

3.3.1 Détection des AU . . . 23

3.3.2 Localisation des points caractéristiques . . . 24

4 Analyse des expressions faciales par apprentissage de métriques **27**

4.1 Estimateur de Nadaraya-Watson et apprentissage de métriques . . . 27

4.2 Améliorations proposées . . . 29

4.2.1 Sélection des descripteurs . . . 29

4.2.2 Descente de gradient stochastique . . . 30

4.2.3 Régularisation de la fonction de coût . . . 30

4.2.4 MLKR multi-labels . . . 31

4.3 Application à l'analyse des expressions faciales . . . 33

4.3.1 Localisation de points caractéristiques . . . 33

4.3.2 Estimation de l'intensité des *Action Units* . . . 34

5 Forêts Aléatoires pour l'analyse en environnement non contraint **39**

5.1 Forêts aléatoires pour l'analyse faciale : intérêt et limitations 39
5.2 Forêts aléatoires conditionnelles 40
5.2.1 Forêts aléatoires conditionnées par paires (PCRF) 40
5.2.2 PCRF Multi-vues (MV-PCRF) 42
5.2.3 Résultats expérimentaux 43
5.3 Forets aléatoires à sous-espaces locaux (LSRF) 44
5.3.1 Prédiction locale de l'expression 44
5.3.2 Prédiction des expressions robustes aux occultations 46
5.3.3 Reconnaissance des AU 47
5.4 Evaluation gloutonne des Forêts Neuronales 48
5.4.1 Principe général 48
5.4.2 Procédure d'évaluation gloutonne 49
5.4.3 Algorithme efficace d'apprentissage 49
5.4.4 Applications 50
5.5 Conclusion 52

6 Applications **55**
6.1 Le projet JEMImE 55
6.1.1 La base de données JEMImE 56
6.1.2 Quelques résultats cliniques 56
6.1.3 Le jeu sérieux 57
6.2 Le projet SMART SeNSE 57

7 Conclusion et perspectives **61**

Annexes **64**

Bibliographie **68**

Chapitre 1
Introduction

Dans ce chapitre, nous introduirons la problématique de l'analyse des expressions faciales (1.1) en décrivant les moyens de les caractériser (1.1.1) et les défis à relever pour les extraire de manière automatique à partir d'une vidéo (1.1.2). Nous présenterons ensuite les jeux de données et les métriques d'évaluation utilisés (1.2). La dernière partie du chapitre décrira la structure du document (1.3).

1.1 Contexte et objectifs

Le visage d'un être humain est constitué d'une quarantaine de muscles qui permettent de produire quelques milliers d'expressions faciales. Ces expressions sont porteuses d'informations relatives à l'état cognitif et aux intentions sociales d'une personne. L'analyse des expressions faciales vise à extraire de manière automatique ces indices dans une image ou un flux vidéo. Il s'agit d'un domaine de recherche très actif à l'interface de la vision par ordinateur et de l'apprentissage artificiel avec de nombreuses applications dans des domaines variés tels que la robotique sociale, le marketing, les études cliniques ou encore l'animation d'agents virtuels.

1.1.1 Caractérisation des expressions faciales

Il existe de nombreuses manières de caractériser des expressions faciales [25] que nous avons choisi de représenter suivant un axe caractérisant le niveau d'interprétation et de subjectivité de la représentation (*cf.* figure 1.1). À une extrémité de l'axe, nous trouvons le suivi et la localisation de points caractéristiques tels que les commissures des lèvres et des yeux ou le bout du nez par exemple. Les représentations actuelles s'intéressent généralement à 68 ou 51 points, suivant que l'on considère ou non le contour de la mâchoire. Il s'agit d'une représentation très bas niveau (i.e. pauvre d'un point de vue sémantique) et qui mélange des facteurs de variation autres que l'expression faciale (la morphologie et l'orientation du visage par exemple). Mais elle est objective (la position d'un point sur le visage est très peu sujette à interprétation) et elle constitue une étape préliminaire à la plupart des tâches d'analyse faciale actuelles, au delà même de l'analyse des expressions (reconnaissance de l'identité, de l'âge, du sexe...).

A l'autre extrémité de l'axe, nous trouvons les représentations qui nécessitent une interprétation importante de ce que l'on perçoit sur le visage. Par exemple, une représentation couramment utilisée dans le domaine de l'informatique affective découle des travaux d'Ekman sur les émotions prototypiques [45]. Celui-ci a proposé une liste d'émotions universellement reconnues : joie, colère, tristesse, peur, dégoût et surprise. Le principal problème de cette représentation est qu'elle n'est pas adaptée pour caractériser les expressions faciales spontanées car la plupart de nos comportements affectifs ne peuvent pas se traduire en terme d'émotions prototypiques. Le processus d'annotation est toutefois relativement simple et intuitif, et il existe de nombreuses bases de données annotées [61] (cf. section 1.2). On trouve également à cette extrémité de l'axe d'autres représentations très répandues telles que la représentation continue des affects. Elle consiste à projeter les expressions sur un nombre restreint de dimensions latentes. Une expression spécifique telle que la joie peut alors être décrite pas sa position dans un espace de faible

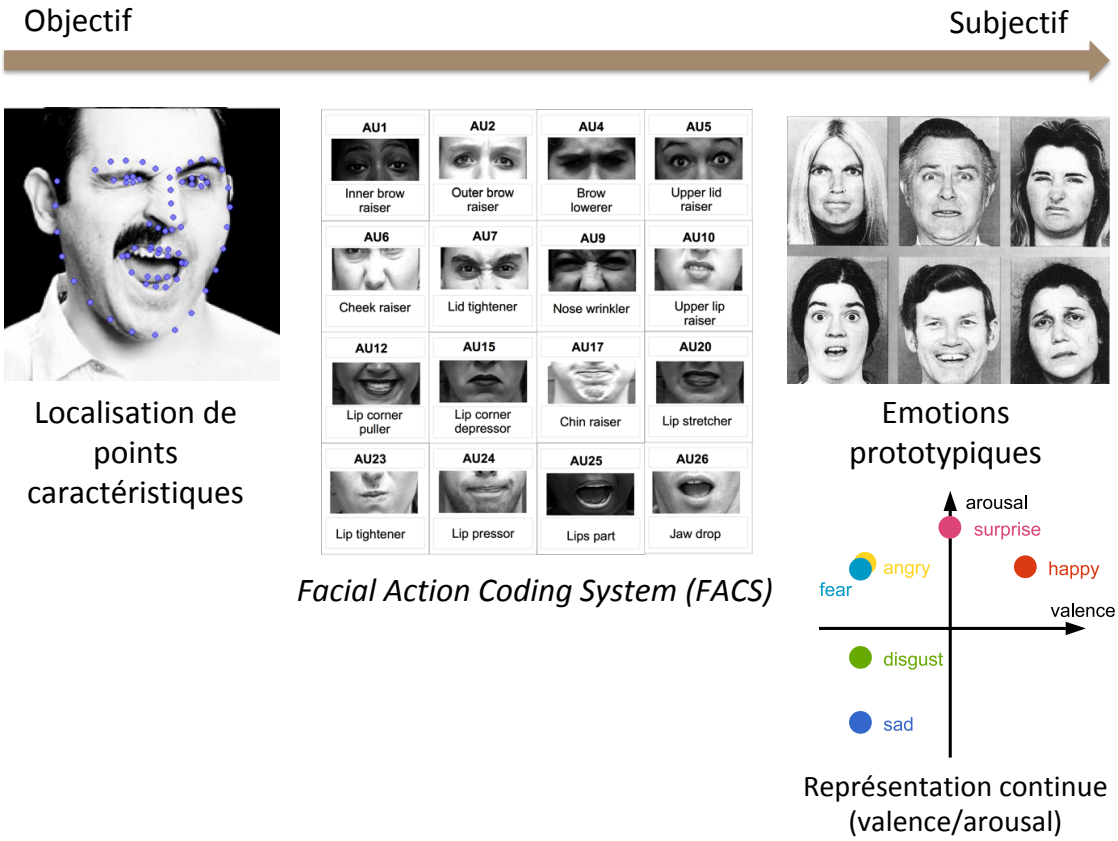


FIGURE 1.1 – Exemples de caractérisation des expressions faciales classés par degrés de subjectivité

dimension. Un exemple simple d'un tel modèle est la représentation de la valence (niveau de (dé)plaisir associé à l'émotion) et de l'arousal (degré d'activation physiologique) dans un espace à deux dimensions. Ainsi, dans cet espace, la joie sera caractérisée par un niveau élevé de valence et d'arousal. À l'inverse la tristesse est définie par un arousal et une valence faibles. Cet espace de représentation est plus riche que celui des émotions prototypiques (espace continu), mais la projection de l'émotion dans un espace de faible dimension peut entraîner une perte d'information. Certaines émotions telles que la peur et la colère ne peuvent pas facilement être différenciées dans cet espace puisque elles se caractérisent toutes deux par une valence négative et une activation physiologique importante. De plus, le processus d'annotation est moins intuitif et précis que celui des émotions prototypiques, ce qui limite le nombre de données annotées exploitables.

Il existe une représentation intermédiaire que l'on pourrait donc situer au centre de cet axe. Elle consiste à décrire une expression faciale par une combinaison d'activations des 44 muscles faciaux. Cette représentation a été codifiée par Ekman au sein du *Facial Action Coding System (FACS)* [46]. Elle est très intéressante car elle est moins subjective (et donc plus fiable) que des représentations de plus haut niveau, tout en ne contenant que de l'information relative aux expressions faciales (contrairement à la représentation par points caractéristiques). Toutefois, l'annotation en activations musculaires (*Action Units*, AU) est fastidieuse et requiert des experts qui ont été spécifiquement formés au *FACS,* ce qui limite le nombre de bases de données annotées disponibles.

1.1.2 Les défis de l'analyse automatique des expressions faciales

Les méthodes d'analyse reposent traditionnellement sur une chaîne de traitements en trois étapes, illustrée dans la figure 1.2 :

— **Prétraitements :** cette étape consiste à détecter le visage que l'on souhaite analyser et à localiser un ensemble de points caractéristiques. A partir de ces positions de points caractéristiques, il est possible d'appliquer des transformations de l'image afin d'annuler certaines sources de variations telles que les rotations du visage dans le plan image et les variations du facteur d'échelle.

— **Représentation :** lorsque le visage est dans une position de référence, on extrait un ensemble de descripteurs permettant de caractériser l'apparence et la géométrie du visage observé. L'objectif de cette étape est d'accentuer l'information relative à l'expression et d'atténuer les autres sources de variations.

— **Décision :** à partir d'un ensemble de données annotées, un prédicteur apprend à discriminer différentes expressions faciales dans l'espace de représentation précédemment choisi. Ainsi, lorsque l'on présente un nouvel échantillon au prédicteur, il utilise les connaissances extraites de la base de données pour prendre une décision sur le type d'expressions.

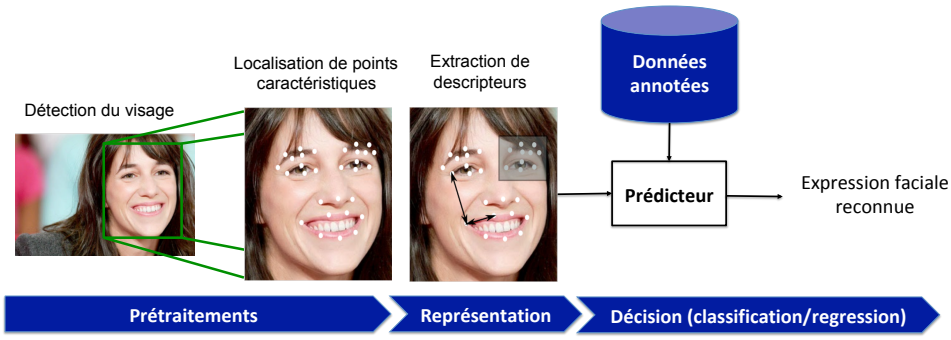


FIGURE 1.2 – Chaîne de traitements traditionnelle d'une méthode de reconnaissance d'expressions faciales

Pour que les systèmes de reconnaissance des expressions faciales soient performants et puissent répondre aux nombreux besoins applicatifs en environnement faiblement ou non contraint, ils doivent répondre à un certain nombre de défis.

La grande variabilité des données : les expressions faciales sont très variables dans leur dynamique et leur intensité, et l'apparence d'un visage est influencée par de nombreux facteurs indépendants de l'expression tels que la pose, l'identité du sujet, son âge, son ethnie ou encore les conditions d'illumination, les occultations et le type de camera utilisé.

L'hétérogénéité des descripteurs : le choix des descripteurs est souvent une étape décisive pour apprendre des modèles prédictifs performants. On distingue souvent les descripteurs géométriques et d'apparence. Les descripteurs géométriques caractérisent les relations entre les différents points clés du visage (par exemple le déplacement des commissures des lèvres par rapport à la position des yeux). Les descripteurs d'apparence caractérisent la texture (pour mettre

en évidence la présence de rides par exemple). De plus, ces descripteurs peuvent-être soit statiques (extraits sur une image) ou dynamiques (calculés à partir de plusieurs images d'une même séquence vidéo). Au final, ces descripteurs peuvent être :

— Peu informatifs (pris séparément, une distance entre deux points clés ou l'intensité d'un contour mesuré dans une zone spécifique du visage ne permet pas de prédire l'expression)

— Très nombreux

— Très hétérogènes : chaque descripteur a une plage de variation et une dynamique qui lui est propre

La quantité et la fiabilité des données d'apprentissage : dans le domaine de l'analyse des expressions faciales, les risques de sur-apprentissage sont particulièrement importants car le nombre de données annotées disponibles pour entrainer les modèles prédictifs est limité (l'annotation manuelle d'images de visage est fastidieuse et certaines tâches telles que la détection des activations musculaires requièrent une expertise spécifique). De plus, les annotations ne sont pas toujours fiables. Dans le cas des représentations les plus subjectives (l'état émotionnel d'une personne au cours du temps), il peut y avoir de fortes disparités entre deux experts annotant la même vidéo. Cette incertitude sur les annotations est une difficulté supplémentaire pour l'apprentissage de modèles prédictifs performants. Pour certaines tâches qui nécessitent une annotation en continue d'une séquence vidéo, il n'est pas rare d'observer un décalage temporel entre le contenu de la vidéo et les annotations.

Les contraintes applicatives et matérielles : dans de nombreux domaines tels que les interactions Hommes/Robots ou la détection de vigilance d'un conducteur, le système devra traiter les données en temps réel. De plus la mémoire vive de certains systèmes embarqués peut être limitée. Ainsi ces contraintes de complexité algorithmique et d'empreinte mémoire ont un impact direct sur le choix des descripteurs et des prédicteurs utilisés.

L'adaptabilité des modèles à de nouveaux domaines et à de nouvelles tâches : les méthodes actuelles ont une capacité de généralisation limitée lorsque les données de test diffèrent sensiblement de celles d'apprentissage. De même, certaines tâches à réaliser peuvent être différentes mais fortement corrélées (la reconnaissance des émotions et la détection des activations musculaires du visage par exemple). Adapter un modèle consiste alors à exploiter les connaissances d'un domaine et une tâche source afin de traiter un nouveau domaine et une nouvelle tâche cible.

1.2 Bases de données et métrique d'évaluation

Dans cette section, nous présentons brièvement les principales bases de données ainsi que les métriques d'évaluation utilisées pour apprendre et évaluer les modèles présentés.

1.2.1 Bases de données disponibles

Le tableau 1.1 répertorie les différentes bases de données utilisées tout au long du manuscrit. Pour chaque base, nous précisons le nombre d'échantillons et d'identités différentes, la nature des données (vidéos vs. images statiques, comportements joués vs. spontanés), la nature des annotations (points caractéristiques, expressions catégoriques, occurence et intensité des AU) ainsi que des informations relatives à la difficulté des données (présence de poses non frontales et d'occultations).

TABLE 1.1 – Vue synthétique des principales bases de données utilisées

Bases	BioID	LFPW	300-W	CK+	FERA 2011	BU-4DFE	FEED	BP4D	SFEW	DISFA	JEM-ImE
	[70]	[14]	[130]	[91]	[145]	[163]	[149]	[171]	[39]	[96]	[34]
#échantillons	1521	1432	3740	309	158	606	716	328	700	27	4632
#identités	23	1432	3740	123	13	101	19	41	95	27	193
Video											
Spontané											
Points caract.	20	29	68								
Catégorique				7	5	6	7	8	7		4
AU (occurence)				14	12			12		12	
AU (intensité)								5		12	
non frontales											
Occultations											

1.2.2 Métrique d'évaluation

Il existe plusieurs métriques d'évaluation pour mesurer les performances d'un système d'analyse automatique des expressions faciales. Nous reportons ici les principales métriques utilisées en fonction de la tâche considérée.

1.2.2.1 Localisation de points caractéristiques

La métrique la plus couramment utilisée pour évaluer les performances d'une méthode de localisation de points caractéristiques est l'erreur euclidienne moyenne point à point entre la forme prédite $\hat{\mathbf{s}} = [\hat{x}_1, \hat{y}_1, \dots \hat{x}_{n_{pts}}, \hat{y}_{n_{pts}}]^T$ et la vérité terrain $\mathbf{s} = [x_1, y_1, \dots x_{n_{pts}}, y_{n_{pts}}]^T$:

$$\varepsilon_{ali} = \frac{1}{d.n_{pts}} \sum_{i=1}^{n_{pts}} \sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2} \tag{1.1}$$

La distance interocculaire d est utilisée comme facteur de normalisation pour comparer les méthodes indépendamment de la taille du visage dans l'image. d correspond soit à la distance entre les deux commissures externes des yeux (d_{outer}), soit à la distance inter-pupilles (d_{inter}). C'est cette dernière qui sera retenue dans le suite de ce document. Lorsque le visage présente de fortes variations de pose, la distance interocculaire n'est plus pertinente et on lui préfèrera la largeur de la fenêtre de détection du visage (d_{head}).

Une seconde manière de comparer les performances des détecteurs de points consiste à afficher la courbe de distribution des erreurs.

1.2.2.2 Evaluation des systèmes de classification

Le taux de reconnaissance est le pourcentage d'exemples dont la prédiction correspond à la vérité terrain. Il est particulièrement utilisé dans les problèmes multi-classes (reconnaissance d'une expression prototypique par exemple). Dans la cas d'un problème à deux classes (la détection de l'activation d'une AU), il est moins souvent utilisé car il dépend fortement de la proportion d'exemples de chaque classe, et on lui préfèrera donc les mesures F1 ou AUC décrites ci-après.

La mesure F1 est adaptée aux problèmes de détection (activation des AU par exemple). Deux indicateurs sont généralement calculés, le rappel ou taux de bonne détection qui compare le nombre de détections correctes (VP) au nombre total d'exemples positifs (VP+FN), et la précision qui compare le nombre de détections correctes (VP) au nombre total de détections (VP+FP). La mesure F1 correspond à la moyenne harmonique de ces deux valeurs :

$$F_1 = 2.\frac{\text{précision . rappel}}{\text{précision + rappel}} \tag{1.2}$$

Cette mesure dépend fortement du point de fonctionnement du système : pour un système fournissant une valeur continue en sortie (la probabilité d'activation d'une AU par exemple), ses performances peuvent varier en fonction du choix du seuil associé à cette valeur.

L'aire sous la courbe ROC (*Area Under Curve*, AUC). La courbe d'efficacité du récepteur (*Receiver Operating Characteristic curve*, ROC) est tracée en reportant le taux de bonne détection (ou rappel) en fonction du taux de fausse alarme pour différents points de fonctionnement du détecteur. L'aire sous cette courbe donne donc une information sur les performances d'un détecteur, indépendamment de son point de fonctionnement.

1.2.2.3 Evaluation des systèmes de régression

L'erreur quadratique moyenne (*Root Mean Square Error, RMSE*) est une métrique communément utilisée pour évaluer des systèmes de régression. La RMSE entre n_{ex} prédictions $\hat{y}_i$ et leur vérité terrain y_i est donnée par :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_{ex}}(y_i - \hat{y}_i)^2}{N_{ex}}} \tag{1.3}$$

Cette métrique est souvent combinée à une mesure de corrélation.

Le Coefficient de Corrélation (CC) de Pearson est une mesure de corrélation centrée réduite :

$$CC = \frac{1}{n_{ex}-1}\sum_{i=1}^{n_{ex}}\left(\frac{y_i - \bar{y}}{s_y}\right)\left(\frac{\hat{y}_i - \bar{\hat{y}}}{s_{\hat{y}}}\right) \tag{1.4}$$

avec

$$s_y = \sqrt{\frac{1}{n_{ex}-1}\sum_{i=1}^{n_{ex}}(y_i - \bar{y})}$$

l'écart type des valeurs de la vérité terrain et $s_{\hat{y}}$ l'écart type des valeurs prédites (calculé de manière similaire).

Le Coefficient de Corrélation Intraclasse (ICC) est également communément utilisé. L'ICC entre deux signaux est donné par :

$$ICC = \frac{1}{s.(n_{ex}-1)}\sum_{i=1}^{n_{ex}}(y_i - \bar{y}^*)(\hat{y}_i - \bar{y}^*) \tag{1.5}$$

avec

$$\bar{y}^* = \frac{1}{2n}\sum_{i=1}^{n_{ex}}(y_i + \hat{y}_i)$$

et

$$s^2 = \frac{1}{2(n_{ex}-1)}\left(\sum_{i=1}^{n_{ex}}(y_i - \bar{y}^*)^2 + \sum_{i=1}^{n_{ex}}(\hat{y}_i - \bar{y}^*)^2\right)$$

Contrairement à la corrélation de Pearson, l'ICC est normalisé par rapport à une moyenne et un écart type calculés à partir du regroupement de la vérité terrain et des prédictions. Ainsi deux signaux très proches, à une transformation linéaire près, auront une corrélation de Pearson élevée mais une corrélation intra-classe plus faible.

1.3 Plan du mémoire

Le rapport d'Habilitation se présente sous la forme de 7 chapitres : le présent chapitre d'introduction suivi de 5 autres chapitres, centrés autour des défis de l'analyse faciale, et d'un dernier chapitre de conclusion.

Le chapitre 2 vise à contextualiser et à positionner nos recherches par rapport à l'état de l'art et permet d'avoir une vue d'ensemble de nos contributions dans le domaine de l'analyse des expressions faciales.

Le chapitre 3 est consacré au problème de l'hétérogénéité des descripteurs. Nous commençons par évaluer la meilleure combinaison "descripteur / fonction noyaux" pour un classifieur SVM (*Support Vector Machine*) avant de proposer une nouvelle fonction adaptée à la prise en compte de l'identité de la personne. Pour clore ce chapitre, nous proposons d'utiliser un classifieur SVM multi-noyaux pour combiner des descripteurs hétérogènes ou multi-échelles et montrons la pertinence de l'approche pour détecter des AU et localiser des points caractéristiques sur un visage.

Le chapitre 4 aborde la problématique du manque de données d'apprentissage et propose des améliorations d'une méthode de régression par noyaux qui prennent en compte ces contraintes. En particulier, nous proposons une nouvelle formulation multi-tâches fortement contrainte qui limite le nombre de paramètres à estimer tout en conservant un pouvoir d'expressivité fort. Cette stratégie s'est révélée particulièrement efficace pour estimer l'intensité des AU.

Le chapitre 5 s'intéresse à la reconnaissance en temps réel d'expressions faciales en environnement faiblement contraint. Les approches proposées reposent sur des extensions de l'algorithme des forêts aléatoires qui permettent une prise en compte de l'information temporelle et assurent une robustesse aux variations de poses et aux occultations. Nous proposons également une nouvelle procédure d'apprentissage et d'évaluation d'un modèle à l'interface des forêts aléatoires et des réseaux de neurones.

Le chapitre 6 illustre l'intérêt des recherches présentées dans ce manuscrit au travers de deux projets collaboratifs. Le premier projet intègre un module de reconnaissance d'expressions prototypiques dans un jeu sérieux pour des enfants avec autisme et le second exploite la méthode d'estimation de l'intensité des AU pour analyser des corpus et apprendre des modèles d'attitudes sociales implantés sur des agents conversationnels animés .

Le chapitre 7 résume nos contributions et propose quelques directions de recherche pour nos travaux futurs.

Chapitre 2

Positionnement et contributions

L'analyse automatique des visages est un domaine de recherche très actif avec de nombreuses équipes nationales et internationales travaillant sur le sujet. En une décennie, la recherche publique mais également privée —portée à la fois par les grands acteurs du numérique tels que les GAFAM et par de nombreuses start-up spécialisées dans le domaine telles que Affectiva, Real Eyes, Datakalab, SenseTime, ou Megvii— ont accentué l'effort de recherche. Et des avancées majeures à tous les niveaux de la chaîne de traitement ont eu un impact déterminant sur les résultats actuels. L'objectif ici n'est pas tant de faire un état de l'art de l'analyse des expressions faciales (le lecteur pourra se référer par exemple à [122], [93] et [160]), que de positionner nos travaux et nos contributions en lien avec les défis précédemment cités et l'état d'avancement du domaine au moment de cette recherche.

Par ailleurs, nous avons fait le choix de présenter conjointement les avancées de toute la chaine de traitements car chaque étape de cette chaine a permis de proposer de nouvelles méthodes et des améliorations significatives des maillons suivants. Les méthodes que nous avons développées et qui seront présentées dans ce document sont, pour la plupart, génériques et ont d'ailleurs été évaluées sur plusieurs tâches afin de mettre en lumière leur capacité à opérer sur des données et des problèmes de types différents.

2.1 2000-2008 : les premiers grands succès de l'analyse faciale

Le début des années 2000 a été marqué par l'émergence des premiers systèmes de détection de visages performants [148, 51, 110]. En particulier le détecteur de Viola et Jones [148] dont le succès a été en partie impulsé par son intégration dans la librairie OpenCV, reste le détecteur de visage le plus couramment utilisé même si ses performances sont en deçà des meilleures méthodes actuelles [167].

Ces détecteurs ont alors ouvert la voie à une analyse plus fine du visage et en particulier à la localisation des points caractéristiques. A cette période, ce sont les méthodes dites génératives qui s'imposent. Elles visent à modéliser les variations d'apparence et de forme d'un visage à partir d'une base d'apprentissage puis, en phase de test, l'objectif est d'estimer automatiquement le jeu de paramètres qui décrit au mieux ce que l'on observe dans l'image. En particulier les Modèles Actifs de Forme (ASM, Active Shape Models, [27]) et les Modèles Actifs d'Apparence (AAM, Active Appearance Models, [26]) font l'objet de nombreuses recherches [95, 44, 99, 18, 50, 142]. Mais les performances de l'époque restreignent le cadre d'utilisation de ces méthodes : les modèles sont appris spécifiquement sur une personne dont le visage doit être filmé de face avec de bonnes conditions d'éclairage [56]. Ces performances s'expliquent en grande partie par un manque de données : les images étaient souvent sans annotations [132] ou annotées avec un nombre restreint de points sur le visage [70] et présentaient très peu de variations (par exemple [108]). Des premiers modèles 3D inspirés des AAM font également leur apparition mais leur utilisation dans un cadre d'analyse faciale reste limitée car l'optimisation des paramètres nécessite beaucoup de ressources et souvent une intervention manuelle pour l'initialisation des paramètres [118].

Les premiers systèmes de reconnaissance émotionnelle voient également le jour à la fin des années 90 [42, 92] et se développent dans les années 2000. Les systèmes de localisation de points caractéristiques étant encore peu fiables et peu répandus, la chaine de traitement se résume le plus souvent à prédire l'émotion à partir des pixels extraits de la fenêtre de détection. De nombreuses recherches se concentrent alors sur l'étape de représentation. Elles évaluent la pertinence des descripteurs d'apparence tels que les ondelettes de Haar [156], les filtres de Gabor [13, 10], les histogrammes de motifs binaires locaux [128, 65] (LBP, *Local Binary Patterns*) ou les histogrammes de gradients orientés [66] (HOG, *Histogram of Oriented Gradients*). Cette description de l'apparence locale s'accompagne le plus souvent d'une étape de réduction de dimension : Analyse en Composante Principale [13] (PCA, *Principal Component Analysis*), Analyse en Composante Indépendante [24] (ICA, *Independant Component Analysis*) ou Analyse Linéaire Discriminante [11] (LDA, *Linear Discriminant Analysis*). Quelques travaux s'intéressent également à la prise en compte de l'information dynamique au niveau des descripteurs [174, 143] ou du prédicteur [179, 58, 172]. Les méthodes d'analyse émotionnelle sont optimisées et évaluées sur des bases de données avec un nombre restreint d'échantillons et une faible variabilité (environnement fortement contraint, émotions non spontanées, peu d'individus différents) [73, 112, 123] ou des tâches relativement simples telles que la détection du sourire [155].

2.2 2008-2012 : l'essor de l'analyse des expressions faciales

Progressivement, la communauté scientifique se structure et intensifie ses efforts pour mettre à disposition de nouvelles bases de données et des outils plus performants :

- Les bases de données présentent plus de variabilité et des annotations plus riches. Par exemple, les bases de localisation de points caractéristiques ont un nombre accru d'images [14] et de points annotés par image [82]. Les bases de données d'analyse émotionnelle sont moins prototypiques et les tâches abordées sont souvent plus complexes et variées (reconnaissance des *Actions Units* [146], analyse fine des sourires [3, 41], analyse de la douleur [5]).
- La diffusion de codes sources et de fichiers exécutables pour la localisation de points caractéristiques [99, 121, 159] ou l'analyse des expressions faciales [86] permet de concentrer les efforts de recherche sur des aspects précis de la chaine de traitement.
- L'organisation de campagnes d'évaluation internationales telles que FERA (*Facial Expression Recognition and Analysis challenge*) ou AVEC (*Audio/Visual Emotion Challenge*) offre la possibilité de confronter équitablement des systèmes d'analyse faciale.

D'un point de vue méthodologique, les descripteurs d'apparence décrits précédemment sont encore majoritairement utilisés [38, 133, 29, 162] mais les méthodes intègrent progressivement l'information issue des systèmes de localisation de points caractéristiques devenus de plus en plus performants [99, 121, 159]. La position de ces points est utilisée, soit pour atténuer les variations d'apparence (par exemple, extraction de zones d'intérêts définies à partir de la position des points [133], ou transfert de la texture vers une forme canonique [21]), soit pour extraire de nouvelles informations ayant trait à la géométrie du visage (distance entre points, paramètres du modèle de forme....).

Pour l'étape de décision, plusieurs classifieurs ont été étudiés tels qu'Adaboost [162, 161] et les réseaux de neurones [164] ; mais ce sont les classifieurs SVM (Machines à Vecteurs Supports, *Support Vector Machine*) qui sont de loin les plus utilisés pour détecter les expressions faciales car ils sont particulièrement performants sur des jeux de données contenant un nombre restreint d'exemples de grande dimension.

Quelque soit le prédicteur utilisé, se pose la question de la sélection et de la combinaison

des descripteurs. Dans une stratégie de fusion *a priori*, plusieurs types de descripteurs peuvent être concaténés en un unique vecteur [135]. Cette étape peut également s'accompagner d'une réduction de dimension (à l'aide, par exemple, d'une PCA [38]). Le principal inconvénient est que les différences de dynamique entre les descripteurs ne sont pas prises en compte. Certains travaux privilégient alors les stratégies de fusion *a posteriori*. Par exemple, Srivastava *et al.* [133] apprennent deux SVM distincts et fusionnent leurs sorties en fonction d'un indice de confiance lié à la distance d'un exemple à l'hyperplan de chaque SVM. Ainsi les prédictions sont moins sensibles aux différences de dynamique mais elles n'exploitent pas les corrélations entre les types de descripteurs. Dans une stratégie intermédiaire, Meng *et al.* [98] utilisent un classifieur SVM-2K permettant d'apprendre conjointement deux SVM tout en cherchant à maximiser les corrélations entre les sorties de ces deux SVM.

Positionnement : c'est dans ce contexte que se sont inscrits les travaux des thèses de Thibaud Sénéchal (2008–2011) et Vincent Rapp (2009–2013) que j'ai co-encadrés avec Lionel Prevost. Ces travaux seront développés plus en détail au chapitre 3. Nous avons focalisé nos recherches sur le choix de descripteurs performants et la manière de les combiner afin d'exploiter au mieux leur complémentarité tout en prenant en compte les différences de dynamique.

Ainsi, nous avons commencé par comparer les performances de différents descripteurs images et de différentes fonctions noyaux pour reconnaitre des émotions. Nous avons proposé une combinaison descripteur / fonction noyau particulièrement performante et nous avons également proposé une nouvelle fonction noyau permettant de prendre en compte l'information d'identité de la personne.

Afin de compléter l'information issue des descripteurs d'apparence, nous avons proposé l'utilisation d'un SVM à noyaux multiples (MK-SVM, *Multi Kernel SVM*). Cette approche permet ainsi de prendre en compte les correlations entre descripteurs tout en respectant les dynamiques propres à chaque catégorie de descripteurs. Nous avons notamment appliqué cette stratégie à deux tâches distinctes de l'analyse faciale :

— La reconnaissance d'émotions et d'Action Units : nous utilisons un SVM multi-noyaux pour combiner l'information des descripteurs géométriques et d'apparence. Cette méthode a obtenue les meilleures performances lors de la première campagne d'évaluation internationale FERA'11 [146].

— La localisation de points caractéristiques : nous avons proposé une architecture de localisation de points en deux étapes. Dans un premier temps, on sélectionne très rapidement les positions potentielles des points caractéristiques à l'aide d'un SVM à noyaux linéaires multiples qui combine l'information des niveaux de gris de l'image à différentes échelles spatiales. On affine la recherche de ces points, dans un second temps, à l'aide de descripteurs d'apparence extraits dans le voisinage des positions potentielles. Les temps de traitements sont plus longs (extraction de descripteurs d'apparence et utilisation de noyaux non-linéaires pour le classifieur SVM) mais concentrés sur les quelques candidats sélectionnés à la première étape.

Par ailleurs, dans le cadre du stage de Master de Jérémie Nicolle, nous nous sommes intéressés à la prédiction de quatre dimensions affectives : la valence – caractère positif ou négatif de l'émotion –, l'arousal –niveau d'excitation ou de calme de la personne –, le contrôle – niveau de dominance ou de soumission – et la spontanéité de l'émotion. La nature du problème est alors bien différente de la reconnaissance des AU. Il ne s'agit plus d'un problème de classification (l'AU est-elle activée ?) mais de régression (quelle est l'intensité pour chaque dimension affective ?) à partir de signaux multi-modaux, le visage et la voix, et pour lesquels la prise en compte de l'information dynamique est primordiale. il est en effet extrêmement difficile de pré-

dire certaines dimensions affectives à partir d'une seule image. La méthode proposée [107]) a obtenu les meilleurs résultats lors de la campagne d'évaluation AVEC 2012 [124]. Nous retrouvons dans ces travaux un exemple caractéristique d'une recherche portant sur la conception de descripteurs adaptés à la nature du problème ainsi que d'une stratégie de combinaison de descripteurs hétérogènes. Nous ne détaillerons pas plus cette méthode dans la suite de ce manuscrit (le lecteur pourra toutefois se référer à [107]) car, à la suite des ces travaux, nous avons fait le choix de recentrer nos efforts de recherche sur l'analyse des expressions faciales (en particulier la prédiction des AU et des expressions prototypiques) car cette représentation nous semblait plus pertinente pour les applications visées, et en particulier celles que nous avons développées dans les projets collaboratifs dans lesquels nous étions impliqués (cf. chapitre 6). Pour autant, ces travaux furent précurseurs d'un point de vue méthodologique de ce que nous avons développé par la suite pour estimer l'intensité des AU (chapitre 4).

2.3 2012-2018 : vers l'analyse faciale "*in the wild*"

Malgré les avancées significatives des systèmes présentés précédemment, ces derniers ne permettaient pas encore de répondre pleinement aux besoins applicatifs en environnement faiblement contraint. En particulier, nous nous sommes intéressés à cette période, à deux défis majeurs :

— le faible nombre de données annotées,

— la grande variabilité des données, liée en particulier aux variations de poses et aux occultations.

Bien évidemment, ces deux défis sont intrinsèquement liés puisque le nombre de données qui est nécessaire pour apprendre des modèles prédictifs augmente à mesure que la variabilité des données augmente. Mais nous avons choisi de les traiter séparément dans ce document car ils correspondent à des travaux distincts : les travaux de thèse de Jérémie Nicolle (2012–2015) d'une part, qui présentent des stratégies d'apprentissage permettant de limiter les risques de sur-apprentissage [106, 104] et les travaux de thèse d'Arnaud Dapogny (2013–2016), d'autre part, dans lesquels nous proposons des méthodes explicitement robustes aux variations de poses [33, 31] et à la présence d'occultations [37].

2.3.1 Robustesse au faible nombre de données d'apprentissage

Une première solution consiste à agir en amont dans la chaine de traitement, en construisant un espace de representation adapté. De nombreuses méthodes s'appuyant sur des techniques de réduction de dimension (ACP [13], Factorisation en Matrices non Négatives [69]...) ou de sélection de descripteurs [55] ont été proposées, dont certaines intègrent la tâche [11, 152] (i.e. les annotations de la base d'apprentissage) et le prédicteur [111].

Pour augmenter la robustesse au sur-apprentissage, le deuxième levier sur lequel on peut agir est le choix du modèle prédictif. Une première solution consiste à utiliser un modèle d'une complexité limitée, et réputé peu sensible au sur-apprentissage. C'est le cas par exemple des SVM et des SVR (*Support Vector Regression*) à noyaux linéaires [53]. Lorsque le prédicteur utilisé est instable (e.g, des arbres de décision ou des réseaux de neurones), il est alors recommandé de combiner les prédictions issues de plusieurs modèles (via une stratégie de *bagging* [16] par exemple) [154, 2].

Enfin, il est également possible d'injecter des connaissances a priori dans la modélisation des données. Certains travaux s'intéressent à la modélisation dynamique des relations inter-AU. Par exemple, Tong *et al.* [138] et Li *et al.* [83] utilisent des réseaux bayésiens dynamiques pour capturer cette connaissance, alors que Zhao *et al.* [176] exploitent conjointement les dépendances

entre les descripteurs et les fortes corrélations entre les AU. De telles stratégies améliorent les taux de reconnaissance, mais introduisent un biais qui n'est pas souhaitable pour certaines applications (par exemple, dans le cas d'une rééducation faciale à la suite d'un accident vasculaire cérébral, certains muscles doivent être entrainés indépendamment et par conséquent reconnus séparément par le système).

Positionnement : portés par les très bonnes performances au challenge AVEC 2012, nous nous sommes intéressés aux algorithmes d'apprentissage de métrique adaptés aux méthodes de régression par noyaux (MLKR, *Metric Learning for Kernel Regression* [153]) car elles sont à la fois intelligibles (i.e. les résultats peuvent s'interpréter facilement) et capables de modéliser des fonctions complexes. Ces méthodes ont toutefois une complexité algorithmique élevée et sont sensibles à la fois à la dimension de l'espace d'origine et au nombre d'exemples d'apprentissage. Ce qui nous a amenés à proposer, lors de la thèse de Jérémie Nicolle, une nouvelle formulation multi-tâches qui permet de réduire les risques de suraprentissage en entrainant un modèle avec moins de paramètres que dans une formulation multi-tâche classique tout en conservant son pouvoir d'expressivité. Ce modèle a été évalué sur une tâche d'alignement de points caractéristiques et d'estimation de l'intensité des AU et a obtenu les meilleures performances lors de la seconde campagne d'évaluation internationale FERA'15. Ces travaux seront détaillés au chapitre 4.

2.3.2 Robustesse aux fortes variations d'apparence

2.3.2.1 Variations intrinsèques

Les premières sources de variations de l'apparence d'un visage sont propres à la personne qui produit l'expression faciale. On parle alors de variations intrinsèques. Par exemple, lorsque l'on observe le visage d'une personne sur une photo, il n'est pas facile d'en déduire son expression faciale. Les raisons en sont multiples : nous ne connaissons ni l'apparence du visage de la personne lorsqu'elle ne produit pas d'expression (on parle alors de visage neutre) ni la manière dont elle produit une expression. Ce problème est alors atténué lorsque l'on dispose du visage neutre de la personne et que l'on s'intéresse aux variations d'apparence de ce visage (i.e. à l'information relative entre le visage expressif et le visage neutre). Ainsi Khademi et Morency [75] apprennent des détecteurs de transition d'AU et Chu *et al.* [23] appliquent une stratégie d'apprentissage par transfert pour adapter la fonction de prédiction à une personne en particulier. Dans la plupart des cas, il n'est pas possible d'avoir une représentation de l'état neutre de la personne. Il est alors très utile d'utiliser l'information dynamique, i.e. reconnaitre l'expression à l'aide d'une vidéo plutôt que d'images traitées indépendamment. L'information temporelle est alors exploitée, soit au niveau de la représentation soit au niveau du prédicteur. Dans le premier cas, on trouve les méthodes qui utilisent des descripteurs spatio-temporels [64, 131, 72]. Le principal inconvénient est que ces descripteurs sont extraits à partir de fenêtres de taille fixe, parfois à différentes résolutions temporelles, alors qu'une expression faciale est un phénomène très dynamique avec de fortes disparités intra- et inter-personnelles dans la vitesse d'exécution de chaque phase de la production émotionnelle. Dans le second cas, les méthodes cherchent à établir un lien entre l'apparence observée et des états latents de la séquence, à l'aide par exemple de Chaînes de Markov Cachées [150] ou de Réseau Bayésiens Dynamiques [151]. Dans ce type d'approches les variations sont intrinsèquement prises en compte, mais l'espace de représentation doit généralement être de faible dimension et l'apprentissage s'appuie sur des séquences annotées, ce qui réduit le nombre de données disponibles par rapport aux méthodes de prédiction image par image puisqu'il n'y a qu'une seule annotation par vidéo.

Positionnement Durant la thèse d'Arnaud Dapogny, nous nous sommes intéressés aux forêts aléatoires (RF, *Random Forests*) car elles sont à la fois robustes au faible nombre de données d'apprentissage (en partie grâce au *bagging*) et capables de modéliser des fonctions complexes en utilisant une grande variété de descripteurs hétérogènes. Afin d'intégrer l'information temporelle, nous avons proposé le modèle *Pairwise Conditional Random Forests* (PCRF) qui entraîne des arbres de décision à reconnaitre des transitions d'émotions à partir de descripteurs statiques et dynamiques entre une image courante et des images prises précédemment dans la séquence. Le choix du classifieur de transition est conditionné par l'émotion reconnue sur la première image de la paire. Ainsi la tâche de chaque classifieur est simplifiée puisqu'il doit apprendre à reconnaitre la transition entre une émotion spécifique et toutes les autres. Les décisions de ces classifieurs de transition prises à plusieurs instants de la séquence sont ensuite recombinées pour obtenir une décision robuste pour l'image courante qui prend en compte l'information temporelle de manière très flexible (i.e. indépendant de la vitesse de production de l'émotion) et sans imposer de contraintes sur la cohérence temporelle de la séquence aussi bien en apprentissage qu'en test. Ces travaux constituent la première partie du chapitre 5

2.3.2.2 Variations extrinsèques

Les secondes sources de variations de l'apparence d'un visage ne dépendent plus de la personne elle-même mais de son environnement, par exemple des conditions d'illumination, du placement de la caméra par rapport au visage ou des éléments qui peuvent partiellement occulter ce visage. Pour améliorer la robustesse aux variations d'illumination, la plupart des méthodes reposent sur des descripteurs géométriques (qui ne sont pas sensibles à ce type de variations à condition que les points caractéristiques aient été correctement localisés) et des descripteurs d'apparence telles que SIFT [90] ou HOG [30] qui effectuent des normalisations locales. Il n'y a, par contre, pas de consensus dans la manière d'aborder les autres sources de variations extrinsèques. Les approches robustes aux variations de poses d'un visage peuvent se décomposer en trois grandes catégories. La première catégorie regroupe les méthodes qui tentent d'apprendre un unique classifieur capable d'appréhender les fortes variations d'apparence, quelque soit le point de vue du visage [177, 136, 47]. Mais de telles approches peuvent avoir des difficultés à capturer la variabilité des expressions faciales lorsque le nombre d'exemples d'apprentissage devient important. La deuxième catégorie de méthodes cherche alors à réduire cette variabilité en apprenant à transformer une vue non-frontale en une image de visage de face [147, 119]. Il s'agit d'un problème complexe [63, 178] nécessitant des données d'apprentissage avec des couples d'images d'une même personne avec la même expression dans différentes poses ou des données 3D qui ne sont pas toujours accessibles en fonction de l'application visée [147]. La dernière catégorie de méthodes cherche à apprendre un classifieur spécifique à chaque point de vue. En test, une étape d'estimation de pose permet de sélectionner le classifieur adapté [101]. Ainsi, chaque classifieur est appris sur des jeux de données plus homogènes et donc, chaque classifieur pourra en théorie capturer plus efficacement les déformations liées aux expressions. De plus, les temps de traitements seront comparables à ceux d'un classifieur frontal, sans nécessiter une étape parfois coûteuse de transformation de l'image vers une vue frontale. Enfin, cela réduit l'empreinte mémoire au moment de l'apprentissage, puisque chaque classifieur est entrainé sur une portion de la base d'apprentissage. En contrepartie, ce type de méthodes nécessite des données d'expression pour chaque classe de pose (ce que l'on peut obtenir en utilisant des bases de données issues d'un scanner 3D [163]) et une estimation robuste de l'orientation de la tête (les systèmes actuels de localisation de points caractéristiques fournissent d'excellents résultats pour des visages présentant $\pm 45°$ de rotation horizontale et $\pm 30°$ de rotation verticale). Les dernières grandes sources de variations extrinsèques proviennent des occultations liées par exemple

à la présence d'éléments au premier plan (autres objets de scène, mains de la personne...) ou d'accessoires sur le visage (lunettes de soleil, foulards, mèches de cheveux...). Kotsia *et al.* [79] ont montré que l'occultation partielle du visage, et en particulier de la zone de la bouche, a un impact très important sur les capacités de reconnaissance des expressions prototypiques, aussi bien pour les humains que pour les systèmes automatiques. En l'absence de bases de données d'expressions faciales avec occultations, certains travaux génèrent des motifs synthétiques sur les images d'apprentissage [28, 52], mais les capacités de généralisation sont souvent limitées en présence d'occultations réalistes. A l'inverse, Huang *et al.* [67] cherchent à détecter automatiquement les parties occultées, mais cette approche reste encore très peu flexible car la détection est binaire et limitée à trois zones du visage. Enfin, il est également possible d'apprendre des modèles génératifs de visages non-occultés qui sont ensuite utilisés en test pour générer une vue synthétique d'un visage initialement occulté. Cela implique d'être capable d'apprendre un tel générateur, tâche à la fois ardue et très couteuse en temps de calcul [84].

Positionnement : Dans la continuité des premiers travaux de thèse d'Arnaud Dapogny, nous avons étendu la méthode PCRF en conditionnant le choix des arbres de décision, non plus uniquement en fonction de l'émotion prédite dans la première image, mais également en fonction de l'orientation du visage. Le modèle multi-vue PCRF (MVPCRF) est alors robuste aux fortes variations de poses.

Dans un second temps, nous nous sommes intéressés à la robustesse aux occultations, et nous avons proposé de combiner des arbres de décision aléatoires définis sur des sous-espaces locaux du visage formant des prédictions d'expressions définies localement (prédictions locales d'expression). Nous avons également appris un réseau hiérarchique de mémoires auto-associatives, lequel est entrainé à reconstruire l'apparence du visage non-occulté. Ces deux éléments permettent, étant donnée une image potentiellement occultée, de définir une mesure de confiance dans chaque partie du visage, laquelle peut être utilisée pour pondérer les prédictions locales d'expression, et ainsi fournir une classification d'expression robuste aux occultations. Les AU étant par nature des phénomènes locaux en lien avec les expressions faciales, nous avons également montré que les prédictions locales d'expression sont des descripteurs pertinents pour détecter les AU. Cette méthode est décrite à la section 5.3 du chapitre 5

2.4 L'émergence de l'apprentissage profond (*Deep Learning*)

Il est de nos jours impossible d'ignorer la prédominance des réseaux de neurones profonds dans le domaine de l'apprentissage statistique, et en particulier pour les applications de vision par ordinateur. Ce succès remarquable s'explique par le concours de plusieurs facteurs :

- L'augmentation des capacités de stockage et de calcul des ordinateurs, avec notamment l'utilisation intensive des processeurs graphiques (GPU, *Graphics Processing Unit*) et plus récemment des processeurs dédiés (par exemple, les TPU, *Tensor Processing Unit*).
- L'augmentation des données collectées et la possibilité de sous traiter à faible coût l'annotation de ces données, via des services tels que *Mechanical Turk* ou *Crowdflower* par exemple.
- La mise à disposition de nouveaux modèles et d'outils logiciels adaptés à une grande quantité de données et nécessitant d'importantes ressources de calcul (Tensorflow, (py)Torch, Caffee, CNTK, Theano...)
- La flexibilité des approches permettant de concevoir très facilement des architectures complexes et adaptées à chaque problème (en combinant notamment des modules de trai-

tement et en définissant des fonctions de coût spécifiques et parfois intermédiaires pour guider l'apprentissage)

— L'adaptabilité des modèles à de nouveaux domaines (en ajustant les paramètres du réseau sur de nouvelles données qui peuvent avoir été capturées dans un environnement différent) ou de nouvelles tâches (en conservant les premières couches du réseau pré-entrainé et en redéfinissant l'architecture des dernières couches).

Dans le domaine de l'analyse des expressions faciales, c'est la tâche de localisation des points caractéristiques qui a profité la première de l'émergence du *deep learning*, car il s'agit de la tâche pour laquelle nous disposons de plus de données d'apprentissage et dont la vérité terrain est la plus fiable. Les premières méthodes proposées ont repris le principe de l'alignement par cascade qui donnait de très bons résultats [20, 159, 7] : en partant d'une estimation grossière de la position de ces points (typiquement la position moyenne estimée à partir de la base d'apprentissage), l'objectif est d'estimer itérativement le déplacement des points qui améliore cette estimation initiale. Cette estimation s'appuie sur une extraction de l'apparence autour de la position courante du modèle et d'une fonction de régression qui prédit le déplacement à partir de cette apparence. Sun *et al.* [134] ont substitué les étapes d'extraction de caractéristiques et de régression de chaque étage de la cascade par des réseaux de neurones à convolution (CNN, *Convolutionnal Neural Networks*) alors que Zhang *et al.* [169] proposent une cascade d'auto-encodeurs profonds. Dans l'approche proposée par Trigeorgis *et al.* [139], des réseaux à convolution récursifs sont appris de bout en bout. Le réseau peut ainsi mémoriser les déplacements précédents et proposer une stratégie de déplacement plus cohérente tout au long des itérations de cette cascade. Zhang *et al.* [173] s'affranchissent de la cascade et cherchent à apprendre un réseau capable de prédire la position des points caractéristiques en une passe, limitant ainsi les temps de traitement par rapport aux approches utilisant de multiples CNN. Compte tenu de la complexité du modèle à apprendre, cette approche nécessite une large base de données annotées avec des attributs auxiliaires (pose, genre, âge...) pour aider le réseau à apprendre une représentation pertinente. Les travaux les plus récents vont dans ce sens, en ajoutant pendant la phase d'apprentissage des tâches supplémentaires (e.g. génération d'une image de contour du visage [158], estimation de cartes de probabilité de présence d'un point caractéristique [80]) ou de nouveaux domaines qui peuvent être synthétisés à l'aide de réseaux antagonistes génératifs (GAN, Generative Adversarial Networks) par exemple [43]. Ce type de stratégie a pour effet de régulariser l'apprentissage des réseaux qui deviennent de plus en plus complexes.

Le succès des réseaux profonds est toutefois plus mitigé en ce qui concerne les autres tâches d'analyse des expressions. Les bases de données proposent des tâches de plus en plus complexes (intensité des AU [144]) et des données capturées dans des environnements non contraints avec des annotations souvent peu fiables [40], mais le nombre de données est insuffisant pour que les méthodes d'apprentissage profond soient pleinement opérationnelles. On pourra par exemple citer les résultats très mitigés obtenus lors du Challenge FERA [59]. Plusieurs stratégies ont toutefois permis de profiter en partie des avantages de l'apprentissage profond :

— En apprenant des modèles peu profonds [76, 68] ou pré-entrainés sur d'autres jeux de données [113, 49, 103].

— En utilisant des descripteurs images *ad hoc* (*handcrafted*) et la position des points caractéristiques du visage pour simplifier la tâche du réseau [71, 62, 68].

— En combinant les sorties de plusieurs prédicteurs [77, 165, 89]

— En apprenant conjointement, tout comme pour la localisation de points caractéristiques, plusieurs tâches pour régulariser l'apprentissage [88, 175, 85]

Les récentes bases de données disposent toutefois de plus en plus de données avec des annotations issues de plusieurs annotateurs [166] et permettent désormais d'entrainer des méthodes

d'apprentissage profond qui atteignent des résultats compétitifs.

Positionnement : Les derniers travaux de thèse d'Arnaud Dapogny, actuellement poursuivis dans le cadre de la thèse d'Estèphe Arnaud, portent sur l'adaptation d'un modèle d'apprentissage récent, les Neural Forests [78], pour l'analyse du visage. Ces algorithmes sont des modèles hybrides entre les *Random Forests* et les réseaux de neurones, et offrent une forte expressivité au prix d'un temps de calcul important. Nous avons proposé des modifications à la fois dans l'algorithme d'apprentissage, permettant notamment un apprentissage complètement incrémental pouvant être réalisé via un algorithme d'optimisation quelconque (descente de gradient stochastique, RMSProp, ADAM...), et à la fois dans l'évaluation de ces modèles, permettant un traitement temps-réel pour l'analyse du visage. Plus particulièrement, cette méthode se base sur l'évaluation "gloutonne" (modèle GNF ou *Greedy Neural Forest*) de ces arbres de décision probabilistes.

Nous avons proposé une application des GNF à la reconnaissance des émotions, permettant notamment de connecter des réseaux de neurones à convolution, afin d'apprendre l'extraction de caractéristiques pertinentes, simultanément aux paramètres d'un prédicteur GNF, avec un temps de calcul divisé par 300 par rapport à une NF traditionnelle. Nous avons également proposé d'utiliser les GNF dans le cadre de l'alignement de points caractéristiques par cascade de régression. La méthode proposée améliore significativement l'état de l'art et permet un traitement largement temps réel. La section 5.4 présente cette méthode et les premiers résultats obtenus.

Chapitre 3

SVM Multi-Noyaux pour la combinaison de descripteurs hétérogènes

L'objectif des recherches présentées de ce chapitre est d'exploiter au mieux les performances d'un classifieur SVM qui était de loin le modèle prédictif le plus performant et le plus utilisé par la communauté pour reconnaitre des expressions faciales. Ainsi nous nous sommes tout d'abord interrogés sur le choix du meilleur couple descripteur / fonction noyau pour une tâche de reconnaissance des expressions faciales (3.1). Nous nous sommes ensuite intéressés à la manière d'intégrer des connaissances *a priori* telles que l'identité de la personne (3.2). Pour finir, ces recherches nous ont amenés à proposer des méthodes orientées SVM multi-noyaux pour combiner des descripteurs hétérogènes ou multi-échelles (3.3), que nous avons appliquées à la reconnaissance des AU (3.3.1) et à la localisation de points caractéristiques (3.3.2).

3.1 Choix du descripteur et de la fonction noyau

3.1.1 Les descripteurs d'apparence

L'utilisation de descripteurs d'apparence vise à extraire l'information pertinente par rapport au phénomène étudié. Par exemple, des ondelettes de Haar sont bien adaptées pour détecter un visage dans une image, mais ne sont pas forcément pertinentes pour caractériser des expressions subtiles. Dans notre cas, nous souhaitons une représentation qui conserve l'information relative aux expressions faciales tout en la dissociant d'autres facteurs tels que l'identité ou l'illumination. Nous avons évalué deux représentations classiquement utilisées en analyse d'expressions faciales, les images de Gabor et les motifs binaires locaux (LBP), et nous avons montré l'intérêt de l'utilisation des motifs binaires locaux de Gabor (LGBP) pour aborder cette tâche.

Les images de Gabor En deux dimension, les filtres de Gabor sont des ondes sinusoïdales planes modulées par un noyau gaussien. L'écriture complexe des filtres de Gabor est de la forme :

$$G_{\mathbf{k}}(\mathbf{z}) = \frac{\mathbf{k}^2}{\sigma^2} e^{(-\frac{\mathbf{k}^2}{2\sigma^2}\mathbf{z}^2)}(e^{i\mathbf{kz}} - e^{-\frac{\sigma^2}{2}}) \tag{3.1}$$

avec σ, l'ecart type de la fonction gaussienne et $\mathbf{k} = \nu e^{i\theta}$, le vecteur d'onde caractéristique dont ν correspond à la fréquence spatiale et θ à l'orientation. L'image de Gabor $\mathbf{I}_G$ est le module de la convolution de l'image d'origine par le filtre de Gabor pour une orientation et une fréquence donnée.

Les motifs binaires locaux (LBP) Cet opérateur permet de caractériser les textures et les motifs réguliers dans une image indépendamment des changements de luminosité. Chaque pixel de l'image est codé par un nombre sur 8 bits en seuillant son voisinage 3×3 par sa propre valeur.

Ainsi la valeur du LBP associé à un pixel $\mathbf{p}$ ayant pour voisinage $\{f_i, i = 0 \dots 7\}$, sera :

$$LBP(\mathbf{p}) = \sum_{i=0}^{7} \delta(f_i - f_p) 2^i \tag{3.2}$$

$$\delta(x) = \begin{cases} 1 & \text{si} \quad x \geq 0 \\ 0 & \text{si} \quad x < 0 \end{cases}$$

Motifs binaires locaux de Gabor (LGBP) Les cartes LGBP sont obtenues en appliquant l'opérateur LBP sur les cartes de Gabor, exploitant ainsi les liens entre les pixels pour plusieurs résolutions et orientations.

Représentation par histogramme Lorsque la localisation du visage n'est pas très précise, la représentation par histogramme peut-être tout à fait adaptée puisqu'elle permet d'attester la présence de certains motifs (une ride d'expression par exemple) sans considérer sa position dans l'image. Il s'agit donc d'une représentation robuste aux erreurs de localisation. Si, toutefois, on souhaite conserver une partie de l'information spatiale (l'apparition d'une ride au niveau du front n'aura pas le même sens qu'une ride aux coins des yeux), le calcul des histogrammes peut se faire sur des régions spécifiques de l'image. Ainsi, l'histogramme $\mathbf{h}^{\mathbf{r}\theta\nu}$ d'une région r de la carte LGBP $\mathbf{I}^{\theta\nu}_{\text{LGBP}}$:

$$\mathbf{h}^{\mathbf{r}\theta\nu}(i) = \sum_{\mathbf{p} \in \mathcal{R}(r)} L(i \leq \mathbf{I}^{\theta\nu}_{\text{LGBP}}(\mathbf{p}) < i + 1), \quad i = 0 \dots 255 \tag{3.3}$$

$$L(A) = \begin{cases} 1 & \text{si A est vrai} \\ 0 & \text{sinon} \end{cases}$$

La figure 3.1 illustre un exemple d'extraction des histogrammes de LGBP. Dans les expériences

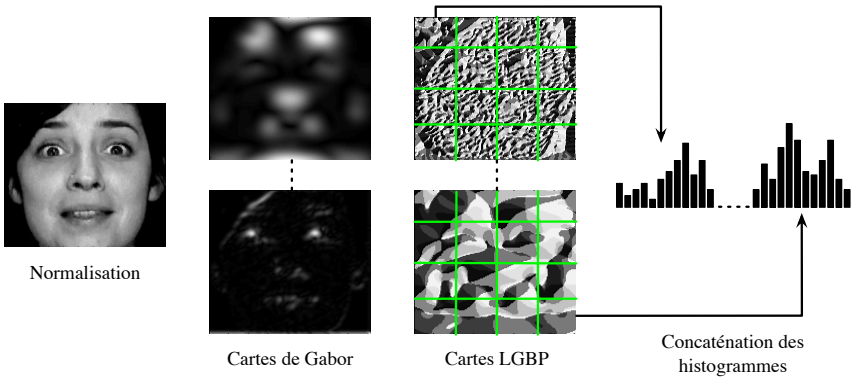


FIGURE 3.1 – Extraction des histrogrammes LGBP

que nous avons menées (section 3.1.3), nous utilisons trois fréquences spatiales $\nu = (\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{8})$, 6 orientations $\theta = \left(\frac{k\pi}{6}, k \in \{0 \dots 5\}\right)$ et 16 régions. Ainsi le vecteur de description final est de dimension $3 \times 6 \times 16 \times 256 = 73728$.

Réduction de la dimension des histogrammes Afin de réduire les temps de calcul en apprentissage et en test, nous avons proposé une nouvelle méthode de réduction de la dimension des histogrammes de LGBP qui consiste à regrouper les fréquences d'apparition de plusieurs motifs dans la même classe. Ce *clusturing* s'appuie sur une distance de Hamming (i.e. le nombre de bits

qui diffèrent entre deux séries binaires) entre un LBP et chaque "représentant" de classe. Ces représentants correspondent aux motifs LBP uniformes (i.e. n'ayant au plus que deux transitions) [109] qui ne contiennent qu'un nombre pair de 1. Nous avons choisi ce sous-ensemble car tous les motifs uniformes sont à une distance de 1 avec au moins un motif de ce sous-ensemble et il n'y a aucune paire de motifs de ce sous-ensemble dont la distance de Hamming est inférieure à 2. Les fréquences d'apparition des autres motifs sont ensuite comptabilisées dans la classe du représentant qui est le plus proche d'eux. Si un motif a plusieurs plus proches voisins, alors sa fréquence d'apparition est équitablement répartie entre les classes des motifs racines proche de lui. La dimension finale du vecteur de description est 7488.

3.1.2 Les fonctions noyaux

Le choix de la fonction noyau a un impact fort sur les performances et les capacités de généralisation du SVM. Les travaux antérieurs en reconnaissance des expressions faciales se sont souvent limités à des fonctions linéaires, polynomiales ou gaussiennes [87], même dans le cas d'approches basées sur des histogrammes [102]. Nous avons proposé d'utiliser le noyau intersection d'histogrammes, qui a obtenu de bon résultats dans le domaine de la reconnaissance d'objets et n'a pas d'hyperparamètres à optimiser, ce qui est un avantage important lorsque l'on dispose de peu d'exemples d'apprentissage. Dans ces travaux, nous avons évalué les performances des fonctions noyaux suivantes :

— Fonction linéaire :

$$k(\mathbf{h}_i, \mathbf{h}_j) = \mathbf{h}_i^T.\mathbf{h}_j \tag{3.4}$$

— Fonction polynomiale :

$$k(\mathbf{h}_i, \mathbf{h}_j) = (\mathbf{h}_i^T.\mathbf{h}_j)^\gamma \tag{3.5}$$

— Fonction gaussienne :

$$k(\mathbf{h}_i, \mathbf{h}_j) = e^{-\frac{\left\|\mathbf{h}_i - \mathbf{h}_j\right\|_2^2}{2\sigma^2}} \tag{3.6}$$

— Fonction d'intersection d'histogrammes :

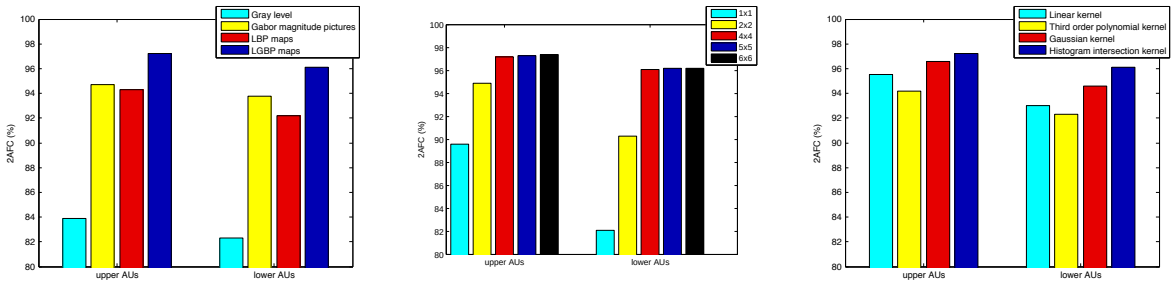
$$k(\mathbf{h}_i, \mathbf{h}_j) = \sum_k min(h_{i,k}, h_{j,k}) \tag{3.7}$$

3.1.3 Résultats expérimentaux

Les résultats expérimentaux menés sur la base de données CK [73], en respectant le protocole proposé par Bartlett *et al.* [12], nous ont amené aux conclusions suivantes [125] :

— **Influence du descripteur :** quelque soit la partie du visage, haute (AU 1, 2, 4, 5, 6, 7 et 9) ou basse (AU 11, 12, 15, 17, 20, 23, 24, 25 et 27), les histogrammes LBP et des images de Gabor donnent de bien meilleurs résultats que les histogrammes des niveaux de gris mais c'est la combinaison des deux, les histogrammes LGBP, qui aboutit aux meilleurs résultats (cf. figure 3.2a). Par ailleurs, les expériences sur la subdivision de l'image 128×128 en régions régulières ont montré qu'une grille 4×4 offrait le meilleur compromis entre performance et temps de traitement (cf. figure 3.2b).

— **Influence du noyau :** la fonction polynomiale d'ordre 3 (l'ordre du modèle a été choisi par validation croisée) et la fonction linéaire obtiennent les moins bonnes performances. La fonction gaussienne donne des résultats légèrement inférieurs à ceux de la fonction intersection d'histogrammes et nécessite l'optimisation de l'hyper paramètre σ (écart-type de la gaussienne). La fonction noyau d'intersection d'histogrammes semble donc la plus adaptée pour reconnaître les AU quand nous utilisons des descripteurs de type histogramme (cf. figure 3.2c).



(a) (b) (c)

FIGURE 3.2 – Evaluation des performances de la reconnaissance des AU de la partie haute et basse du visage en fonction du choix du descripteur 3.2a, du nombre de régions 3.2b et du choix du noyau 3.2c

3.2 Prise en compte de l'identité

Les bonnes performances obtenues avec des histogrammes LGBP peuvent s'expliquer par leur faculté à décorréler l'expression de l'identité (ce qui expliquerait également les performances en reconnaissance faciale). Une manière de rendre le système moins sensible à l'identité est de lui fournir une information propre au sujet que l'on cherche à analyser. Nous avons proposé pour cela de calculer un nouveau descripteur correspondant à la différence entre l'histogramme LGBP $\mathbf{h}$ de l'image du visage dont on souhaite analyser l'expression et l'histogramme $\mathbf{h}^{\text{neutre}}$ d'une image du même sujet mais avec une expression neutre :

$$\mathbf{\Delta h} = \mathbf{h} - \mathbf{h}^{\text{neutre}} \tag{3.8}$$

Comme $\mathbf{\Delta h}$ comporte des valeurs négatives et que la somme des valeurs de chaque classe n'est pas constante, nous ne pouvons pas appliquer directement la fonction noyau d'intersection d'histogramme. Nous avons alors proposé une nouvelle fonction noyau, la fonction HDI (pour *Histogram Difference Intersection*) :

$$k(\mathbf{\Delta h}_i, \mathbf{\Delta h}_j) = \sum_{r,\theta,\nu} \frac{\sum_n \text{minabs}\left(\Delta h_{i,n}^{r\theta\nu}, \Delta h_{j,n}^{r\theta\nu}\right)}{\sqrt{\sum_n \left|\Delta h_{i,n}^{r\theta\nu}\right| . \left|\Delta h_{j,n}^{r\theta\nu}\right|}} \tag{3.9}$$

$$minabs(x,y) = \begin{cases} \min(|x|,|y|) & \text{si} \quad xy > 0 \\ 0 & \text{sinon} \end{cases}$$

avec $\mathbf{h}_i^{r\theta\nu}$ la sous partie de l'histogramme $\mathbf{h}_i$ calculée sur la région r de la carte des LGBP de fréquence spatiale ν et d'orientation θ. Ainsi, le score de la fonction noyau HDI entre deux différences d'histogrammes est élevé si les différences d'histogrammes varient dans le même sens et avec la même amplitude (normalisation par la somme des valeurs absolues de chaque différence d'histogramme). Les résultats expérimentaux ont montré une amélioration des performances par rapport à l'utilisation d'histogrammes LGBP calculés uniquement à partir du visage expressif.

Si nous ne disposons pas de l'image du visage neutre (ce qui est assez fréquent dans de nombreux cas d'utilisation réalistes), nous avons également montré qu'il était possible d'en créer une version synthétique en projetant l'histogramme du visage expressif dans un espace propre construit à partir d'histogrammes de visages neutres. Les performances obtenues avec cet histogramme neutre synthétisé sont proches de celles obtenues en utilisant l'histogramme authentique,

et surpasse l'état de l'art [12, 138]. Ainsi, il est possible de s'affranchir de la contrainte du visage neutre, sans dégrader significativement les résultats.

3.3 Combinaisons de descripteurs hétérogènes

Dans les travaux précédents, nous avons concentré nos efforts sur le choix du meilleur couple descripteur/fonction noyau. Toutefois, en faisant le choix d'un seul type de descripteur, il se peut qu'une partie de l'information importante soit perdue. Par exemple, dans le cas des histogrammes LGBP, une partie de l'information spatiale est supprimée, ce qui rend le descripteur plus robuste aux erreurs d'alignement. Mais en contrepartie, ce descripteur peut difficilement capturer des informations de distances telle que la position du sourcil par rapport à l'œil. Concaténer directement cette nouvelle information au vecteur de caractéristique $\mathbf{h}$ ne semble pas non plus pertinent car l'étendue et la dynamique de ces valeurs et le type de noyau adapté à chaque représentation peuvent être très différents. Nous avons donc choisi d'utiliser l'apprentissage multi-noyaux pour combiner l'information issue de descripteurs hétérogènes [54]. Il s'agit d'un mode de combinaison intermédiaire [15], à mi-chemin entre la fusion a priori (les descripteurs sont concaténés et fournis en entrée d'un unique classifieur) et a posteriori (chaque catégorie de descripteur est associée à un classifieur spécifique et la décision finale résulte de la combinaison des sorties de chaque classifieur). Dans le cas d'un SVM multi-noyaux, l'objectif est d'apprendre conjointement l'hyperplan séparateur et la pondération β_j associée à chaque fonction noyau. Ainsi le score de similitude entre deux exemples décrit par un ensemble de descripteurs hétérogènes est donné par la combinaison convexe de K noyaux :

$$k = \sum_{j=1}^{K} \beta_j k_j \quad \text{avec } \beta_j \geq 0, \sum_{j=1}^{K} \beta_j = 1 \tag{3.10}$$

Nous avons appliqué l'apprentissage multi-noyaux à deux cas d'utilisation : à la détection des AU (3.3.1) et à la localisation de points caractéristiques (3.3.2).

3.3.1 Détection des AU

Afin de remédier à la perte d'informations spatiales inhérente aux histogrammes de LGBP, nous avons choisi d'utiliser les coefficients d'un modèle actif d'apparence (AAM) qui apportent cette information spatiale importante mais qui, contrairement aux histogrammes LGBP, dépendent d'un alignement précis du modèle. Ainsi, l'objectif de la fusion est de combiner la précision des AAM à la robustesse des histogrammes LGBP.

Les travaux présentés dans cette thèse sont le fruit d'une collaboration (projet ANR Immemo) entre l'ISIR et l'équipe de Renaud Séguier à Centrale-Supélec qui a développé le modèle AAM 2.5D utilisé dans ces travaux [1]. Deux AAM locaux, le premier centré sur la zone des yeux et le second sur la zone de la bouche, ont été entrainés sur la base Bosphorus pour laquelle la position 3D des points caractéristiques du visage est connue [123]. Le vecteur de caractéristiques $\mathbf{c}$ est obtenu par concaténation des 29 composantes du modèle de la bouche et des 41 composantes du modèle des yeux.

La fusion des descripteurs est réalisée par un SVM multi-noyaux dont le score de similitude entre deux exemples $\mathbf{x} = (\mathbf{h}, \mathbf{c})$ et $\mathbf{x}_i = (\mathbf{h}_i, \mathbf{c}_i)$ est :

$$k(\mathbf{x}_i, \mathbf{x}) = \beta_1 k_{\text{LGBP}}(\mathbf{h}_i, \mathbf{h}) + \beta_2 k_{\text{AAM}}(\mathbf{c}_i, \mathbf{c}) \quad \text{avec } \beta_j \geq 0, \beta_1 + \beta_2 = 1 \tag{3.11}$$

avec k_{LGBP}, une fonction intersection d'histogrammes et k_{AAM}, une fonction gaussienne. Pour

ajouter une cohérence temporelle, les prédictions image par image sont ensuite lissées à l'aide d'un filtre moyenneur, dont la taille est optimisée pour chaque AU par validation croisée.

Les résultats expérimentaux présentés dans [126] et [127] ont montré l'apport de la fusion des descripteurs hétérogènes. Cette méthode a été testée dans le cadre de la campagne d'évaluation internationale FERA 2011 [146]. Comme le montrent les résultats présentés dans la Figure 3.3, ce système s'est avéré plus performant que les autres méthodes pour détecter les AU.

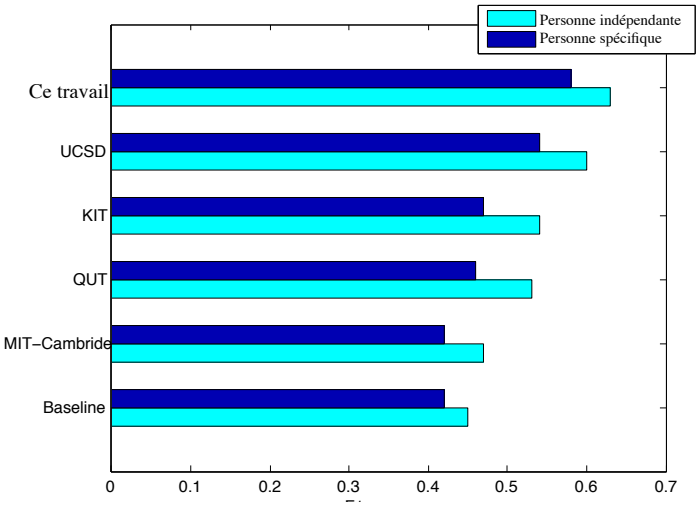


FIGURE 3.3 – Résultats officiels (scores F1) de la compétition FERA'11. USCD : University of San Diego [157]. KIT : Karlsruhe Institute of Technology. QUT : Queensland University of Technology [22]. MIT-Cambridge : Massachusetts Institute of Technology et Cambridge University [9]

3.3.2 Localisation des points caractéristiques

Dans les méthodes de type *Contrained Local Model* (CLM), on distingue deux étapes principales : une première étape de détections indépendantes des points caractéristiques suivie d'une seconde étape d'alignement d'un modèle déformable qui cherche un compromis entre les positions probables des points dans l'image et les contraintes internes du modèle de forme. Dans Saragih *et al.* [121], la première étape de détection est assurée par un SVM linéaire appliqué sur des images en niveaux de gris. Dans la formulation primale du SVM, cela revient à apprendre un motif discriminant (encodé par les paramètres de l'hyperplan séparateur du SVM) et à effectuer une convolution de l'image avec ce motif pour localiser un point dans une région de l'image. Cette opération est très rapide, mais la carte de probabilité des positions des points caractéristiques est très bruitée, ce qui complexifie l'étape d'alignement du modèle.

Dans nos travaux, nous avons cherché à améliorer l'étape de localisation pour combiner plusieurs descripteurs à plusieurs échelles spatiales. Afin de ne pas sacrifier les temps de traitement aux performances, nous avons proposé une méthode d'estimation des cartes de probabilité de présence des points caractéristiques en deux temps.

Dans un premier temps (étage $\mathcal{S}_1$ de la figure 3.4), nous apprenons un SVM linéaire multi-noyaux. La sortie non-seuillée de ce classifieur pour chaque pixel de la zone est donnée par la combinaison linéaire des cartes issues de la convolution de l'image d'origine, avec les motifs encodés dans les hyperplans associés à chaque échelle. De cette carte de réponses dense, on ne sélectionne que les maxima locaux correspondant aux positions les plus probables de présence du point recherché.

Dans un second temps (étage $\mathcal{S}_2$), nous apprenons à prédire la probabilité que chaque candi-

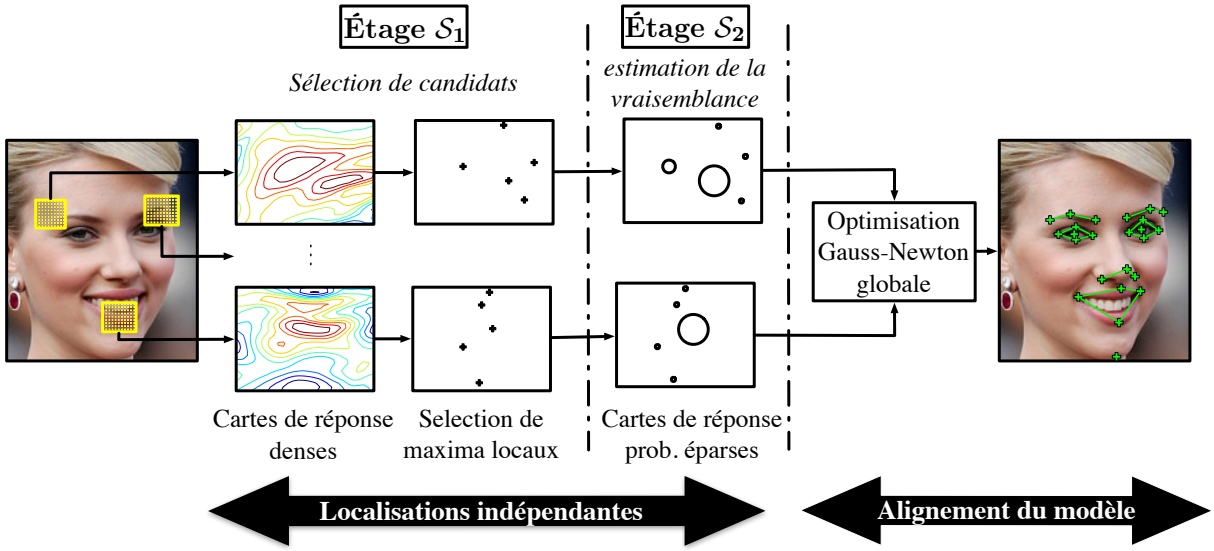


FIGURE 3.4 – Vue d'ensemble de la méthode d'alignement de points caractéristiques

dat sélectionné soit le point recherché en apprenant à un SVM à discriminer les bonnes positions des mauvaises, et en calibrant la sortie de ce SVM [114]. Compte tenu du nombre restreint de candidats à analyser, le classifieur SVM peut s'appuyer sur des descripteurs image variés et sur des fonctions noyaux plus élaborées (cf. figure 3.5).

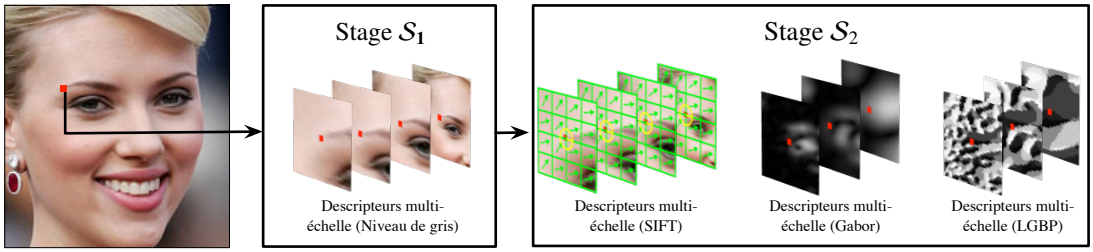


FIGURE 3.5 – Illustration des descripteurs multi-échelles utilisés pour chaque étage

A l'issue de l'étape de localisation indépendante des points, certaines détections peuvent être fausses (e.g. lorsque le point est occulté) ou imprécises (e.g. l'apparence locale de certains points le long de la mâchoire ne permet pas leur localisation précise et, seule leur position relative par rapport aux autres points, permet de lever cette ambiguïté). Une étape d'alignement d'un modèle déformable permet d'apporter une cohérence spatiale entre tous les points du visage. L'estimation des paramètres du modèle de forme est obtenue par minimisation d'une fonction de coût composée de l'erreur d'alignement (i.e. la somme des distances de chaque point aux pixels candidats, pondérée par la probabilité d'être le point recherché) et d'un terme de régularisation qui pénalise les déformations peu probables.

Les expériences menées sur plusieurs bases de données de la littérature ont permis de montrer les bonnes performances en généralisation de notre méthode, ainsi que l'apport de chaque étape du processus de localisation. En particulier, cette méthode est capable de suivre les déformations du visage liées aux variations d'expressions pour des visages frontaux ou présentant de faibles variations de pose ($\pm 35°$) et obtient des résultats équivalents à l'état de l'art [116].

Publications en lien avec le chapitre

(T1) T. Senechal. *Ce que le visage révèle : Analyse des mouvements faciaux pour l'interprétation émotionnelle*, **Thèse de doctorat.** 2011

(T2) V. Rapp. *Analyse du visage pour l'interprétation de l'état émotionnel*, **Thèse de doctorat.** 2013

(J1) T. Senechal, K. Bailly et L. Prevost. *Impact of Action Unit Detection in Automatic Emotion Recognition*, **Pattern Analysis and Applications**, **17** (1) : 51-67, 2014.

(J2) V. Rapp, K. Bailly, T. Senechal, L. Prevost. *Multi-Kernel Appearance Model*, **Image and Vision Computing**, **31** (8) : 542-554, 2013.

(J3) T. Sénéchal, V. Rapp, H. Salam, R. Seguier, K. Bailly et L. Prevost. *Facial Action Recognition Combining Heterogeneous Features via Multi-Kernel Learning*, **IEEE Transactions on Systems, Man, and Cybernetics–Part B**, **42** (4) : 993-1005, 2012.

(C1) V. Rapp, T. Senechal, K. Bailly, L. Prevost. *Multiple Kernel Learning SVM and Statistical Validation for Facial Landmark Detection*, **Automatic Face and Gesture Recognition (FG'2011)**.

(C2) T. Sénéchal, V. Rapp, H. Salam, R. Seguier, K. Bailly, L. Prevost. *Combining LGBP Histograms with AAM coefficients in the Multi-Kernel SVM framework to detect Facial Action Units*, **Facial Expression Recognition and Analysis Challenge (FERA'2011)**.

(C3) T. Senechal, K. Bailly, L. Prevost. *Automatic Facial Action Detection Using Histogram Variation Between Emotional States*, **Proc. of Int'l Conference on Pattern Recognition (ICPR'2010)**.

Encadrement doctoral

— Thibaud Sénéchal [9/2008–11/2011] (dir. Lionel Prevost)

— Vincent Rapp [12/2009–7/2013] (dir. Lionel Prevost)

Projet collaboratif

— Projet ANR **IMMEMO** (2009–2012) www.rennes.supelec.fr/immemo/

Chapitre 4

Analyse des expressions faciales par apprentissage de métriques

Dans la continuité de nos travaux sur la reconnaissance des AU présentés précédemment, les recherches décrites dans ce chapitre visent à caractériser plus finement les expressions faciales, en prédisant l'intensité des activations musculaires. Il s'agit d'un problème de régression particulièrement propice au sur-apprentissage, car nous ne disposons que de peu de données au regard de la complexité de la tâche. Ainsi, nous avons proposé des extensions de la méthode MLKR (*Metric Learning for Kernel Regression*, 4.1) qui ont permis de réduire la complexité algorithmique et les risques de sur-apprentissage (4.2). Nous avons également étendu la méthode aux problèmes de régression multi-labels (4.2.4), et la formulation H-MT-MLKR (4.2.4.3) a notamment permis de réduire le nombre de paramètres à estimer, tout en conservant un fort pouvoir d'expressivité du modèle. Pour finir, nous illustrons la généricité et la pertinence de cette méthode pour localiser des points caractéristiques sur le visage (4.3.1) et estimer l'intensité des AU (4.3.2).

4.1 Estimateur de Nadaraya-Watson et apprentissage de métriques

L'estimateur de Nardaraya-Watson est l'un des estimateurs les plus simples et les plus intuitifs pour une tâche de régression. Le label d'un exemple de test est la moyenne des labels de tous les exemples d'apprentissage, pondérée par la ressemblance de l'exemple de test aux exemples d'apprentissage. Soit $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_s}\}$, n_s exemples d'apprentissage et $\{y_1, y_2, \ldots, y_{n_s}\}$ les annotations correspondantes, le label associé à un nouvel exemple de test est donné par :

$$\hat{y}_t = \frac{\sum_{i=1}^{n_s} y_i k_{i,t}}{\sum_{i=1}^{n_s} k_{i,t}} \tag{4.1}$$

Le noyau $k_{i,t} = k(\mathbf{x}_i, \mathbf{x}_t)$ est une mesure de similarité entre les exemples $\mathbf{x}_i$ et $\mathbf{x}_t$. Le noyau le plus couramment utilisé est le noyau Gaussien :

$$k_{i,j} = \frac{1}{2\sigma\sqrt{2\pi}} e^{-\frac{d_{i,j}^2}{\sigma^2}} \tag{4.2}$$

avec σ, l'écart type de la Gaussienne et $d_{i,j} = d(\mathbf{x}_i, \mathbf{x}_j)$ la distance euclidienne entre deux exemples $\mathbf{x}_i$ et $\mathbf{x}_j$. Le seul hyper-paramètre de cette méthode est l'écart-type de la Gaussienne, qui peut être optimisé par une procédure de validation croisée, ce qui rend la méthode peu sensible au sur-apprentissage. Il est, de plus, possible d'approximer des fonctions fortement non linéaires, car la prédiction du label d'un exemple est principalement définie par son voisinage. Toutefois, les performances de l'estimateur de Nadaraya-Watson dépendent de l'espace de représentation des échantillons, qui doit être pertinent pour une tâche donnée. L'apprentissage de métrique pour la régression par noyaux (*Metric Learning for Kernel Regression*, MLKR), proposée par Weinberger et Tesauro [153], vise à apprendre un espace de représentation pertinent pour l'estimateur de Nadaraya-Watson. L'objectif est d'estimer le sous-espace linéaire optimal

qui minimise l'erreur quadratique de l'estimateur de Nadaraya-Watson. Soit n_d la dimension de l'espace d'origine et n_r la dimension de l'espace d'arrivée, la méthode MLKR estime la matrice de projection $\mathbf{A} \in \mathcal{M}_{n_r,n_d}(\mathbb{R})$ qui minimise l'erreur suivante :

$$\mathcal{L}(\mathbf{A}) = \sum_{i=1}^{n_s} (\hat{y}_i - y_i)^2 \tag{4.3}$$

avec

$$\hat{y}_i = \frac{\sum_{i \neq j} y_j k_{j,i}(\mathbf{A})}{\sum_{i \neq j} k_{j,i}(\mathbf{A})}$$

et

$$k_{i,j}(\mathbf{A}) = \frac{1}{2\sigma\sqrt{2\pi}} e^{-\frac{d_{i,j}(\mathbf{A})^2}{\sigma^2}}$$
$$d_{i,j}(\mathbf{A})^2 = \|\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\|^2 = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A}^T \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j)$$

la distance dans l'espace de projection de dimension n_r. Le minimum est obtenu par descente de gradient :

$$\frac{\partial \mathcal{L}(\mathbf{A})}{\partial \mathbf{A}} = 4\mathbf{A} \sum_i \frac{(\hat{y}_i - y_i)}{\sum_{i \neq j} k_{ij}} \sum_j (\hat{y}_i - y_j) k_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \tag{4.4}$$

La méthode MLKR a deux hyper paramètres : l'écart type de la gaussienne σ et la dimension de l'espace de projection n_r. Le label d'un exemple de test est alors donné par l'estimateur de Nadaraya-Watson, après projection des exemples dans le sous-espace défini par $\mathbf{A}$.

Outre sa métrique optimisée pour une tâche donnée, la méthode MLKR présente les mêmes avantages que l'estimateur de Nadaraya-Watson : les résultats sont facilement interprétables, elle permet d'approximer des fonctions complexes et elle a une bonne capacité d'extrapolation dans les régions de l'espace très faiblement peuplées (par opposition aux SVR par exemple).

Toutefois, la méthode MLKR a des limitations importantes qui doivent être levées pour être pleinement exploitable dans une application d'analyse faciale.

La complexité : le calcul du gradient pour une projection de n_s exemples de dimension n_d, a une complexité en $O(n_s^2 * n_d^2)$, ce qui ne permet pas de manipuler des exemples dans des espaces de grande dimension. Ce qui est le cas lorsqu'on analyse un visage, soit directement à partir des niveaux de gris, soit à partir de vecteurs de caractéristiques, comme nous l'avons vu au chapitre précédent. L'empreinte mémoire est également importante car, tout comme les autres méthodes à base de noyaux telles que les SVR, elle repose sur le calcul d'une matrice contenant l'ensemble des mesures de similarité entre toutes les paires d'exemples de la base d'apprentissage.

Les risques de sur-apprentissage : nous avons empiriquement montré que certaines configurations étaient critiques pour l'apprentissage du MLKR.

— Choix de l'écart type des noyaux gaussiens : si la variance est trop petite, le calcul du gradient est faux et instable, car il repose sur un nombre trop restreint d'échantillons (chaque exemple d'apprentissage a un nombre très limité de voisins). A l'inverse, si l'écart type est trop grand, tous les exemples seront approximativement à la même distance les uns des autres et l'estimation du gradient est alors constante à chaque itération. Il est donc essentiel d'estimer la valeur de ce paramètre par validation croisée.

— Influence du bruit : Lorsque les valeurs des descripteurs qui composent le vecteur de caractéristique est bruité, les performances du MLKR sont fortement impactées, car la mesure de similarité entre les exemples n'est plus pertinente et le calcul du gradient devient incohérent. Il est donc essentiel de limiter le nombre de descripteurs non informatifs pour améliorer les performances.

— Influence du nombre d'exemples : les résultats expérimentaux ont également montré que les performances étaient liées au nombre d'exemples d'apprentissage. Plus ce nombre est élevé, plus l'algorithme tolère de bruit sur les données.

— Influence du nombre de paramètres du modèle : le nombre de paramètres à estimer dépend du nombre d'éléments de la matrice $\mathbf{A}$ et augmente donc linéairement avec la dimension de l'espace de représentation n_d et de l'espace de projection n_r. La dimension de l'espace de projection est un hyper-paramètre qui doit être choisi avec soin par cross-validation car nous avons observé des risques de sous-apprentissage et de sur-apprentissage lorsque la dimension était respectivement trop faible ou trop grande.

4.2 Améliorations proposées

Nous avons évoqué, dans la section précédente, les limitations de la méthode MLKR, en terme de complexité algorithmique et de sur-apprentissage, lorsque les descripteurs sont très bruités ou que le nombre de paramètres est trop important au regard du nombre de données disponibles. Nous avons alors proposé plusieurs améliorations afin que MLKR puisse être utilisé dans un contexte d'analyse faciale. Nous avons tout d'abord proposé une étape de sélection de descripteurs (4.2.1) ainsi qu'une méthode de descente de gradient stochastique (4.2.2). Nous avons ensuite proposé une version régularisée de l'algorithme (4.2.3). Pour finir nous avons proposé trois extensions de la méthode MLKR pour la régression de labels multi-dimensionnels (4.2.4).

4.2.1 Sélection des descripteurs

Compte tenu de sa complexité algorithmique en $O(n_s^2 * n_d^2)$ et des problèmes d'apprentissage lorsque la dimension des exemples d'apprentissage est grande et comporte des dimensions pas ou peu informatives, nous avons proposé de sélectionner un sous-ensemble de descripteurs en amont de la phase d'apprentissage du MLKR. L'estimateur de Nadaraya-Watson étant fortement non linéaire, nous avons choisi d'utiliser l'entropie conditionnelle pour quantifier la relation fonctionnelle non-linéaire entre le label l et la valeur du descripteur f :

$$H(l|f) = -\sum_{f\in\mathcal{F}} p(f) \sum_{l\in\mathcal{L}} p(l|f) \log\left(p(l|f)\right) \tag{4.5}$$

avec $\mathcal{F}$ et $\mathcal{L}$, les ensembles de définition des descripteurs et des labels. Une estimation fine des probabilités conditionnelles pouvant être très longue lorsque le nombre d'échantillons est important, nous calculons une valeur approchée en discrétisant l'espace des descripteurs et des labels. Le choix des pas de quantification est important, puisqu'un pas trop grand ne permettra pas de capturer des relations non linéaires fines entre les descripteurs et les labels, et à l'inverse, un pas trop petit donnera une mauvaise estimation de l'entropie conditionnelle, car le nombre d'échantillons par région de l'espace sera trop faible. Ces hyper-paramètres seront estimés par validation croisée.

4.2.2 Descente de gradient stochastique

Le coût de calcul du gradient est quadratique par rapport au nombre d'échantillons considérés. Nous avons donc choisi d'utiliser la descente de gradient stochastique par lot (*batch stochastic gradient descent*, BSGD). Cette méthode consiste à n'utiliser qu'un sous-ensemble des exemples d'apprentissage (un *batch*) pour calculer la valeur du gradient, à chaque étape de la descente. Appliquée au MLKR, nous avons montré que la descente de gradient stochastique par lot converge plus rapidement et également vers une meilleure solution que celle obtenue par une descente de gradient classique. Cette amélioration s'explique par une réduction du sur-apprentissage induite par la sélection aléatoire d'un nouveau lot à chaque étape.

4.2.3 Régularisation de la fonction de coût

La version originale du MLKR estime les coefficients de la matrice de projection $\mathbf{A}$ qui minimisent l'erreur de prédiction. Afin de réduire les risques de sur-apprentissage, nous avons évalué l'intérêt d'ajouter un terme de pénalisation à la fonction de coût originale, et testé deux stratégies très couramment utilisées : la pénalisation Lasso et la norme $L2$.

4.2.3.1 Régularisation Lasso

Nous ajoutons à la fonction de coût un terme de pénalité Lasso correspondant à la norme $L1$ de la matrice $\mathbf{A}$ (la somme des valeurs absolues des coefficients de la matrice). Tibshirani [137] a montré que cette pénalité induisait de la parcimonie dans le vecteur de paramètres et réduisait les risques de sur-apprentissage. La nouvelle fonction de coût devient :

$$\mathcal{L}(\mathbf{A}) = \sum_{i=1}^{n_s} (\hat{y}_i - y_i)^2 + \gamma . L_1(\mathbf{A}) \tag{4.6}$$

avec γ le paramètre de régularisation qui contrôle le niveau de régularisation et peut être optimisé par validation croisée. Le gradient associé devient :

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 4\mathbf{A} \sum_{i=1}^{n_s} (\hat{y}_i - y_i) \sum_{j=1}^{n_s} (\hat{y}_j - y_j) k_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top + \gamma . \mathrm{sgn}(\mathbf{A}) \tag{4.7}$$

avec sgn(.) la fonction signe.

4.2.3.2 Régularisation $L2$

Nous pouvons également pénaliser la fonction de coût par la norme $L2$ de la matrice $\mathbf{A}$. Dans ce cas la fonction de coût s'écrit :

$$\mathcal{L}(\mathbf{A}) = \sum_{i=1}^{n_s} (\hat{y}_i - y_i)^2 + \gamma . \|\mathbf{A}\|^2 \tag{4.8}$$

avec $\|.\|$ la norme de Frobenius. Le gradient correspondant à cette fonction de coût s'écrit :

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}} = 4\mathbf{A} \sum_{i=1}^{n_s} (\hat{y}_i - y_i) \sum_{j=1}^{n_s} (\hat{y}_j - y_j) k_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top + 2\gamma \mathbf{A} \tag{4.9}$$

Dans le cadre des applications visées, les deux régularisations, Lasso et $L2$, aboutissent à des

résultats équivalents en terme de représentation parcimonieuse de la matrice de projection et de performances. Nous avons privilégié la régularisation $L2$ dans la suite du document.

4.2.4 MLKR multi-labels

Dans cette partie, nous nous intéressons aux problèmes de régression multi-labels, i.e. lorsqu'on doit prédire plusieurs labels associés à un même exemple. Chaque label à prédire peut-être considéré comme une tâche spécifique, par exemple, une image de visage peut-être annotée avec l'intensité de plusieurs AU. Si l'on fait l'hypothèse que les tâches sont liées et ont une forte probabilité de partager un espace de représentation commun, alors un apprentissage conjoint de ces différentes tâches aboutira probablement à de meilleurs résultats que ceux obtenus par des modèles appris à partir de chaque tâche indépendamment.

4.2.4.1 MLKR à espace commun (CS-MLKR)

La première extension du MLKR multi-labels que nous avons proposée, consiste à apprendre un unique espace de représentation, dans lequel les exemples sont projetés pour prédire l'ensemble des labels. Soit $\mathcal{L}_t$ la fonction de coût associée à une tâche t, la fonction de coût associée au MLKR à espace commun (*Common-Space MLKR*, CS-MLKR) pour un problème à T tâches est :

$$\mathcal{L}_{CS} = \sum_{t=1}^{T} \mathcal{L}_t(\mathbf{A}) + \gamma \left\|\mathbf{A}\right\|^2 \tag{4.10}$$

Pour simplifier l'expression, on introduit :

$$\mathbf{D}_t = \sum_{i=1}^{n_s} \frac{(\hat{y}_i - y_i)}{\sum\limits_{j \neq i} k_{ij}} \sum_{j=1}^{n_s} (\hat{y}_i - y_j) k_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{4.11}$$

On obtient alors :

$$\frac{\partial \mathcal{L}_{CS}}{\partial \mathbf{A}} = 4 \sum_{t=1}^{T} \mathbf{A}\mathbf{D}_t + 2\gamma\mathbf{A} \tag{4.12}$$

La formulation CS-MLKR fait l'hypothèse d'une corrélation très forte entre les tâches, puisque l'espace de projection est le même, quelque soit la tâche. Si les tâches sont plus faiblement corrélées, l'information spécifique à chaque tâche sera difficilement capturée. Nous avons proposé une formulation multi-tâches pour répondre à ce problème spécifique.

4.2.4.2 MLKR multi-tâches (MT-MLKR)

À la manière des SVM mulit-tâches proposés par Evgeniou et Pontil [48], l'objectif du MLKR multi-tâches (*Multi-Task MLKR*, MT-MLKR) est de découvrir un espace de représentation commun à l'ensemble des tâches $\mathbf{B}_0$, tout en capturant les spécificités propres à chaque tâche t. On peut introduire cette représentation commune en décomposant la matrice de projection associée à une tâche :

$$\mathbf{A}_t = \mathbf{B}_0 + \mathbf{B}_t \tag{4.13}$$

La fonction de coût du MT-MLKR s'écrit alors :

$$\mathcal{L}_{MT} = \sum_{t=1}^{T} \mathcal{L}_t(\mathbf{B}_0 + \mathbf{B}_t) + \gamma \sum_{t=0}^{T} \left\|\mathbf{B}_t\right\|^2 \tag{4.14}$$

Un terme de pénalité est ajouté à la fonction de coût pour encourager $\mathbf{B}_0$ à contenir une représentation partagée par les différentes tâches et à conserver dans $\mathbf{B}_t$, uniquement les informations spécifiques à la tâche t.
Le gradient de cette fonction de coût s'écrit :

$$\frac{\partial \mathcal{L}_{MT}}{\partial \mathbf{B}_0} = 4\sum_{t=1}^{T}(\mathbf{B}_0 + \mathbf{B}_t)\mathbf{D}_t + 2\gamma\mathbf{B}_0 \tag{4.15}$$
$$\frac{\partial \mathcal{L}_{MT}}{\partial \mathbf{B}_t} = 4(\mathbf{B}_0 + \mathbf{B}_t)\mathbf{D}_t + 2\gamma\mathbf{B}_t$$

4.2.4.3 MLKR multi-tâches fortement contraint (H-MT-MLKR)

Nous avons proposé une régularisation multi-tâches plus contrainte que la formulation précédente. Cette méthode, appelée MLKR multi-tâches fortement contraint (*Hard Multi-Task MLKR, H-MT-MLKR*), force un nombre n_c d'axes à être partagés par les différents espaces de projection de taille n_r. Cela ce traduit dans la définition de l'espace de projection $\mathbf{A}_t$ par une concaténation à la place d'une somme :

$$\mathbf{A}_t = \left[\frac{\mathbf{B}_0}{\mathbf{B}_t}\right] \tag{4.16}$$

avec $\mathbf{B}_0 \in \mathcal{M}_{n_c,n_d}(\mathbb{R})$ and $\mathbf{B}_t \in \mathcal{M}_{n_r-n_c,n_d}(\mathbb{R})$.
Le gradient s'ecrit :

$$\frac{\partial \mathcal{L}_{HMT}}{\partial \mathbf{B}_0} = 4\mathbf{B}_0\sum_{t=1}^{T}\mathbf{D}_t + 2\gamma\mathbf{B}_0 \tag{4.17}$$
$$\frac{\partial \mathcal{L}_{HMT}}{\partial \mathbf{B}_t} = 4\mathbf{B}_t\mathbf{D}_t + 2\gamma\mathbf{B}_t$$

Le principal intérêt de la régularisation multi-tâches fortement contraint, est de réduire les risques de sur-apprentissage en limitant le nombre de paramètres à apprendre, tout en conservant la même dimension de l'espace de projection. Pour MT-MLKR, le nombre de paramètres est :

$$n_{par}^{MT} = n_d.n_r.(T+1) \tag{4.18}$$

Alors que le nombre de paramètres pour H-MT-MLKR est :

$$n_{par}^{HMT} = n_d.(n_c + T.(n_r - n_c)) \tag{4.19}$$

Dans le cas, par exemple, d'un problème à $T = 5$ tâches, la méthode MT-MLKR nécessite l'optimisation de $n_{par}^{MT} = 2400$ paramètres pour projeter $n_d = 80$ descripteurs de l'espace d'origine vers un espace de dimension $n_r = 5$. Pour la même configuration, la méthode H-MT-MLKR requière $n_{par}^{HMT} = 1040$ paramètres si $n_c = 3$ axes sont partagés. A complexité équivalente, la formulation fortement contrainte nécessite donc deux fois moins de paramètres que la formulation "classique", ce qui réduit potentiellement les risques de sur-aprentissage.

4.3 Application à l'analyse des expressions faciales

Nous allons à présent montrer l'intérêt de la formulation H-MT-MLKR dans deux applications d'analyse des expressions faciales : la localisation de points caractéristiques et l'estimation de l'intensité des *Action Units*. Les autres apports présentés au chapitre précédent ont également été évalués dans le cadre d'une application d'analyse d'expressions faciales dont les résultats sont reportés dans [106].

4.3.1 Localisation de points caractéristiques

La méthode proposée s'inscrit dans la lignée des méthodes par cascade (2.4) : après une initialisation grossière du modèle, la position des points caractéristiques du visage est affinée conjointement et itérativement. A chaque étape, un ensemble de caractéristiques est extrait dans le voisinage de chaque point, puis fourni en entrée du régresseur.

Extraction des descripteurs Des descripteurs de type histogramme de gradients orientés (HOG) à 2×2 cellules sont extraits à partir de fenêtres centrées sur chaque point caractéristique du visage. La taille de ces fenêtres dépend de l'itération courante, car à mesure que le modèle se rapproche de la vérité terrain, l'information importante est très locale. Ainsi nous réduisons la taille de fenêtre à chaque étape de la cascade.

Méthode de régression L'étape de régression s'appuie sur la méthode H-MT-MLKR que nous avons présentée précédemment. Nous avons défini 5 groupes de points caractéristiques (cf. Figure 4.1) qui constituent les 5 tâches de notre méthode de régression. L'objectif est alors d'apprendre

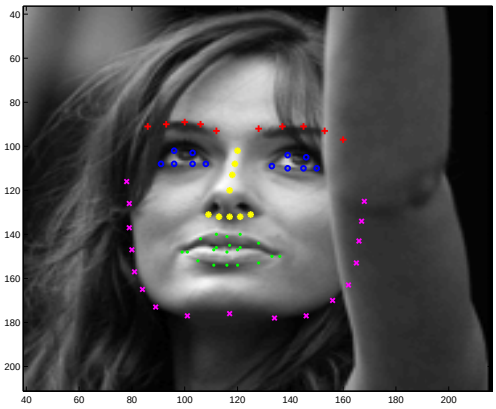
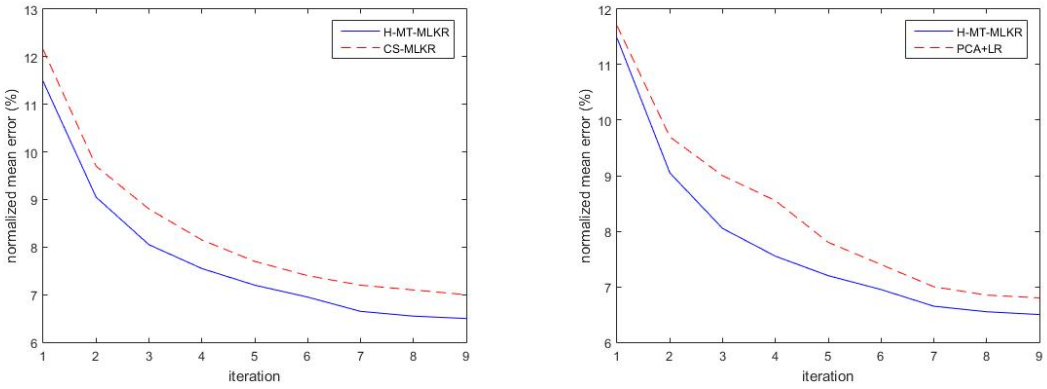


FIGURE 4.1 – Définition des 5 groupes de points caractéristiques

5 espaces de projection (un par groupe de points caractéristiques), chacun étant défini comme la concaténation d'axes communs (partagés par les 5 espaces) et d'axes spécifiques. Le fait d'imposer à chaque groupe un espace commun (et donc les mêmes descripteurs image) impose une contrainte spatiale entre les points. De plus, la cohérence spatiale est renforcée, car le déplacement de chaque point est une somme pondérée des déplacements de la base d'apprentissage (principe de l'estimateur de Nadaray-Watson). Les groupes ont été choisis de manière à rassembler les points spatialement proches et qui partagent des mouvements très corrélés (les deux yeux par exemple) et à séparer les points dont les mouvements sont souvent indépendants (les points des sourcils de ceux de la bouche par exemple).

Résultats expérimentaux Les résultats, présentés en détail dans la thèse de Nicolle [105], ont montré que l'approche H-MT-MLKR donne de meilleurs résultats d'alignement, quelque soit l'étape de la cascade, que l'approche CS-MLKR, dans laquelle les 5 espaces de projection ont été appris séparément (cf. Figure 4.2a). Ceci s'explique par le manque de contraintes géométriques entre les 5 groupes, induit par l'approche CS-MLKR. Lorsque l'on compare notre méthode à la méthode SDM (cf. Figure 4.2b), qui combine une projection par PCA et une régression linéaire (RL), on remarque que la méthode qui utilise H-MT-MLKR donne de meilleurs résultats, en particulier sur les dernières étapes de la cascade, où l'information extraite par notre méthode est plus locale, et donc plus utile pour un alignement précis. Pour finir le tableau 4.1 montre



(a) H-MT-MLKR vs CS-MLKR (b) H-MT-MLKR vs PCA+RL

FIGURE 4.2 – Comparaison de H-MT-MLKR par rapport à CS-MLKR 4.2a, et PCA+RL 4.2b. L'erreur d'alignement est donnée pour chaque étape de la cascade

que notre méthode offre des performances proches des meilleures méthodes de l'état de l'art au moment de ces travaux.

TABLE 4.1 – Comparaison entre la méthode orientée H-MT-MLKR et trois méthodes de l'état de l'art sur le jeu de données 300W

Méthode	Erreur moyenne normalisée (%)
ESR [20] (reported in [117])	7.58
SDM [159] (reported in [117])	7.52
LBF [117]	**6.32**
H-MT-MLKR	6.50

4.3.2 Estimation de l'intensité des *Action Units*

L'estimation de l'intensité des AU constitue un cadre d'étude tout à fait adapté à notre méthode H-MT-MLKR car :

— Les descripteurs sont nombreux et hétérogènes : comme nous l'avons vu au chapitre 3, il est pertinent d'utiliser des descripteurs géométriques (pour caractériser par exemple l'ouverture de la bouche) et des descripteurs d'apparence (pour caractériser par exemple l'apparition de rides d'expression).

— La relation entre ces descripteurs et l'intensité de l'AU est probablement non linéaire : par exemple, si l'on considère les rides qui apparaissent entre les sourcils lors de l'activation de l'AU4 (froncement de sourcils), l'extraction de contours verticaux est tout à fait pertinente et la relation entre l'intensité de l'AU et ce descripteur est certainement monotone, mais probablement pas linéaire. En effet, l'intensité des gradients peut varier très faiblement

lors des premiers niveaux d'activation de l'AU, puis l'apparition de rides marquées pour des niveaux d'activation élevés, peut provoquer une augmentation rapide de cette valeur.

— Les risques de sur-apprentissage liés au manque de données : l'annotation de l'intensité des AU dans une vidéo est un processus très coûteux et certaines AU sont très rarement activées lors de comportements naturels. À cela s'ajoutent les fortes variations interpersonnelles, qui font de la prédiction d'AU, un domaine très sensible au sur-apprentissage.

4.3.2.1 Extraction des descripteurs

Nous calculons deux types de descripteurs géométriques à partir de triplets de points. Pour chaque triplet $\mathbf{t}_{k_1k_2k_3} = (\mathbf{p}_{k_1}, \mathbf{p}_{k_2}, \mathbf{p}_{k_3})$, on calcule le ratio entre deux vecteurs

$$\mathbf{v}_{k_2k_3} = \mathbf{p}_{k_3} - \mathbf{p}_{k_2} = (\mathbf{p}^x_{k_3} - \mathbf{p}^x_{k_2}) + i.(\mathbf{p}^y_{k_3} - \mathbf{p}^y_{k_2})$$

et

$$\mathbf{v}_{k_2k_1} = \mathbf{p}_{k_1} - \mathbf{p}_{k_2} = (\mathbf{p}^x_{k_1} - \mathbf{p}^x_{k_2}) + i.(\mathbf{p}^y_{k_1} - \mathbf{p}^y_{k_2})$$

pour former

$$f(\mathbf{t}_{k_1k_2k_3}) = \frac{\mathbf{v}_{k_2k_1}}{\mathbf{v}_{k_2k_3}} = \frac{\|\mathbf{v}_{k_2k_1}\|}{\|\mathbf{v}_{k_2k_3}\|}.e^{i(\widehat{\mathbf{v}_{k_2k_3}, \mathbf{v}_{k_2k_1}})}$$

qui indique la position de $\mathbf{p}_{k_1}$ par rapport à $\mathbf{p}_{k_2}$ et $\mathbf{p}_{k_3}$. On utilise alors la norme et l'angle de $f(\mathbf{t}_{k_1k_2k_3})$ comme descripteurs géométriques. Il s'agit de descripteurs locaux, invariants aux changements d'échelle et aux rotations dans le plan image.

Les descripteurs d'apparence sont des HOG extraits à partir deux types de fenêtres : certaines sont centrées sur les points caractéristiques (pour capturer des rides d'expressions), alors que d'autres sont issues d'une division régulière de l'image du visage, afin de pouvoir récupérer des informations utiles, même lorsque l'étape d'alignement du modèle a échoué. Les images du visage ont été préalablement normalisées en les pivotant et en les redimensionnant par rapport à la position des yeux.

4.3.2.2 Résultats expérimentaux

Sur un total de 2768 descripteurs, nous sélectionnons $n_d = 80$ descripteurs en utilisant la somme des entropies conditionnelles (4.5) calculée pour chaque tâche (i.e. chaque AU). L'espace de projection est de dimension $n_r = 5$ donc $n_c = 3$ axes sont partagés. L'hyper-paramètre de régularisation $\gamma = 0,9$ a été obtenu par validation croisée. L'apprentissage utilise 10000 exemples choisis aléatoirement dans la base d'apprentissage.

Le tableau 4.2 présente les résultats obtenus par H-MT-MLKR, comparé aux modèles MLKR et MT-MLKR, en terme de corrélation de Pearson entre la prédiction et la vérité terrain. On remarque une amélioration significative des résultats pour 4 des 5 AU testées. L'amélioration est d'autant plus importante que l'AU est difficile à prédire (AU14 par exemple).

Le tableau 4.3 présente les performances obtenues par les participants à la campagne d'évaluation FERA 2015. Nous pouvons voir que notre approche (ISIR) a surpassé les autres méthodes quelque soit l'AU. L'ICC moyenne pour les 5 AU est supérieure de 13% à celle obtenue par la deuxième meilleure équipe (72% vs 64%).

Pour finir, la Figure 4.3 illustre les principaux descripteurs, qui forment les axes du sous-espace commun et des 5 sous-espaces spécifiques à chaque AU. Pour chaque sous-espace, nous

TABLE 4.2 – Comparaison entre MLKR et les deux extensions multi-tâches proposées en terme de corrélation de Pearson (en %)

AU	MLKR	MT-MLKR	H-MT-MLKR
6	74.2	75.4	**76.3**
10	70.9	72.8	**75.2**
12	**86.6**	86.4	86.5
14	41.4	44.3	**47.7**
17	52.6	52.7	**54.8**
Mean	65.1	66.3	**68.1**

TABLE 4.3 – Résultats officiels du challenge "fully continuous" de FERA'2015 en terme d'ICC

AU	ISIR [104]	Cambridge [8]	KIT	VicarVision [59]	LaBRI [100]
6	**78.7**	71.9	67.8	66.4	72
10	**80.2**	71.8	73.1	73.4	72
12	**86.1**	82.8	82.6	78.8	78.4
14	**71.1**	54.6	53.3	54.9	27.7
17	**44.3**	37.7	30.8	32.9	26.8
Mean	**72.1**	63.8	61.5	61.3	55.4

n'avons présenté que les 4 plus importants. Les lignes blanches représentent les angles sélectionnés, et les flèches noires correspondent à la position et à l'orientation du descripteur HOG sélectionné. On remarque, par exemple, que l'angle entre l'extrémité de l'œil, le centre de la bouche et la commissure des lèvres est un descripteur pertinent pour l'ensemble des AU, ce qui est tout à fait cohérent, puisque cet angle varie lorsque les AU12 (étirement du coin des lèvres), AU10 (remontée de la partie supérieure de la lèvre) et 17 (élévation du menton) sont activées. Pour l'AU 14 (le plissement externe des lèvres), on remarque, par exemple, que H-MT-MLKR a sélectionné un gradient sur la joue droite, dans une zone où apparaissent les fossettes lorsque l'AU est activée.

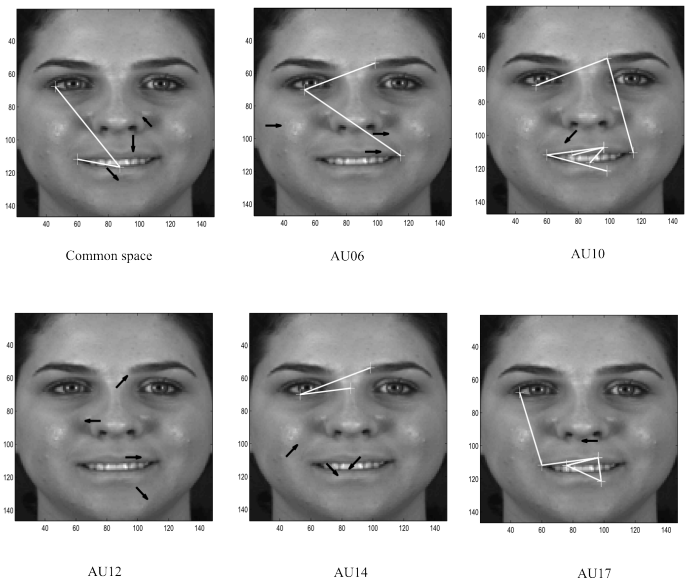


FIGURE 4.3 – Principaux descripteurs qui forment les axes du sous-espace commun et des 5 sous-espaces spécifiques à chaque AU

Publications en lien avec le chapitre

(T1) J. Nicolle. *Reading Faces. Using Hard Multi-Task Metric Learning for Kernel Regression.* , **Thèse de doctorat.** 2016

(J1) J. Nicolle, K. Bailly et M. Chetouani. *Real-time facial action unit intensity prediction with regularized metric learning*, **Image Vision Computing**, **52** (1) : 1-14, 2016.

(C1) J. Nicolle, K. Bailly et M. Chetouani. *Facial Action Unit intensity prediction via Hard Multi-Task Metric Learning for Kernel Regression*, **Facial Expression Recognition and Analysis Challenge (FERA'2015)**.

(C2) J. Nicolle, V. Rapp, K. Bailly, L. Prevost, M. Chetouani : *Combining LGBP Histograms with AAM coefficients in the Multi-Kernel SVM framework to detect Facial Action Units*, **Facial Expression Recognition and Analysis Challenge (FERA'2011)**.

(C3) T. Senechal, K. Bailly, L. Prevost. *Robust continuous prediction of human emotions using multiscale dynamic cues*, **The Continuous Audio/Visual Emotion Challenge (AVEC'2012)**.

Encadrement doctoral

— Jérémie Nicolle [9/2012–3/2016] (dir. Mohamed Chetouani)

Projets collaboratifs

— Projet FUI **A1:1** (2012–2015)

— Projet ANR **JEMImE** (2013–2018) jemime.isir.upmc.fr/

Chapitre 5

Forêts Aléatoires pour l'analyse en environnement non contraint

Les Forets Aléatoires sont des modèles simples, avec un fort pouvoir expressif et peu sensibles aux sur-apprentissage. Dans les recherches que nous présentons dans ce chapitre, nous avons donc souhaité explorer les capacités de ces modèles à œuvrer dans un environnement non contraint. Les limitations inhérentes aux Forêts Aléatoires, que nous présentons dans la première partie (5.1), nous ont amenées à proposer des extensions de ces modèles pour intégrer l'information temporelle (5.2.1), et pour augmenter la robustesse aux variations de pose du visage (5.2.2) et aux occultations (5.3). Pour finir ce chapitre, nous présentons les récents travaux que nous avons développés, à l'interface des Forêts Aléatoires et des réseaux profonds (5.4). Comme pour les précédents modèles, nous montrons leurs capacités sur des tâches variées de reconnaissance d'expressions faciales et de localisation de points caractéristiques.

5.1 Forêts aléatoires pour l'analyse faciale : intérêt et limitations

Les Forêts Aléatoires (*Random Forests*, RF) constituent une famille de méthodes d'apprentissage proposée par Breiman [17]. Elles consistent à apprendre un ensemble de T arbres de décision ayant été construits à partir de T échantillons *bootstrap* issus de la base d'apprentissage. Chaque arbre est construit par une procédure gloutonne (*greedy*) qui estime, pour chaque nœud de l'arbre, les paramètres de la fonction de séparation (ϕ, la valeur du descripteur et θ le seuil associé) qui aiguillera un exemple vers son nœud fils droit ou gauche. Le meilleur jeu de paramètres de la fonction de séparation est estimé à l'aide d'une mesure de gain d'information $H_{\phi,\theta}$ (par exemple la mesure de l'entropie de Shannon ou le critère d'impureté de Gini). Cette procédure est appliquée récursivement aux sous-arbres de droite et de gauche avec les exemples aiguillés de part et d'autre, jusqu'à ce que la répartition des exemples dans chaque nœud soit homogène (i.e. que tous les exemples soient de la même classe $l \in \mathcal{L}$).

En phase de test, une image x est aiguillée successivement à droite ou à gauche dans un arbre t en fonction des tests binaires, jusqu'à ce qu'elle atteigne une feuille de l'arbre. L'arbre peut alors retourner une probabilité $p_t(l|x)$ qui prend la valeur 1 pour la classe représentée dans ce nœud et 0 sinon. La probabilité finale est obtenue en calculant la moyenne des probabilités retournées par les T arbres :

$$p(l|x) = \frac{1}{T}\sum_{i=1}^{T} p_t(l|x) \tag{5.1}$$

La robustesse de l'algorithme est assurée par la force individuelle de chaque arbre et la décorrélation entre ces arbres. En construisant des arbres à partir de sous ensembles différents et indépendants, et en utilisant un algorithme de sous-espace aléatoire (en n'examinant, par exemple, qu'un sous-ensemble de descripteurs pour apprendre chaque fonction de séparation), le pouvoir de prédiction de chaque arbre est plus faible, mais celui-ci se trouve très décorrélé des autres arbres de la forêt.

Les méthodes par RF sont très répandues dans le domaine de la vision par ordinateur en général, et dans le domaine de l'analyse faciale en particulier, car elles sont capables d'apprendre des fonctions complexes, à partir d'un nombre restreint de données de grande dimension et potentiellement bruitées. De plus, elles sont robustes, ont une bonne capacité de généralisation et les temps d'exécution sont faibles, car les fonctions de séparations sont simples et, seul un sous-ensemble de l'arbre est exploré pour classifier un exemple.

Toutefois, les RF présentent certaines limitations pour l'analyse faciale, en particulier :

— L'information dynamique n'est pas exploitée.

— La robustesse aux variations de poses n'est pas assurée, car la valeur des descripteurs peut varier lorsque le visage effectue des rotations en dehors du plan image.

— Les performances peuvent fortement diminuer en présence d'occultations, car les prédictions de tous les arbres utilisant des descripteurs extraits dans la zone occultée du visage peuvent être fausses.

Les travaux, entrepris dans le cadre de la thèse d'Arnaud Dapogny, ont cherché à dépasser ces limitations pour améliorer les performances et la robustesse des méthodes d'analyse des expressions faciales.

5.2 Forêts aléatoires conditionnelles

5.2.1 Forêts aléatoires conditionnées par paires (PCRF)

L'idée directrice des forêts aléatoires conditionnées par paires (*Pairwise Conditional Random Forests*, PCRF) est d'intégrer l'information spatio-temporelle sous la forme de classifieurs de transition d'émotion. Comme le montre la Figure 5.1, un classifieur statique utilise uniquement l'information statique (flèche bleue) contenue dans l'image courante n pour prédire l'expression faciale. A l'inverse, la sortie de notre méthode PCRF est la combinaison de plusieurs classifications de transitions (flèches jaunes). Ces transitions sont évaluées sur des paires d'images extraites à différents intervalles dans la séquence. Du point de vue de la RF, cela revient à étendre la force de chaque arbre (en augmentant l'ensemble des descripteurs qui contient, dans ce cas, des descripteurs statiques et dynamiques) tout en augmentant la décorrélation entre les arbres, car les prédictions sont issues de paires échantillonnées à différents moments de la séquence. Cette méthode est très flexible car elle est indépendante de la vitesse d'exécution de

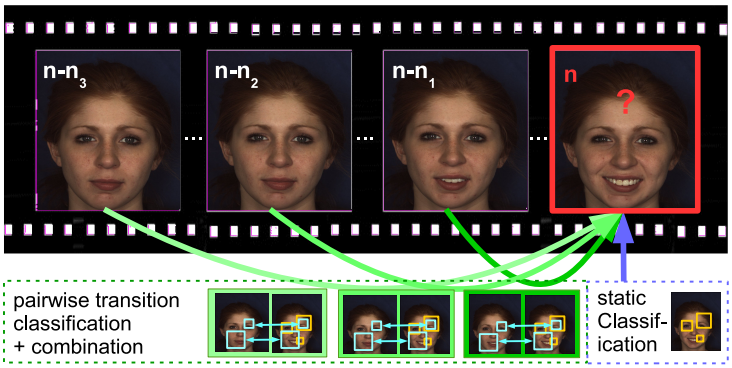


FIGURE 5.1 – Clasiffication statique (bleu) et des transitions (vert)

chaque phase de la production émotionnelle (contrairement aux méthodes qui utilisent des descripteurs spatio-temporels) et ne nécessite aucune cohérence temporelle dans la séquence, aussi bien en apprentissage qu'en test (contrairement aux chaines de Markov cachées par exemple).

De plus, nous proposons de conditionner le choix des arbres de décision par l'expression du visage dans la première image de la paire. Ainsi la tâche assignée à chaque arbre est simplifiée, puisqu'il est spécialisé dans la classification de la transition d'une classe spécifique vers toutes les classes (y compris la même que celle de la première image, puisque la personne peut rester ou retrouver la même expression après un certain laps de temps).

Plus formellement, si l'on considère des paires d'images $(\mathcal{I}', \mathcal{I})$, on apprend un arbre t, dont l'objectif est de prédire la probabilité $p_t(c|\mathcal{I}, \mathcal{I}', c')$ que le label de l'image $\mathcal{I}$ soit c, connaissant la première image $\mathcal{I}'$ et son label associé c'. Chaque arbre n'encode pas l'évolution temporelle de l'émotion mais une information différentielle entre des paires d'images. Ainsi, deux paires d'images doivent appartenir à la même personne, mais pas forcément à la même vidéo. Cela permet, entre autre, de créer des paires d'images échantillonnées à partir de différentes séquences pour couvrir toutes les transitions possibles et équilibrer les classes. Cette stratégie peut donc également être utilisée pour "calibrer" le détecteur d'expression faciale en fonction d'une personne, si l'on dispose, par exemple, d'une image du visage de la personne sans expression.

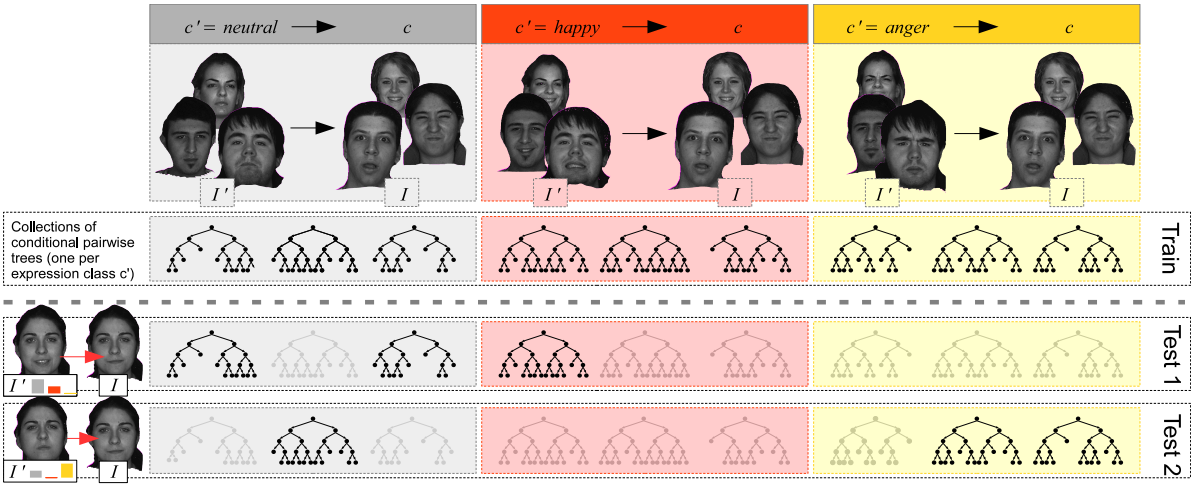


FIGURE 5.2 – Exemple d'une collection d'arbres pour 3 classes d'expressions basiques. Les probabilités de chaque classe d'expression estimées dans la première image de la paire, sont utilisées pour sélectionner les arbres qui permettront de prédire la transition d'expression de l'image $\mathcal{I}'$ vers $\mathcal{I}$

Les prédictions individuelles sont ensuite combinées de manière à obtenir $p^n(c)$ la probabilité que le label c soit associé à l'image $\mathcal{I}^n$ de la vidéo. Dans le cas d'une RF statique (formulation classique), cette probabilité est donnée par la moyenne des prédictions de chaque arbre :

$$p^n(c) = \frac{1}{T} \sum_{t=1}^{T} p_t(c|\mathcal{I}^n) \tag{5.2}$$

Pour utiliser l'information spatio-temporelle, on applique un modèle RF sur des paires d'images $(\mathcal{I}^m, \mathcal{I}^n)$ avec $\{\mathcal{I}^m\}_{m=,n-N,\ldots,n-1}$ les images précédentes de la vidéo. On calcule la moyenne des prédictions au cours du temps, pour obtenir une nouvelle estimation de p^n, qui prenne en compte les observations précédentes jusqu'à l'image n. Ainsi, si nous n'avons pas d'informations a priori sur ces images, la probabilité p^n devient :

$$p^n(c) = \frac{1}{NT} \sum_{m=n-N}^{n-1} \sum_{t=1}^{T} p_t(c|\mathcal{I}^n, \mathcal{I}^m) \tag{5.3}$$

Les modèles *static* et *full dynamic* feront respectivement référence aux équations (5.2) et (5.3) dans la suite du document. Les arbres du model *full dynamic* sont potentiellement plus forts, car ils sont appris à partir d'un ensemble de descripteurs étendu par rapport à celui du modèle *static*. La prédiction à partir de paires d'images différentes, combinée à un ensemble de descripteurs étendu, contribue à décorréler les prédictions des arbres, puisque chaque arbre dispose d'une information plus variée que dans le cas statique. Toutefois les performances peuvent ne pas être améliorées si la tâche à prédire est trop complexe (i.e. la variabilité de l'apparence des transitions de n'importe quelle expression vers n'importe quelle expression est très forte).

Pour réduire cette variabilité, on considère qu'il existe une distribution de probabilité $p_0^m(c')$ d'observer le label d'expression c' à l'image m. Ces probabilités peuvent être estimées à l'aide d'un classifieur statique (c'est le cas par exemple de la première image de la vidéo) ou à l'aide de prédictions dynamiques estimées dans les images précédentes de la vidéo. Dans ce cas, pour l'image m, les arbres de prédiction sont échantillonnés par rapport à la distribution $p_0^m(c')$. Pour chaque image précédente m et chaque label d'expression c', on sélectionne aléatoirement $\mathcal{N}^m(c')$ arbres à partir du modèle PCRF dédié aux transitions, dont l'expression de l'image initiale est c'. Ainsi l'équation (5.3) devient :

$$p^n(c) = \frac{1}{NT} \sum_{m=n-N}^{n-1} \sum_{c' \in \mathcal{C}} \sum_{t=1}^{\mathcal{N}^m(c')} p_t(c|\mathcal{I}^n, \mathcal{I}^m, c') \tag{5.4}$$

avec $\mathcal{N}^m(c') \approx T p_0^m(c')$ et $T = \sum_{c' \in \mathcal{C}} \mathcal{N}^m(c')$ le nombre d'arbres dédiés à la classification de chaque transition, qui pourra être choisi en fonction des ressources disponibles. Le model *conditionnal* fera référence à l'équation (5.4).

5.2.2 PCRF Multi-vues (MV-PCRF)

Dans le même esprit que le modèle PCRF, il est possible d'augmenter la robustesse aux variations de pose en conditionnant le modèle par rapport à une estimation de la pose de la tête $\omega(\mathcal{I}^n)$ pour l'image n. Pour cela, nous quantifions l'espace des poses Ω en $k = \Gamma \times B$ classes de poses $\{\Omega_i = \Omega_{\gamma_i,\beta_i}\}_{i=1,\dots,k}$ définies respectivement par les angles de lacet (*yaw*) et tangage (*pitch*). L'extension multi-vue du modèle statique (MVRF) s'écrit alors :

$$p^n(c) = \frac{1}{T} \sum_{\Omega_i \in \Omega} \sum_{t=1}^{\mathcal{N}(\Omega_i)} p_t(c|\mathcal{I}^n, \Omega_i) \tag{5.5}$$

Il est, de plus, possible de combiner le modèle MVRF au modèle PCRF pour obtenir une méthode de reconnaissance d'expressions faciales robuste aux variations de pose, et capable d'intégrer l'information dynamique de la séquence :

$$p^n(c) = \frac{1}{T} \sum_{m=n-N}^{n-1} \sum_{\Omega_i \in \Omega} \sum_{c' \in \mathcal{C}} \sum_{t=1}^{\mathcal{N}^m(c',\Omega_i)} p_t(c|\mathcal{I}^n, \mathcal{I}^m, \Omega_i, c') \tag{5.6}$$

que nous appellerons le modèle multi-vue PCFR (MV-PCRF) dans la suite du document. La méthode implique l'apprentissage de modèles spécifiques à une pose et une expression dans la première image de la paire, symbolisés par les éléments du tenseur de la Figure 5.3. Afin que chaque modèle soit appris avec suffisamment d'exemples, nous avons utilisé une base de données de modèles 3D pour synthétiser l'apparence de visages avec des expressions et des poses variées.

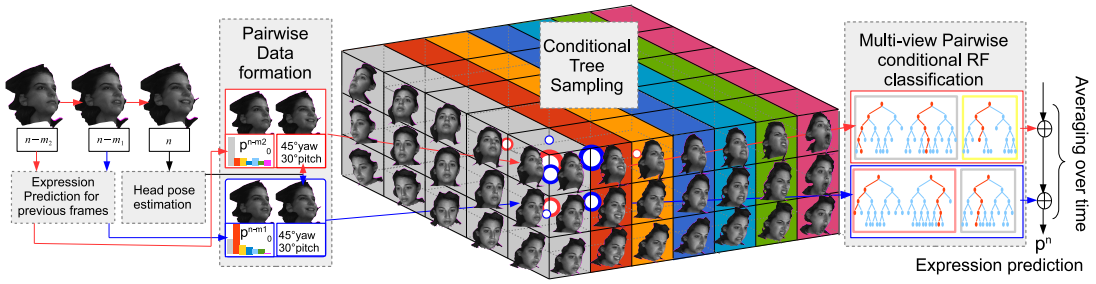


FIGURE 5.3 – Chaine de traitement de la méthode MVPCRF

5.2.3 Résultats expérimentaux

Pour valider la pertinence du modèle PCRF, nous l'avons comparé au modèle *static* et *full dynamic* avec différents paramètres d'intégration dynamique (la longueur de la fenêtre temporelle et le pas entre deux images successives) sur la base de données BU-4DFE. Nous avons également évalué la pertinence d'utiliser une prédiction dynamique pour les images précédentes de la séquence (i.e. les prédictions précédentes du modèle PCRF), plutôt que les prédictions issues d'un classifieur statique. Nous avons reporté les performances dans la figure 5.4. Nous pouvons remarquer que le modèle dynamique surpasse le modèle statique, grâce aux descripteurs dynamiques qui apportent plus de robustesse et de décorrélation aux arbres individuels, et que le conditionnement des arbres, en fonction de l'expression dans la première image de la paire – en particulier à l'aide des prédictions dynamiques précédentes – augmente significativement les performances. On peut également remarquer qu'il est préférable de remonter plus en amont dans la séquence ($N = 60$) avec moins de corrélations entre les images ($Step = 3$ ou 6).

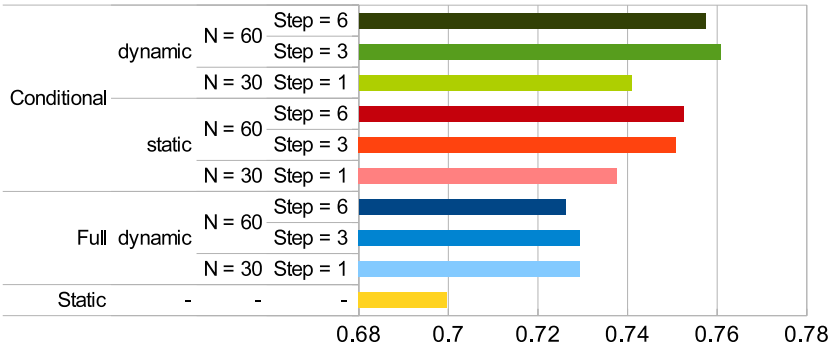


FIGURE 5.4 – Précisions moyennes obtenues pour différents paramètres d'intégration temporelle sur la base BU-4DFE

La figure 5.5 illustre les précisions par classe de poses pour les 6 expressions. On remarque, d'une part, que les performances du modèle RF s'effondrent dès que l'on s'éloigne d'une position frontale, et que le modèle PCRF résiste mieux aux variations de pose. Et d'autre part, ce sont les modèles multi-vues, et en particulier le modèle MVPCRF, qui supportent mieux ce type de variations.

Pour les détails de conception et d'implémentation de la méthode, ainsi qu'une étude approfondie des performances, le lecteur pourra se reporter à la thèse de Dapogny [35].

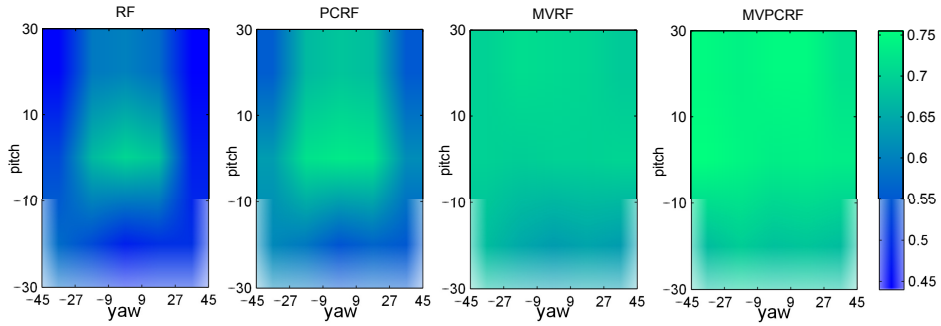


FIGURE 5.5 – Précisions par classe de pose (moyenne pour toutes les expressions)

5.3 Forets aléatoires à sous-espaces locaux (LSRF)

Lors de l'apprentissage d'une RF traditionnelle, le sous-ensemble de descripteurs sélectionnés, pour apprendre la fonction de séparation, est issu d'un tirage aléatoire uniforme sur l'ensemble du visage (*Random Subspace Random Forests*, RS-RF). Il est donc assez naturel que les arbres sélectionnent préférentiellement des descripteurs dans la zone de la bouche, puisqu'il s'agit d'une partie très informative pour l'analyse émotionnelle (cf. Figure 5.6). En particulier, dans le cas

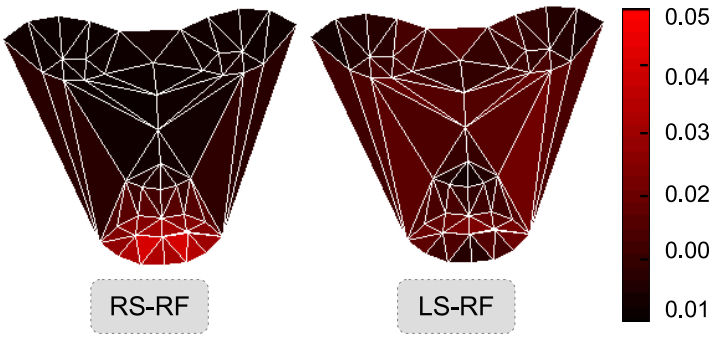


FIGURE 5.6 – Proportion par triangle de descripteurs sélectionnés dans le premier niveau des arbres par une RF classique (RS-RF) et par notre méthode (LS-RF)

d'expressions faciales prototypiques, il est très facile de discriminer la joie, la tristesse ou encore la surprise uniquement en observant les mouvements des lèvres. Pour autant, ce résultat est préjudiciable du point de vue de la RF pour deux raisons. D'une part les sorties des arbres risquent d'être fortement corrélées, diminuant ainsi les performances de la RF. D'autre part, la robustesse aux occultations n'est pas assurée puisque, en cas d'occultation de la zone de la bouche, les prédictions de la plupart des arbres de la RF seront fausses.

Afin d'obtenir une meilleure répartition des descripteurs sélectionnés, nous avons proposé une nouvelle méthode, dont l'objectif est d'apprendre chaque arbre à partir d'un ensemble spatialement contraint de descripteurs (*cf.* Figure 5.7(a,b)). Nous proposons d'exploiter ce modèle dans une méthode d'analyse d'expressions faciales robuste aux occultations (e-g) en pondérant les prédictions locales d'expression (c) par la probabilité d'occultation des différentes parties du visage fournie par un ensemble de mémoires associatives (d). Nous montrons également comment la représentation apprise par chaque arbre peut-être exploitée pour reconnaitre des AU (h).

5.3.1 Prédiction locale de l'expression

La prédiction locale s'effectue en deux temps. Dans un premier temps, des arbres de décision sont entraînés à prédire l'émotion à partir d'un ensemble de descripteurs extraits dans une zone délimitée par un masque de taille fixe, et dont la position a été définie aléatoirement sur le visage.

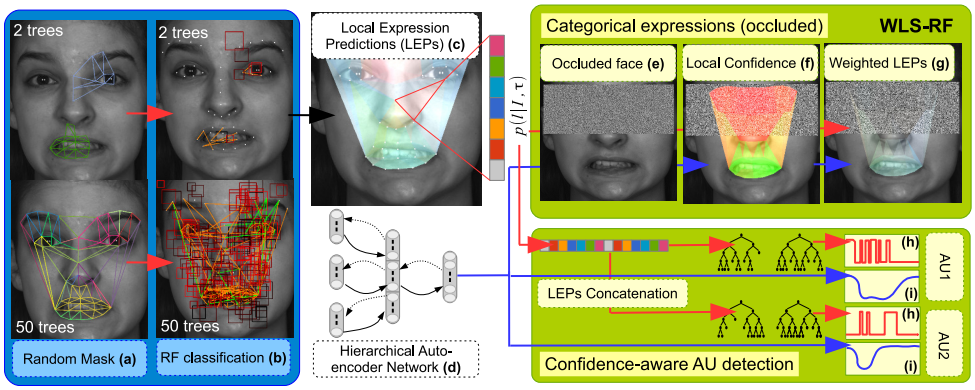


FIGURE 5.7 – Principe général de la méthode proposée.

Dans un second temps, les prédictions de chaque arbre sont combinées de manière à obtenir une prédiction associée à chaque triangle qui compose le modèle du visage.

Apprentissage des arbres Après avoir calculé une forme moyenne $\bar{f}$ du visage et la surface normalisée $s(\tau(\bar{f}))$ couverte par chaque triangle τ, on génère un masque M_t, en sélectionnant aléatoirement un triangle τ_i, puis en ajoutant des triangles de son voisinage, jusqu'à obtenir une surface supérieure à un hyper-paramètre R. Ce dernier représente la surface qui devra être couverte par chaque masque, comme illustré sur la figure 5.8. Du point de vue de la RF, il permet de trouver un compromis entre la force individuelle des arbres et leur niveau de corrélation. La

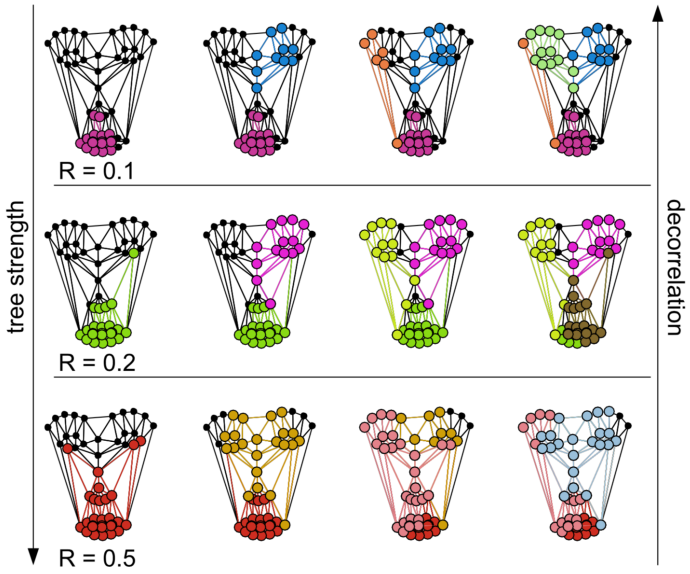


FIGURE 5.8 – Masques generés avec $R = 0.1$, 0.2 et 0.5 pour 1,2,3 et 4 arbres.

procédure d'apprentissage d'un arbre t est alors la même que pour une RF classique, à l'exception du choix des descripteurs qui ne peuvent provenir que de la zone délimitée par le masque M_t.

Predictions locales Tout comme une RF classique, la probabilité de chaque classe est obtenue en calculant la moyenne des probabilités de chaque arbre :

$$p(c|\mathcal{I}) = \frac{1}{T}\sum_{i=1}^{T} p_t(c|\mathcal{I}) \tag{5.7}$$

Mais dans le cas de l'approche LS-RF, les arbres ont capturé une information locale et il est alors possible de ré-écrire l'équation (5.7) en fonction de probabilités définies pour chaque triangle :

$$p(c|\mathcal{I}) = \frac{1}{T}\sum_{\tau} Z_\tau p(c|\mathcal{I},\tau) \tag{5.8}$$

avec $p(c|\mathcal{I},\tau)$ le vecteur de probabilité de la prédiction locale d'expression (*Local Expression Prediction*, LEP) du triangle τ :

$$p(c|\mathcal{I},\tau) = \frac{1}{Z_\tau}\sum_{t=1}^{T} \frac{\delta(\tau \in M_t)p_t(c|\mathcal{I})}{|M_t|} \tag{5.9}$$

La fonction $\delta(\tau \in M_t)$ retourne 1 si le triangle τ appartient au masque M_t, et 0 sinon. $|M_t|$ est le nombre de fois que l'arbre t est utilisé dans l'équation (5.8), et le coefficient de normalisation Z_τ est la somme des prédictions pour toutes les classes d'expression c et tous les triangles τ. A noter que le vecteur LEP $p(c|\mathcal{I},\tau)$ n'est pas issu d'une prédiction limitée au triangle τ, mais est défini à partir de son voisinage, dont la taille dépend de l'hyperparametre R.

5.3.2 Prédiction des expressions robustes aux occultations

En forçant les arbres de décision à utiliser des descripteurs issus de zones restreintes et en sélectionnant ces zones aléatoirement sur tout le visage, on améliore la robustesse, puisque une occultation partielle du visage aura un impact limité aux seuls arbres appris à partir de cette région. Nous proposons d'améliorer cette robustesse en détectant les zones occultées du visage à l'aide de mémoires auto-associatives définies dans le voisinage de chaque point caractéristique. Lorsque ce point est occulté, l'erreur de reconstruction est importante. Ainsi, nous pouvons pondérer les prédictions locales d'expression en fonction d'un indice de confiance $\alpha^{(t)}$ associé à chaque triangle. Le modèle LS-RF pondéré (Weighted LS-RF) est défini par :

$$p(c|\mathcal{I}) = \frac{\sum_\tau \alpha^{(t)} Z_\tau p(c|\mathcal{I},\tau)}{\sum_\tau \alpha^{(t)} Z_\tau} \tag{5.10}$$

La Figure 5.9 montre les variations de précision sur la base CK+ en fonction de l'hyperparamètre R, en cas d'occultation des yeux et de la bouche. Les performances d'un RF classique chutent fortement lorsque la bouche est occultée (de $91,5\%$ à 25.4%), car ce modèle s'appuie essentiellement sur les descripteurs issus de cette zone pour prédire l'expression faciale. En forçant les arbres à être plus locaux ($R = 0,1$ ou $0,2$), le modèle LS-RF résiste mieux aux occultations de cette partie du visage. On observe également que, quelque soit le scénario, le modèle WLS-RF donne de bien meilleurs résultats que les versions non pondérées sur les bases CK+ et BU-4DFE

Nous avons également comparé notre méthode à l'état de l'art en suivant le protocole défini dans [170]. Les résultats reportés dans le tableau 5.1 montrent l'intérêt de notre méthode, en particulier lorsque le visage est fortement occulté ($R16$ et $R24$, avec RS correspondant à un masque de taille $S \times S$ placé aléatoirement dans l'image du visage de taille 64×64).

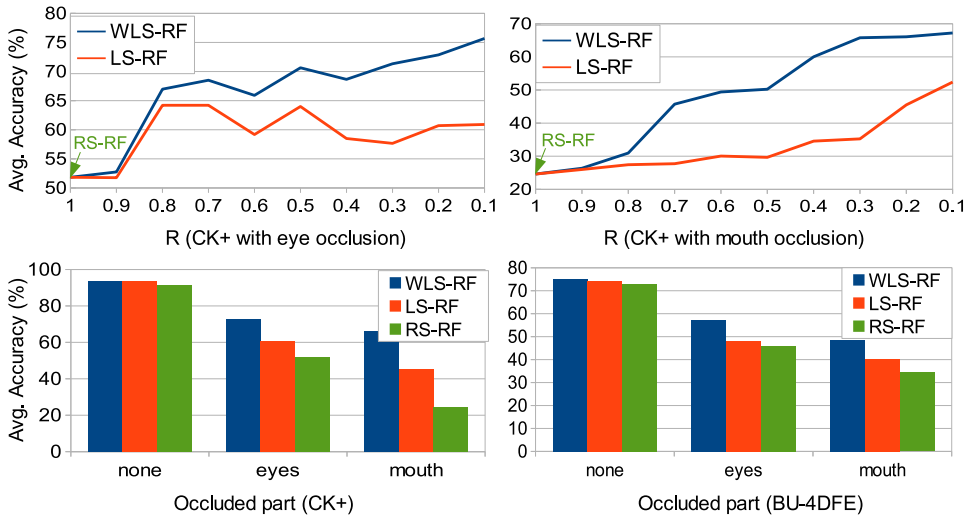


FIGURE 5.9 – Précision des modèle LS-RF et WLS-RF par rapport à RS-RF

TABLE 5.1 – Comparison du modèle WLS-RF par rapport à [170]

Protocole	WLS-RF	[170]
R8	**92.2**	**92**
R16	**86.4**	82
R24	**74.8**	62.5
Yeux occultés	**87.9**	**88**
Bouche occultée	**72.7**	30.3

5.3.3 Reconnaissance des AU

Notre modèle de prédiction locale d'expression a un deuxième intérêt : il peut être utilisé comme représentation pour prédire des phénomènes expressifs plus locaux que les émotions basiques, telles que les activations musculaires. L'intérêt pour prédire les AU est évident car nous disposons de beaucoup plus de données annotées en émotions basiques qu'en AU. Ainsi, la représentation peut être apprise sur un grand nombre de données, puis utilisée pour entrainer des RF à détecter les AU.

Au chapitre précédent, nous avons montré l'intérêt d'apprendre à estimer conjointement plusieurs AU, afin de prendre en compte les dépendances inter-AU. De manière similaire, nous avons exploré des stratégies de prédiction multi-labels adaptées aux forêts aléatoires. En particulier, les stratégies diffèrent sur deux aspects :

— Le choix de ou des AU associées à chaque arbre de décision

— Le choix de ou des AU utilisées pour apprendre la fonction de séparation de chaque nœud

Les résultats expérimentaux [32] ont montré que la stratégie qui donne les meilleurs résultats consiste à construire des arbres qui prédisent simultanément toutes les AU, mais à ne choisir aléatoirement qu'une seule AU pour apprendre la fonction de séparation de chaque nœud.

Le tableau 5.2 présente les résultats obtenus par notre méthode et deux autres approches de l'état de l'art. Notre approche donne des résultats significativement meilleurs que ceux de Zhao *et al.* [175] et légèrement supérieurs à [168], en particulier pour les AU les plus couramment utilisées (AU1, AU6, AU12 and AU25).

TABLE 5.2 – Comparaison de notre méthode de détection des AU par rapport à iCPM [168] et DRML [175]

	F1			AUC	
AU	méthode proposée	DRML [175]	iCPM [168]	méthode proposée	DRML [175]
1	**31.1**	17.3	29.5	**72.8**	53.3
2	24.1	17.7	**24.8**	**61.9**	53.2
4	52.9	37.4	**56.8**	**77.8**	60.0
6	**49.6**	29.0	41.7	**90.5**	54.9
9	25.1	10.7	**31.5**	**84.9**	51.5
12	**74.8**	37.7	71.9	**96.2**	54.6
25	**85.8**	38.5	81.6	**95.8**	45.6
26	49.7	20.1	**51.3**	**79.3**	45.3
Moy.	**49.1**	26.7	48.6	**82.4**	52.3

5.4 Evaluation gloutonne des Forêts Neuronales

Les travaux présentés dans ce chapitre ont mis en avant l'intérêt des RF dans un contexte d'analyse faciale. Toutefois, ces modèles présentent un inconvénient majeur : ils ne sont pas différentiables. Ils n'offrent donc pas la même flexibilité que les réseaux de neurones pour l'apprentissage en ligne et l'apprentissage de représentation. Nous avons alors proposé un modèle hybride, qui s'appuie sur les Forêts Neuronales (*Neural Forest*, NF), et évalué ses performances pour l'analyse de visages.

5.4.1 Principe général

Quel que soit la nature de la forêt de décision, la probabilité qu'un arbre de décision t associe la classe c à un exemple $\mathbf{x}_i$ peut s'écrire $p_t(c|\mathbf{x}_i) = \sum_l \mu^l(\mathbf{x}_i) p_t^l(c)$, avec $p_t^l(c)$ la prédiction associée à chaque feuille l, et $\mu^l(\mathbf{x}_i)$ la probabilité d'atteindre le nœud l. Dans le cas d'une RF classique, il n'y a qu'un seul coefficient non nul $\mu^l(\mathbf{x}_i)$, en fonction du chemin suivi par l'exemple $\mathbf{x_i}$ dans l'arbre. Formellement, pour chaque feuille l, il est possible de définir le chemin pour atteindre ce nœud, par deux ensembles $\mathcal{N}_l^{left}$ et $\mathcal{N}_l^{right}$ de nœuds, suivant que la feuille l est respectivement issue du sous-arbre de gauche ou de droite pour chaque nœud n. Le chemin dit "dur" (*hard path*) jusqu'à la feuille l est défini par le produit de fonctions de Kronecker $\delta^n(\mathbf{x_i})$ pour chaque nœud $n \in \mathcal{N}_l^{right}$, et $1 - \delta^n(\mathbf{x_i})$ pour chaque nœud $n \in \mathcal{N}_l^{left}$:

$$p_t(c|\mathbf{x_i}) = \sum_l \prod_{n \in \mathcal{N}_l^{right}} \delta^n(\mathbf{x_i}) \prod_{n \in \mathcal{N}_l^{left}} (1 - \delta^n(\mathbf{x_i})) p_t^l(c) \tag{5.11}$$

Les Forêts Neuronales (*Neural Forest*, NF) sont des modèles hybrides à l'interface des RF et des réseaux de neurones [78]. Dans le cas d'une NF, le résultat de la fonction de séparation n'est plus binaire (l'exemple est aiguillé vers le sous arbre de droite ou de gauche), mais continu (l'exemple est aiguillé à gauche et à droite avec une probabilité paramétrée par une variable de Bernouilli $d^n \in [0, 1]$). En considérant l'espérance de chaque nœud n (qui correspond à un nombre infini de réalisation pour l'arbre t), un exemple $\mathbf{x}_i$ sera aiguillé à droite du nœud n, avec une probabilité donnée par l'activation $d^n(\mathbf{x}_i)$, et à gauche avec la probabilité $1 - d^n(\mathbf{x}_i)$.

$$p_t(c|\mathbf{x_i}) = \sum_l \prod_{n \in \mathcal{N}_l^{right}} d^n(\mathbf{x_i}) \prod_{n \in \mathcal{N}_l^{left}} (1 - d^n(\mathbf{x_i})) p_t^l(c) \tag{5.12}$$

L'activation $d^n(\mathbf{x_i})$ du nœud n correspond à une fonction sigmoïde paramétrée par le vecteur β^n et le biais $-\theta^n$:

$$d^n(\mathbf{x_i}) = \sigma(\sum_{j=1}^{k} \beta_j^n x_{i,j} - \theta^n) \tag{5.13}$$

Par rapport à la version originale des NF [78], nous avons proposé des adaptations de l'algorithme, tant au niveau de la procédure d'évaluation (5.4.2), que celle d'apprentissage (5.4.3). Nous avons, de plus, évalué la pertinence de cette approche, dans le cas d'un problème de classification d'une part (reconnaissance des émotions, 5.4.4.1), et de régression d'autre part (localisation de points caractéristiques 5.4.4.2).

5.4.2 Procédure d'évaluation gloutonne

Dans le cas où $d^n(\mathbf{x}_i) \rightarrow \delta^n(\mathbf{x}_i)$ dans l'équation (5.12), la NF devient une RF à fonction de séparation oblique. Du point de vue de la NF, cela consiste à choisir le meilleur chemin à travers l'arbre de manière gloutonne, nœud après nœud. Nous ferons alors référence à ce modèle sous le terme de *Greedy Neural Forest (GNF)*. Pour T arbres de profondeur $\mathcal{D}$, la complexité algorithmique pour évaluer un exemple $\mathbf{x}$ de dimension k avec une NF est $T.k.(2^{\mathcal{D}+1} - 1)$. Elle est donc exponentielle par rapport à la profondeur des arbres. Dans le cas d'une GNF, seul le "meilleur" chemin à travers chaque arbre est évalué. Ainsi la complexité est égale à $T.k.\mathcal{D}$. Elle est donc linéaire par rapport à $\mathcal{D}$.

5.4.3 Algorithme efficace d'apprentissage

Kontschieder *et al.* [78] proposent une procédure d'optimisation en deux étapes. La propagation d'un exemple d'apprentissage $\mathbf{x}_i$ à travers les arbres fournit les valeurs des activations $d^n(\mathbf{x}_i)$ et des probabilités $\mu^l(\mathbf{x}_i)$. L'erreur est ensuite récursivement rétro-propagée des feuilles jusqu'à la racine. Après un certain nombre d'*epochs*[1], les probabilités contenues dans les feuilles sont mises à jour à l'aide d'une procédure d'optimisation convexe. Cette mise à jour du contenu des feuilles est toutefois coûteuse en temps, implique de nouveaux hyper-paramètres, et nécessite l'utilisation de toutes les données d'apprentissage. Cette procédure ne peut donc pas être effectuée en ligne.

Afin de répondre à ces limitations, nous avons proposé une procédure d'apprentissage alternative, qui s'appuie sur des feuilles dont le contenu reste fixe tout au long de l'apprentissage. Dans un premier temps, nous initialisons aléatoirement les arbres (i.e. nous générons aléatoirement des poids β_j^n et des seuils θ^n pour chaque nœud). Afin de maximiser le gain d'information par rapport à l'attribution d'un label de classe, les feuilles doivent contenir des distributions pures. On assigne alors des prédictions à chaque feuille, qui resteront inchangées tout au long du processus d'apprentissage. Dans le cas d'un problème de classification, une classe est choisie aléatoirement pour chaque nœud. Dans le cas d'une régression, la feuille est initialisée avec une valeur tirée aléatoirement dans l'intervalle de valeurs défini par la vérité terrain. A noter que ce tirage aléatoire peut suivre une distribution spécifique qui dépend de cette vérité terrain.

Ainsi, seuls les paramètres des fonctions de séparation sont mis à jour à chaque itération de la descente de gradient stochastique. Ce schéma d'apprentissage n'est certes pas intuitif, mais il permet une optimisation plus rapide et totalement en ligne, tout en réduisant le nombre d'hyper-paramètres.

A noter que, en classification, la profondeur de l'arbre $\mathcal{D}$ doit être choisie en fonction du

1. une *epoch* correspond à itération de l'algorithme d'apprentissage sur l'ensemble de la base de données

nombre de classes afin que chaque classe soit représentée au moins une fois dans chaque arbre. Nous avons montré que choisir $\mathcal{D}$ en fonction de l'équation (5.14), nous assurait avec une probabilité supérieure à 99% que cette condition soit respectée. Nous avons, de plus, montré [35] que cette borne inférieure évolue logarithmiquement par rapport au nombre de classes :

$$\mathcal{D} > \mathcal{D}_0 = \frac{1}{ln(2)} ln(\frac{ln(1-(1-0.99)^{1/\mathcal{C}})}{ln(1-1/\mathcal{C})}) \tag{5.14}$$

Nous avons également proposé une formule équivalente pour la régression [36].

De plus, l'apprentissage des arbres peut s'effectuer en parallèle (et sur des échantillons d'apprentissage différents pour augmenter la décorrélation entre les arbres). A l'issue de l'apprentissage, la NF est convertie en GNF en convertissant les fonctions de séparation douces d^n en fonctions de séparation dures δ^n pour chaque nœud n. Les résultats expérimentaux ont donné des performances similaires en terme de précision pour un temps de traitement bien inférieur.

5.4.4 Applications

5.4.4.1 GNF pour la reconnaissance des expressions

Nous avons, dans un premier temps, évalué les performances de notre modèle GNF pour un problème de classification, la reconnaissance des expressions prototypiques. L'architecture proposée combine, en entrée, des descripteurs géométriques (distances entre deux points caractéristiques normalisées par la distance inter-oculaire) et des descripteurs d'apparence issus d'un réseau CNN multicouches. La première couche du réseau CNN est composée de 40 filtres 5×5 appliqués à une image cadrée du visage de taille 48×48. On applique ensuite une étape de sous-échantillonnage (*max-pool* 2×2) avant une seconde couche composée de 40 filtres 3×3. Les sorties des deux couches sont concaténées et utilisées en entrée de la GNF.

Nous avons mené plusieurs évaluations qui ont abouti aux conclusions suivantes :

— **Profondeur de l'arbre :** l'influence de la profondeur de l'arbre n'est pas critique, à condition qu'elle soit supérieure à 4, afin que toutes les classes soient correctement représentées dans chaque arbre (les résultats sont donc en accord avec l'équation (5.14)).

— **Comparaison à une RF classique :** pour un nombre équivalent d'arbres, la GNF obtient des résultats très légèrement supérieurs à ceux de la RF. Ce résultat est d'autant plus encourageant que les autres méthodes d'apprentissage en ligne de RF [81, 120] aboutissent généralement à des modèles moins performants que la version originale.

— **Procédure d'évaluation gloutonne :** la précision des GNF est équivalente (voire légèrement meilleure) que celle obtenue avec une NF pour des temps de traitement largement inférieurs (d'un facteur 30 environ). Cela peut s'expliquer par une augmentation de la décorrélation des arbres, qui compense les pertes liées à la procédure d'évaluation gloutonne.

— **Apprentissage de la représentation :** les résultats obtenus à partir d'une combinaison de descripteurs géométriques et de descripteurs d'apparence appris sont bien meilleurs que ceux obtenus à partir de descripteurs géométriques seuls. Les résultats sont également meilleurs qu'avec une RF combinant descripteurs géométriques et HOG.

Pour finir, nous présentons dans le tableau 5.3 une comparaison de notre méthode par rapport à l'état de l'art sur 3 bases de données de difficulté croissante.

TABLE 5.3 – Résultats obtenus par notre méthode (précision) sur 3 bases de données et comparaison avec l'état de l'art. † : La prédiction est considérée comme valide si la vérité terrain correspond à l'une des deux prédictions les plus probables.

	CK+		BU-4DFE		FEED	
features	NF	GNF	NF	GNF	NF	GNF
geo	87.3	87.3	71.6	71.8	46.6	47.9
geo+CNN	**92.2**	**92.2**	**74.0**	**74.0**	51.8	52.0
LBP/SVM [129]	88.9		-		-	
RF/SVM [97]	-		73.1†		**53.7**†	
MRF/DBN [115]	90.1		-		-	
geo+HOG/RF	91.1		72.8		50.3	

5.4.4.2 GNF pour la localisation de points caractéristiques

Nous avons proposé un méthode de localisation de points caractéristiques, qui s'appuie sur le modèle GNF présenté précédemment. Cette méthode, illustrée dans la figure 5.10, se compose principalement de deux cascades de régression. La première consiste à estimer les paramètres $\mathbf{p}$ d'un modèle de forme de type *Point Distribution Model* [27] . La seconde cascade affine la position du modèle en prédisant directement les déplacements $\delta\mathbf{s}$ de chaque point dans l'image. Pour chaque étape des deux cascades, les déplacements sont prédits à l'aide d'une GNF. Cette forêt reçoit en entrée une concaténation de descripteurs SIFT extraits dans le voisinage des positions courantes des points du modèle et dont la taille est réduite, à l'aide d'une couche de neurones totalement connectés.

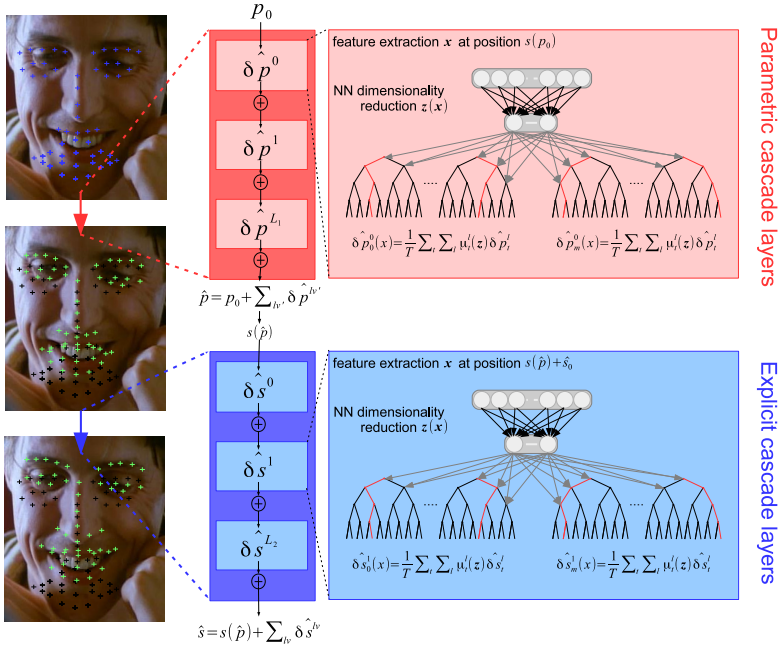


FIGURE 5.10 – Illustration de la méthode d'alignement de points caractéristiques proposée

Le tableau 5.4 montre que notre méthode obtient de meilleurs résultats que l'état de l'art sur les 3 bases de données les plus couramment utilisées pour cette tâche. De plus, la procédure d'évaluation gloutonne permet à cette méthode de s'exécuter largement en temps réel sur un simple cœur Intel I5. Pour finir, il est intéressant de noter un dernier intérêt à utiliser une méthode d'ensemble : il est possible d'estimer une confiance associée à la prédiction. Ce point est crucial, puisqu'il permet d'identifier lorsque le modèle dégénère pendant le *tracking*, afin de relancer une procédure de détection. Pour ce faire, il suffit de définir un score en calculant

une moyenne des prédictions de chaque arbre. L'écart type de ces prédictions donne alors une indication sur la difficulté rencontrée par la GNF, à prédire le déplacement des points, et peut ainsi être utilisé comme score d'alignement.

TABLE 5.4 – Comparaison par rapport à l'état de l'art en terme d'erreur moyenne normalisée

	LFPW		HELEN		IBUG	
Notre méthode	51 pts	68 pts	51 pts	68 pts	51 pts	68 pts
SDM [159]	4.47	5.67	4.25	5.50	-	15.40
RCPR [19]	5.48	6.56	4.64	5.93	-	17.26
IFA [7]	6.12	-	5.86	-	-	-
DRMF [6]	4.40	5.80	4.60	5.80	-	19.79
CFAN [169]	-	5.44	-	5.53	-	-
L21-Cascade [94]	3.80	-	4.1	-	16.3	-
GN-DPM [141]	4.43	5.92	4.06	5.69	-	-
PO-CR [140]	4.08	-	3.90	-	-	-
CSP-dGNDF	**3.76**	**4.84**	**3.87**	**5.16**	**10.45**	**12.74**

5.5 Conclusion

Les travaux présentés dans ce chapitre ont montré l'intérêt des méthodes d'ensemble de type RF pour l'analyse faciale. En particulier, les derniers travaux portant sur les Forêts Neuronales ouvrent des perspectives intéressantes, combinant les forces des RF et des réseaux neuronaux. Toutefois, l'architecture proposée reste relativement simple et de nombreuses améliorations mériteraient d'être explorées. Par exemple, les descripteurs d'apparence utilisés (HOG et CNN) pourraient également être remplacés par une structure arborescente dans laquelle chaque fonction de séparation pourrait être vue comme un motif discriminant. Un chemin particulier dans cet arbre correspondrait alors à une combinaison particulière de ces motifs. Par ailleurs, l'utilisation de feuilles fixes tout au long de l'apprentissage assure de la stabilité dans le processus d'apprentissage, mais apporte également de la rigidité. De nouvelles stratégies d'initialisation ou de mise à jour des feuilles pourraient être évaluées. Pour finir, nous pourrions exploiter la nature différentiable du modèle GNF pour calibrer les prédictions par rapport à un individu ou un contexte particulier.

Publications en lien avec le chapitre

(T1) A. Dapogny *A walk through randomness for face analysis in unconstrained environments*, **Thèse de doctorat.** 2016

(J1) A. Dapogny, K. Bailly et S. Dubuisson. *Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection.*, **International Journal of Computer Vision (IJCV)**, **126** (2-4) : 255–271., 2018.

(J2) A. Dapogny et K. Bailly *Face Alignment with Cascaded Semi-Parametric Deep Greedy Neural Forests*, **Pattern Recognition Letters (PRL)**, **102** (1) : 75–81, 2018.

(J3) A. Dapogny, K. Bailly et S. Dubuisson. *CDynamic Pose-Robust Facial Expression Recognition by Multi-View Pairwise Conditional Random Forests*, **IEEE Transactions on Affective Computing (TAC)**, **126** (2-4) : 255–271., 2018.

(C1) A. Dapogny, K. Bailly. *Investigating Deep Neural Forests for Facial Expression Recognition*, **Automatic Face and Gesture Recognition (FG'2018)**.

(C2) A. Dapogny, K. Bailly et S. Dubuisson. *Multi-Output Random Forests for Facial Action Unit Detection*, **Automatic Face and Gesture Recognition (FG'2017)**.

(C3) A. Dapogny, K. Bailly et S. Dubuisson. *Dynamic facial expression recognition by joint static and multi-time gap transition classification*, **Automatic Face and Gesture Recognition (FG'2015)**.

(C4) A. Dapogny, K. Bailly et S. Dubuisson. *Pairwise Conditional Random Forests for Facial Expression Recognition*, **International Conference on Computer Vision (ICCV'2015)**.

Encadrement doctoral

— Arnaud Dapogny [10/2013–12/2016] (dir. Severine Dubuisson)

— Estephe Arnaud [depuis 11/2017]

Projets collaboratifs

— Projet ANR **JEMImE** (2013–2018) jemime.isir.upmc.fr/

— Projet ANR JCJC **FacIL** (2017-2021)

Chapitre 6
Applications

Ce chapitre présente les principaux résultats obtenus dans le cadre de deux projets collaboratifs auxquels j'ai participé : le projet ANR Contint JEMImE (2013–2018), en tant que coordinateur, et le projet Labex SMART Sense, au travers du co-encadrement avec C. Clavel et G. Richard de la thèse de T. Janssoone (2014–2018).

6.1 Le projet JEMImE

JEMImE (Jeu Educatif Multimodal Imitation Emotionnelle, http://jemime.isir.upmc.fr) est un projet financé par l'ANR et coordonné par l'Institut des Systèmes Intelligents et de Robotique (ISIR, Sorbonne Université). Il associe l'entreprise Genious Healthcare, ainsi que l'équipe Cognition – Behaviour – Technologie (CoBTek, Université de Nice) et le Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS, Ecole Centrale de Lyon). L'objectif du projet était de concevoir un "jeu sérieux" (*serious game*) pour apprendre à des enfants avec autisme à produire des expressions faciales adaptées à un contexte social donné. La figure 6.1 en illustre le principe général.

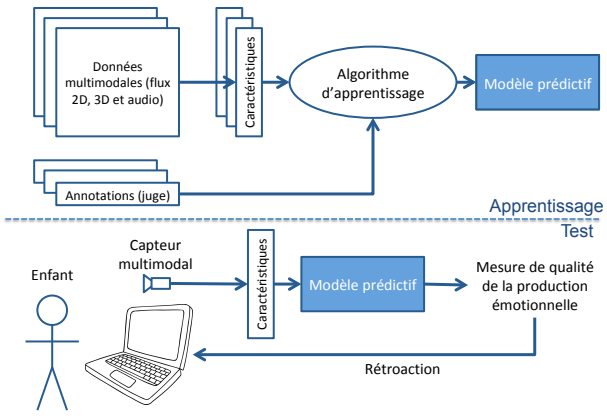


FIGURE 6.1 – Principe général du projet JEMImE

Dans un premier temps, nous avons dû collecter et annoter une grande base de données d'enfants typiques et d'enfants présentant des troubles du spectre autistique (TSA). L'objectif de cette base de données était double : d'une part, entrainer les algorithmes d'analyse automatique des expressions faciales que nous avons présentés précédemment (cf. chapitre 5), et d'autre part, mener des études cliniques "assistées par ordinateur".

Une fois appris, le modèle prédictif est intégré dans le jeu sérieux, pour analyser en temps réel les productions émotionnelles, et fournir un retour d'information (*feedback*) pour l'enfant (qui apprend à se corriger) et pour le jeu (qui enregistre les performances et valide ou non l'exercice en cours).

6.1.1 La base de données JEMImE

La base de données JEMImE contient au total 193 enfants volontaires âgés de 6 à 11 ans dont 157 sont des enfants typiques et 36 sont des enfants TSA. Plusieurs modalités ont été enregistrées : le visage de l'enfant avec un capteur 2D et 3D ainsi que sa voix. Chaque enfant devait produire 4 expressions faciales : neutre, joie, colère et tristesse au cours de deux tâches :

— La production émotionnelle sur demande : l'écran en face de l'enfant affiche explicitement l'émotion que l'enfant doit produire.

— L'imitation : on présente à l'enfant un avatar produisant l'émotion désirée et on demande ensuite à l'enfant de l'imiter.

Chaque expression faciale est produite 6 fois par l'enfant, 2 fois pour la tâche sur demande et 4 fois pour l'imitation de l'avatar (soit avec la modalité audiovisuelle ou visuelle seule et avec un avatar de chaque genre). Le genre de l'avatar et le type de tâche et la modalité sont choisis aléatoirement pour éviter un biais d'apprentissage. Ainsi, la base contient au total plus de 4600 vidéos de 3 secondes en moyenne.

L'objectif du jeu étant d'évaluer si la production émotionnelle de l'enfant est en adéquation avec le contexte social, l'objectif du module d'analyse d'expressions faciales n'est pas uniquement de reconnaitre l'expressions, mais d'évaluer sa crédibilité. Ainsi, 3 juges ont annoté, à l'aveugle, les vidéos en terme de qualité d'expression. Cette qualité est mesurée sur une échelle continue, de 0 à 10 : 0 correspond à une expression non reconnue, 5 à une expression reconnue mais pas crédible, et 10, à une émotion reconnue et crédible.

6.1.2 Quelques résultats cliniques

Etude statistique de la base de données : les premiers résultats de l'étude clinique proviennent d'une analyse multivariée des annotations des enfants typiques [57]. En particulier, cette analyse a révélé que (1) la qualité des productions émotionnelles augmentait avec l'âge de l'enfant, (2) les émotions positives sont plus faciles à produire que les émotions négatives, (3) La qualité des émotions produites sur demande est meilleure que celle obtenue par imitation, et (4) les enfants de Nice réussissent mieux à produire des émotions de qualité que les enfants de Paris, ce qui laisse supposer une influence régionale sur la qualité de la production émotionnelle. Cette première étude a donc montré que la production des expressions faciales était un processus développemental complexe, influencé par plusieurs facteurs.

Analyse des descripteurs pertinents : dans une seconde étude, dont l'ensemble des résultats sont disponibles dans [**?**], nous avons cherché à identifier les descripteurs permettant aux machines de reconnaitre et d'estimer la qualité des différentes expressions faciales. La figure 6.2 illustre un exemple de visualisation permis par cette méthode dans le cas d'une tâche de reconnaissance des expressions faciales produites par des enfants typiques. Par exemple, l'expression de la joie se caractérise principalement par des distances et des angles extraits dans la zone de la bouche, ainsi que par des HOG extraits sur les joues. La colère est plus particulièrement caractérisée par des descripteurs géométriques extraits autour de la région œil / sourcil, ainsi que des informations de texture aux commissures des yeux et dans la zone inter-oculaire (qui correspondent respectivement au plissement des yeux et au froncement des sourcils). La reconnaissance de la tristesse repose principalement sur une information géométrique du coin de la bouche (abaissement des coins externes des lèvres), ainsi que sur des descripteurs d'apparence dans la région des yeux. Ce type de visualisation ouvre des perspectives intéressantes pour comprendre le processus de production des expressions faciales et étudier les singularités des enfants TSA.

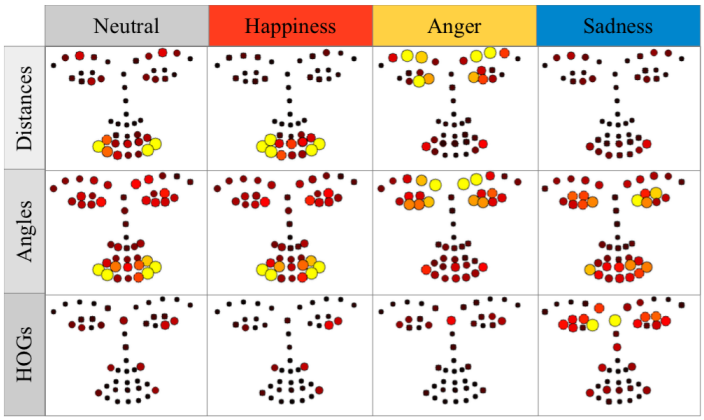


FIGURE 6.2 – Visualisation des descripteurs pertinents pour reconnaitre les expressions faciales d'enfants typiques

6.1.3 Le jeu sérieux

La preuve de concept du jeu sérieux, qui a été développée dans le cadre du projet JEMImE, intègre les méthodes présentées au chapitre 5. Ce jeu se décompose en 2 phases. La première, dite d'entrainement, apprend à l'enfant à produire l'expression faciale par imitation (figure 6.3a) et sur demande (figure 6.3b). L"enfant a un retour visuel de l'émotion produite sous forme de jauges de couleurs correspondant aux différentes expressions analysées figure (6.3c). L'enfant valide l'exercice s'il parvient à produire les expressions attendues avec un certain taux de réussite.



(a) (b) (c)

FIGURE 6.3 – Illustration de la phase d'apprentissage

Dans la seconde phase, dite de jeu, l'enfant est immergé dans un monde virtuel. Il peut se déplacer librement et interagir avec les éléments de l'environnement (6.4a). Dans ce monde, l'enfant est confronté à des situations d'interaction sociale, et il doit mettre en pratique ce qu'il a appris lors de la première phase. Par exemple, dans le scénario présenté dans la figure 6.4, des personnages virtuels qui jouent au ballon, lui demandent s'il souhaite se joindre à eux (6.4b). Soit le personnage virtuel lui tend le ballon (6.4c) et l'enfant doit alors produire une expression de joie, soit il lui explique que le ballon n'est pas pour lui (6.4d). Dans le deuxième cas, l'enfant pourra produire une expression de colère ou de tristesse pour valider le scénario. Si l'émotion attendue est correctement produite, l'enfant reçoit alors une récompense et dans tous les cas il peut reprendre son exploration du monde.

6.2 Le projet SMART SeNSE

SeNSE (Socio Emotional Signals, http://sense.isir.upmc.fr/) est un projet de recherche financé par le Labex SMART. Il s'intéresse aux signaux sociaux émotionnels échangés lors d'une interaction et couvre l'intégralité de la chaîne de traitement, allant de la capture des signaux (vidéo, audio, neurophysiologiques) jusqu'à leur exploitation (agent virtuel, interaction musicale,



(a) (b) (c) (d)

FIGURE 6.4 – Illustration de la phase de jeu en immersion dans le monde virtuel

groupe de personnes).

Plus spécifiquement, la dynamique des signaux sociaux contient une information importante pour l'expression de différents états affectifs. Keltner [74] illustre l'importance de cette dynamique avec l'exemple du sourire : un sourire long montre de l'amusement là où un regard fuyant suivi d'un sourire contrôlé peut signifier de l'embarras. Dans le cadre de la thèse de Thomas Janssoone, nous avons proposé une méthode de fouille de séquence (*sequence mining*) dont l'objectif est d'extraire, à partir d'un corpus d'étude, des règles d'association temporelles caractéristiques de l'expression d'une attitude sociale et d'utiliser ces règles pour contrôler le comportement d'un agent virtuel. La chaine de traitement intitulée Smart (*Social Multimodal Association with Timing*) est illustrée dans la figure 6.5.

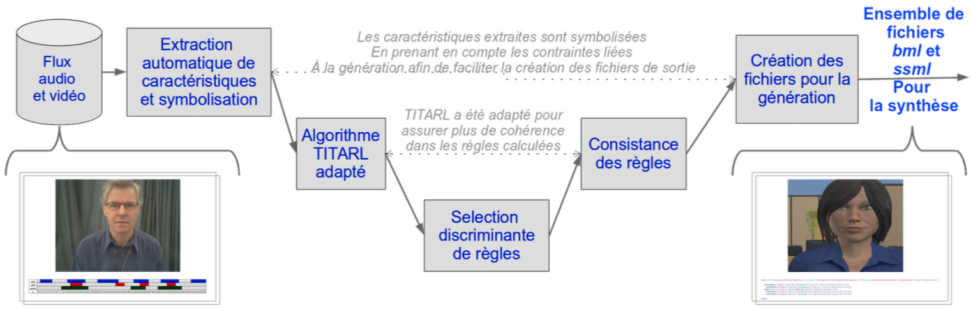


FIGURE 6.5 – Schéma de fonctionnement de SMART

Ces travaux ont donné lieu à plusieurs contributions :

La constitution d'un corpus d'analyse multimodal : ce corpus est constitué des allocutions du président Obama pour Thanksgiving et Pâques entre 2009 et 2015. Ces allocutions ont l'avantage d'être filmées de face (pour faciliter l'extraction automatique des signaux), avec une bonne qualité d'image, et présentent des attitudes variées. Des séquences courtes (environ 15s) ont ensuite été annotées indépendamment et aléatoirement, afin d'observer l'évolution des attitudes sur des pas de temps relativement fins [4].

Une méthode d'analyse bout-en-bout : une originalité forte de ce projet consiste à travailler directement à partir des données multimédia brutes, dont les signaux sociaux sont extraits automatiquement. En particulier, nous avons extrait des informations prosodiques, ainsi que des expressions faciales. Ces dernières sont des AU extraites à partir de la méthode présentée dans le chapitre 4. Cela constitue un défi par rapport aux approches de l'état de l'art, qui utilisent des informations sémantiques de plus haut niveau et annotées manuellement. Mais en contrepartie, une approche totalement automatique permet d'envisager une étude de corpus à plus grande échelle. En sortie de notre chaine de traitement, nous obtenons des fichiers BML (*Behavior Markup Language*) et SSML (*Speech Synthesis Markup Language*) permettant de

contrôler l'agent virtuel, et qui dépendent directement des règles que nous avons apprises à partir du corpus d'étude.

Une méthode d'extraction automatique des règles : inspirée de la méthode TITARL [60] (*Temporal Interval Tree Association Rule Learning*), cette méthode permet d'apprendre des règles d'associations temporelles, i.e. de trouver des associations entre différents signaux sociaux (par exemple, une attitude amicale se caractérise par des sourires et une intonation de voix variée), mais également une information sur la temporalité de ces signaux (durées et écarts temporels).

Des études perceptives ont validé l'approche proposée, et ont montré qu'il était donc possible de trouver de manière automatique, des associations temporelles entre des signaux sous forme de règles, et de les adapter pour synthétiser le comportement d'agents virtuels.

Publications en lien avec le chapitre

(T1) A. Dapogny, *A walk through randomness for face analysis in unconstrained environments*, **Thèse de doctorat.** 2016

(T2) T. Janssoone, *Analyse de signaux sociaux multimodaux : application à la synthèse d'attitudes sociales chez un agent conversationnel animé*, **Thèse de doctorat.** 2018

(J1) C. Grossard, L. Chaby, S. Hun, H. Pellerin, J. Bourgeois, A. Dapogny, H. Ding, S. Serret, P. Foulon, M. Chetouani, L. Chen, K. Bailly, O. Grynszpan et D. Cohen. *Children facial expression production : influence of age, gender, emotion subtype, elicitation condition and culture*, **Frontiers in Psychology**, **9** : 446, 2018.

(J2) C Grossard, O Grynspan, S Serret, AL Jouen, K Bailly et D Cohen. *Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (ASD)*, **Computers & Education** , **113** : 195-211, 2017.

(J3) T Janssoone, C Clavel, K Bailly et G Richard. *Règles d'Associations Temporelles de signaux sociaux pour la synthèse d'Agents Conversationnels Animés*, **Revue d'intelligence artificielle**, : 512-517, 2017.

(J4) C Grossard, S Hun, S Serret, O Grynszpan, P Foulon, A Dapogny, K Bailly, L Chaby et D Cohen. *Rééducation de l'expression émotionnelle chez l'enfant avec trouble du spectre autistique grâce aux supports numériques : le projet JEMImE*, **Neuropsychiatrie de l'Enfance et de l'Adolescence**, **65** (1) : 21–32, 2017.

(C1) A. Dapogny, C. Grossard, S. Hun, S. Serret, J. Bourgeois, H. Jean-Marie, P. Foulon, H. Ding, L. Chen, S. Dubuisson, O. Grynszpan, D. Cohen, K. Bailly. *JEMImE : A Serious Game to Teach Children with ASD How to Adequately Produce Facial Expressions*, **Workshop on Face and Gesture Analysis for Health Informatics (FGAHI'2018)** .

(C2) T. Janssoone, C. Clavel, K. Bailly, G. Richard. *Using temporal association rules for the synthesis of embodied conversational agents with a specific stance*, **International Conference on Intelligent Virtual Agents (IVA'2016)**.

(C3) S. Hun, S. Serret, C. Grossard, J. Bourgeois, O. Grynszpan, D. Cohen, F. Askénazy, A. Dapogny, L. Chen, S. Dubuisson, K. Bailly, *JEMImE, a serious game to teach emotional facial expressiveness for people with Autism Spectrum Disorders*, **International Meeting for Autism Research (IMFAR'2017)**.

Encadrement doctoral

— Arnaud Dapogny [10/2013–12/2016] (dir. Severine Dubuisson)

— Thomas Janssoone [12/2014–2/2018] (dir. Gaël Richard)

Projet collaboratif

— Projet ANR **JEMImE** (2013–2018) jemime.isir.upmc.fr/

— Projet SMART **Sense** (2014–2017) sense.isir.upmc.fr/

Chapitre 7

Conclusion et perspectives

Dans ce manuscrit nous avons présenté nos principales contributions dans le domaine de l'apprentissage automatique et de la vision par ordinateur appliqués à l'analyse automatique des expressions faciales. Nous pouvons les résumer par les trois points suivants.

Machine à Vecteur Support (SVM) Multi-Noyaux pour la combinaison de descripteurs hétérogènes : dans le cadre de la thèse de Thibaud Sénéchal, nous avons cherché à améliorer la robustesse des systèmes de détection d'Action Units (micro-mouvements du visage liés à une activation musculaire) en combinant, via une SVM multi-noyaux, des histogrammes de LGBP (Local Gabor Binary Pattern) extraits à différentes orientations et différentes fréquences spatiales. Cette représentation améliore la robustesse du système aux changements d'illumination. Nous avons également proposé une nouvelle fonction noyau, la fonction HDI, adaptée aux différences d'histogrammes qui permet d'améliorer la robustesse à l'identité. La sélection de descripteurs hétérogènes par SVM Multi-noyaux, combinée à un processus de type cascade attentionnelle, a également été proposée pour l'alignement d'un modèle déformable et la localisation de points caractéristiques du visage (thèse de Vincent Rapp).

Hard Multi-Task Metric Learning for Kernel Regression : dans le domaine de l'apprentissage supervisé en général et dans le domaine de l'analyse des expressions faciales en particulier, l'étape d'acquisition et d'annotation des données d'apprentissage est fastidieuse et peut nécessiter des compétences spécifiques comme c'est le cas par exemple pour annoter les Action Units (il existe très peu d'experts certifiés). Ainsi, les risques de sur-apprentissage sont élevés car le nombre de données d'apprentissage est très faible au regard de la difficulté de la tâche. Pour répondre à ce problème, nous avons proposé dans le cadre de la thèse de Jérémie Nicolle une nouvelle méthode d'apprentissage appelée Hard Multi-Task Metric Learning for Kernel Regression (H-MT-MLKR). Cette méthode introduit notamment une nouvelle formulation de la régularisation multi-tâche qui réduit le nombre de paramètres du modèle à estimer (et donc les risques sur-apprentissage) sans réduire sa complexité. Cette méthode d'apprentissage a également été appliquée avec succès dans une tâche d'alignement d'un modèle déformable.

Forêts aléatoires pour l'analyse dynamique et robuste : dans le cadre du projet ANR JEMImE que j'ai coordonné (2013-2018) et de la thèse d'Arnaud Dapogny, nous avons conçu des méthodes d'analyse émotionnelle (joie, colère, tristesse, peur...). Nous avons proposé une nouvelle méthode d'apprentissage par paires d'images qui s'appuie sur des Forêts Aléatoires Conditionnelles afin d'exploiter d'une part l'information temporelle de la séquence vidéo et d'autre part d'améliorer la robustesse aux variations de pose du visage. Cette méthode fonctionne en temps réel et a pu être intégrée dans le démonstrateur du projet ANR JEMImE. Nous avons également proposé une méthode d'apprentissage des forêts aléatoires qui privilégie des représentations spatialement localisées dans l'image. Combinée à des mémoires auto-associatives, cette méthode permet d'analyser des visages partiellement occultés.

Perspectives

Durant ces 7 années de recherche à l'ISIR, nous nous sommes donc attelés à lever les verrous permettant d'obtenir des méthodes d'analyse faciale, qui soient fonctionnelles en dehors du simple cadre académique. En particulier les travaux, entrepris à l'occasion du projet JEMImE, sont un pas dans cette direction. Pour autant le chemin est encore long avant que les systèmes automatiques atteignent la finesse d'analyse des humains dans des environnements complexes. Pour améliorer les systèmes existants, nous souhaitons avancer dans les prochaines années suivant les trois axes complémentaires présentés ci-après.

Analyse de données subtiles et non prototypiques : il est très difficile d'analyser des émotions subtiles car les bases de données de la littérature présentent souvent des émotions posées, standardisées et de forte intensité. De plus, les variations interpersonnelles sont importantes, ce qui rend la tâche d'annotation et de prédiction délicate. La réponse à cette problématique se trouve certainement à la fois dans les données qui devront être collectées, et dans les algorithmes qui devront être à même d'intégrer une manière d'étalonner les prédictions pour chaque individu. Une collaboration que nous avons initiée avec le service de pneumologie et réanimation de l'hôpital Pitié-Salpêtrière et l'UMRS 1158 Inserm-SU à laquelle ce service est associé, est une illustration de ce type de recherche. Le projet vise à caractériser la souffrance respiratoire (dyspnée) de patients placés dans l'incapacité d'en faire part verbalement à leurs soignants, par analyse des expressions faciales. Une base de données est déjà disponible, consistant en une trentaine de films enregistrés au cours, soit d'inductions expérimentales de dyspnée en laboratoire chez des sujets sains, soit de situations de dyspnée clinique chez des patients placés sous assistance respiratoire en réanimation. Ces films sont couplés aux enregistrements physiologiques correspondants et ont été annotés en AU par 3 experts certifiés FACS. Le défi ici sera donc d'identifier et de quantifier l'intensité de la dyspnée pour un sujet dans un environnement complexe avec, par exemple, une occultation d'une partie du visage par le respirateur artificiel. Une calibration est de plus envisageable car le système s'adresse à des personnes qui sont filmées sur une longue période (personnes généralement hospitalisées). D'autres enregistrements seront effectués au cours du projet, mais il est certain que nous ne pourrons toutefois pas tout attendre des données et espérer que n'importe quel modèle puisse capturer des informations émotionnelles subtiles, tout en étant robustes à de nombreuses sources de variations. La conception de l'architecture du modèle et de sa stratégie d'apprentissage seront donc déterminants.

Architecture des modèles : Les réseaux de neurones profonds et les forêts aléatoires constituent deux catégories de modèles d'apprentissage puissants qui ont connu de nombreux succès applicatifs ces dernières années. Les paradigmes qui sous-tendent ces modèles sont toutefois assez différents : les réseaux de neurones cherchent à apprendre une représentation hiérarchique au travers de projections non linéaires successives. La nature différentiable de ce modèle permet un apprentissage conjoint et de bout-en-bout de l'espace de représentation et de la fonction de prédiction. Les réseaux de neurones présentent cependant deux inconvénients majeurs : les temps de prédiction peuvent être longs, en particulier lors de l'évaluation des couches complètement connectées, et les réseaux sont très sensibles au sur-apprentissage lorsque le nombre de données est faible au regard de la tâche à prédire. A l'inverse, les arbres des forêts aléatoires apprennent un partitionnement hiérarchique de l'espace de représentation. Cette stratégie est particulièrement souhaitable lorsque les données sont morcelées dans l'espace de représentation (configuration fréquente pour des données issues d'environnements non contraints). Et les temps de traitement des forêts aléatoires sont réduits car seule une faible portion de la forêt est explorée à chaque prédiction. De plus, la combinaison des prédictions des modèles appris sur des échan-

tillons de données différents, limite les risques de sur-apprentissage. Notre objectif scientifique dans les années à venir est donc de proposer des architectures hybrides, qui combinent les avantages des deux approches, tout en atténuant leurs limitations respectives. Les premiers résultats sur les *Greedy Neural Forests*, présentés dans la section 5.4, ont montré la pertinence de ce type d'architecture, mais les possibilités sont nombreuses et seront explorées dans les années à venir au travers de la thèse d'Estephe Arnaud et du projet ANR JCJC Facil (*Face Interpretation with deep and ensemble Learning*) que je coordonne (2017-2021).

Stratégie d'apprentissage : compte tenu de la complexité des modèles actuels, l'un des enjeux majeurs est de concevoir des stratégies d'apprentissage qui limitent les risques de sur-apprentissage. Un moyen d'y parvenir consiste à utiliser au maximum l'information dont on dispose. Les architectures totalement différentiables, telles que les réseaux de neurones et les forêts de décisions neuronales, offrent une souplesse d'utilisation permettant la conception d'architectures modulaires. Nous pouvons ainsi imaginer d'exploiter des données faiblement annotées, ou annotées pour d'autres tâches via des formulations multitâches, ou des architectures spécifiques permettant d'intégrer des objectifs intermédiaires dans le réseau. Dans le cas de la reconnaissance des AU, la reconnaissance d'émotions prototypiques (joie, colère...) pourrait, par exemple, servir d'objectif intermédiaire pour améliorer l'espace de représentation des données, et permettrait ainsi d'exploiter les bases de données d'émotions, qui sont plus nombreuses et plus faciles à annoter que les bases d'AU. Ce schéma d'apprentissage reprend l'idée que nous avons exposée dans la section 5.3, mais, appliqué à un modèle neuronal, il permet un apprentissage bout-en-bout du modèle. Nous souhaitons également tester d'autres stratégies d'apprentissage combinant les forces de l'apprentissage profond et des méthodes d'ensemble. Les travaux préliminaires que nous menons dans la thèse d'Estèphe Arnaud, montrent que la substitution d'une couche de perceptrons totalement connectée par un comité de perceptrons multicouches indépendants améliore sensiblement les résultats, tout en conservant la flexibilité des architectures profondes. De plus, les résultats peuvent-être améliorés en apprenant à pondérer les contributions de chaque prédicteur conditionnellement à l'entrée, ce qui permet de spécialiser certaines parties du réseau à un type de données (une certaine orientation du visage par exemple). Dans une perspective à plus long terme, nous pouvons envisager d'apprendre conjointement les paramètres du modèle et son architecture en faisant, par exemple, croître les arbres neuronaux à la manière des arbres de décision, ou en apprenant des structures conditionnelles à partir d'architectures performantes qui existent et qui ont déjà été entrainées, mais qui sont souvent très lourdes (Inception ou ResNet, par exemple).

Publications issues des travaux de recherche (10/2018)

Articles dans des revues internationales avec comité de lecture

[1] Arnaud Dapogny, Kevin Bailly et Séverine Dubuisson : Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *International Journal of Computer Vision (IJCV)*, 126(2-4):255–271, 2018.

[2] Arnaud Dapogny, Kevin Bailly et Severine Dubuisson : Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests. *IEEE Transactions on Affective Computing*, 2018.

[3] Arnaud Dapogny et Kévin Bailly : Face alignment with cascaded semi-parametric deep greedy neural forests. *Pattern Recognition Letters*, 102:75–81, 2018.

[4] Charline Grossard, Laurence Chaby, Stephanie Hun, Hugues Pellerin, Jérémy Bourgeois, Arnaud Dapogny, Huaxiong Ding, Sylvie Serret, Pierre Foulon, Mohamed Chetouani *et al.* : Children facial expression production : influence of age, gender, emotion subtype, elicitation condition and culture. *Frontiers in Psychology*, 9:446, 2018.

[5] Lisa Ouss, Marie Therese Le Normand, Kevin Bailly, Marluce Leitgel Gille, Christelle Gosme, Roberta Simas, Julia Wencke, Xavier Jeudon, Stephanie Thepot, Telma Da Silva *et al.* : Developmental trajectories of hand movements in typical infants and those at risk of developmental disorders : An observational study of kinematics during the first year of life. *Frontiers in Psychology*, 9:83, 2018.

[6] Charline Grossard, Ouriel Grynspan, Sylvie Serret, Anne-Lise Jouen, Kevin Bailly et David Cohen : Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (asd). *Computers & Education*, 113:195–211, 2017.

[7] Jérémie Nicolle, Kévin Bailly et Mohamed Chetouani : Real-time facial action unit intensity prediction with regularized metric learning. *Image and Vision Computing*, 52:1–14, 2016.

[8] Thibaud Senechal, Kevin Bailly et Lionel Prevost : Impact of action unit detection in automatic emotion recognition. *Pattern Analysis and Applications*, 17(1):51–67, 2014.

[9] Vincent Rapp, Kevin Bailly, Thibaud Senechal et Lionel Prevost : Multi-kernel appearance model. *Image and Vision Computing*, 31(8):542–554, 2013.

[10] Thibaud Senechal, Vincent Rapp, Hanan Salam, Renaud Seguier, Kevin Bailly et Lionel Prevost : Facial action recognition combining heterogeneous features via multikernel learning. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 42(4):993–1005, 2012.

[11] Kevin Bailly et Maurice Milgram : Boosting feature selection for neural network based regression. *Neural Networks*, 22(5-6):748–756, 2009.

Articles dans des revues nationales avec comité de lecture

[1] Lisa Ouss, Marie Therese Le Normand, Kevin Bailly, Marluce Leitgel Gille, Christelle Gosme, Roberta Simas, Julia Wencke, Xavier Jeudon, Stephanie Thepot, Telma Da Silva *et al.* : Developmental trajectories of hand movements in typical infants and those at risk of developmental disorders : An observational study of kinematics during the first year of life. *Frontiers in Psychology*, 9:83, 2018.

[2] C Grossard, S Hun, S Serret, O Grynszpan, P Foulon, A Dapogny, K Bailly, L Chaby et D Cohen : Rééducation de l'expression émotionnelle chez l'enfant avec trouble du spectre autistique grâce aux supports numériques : le projet jemime. *Neuropsychiatrie de l'Enfance et de l'Adolescence*, 65(1):21–32, 2017.

[3] Thomas Janssoone, Chloé Clavel, Kévin Bailly et Gaël Richard : Règles d'associations temporelles de signaux sociaux pour la synthèse d'agents conversationnels animés. *Revue d'intelligence artificielle–no*, 511:537, 2017.

Conférences invitées internationales

[1] Kevin Bailly : Random forests for facial expression analysis. *In SMART School on Computational Social and Behavioral Sciences*, 2017.

Articles de conférences internationales avec actes et comité de lecture

[1] Arnaud Dapogny, Charline Grossard, Stephanie Hun, Sylvie Serret, Jeremy Bourgeois, Hedy Jean-Marie, Pierre Foulon, Huaxiong Ding, Liming Chen, Severine Dubuisson *et al.* : Jemime : A serious game to teach children with asd how to adequately produce facial expressions. *In IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018.

[2] Arnaud Dapogny et Kevin Bailly : Investigating deep neural forests for facial expression recognition. *In IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018.

[3] Alex Vásquez, Arnaud Dapogny, Kévin Bailly et Véronique Perdereau : Sequential recognition of in-hand object shape using a collection of neural forests. *In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[4] Arnaud Dapogny, Kevin Bailly et Séverine Dubuisson : Multi-output random forests for facial action unit detection. *In IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2017.

[5] Jonathan Aigrain, Arnaud Dapogny, Kévin Bailly, Séverine Dubuisson, Marcin Detyniecki et Mohamed Chetouani : On leveraging crowdsourced data for automatic perceived stress detection. *In ACM International Conference on Multimodal Interaction (ICMI)*, 2016.

[6] Thomas Janssoone, Chloé Clavel, Kévin Bailly et Gaël Richard : Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. *In International Conference on Intelligent Virtual Agents (IVA)*, 2016.

[7] Arnaud Dapogny, Kevin Bailly et Séverine Dubuisson : Pairwise conditional random forests for facial expression recognition. *In IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] Jeremie Nicolle, Kevin Bailly et Mohamed Chetouani : Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. *In IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[9] Arnaud Dapogny, Kevin Bailly et Séverine Dubuisson : Dynamic facial expression recognition by joint static and multi-time gap transition classification. *In IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[10] Lucas Zamuner, Kevin Bailly et Erwan Bigorgne : A pose-adaptive constrained local model for accurate head pose tracking. *In International Conference on Pattern Recognition (ICPR)*, 2014.

[11] Jérémie Nicolle, Kévin Bailly, Vincent Rapp et Mohamed Chetouani : Locating facial landmarks with binary map cross-correlations. *In IEEE International Conference on Image Processing (ICIP)*, 2013.

[12] Kevin Bailly, Maurice Milgram, Philippe Phothisane et Erwan Bigorgne : Learning global cost function for face alignment. *In International Conference on Pattern Recognition (ICPR)*, 2012.

[13] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost et Mohamed Chetouani : Robust continuous prediction of human emotions using multiscale dynamic cues. *In ACM international conference on Multimodal interaction (ICMI)*, 2012.

[14] Thibaud Senechal, Vincent Rapp, Hanan Salam, Renaud Seguier, Kevin Bailly et Lionel Prevost : Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units. *In IEEE International Conference on Automatic Face & Gesture Recognition(FG)*, 2011.

[15] Vincent Rapp, Thibaud Senechal, Kevin Bailly et Lionel Prevost : Multiple kernel learning svm and statistical validation for facial landmark detection. *In IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2011.

[16] Thibaud Senechal, Kevin Bailly et Lionel Prevost : Automatic facial action detection using histogram variation between emotional states. *In International Conference on Pattern Recognition (ICPR)*, 2010.

[17] Azza Mokadem, Maurice Charbit, Gérard Chollet et Kevin Bailly : Age regression based on local image features. *In Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, 2010.

[18] Kevin Bailly et Maurice Milgram : Head pan angle estimation by a nonlinear regression on selected features. *In IEEE International Conference on Image Processing (ICIP)*, 2009.

[19] Kevin Bailly, Maurice Milgram et Philippe Phothisane : Head pose estimation by a stepwise nonlinear regression. *In International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2009.

[20] Kevin Bailly et Maurice Milgram : Bisar : Boosted input selection algorithm for regression. *In International Joint Conference on Neural Networks (IJCNN)*, 2009.

[21] Kevin Bailly et Maurice Milgram : Recursive shape and pose determination using deformable model. *In Progress in Pattern Recognition, Image Analysis and Applications*, 2008.

[22] Kevin Bailly et Maurice Milgram : Head pose determination using synthetic images. *In International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2008.

Bibliographie

[1] Yasser Aidarouss, Sylvain Le Gallou, Abdul Sattar et Renaud Seguier : Face Alignment using active appearance model optimized by simplex. *In International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 231–234, Barcelona, Spain, 2007.

[2] Ghulam Ali, Muhammad Amjad Iqbal et Tae-Sun Choi : Boosted NNE collections for multicultural facial expression recognition. *Pattern Recognition*, 55:14–27, 2016. ISSN 0031-3203.

[3] Zara Ambadar, Jeffrey F. Cohn et Lawrence Ian Reed : All Smiles are Not Created Equal : Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous. *Journal of Nonverbal Behavior*, 33(1):17–34, 2009. ISSN 0191-5886, 1573-3653.

[4] Nalini Ambady et Robert Rosenthal : Thin slices of expressive behavior as predictors of interpersonal consequences : A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992. ISSN 1939-1455(Electronic),0033-2909(Print).

[5] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F. Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M. Prkachin et Patricia E. Solomon : The painful face – Pain expression recognition using active appearance models. *Image and Vision Computing*, 27(12):1788–1796, 2009. ISSN 0262-8856.

[6] A. Asthana, S. Zafeiriou, S. Cheng et M. Pantic : Robust Discriminative Response Map Fitting with Constrained Local Models. *In 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013.

[7] A. Asthana, S. Zafeiriou, S. Cheng et M. Pantic : Incremental Face Alignment in the Wild. *In 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014.

[8] T. Baltrušaitis, M. Mahmoud et P. Robinson : Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. *In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6, 2015.

[9] T. Baltrušaitis, D. McDuff, N. Banda, M. Mahmoud, R. e Kaliouby, P. Robinson et R. Picard : Real-time inference of mental states from facial expressions and upper body gestures. *In Face and Gesture 2011*, pages 909–914, 2011.

[10] M. S. Bartlett, G. Littlewort, I. Fasel et J. R. Movellan : Real Time Face Detection and Facial Expression Recognition : Development and Applications to Human Computer Interaction. *In 2003 Conference on Computer Vision and Pattern Recognition Workshop*, volume 5, pages 53–53, 2003.

[11] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel et J. Movellan : Recognizing facial expression : Machine learning and application to spontaneous behavior. *In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 568–573 vol. 2, 2005.

[12] Marian Stewart Bartlett, Gwen Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, Javier R Movellan et others : Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006.

[13] J. J. Bazzo et M. V. Lamar : Recognizing facial actions using Gabor wavelets with neutral face average difference. *In Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 505–510, 2004.

[14] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman et Neeraj Kumar : Localizing Parts of Faces Using a Consensus of Exemplars. *In The 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[15] Asa Ben-Hur et William Stafford Noble : Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl_1):i38–i46, 2005. ISSN 1367-4803.

[16] Leo Breiman : Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 0885-6125, 1573-0565.

[17] Leo Breiman : Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565.

[18] N. Brunet, F. Perez et F. De la Torre : Learning good features for Active Shape Models. *In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 206–211, 2009.

[19] X. P. Burgos-Artizzu, P. Perona et P. Dollár : Robust Face Landmark Estimation under Occlusion. *In 2013 IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[20] X. Cao, Y. Wei, F. Wen et J. Sun : Face alignment by Explicit Shape Regression. *In 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894, 2012.

[21] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, I. Matthews et S. Sridharan : In the Pursuit of Effective Affective Computing : The Relationship Between Features and Registration. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1006–1016, 2012. ISSN 1083-4419.

[22] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn et S. Sridharan : Person-independent facial expression detection using Constrained Local Models. *In Face and Gesture 2011*, pages 915–920, 2011.

[23] W. Chu, F. De la Torre et J. F. Cohn : Selective Transfer Machine for Personalized Facial Action Unit Detection. *In 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

[24] Chao-Fa Chuang et Frank Y. Shih : Recognizing facial action units using independent component analysis and support vector machine. *Pattern Recognition*, 39(9):1795–1798, 2006. ISSN 0031-3203.

[25] James A. Coan et John J. B. Allen, éditeurs. *Handbook of Emotion Elicitation and Assessment.* Handbook of emotion elicitation and assessment. Oxford University Press, New York, NY, US, 2007. ISBN 978-0-19-516915-7 (Hardcover).

[26] T. F. Cootes, G. J. Edwards et C. J. Taylor : Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. ISSN 0162-8828.

[27] T. F. Cootes, C. J. Taylor, D. H. Cooper et J. Graham : Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. ISSN 1077-3142.

[28] S. F. Cotter : Sparse Representation for accurate classification of corrupted and occluded facial expressions. *In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 838–841, 2010.

[29] M. Dahmane et J. Meunier : Emotion recognition using dynamic grid-based HoG features. *In Face and Gesture 2011*, pages 884–888, 2011.

[30] N. Dalal et B. Triggs : Histograms of oriented gradients for human detection. *In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[31] A. Dapogny, K. Bailly et S. Dubuisson : Pairwise Conditional Random Forests for Facial Expression Recognition. *In 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3783–3791, 2015.

[32] A. Dapogny, K. Bailly et S. Dubuisson : Multi-Output Random Forests for Facial Action Unit Detection. *In 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 135–140, 2017.

[33] A Dapogny, K. Bailly et S. Dubuisson : Dynamic Pose-Robust Facial Expression Recognition by Multi-View Pairwise Conditional Random Forests. *IEEE Transactions on Affective Computing*, pages 1–1, 2018. ISSN 1949-3045.

[34] A. Dapogny, C. Grossard, S. Hun, S. Serret, J. Bourgeois, H. Jean-Marie, P. Foulon, H. Ding, L. Chen, S. Dubuisson, O. Grynszpan, D. Cohen et K. Bailly : JEMImE - A Serious Game to Teach Children with ASD How to Adequately Produce Facial Expressions. *In 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 723–730, 2018.

[35] Arnaud Dapogny : *A Walk through Randomness for Face Analysis in Unconstrained Environments*. Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, 2016.

[36] Arnaud Dapogny et Kévin Bailly : Face alignment with cascaded semi-parametric deep greedy neural forests. *Pattern Recognition Letters*, 102:75–81, janvier 2018. ISSN 0167-8655.

[37] Arnaud Dapogny, Kevin Bailly et Séverine Dubuisson : Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit Detection. *International Journal of Computer Vision*, 126(2):255–271, 2018. ISSN 1573-1405.

[38] A. Dhall, A. Asthana, R. Goecke et T. Gedeon : Emotion recognition using PHOG and LPQ features. *In Face and Gesture 2011*, pages 878–883, 2011.

[39] A. Dhall, R. Goecke, S. Lucey et T. Gedeon : Static facial expression analysis in tough conditions : Data, evaluation protocol and benchmark. *In 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112, 2011.

[40] Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi et Tom Gedeon : Video and Image Based Emotion Recognition Challenges in the Wild : EmotiW 2015. *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 423–426, New York, NY, USA, 2015. ACM.

[41] Hamdi Dibeklioğlu, Albert Ali Salah et Theo Gevers : Are You Really Smiling at Me ? Spontaneous versus Posed Enjoyment Smiles. *In Computer Vision – ECCV 2012*, Lecture Notes in Computer Science, pages 525–538. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-33711-6 978-3-642-33712-3.

[42] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman et T. J. Sejnowski : Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10): 974–989, 1999. ISSN 0162-8828.

[43] Xuanyi Dong, Yan Yan, Wanli Ouyang et Yi Yang : Style Aggregated Network for Facial Landmark Detection. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] R. Donner, M. Reiter, G. Langs, P. Peloschek et H ; Bischof : Fast Active Appearance Model Search Using Canonical Correlation Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1690–1694, 2006.

[45] Paul Ekman et Wallace V Friesen : Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[46] Paul Ekman et Wallace V. Friesen : Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976. ISSN 1573-3653.

[47] S. Eleftheriadis, O. Rudovic et M. Pantic : Discriminative Shared Gaussian Processes for Multiview and View-Invariant Facial Expression Recognition. *IEEE Transactions on Image Processing*, 24(1):189–204, 2015. ISSN 1057-7149.

[48] Theodoros Evgeniou et Massimiliano Pontil : Regularized Multi–task Learning. *In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 109–117, New York, NY, USA, 2004. ACM. ISBN 978-1-58113-888-7.

[49] Yin Fan, Xiangju Lu, Dian Li et Yuanliu Liu : Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. *In Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 445–450, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4556-9.

[50] X. Gao, Y. Su, X. Li et D. Tao : A Review of Active Appearance Models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(2):145–158, 2010. ISSN 1094-6977.

[51] C. Garcia et M. Delakis : Convolutional face finder : A neural architecture for fast and robust face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1408–1423, 2004. ISSN 0162-8828.

[52] G. Ghiasi et C. C. Fowlkes : Occlusion Coherence : Localizing Occluded Faces with a Hierarchical Deformable Part Model. *In 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1899–1906, 2014.

[53] Jeffrey M. Girard, Jeffrey F. Cohn et Fernando De la Torre : Estimating smile intensity : A better way. *Pattern recognition letters*, 66:13–21, 2015. ISSN 0167-8655.

[54] Mehmet Gönen et Ethem Alpaydin : Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research*, 12(Jul):2211–2268, 2011. ISSN ISSN 1533-7928.

[55] Isabel Gonzalez, Hichem Sahli, Valentin Enescu et Werner Verhelst : Context-Independent Facial Action Unit Recognition Using Shape and Gabor Phase Information. *In Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science, pages 548–557. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-24599-2 978-3-642-24600-5.

[56] Ralph Gross, Iain Matthews et Simon Baker : Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093, 2005. ISSN 0262-8856.

[57] Charline Grossard, Laurence Chaby, Stéphanie Hun, Hugues Pellerin, Jérémy Bourgeois, Arnaud Dapogny, Huaxiong Ding, Sylvie Serret, Pierre Foulon, Mohamed Chetouani, Liming Chen, Kevin Bailly, Ouriel Grynszpan et David Cohen : Children Facial Expression Production : Influence of Age, Gender, Emotion Subtype, Elicitation Condition and Culture. *Frontiers in Psychology*, 9, 2018. ISSN 1664-1078.

[58] H. Gu et Q. Ji : Facial Event Classification with Task Oriented Dynamic Bayesian Network. *In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.(CVPR)*, volume 02, pages 870–875, 2004. ISBN 978-0-7695-2158-9.

[59] A. Gudi, H. E. Tasli, T. M. den Uyl et A. Maroulis : Deep learning based FACS Action Unit occurrence and intensity estimation. *In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–5, 2015.

[60] Mathieu Guillame-Bert et James L. Crowley : Learning Temporal Association Rules on Symbolic Time Sequences. *In Asian Conference on Machine Learning*, pages 159–174, 2012.

[61] Rain Eric Haamer, Eka Rusadze, Iiris Lüsi, Tauseef Ahmed, Sergio Escalera et Gholamreza Anbarjafari : Review on Emotion Recognition Databases. *Human-Robot Interaction-Theory and Application. IntechOpen*, 2018.

[62] B. Hasani et M. H. Mahoor : Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2278–2288, 2017.

[63] T. Hassner, S. Harel, E. Paz et R. Enbar : Effective face frontalization in unconstrained images. *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4295–4304, 2015.

[64] M. Hayat, M. Bennamoun et A. El-Sallam : Evaluation of Spatiotemporal Detectors and Descriptors for Facial Expression Recognition. *In 2012 5th International Conference on Human System Interactions*, pages 43–47, 2012.

[65] Lianghua He, Cairong Zou, Li Zhao et Die Hu : An Enhanced LBP Feature Based on Facial Expression Recognition. *In 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 3300–3303, 2005.

[66] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou et T. S. Huang : Multi-view facial expression recognition. *In 2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, 2008.

[67] Xiaohua Huang, Guoying Zhao, Wenming Zheng et Matti Pietikäinen : Towards a dynamic expression recognition system under facial occlusion. *Pattern Recognition Letters*, 33(16):2181–2191, 2012. ISSN 0167-8655.

[68] S. Jaiswal et M. Valstar : Deep learning the dynamic appearance and shape of facial action units. *In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8, 2016.

[69] L. A. Jeni, J. M. Girard, J. F. Cohn et F. De la Torre : Continuous AU intensity estimation using localized, sparse facial feature space. *In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2013.

[70] Oliver Jesorsky, Klaus J. Kirchberg et Robert W. Frischholz : Robust Face Detection Using the Hausdorff Distance. *In Audio- and Video-Based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 90–95. Springer, Berlin, Heidelberg, 2001. ISBN 978-3-540-42216-7 978-3-540-45344-4.

[71] H. Jung, S. Lee, J. Yim, S. Park et J. Kim : Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. *In 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2983–2991, 2015.

[72] Heechul Jung, Sihaeng Lee, Sunjeong Park, Injae Lee, Chunghyun Ahn et Junmo Kim : Deep Temporal Appearance-Geometry Network for Facial Expression Recognition. *arXiv :1503.01532 [cs]*, 2015.

[73] T. Kanade, J. F. Cohn et Yingli Tian : Comprehensive database for facial expression analysis. *In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

[74] Dacher Keltner : Signs of appeasement : Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68(3):441–454, 1995. ISSN 1939-1315(Electronic),0022-3514(Print).

[75] M. Khademi et L. P. Morency : Relative facial action unit detection. *In IEEE Winter Conference on Applications of Computer Vision*, pages 1090–1095, 2014.

[76] P. Khorrami, T. L. Paine et T. S. Huang : Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition ? *In 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 19–27, 2015.

[77] Bo-Kyeong Kim, Hwaran Lee, Jihyeon Roh et Soo-Young Lee : Hierarchical Committee of Deep CNNs with Exponentially-Weighted Decision Fusion for Static Facial Expression Recognition. *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 427–434, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3912-4.

[78] P. Kontschieder, M. Fiterau, A. Criminisi et S. R. Bulò : Deep Neural Decision Forests. *In 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1467–1475, 2015.

[79] Irene Kotsia, Ioan Buciu et Ioannis Pitas : An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26(7):1052–1067, 2008. ISSN 0262-8856.

[80] M. Kowalski, J. Naruniec et T. Trzcinski : Deep Alignment Network : A Convolutional Neural Network for Robust Face Alignment. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2034–2043, 2017.

[81] Balaji Lakshminarayanan, Daniel M Roy et Yee Whye Teh : Mondrian Forests : Efficient Online Random Forests. *In* Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence et K. Q. Weinberger, éditeurs : *Advances in Neural Information Processing Systems 27*, pages 3140–3148. Curran Associates, Inc., 2014.

[82] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev et Thomas S Huang : Interactive facial feature localization. *In European Conference on Computer Vision*, pages 679–692. Springer, 2012.

[83] Y. Li, J. Chen, Y. Zhao et Q. Ji : Data-Free Prior Model for Facial Action Unit Recognition. *IEEE Transactions on Affective Computing*, 4(2):127–141, 2013. ISSN 1949-3045.

[84] Y. Li, S. Liu, J. Yang et M. H. Yang : Generative Face Completion. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5892–5900, 2017.

[85] Dieu Linh Tran, Robert Walecki, Ognjen (Oggi) Rudovic, Stefanos Eleftheriadis, Bjorn Schuller et Maja Pantic : DeepCoder : Semi-Parametric Variational Autoencoders for Automatic Facial Action Coding. *In 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3190–3199, 2017.

[86] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan et M. Bartlett : The computer expression recognition toolbox (CERT). *In Face and Gesture 2011*, pages 298–305, 2011.

[87] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind et Javier Movellan : Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006. ISSN 0262-8856.

[88] Mengyi Liu, Shaoxin Li, Shiguang Shan et Xilin Chen : AU-aware Deep Networks for facial expression recognition. *In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013.

[89] P. Liu, S. Han, Z. Meng et Y. Tong : Facial Expression Recognition via a Boosted Deep Belief Network. *In 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[90] D. G. Lowe : Object recognition from local scale-invariant features. *In Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.

[91] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar et I. Matthews : The Extended Cohn-Kanade Dataset (CK+) : A complete dataset for action unit and emotion-specified expression. *In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[92] M. J. Lyons, J. Budynek et S. Akamatsu : Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999. ISSN 0162-8828.

[93] B. Martinez, M. F. Valstar, B. Jiang et M. Pantic : Automatic Analysis of Facial Actions : A Survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2017. ISSN 1949-3045.

[94] Brais Martinez et Michel F. Valstar : L2,1-based regression and prediction accumulation across views for robust facial landmark detection. *Image and Vision Computing*, 47:36–44, 2016. ISSN 0262-8856.

[95] Iain Matthews et Simon Baker : Active Appearance Models Revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004. ISSN 0920-5691, 1573-1405.

[96] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh et J. F. Cohn : DISFA A Spontaneous Facial Action Intensity Database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. ISSN 1949-3045.

[97] M. K. Abd El Meguid et M. D. Levine : Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers. *IEEE Transactions on Affective Computing*, 5(2):141–154, 2014. ISSN 1949-3045.

[98] H. Meng, B. Romera-Paredes et N. Bianchi-Berthouze : Emotion recognition by two view SVM_2K classifier on dynamic facial expression features. *In Face and Gesture 2011*, pages 854–859, 2011.

[99] Stephen Milborrow et Fred Nicolls : Locating Facial Features with an Extended Active Shape Model. *In Computer Vision – ECCV 2008*, Lecture Notes in Computer Science, pages 504–513. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-88692-1 978-3-540-88693-8.

[100] Z. Ming, A. Bugeau, J. Rouas et T. Shochi : Facial Action Units intensity estimation by the fusion of features with multi-kernel Support Vector Machine. *In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6, 2015.

[101] S. Moore et R. Bowden : Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011. ISSN 1077-3142.

[102] Stephen Moore et Richard Bowden : The effects of Pose on Facial Expression Recognition. *In Proceedings of the British Machine Vision Conference*, pages 79.1–79.11. BMVA Press, 2009. ISBN 1-901725-39-1.

[103] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis et Stefan Winkler : Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning. *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 443–449, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3912-4.

[104] J. Nicolle, K. Bailly et M. Chetouani : Facial Action Unit intensity prediction via Hard Multi-Task Metric Learning for Kernel Regression. *In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6, 2015.

[105] Jérémie Nicolle : *Reading Faces. Using Hard Multi-Task Metric Learning for Kernel Regression.* Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, 2016.

[106] Jérémie Nicolle, Kévin Bailly et Mohamed Chetouani : Real-time facial action unit intensity prediction with regularized metric learning. *Image and Vision Computing*, 52:1–14, 2016. ISSN 0262-8856.

[107] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost et Mohamed Chetouani : Robust Continuous Prediction of Human Emotions Using Multiscale Dynamic Cues. *In Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 501–508, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1467-1.

[108] Michael M Nordstrøm, Mads Larsen, Janusz Sierakowski et Mikkel B Stegmann : The IMM face database. *environment*, 22(10):1319–1331, 2003.

[109] Timo Ojala, Matti Pietikäinen et David Harwood : A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996. ISSN 0031-3203.

[110] Margarita Osadchy, Yann Le Cun et Matthew L. Miller : Synergistic Face Detection and Pose Estimation with Energy-Based Models. *Journal of Machine Learning Research*, 8(May):1197–1215, 2007. ISSN ISSN 1533-7928.

[111] Ebenezer Owusu, Yongzhao Zhan et Qi Rong Mao : A neural-AdaBoost based facial expression recognition system. *Expert Systems with Applications*, 41(7):3383–3390, 2014. ISSN 0957-4174.

[112] M. Pantic, M. Valstar, R. Rademaker et L. Maat : Web-based database for facial expression analysis. *In 2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, 2005.

[113] O. M. Parkhi, A. Vedaldi et A. Zisserman : Deep Face Recognition. *In British Machine Vision Conference*, 2015.

[114] John C. Platt : Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *In Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[115] M. Ranzato, J. Susskind, V. Mnih et G. Hinton : On deep generative models with applications to recognition. *In CVPR 2011*, pages 2857–2864, 2011.

[116] Vincent Rapp, Kevin Bailly, Thibaud Senechal et Lionel Prevost : Multi-Kernel Appearance Model. *Image and Vision Computing*, 31(8):542–554, 2013. ISSN 0262-8856.

[117] S. Ren, X. Cao, Y. Wei et J. Sun : Face Alignment at 3000 FPS via Regressing Local Binary Features. *In 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[118] S. Romdhani et T. Vetter : Efficient, robust and accurate fitting of a 3D morphable model. *In Proceedings Ninth IEEE International Conference on Computer Vision*, pages 59–66 vol.1, 2003.

[119] O. Rudovic, M. Pantic et I. Patras : Coupled Gaussian processes for pose-invariant facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1357–1369, 2013. ISSN 0162-8828.

[120] A. Saffari, C. Leistner, J. Santner, M. Godec et H. Bischof : On-line Random Forests. *In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1393–1400, 2009.

[121] Jason M. Saragih, Simon Lucey et Jeffrey F. Cohn : Deformable Model Fitting by Regularized Landmark Mean-Shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. ISSN 0920-5691, 1573-1405.

[122] E. Sariyanidi, H. Gunes et A. Cavallaro : Automatic Analysis of Facial Affect : A Survey of Registration, Representation, and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, 2015. ISSN 0162-8828.

[123] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur et Lale Akarun : Bosphorus Database for 3D Face Analysis. *In Biometrics and Identity Management*, Lecture Notes in Computer Science, pages 47–56. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-89990-7 978-3-540-89991-4.

[124] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie et Maja Pantic : AVEC 2012 : The Continuous Audio/Visual Emotion Challenge. *In Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 449–456, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1467-1.

[125] T. Senechal, K. Bailly et L. Prevost : Automatic Facial Action Detection Using Histogram Variation Between Emotional States. *In 2010 20th International Conference on Pattern Recognition*, pages 3752–3755, 2010.

[126] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly et L. Prevost : Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):993–1005, 2012. ISSN 1083-4419.

[127] Thibaud Senechal : *Ce Que Le Visage Révèle : Analyse Des Mouvements Faciaux Pour l'interprétation Émotionnelle*. Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, 2011. 2011PA066586.

[128] Caifeng Shan, Shaogang Gong et P. W. McOwan : Robust facial expression recognition using local binary patterns. *In IEEE International Conference on Image Processing 2005*, volume 2, pages II–370–3, 2005.

[129] Caifeng Shan, Shaogang Gong et Peter W. McOwan : Facial expression recognition based on Local Binary Patterns : A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. ISSN 0262-8856.

[130] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos et M. Pantic : The First Facial Landmark Tracking in-the-Wild Challenge : Benchmark and Results. *In 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 1003–1011, 2015.

[131] Seyedehsamaneh Shojaeilangari, Wei-Yun Yau, Jun Li et Eam-Khwang Teoh : Multiscale analysis of local phase and local orientation for dynamic facial expression recognition. *Journal ISSN*, 1(1), 2014.

[132] Terence Sim, Simon Baker et Maan Bsat : The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces. Rapport technique CMU-RI-TR-01-02, Carnegie Mellon University, Pittsburgh, PA, 2001.

[133] R. Srivastava, S. Roy, S. Yan et T. Sim : Accumulated motion images for facial expression recognition in videos. *In Face and Gesture 2011*, pages 903–908, 2011.

[134] Yi Sun, Xiaogang Wang et Xiaoou Tang : Deep Convolutional Network Cascade for Facial Point Detection. *In 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[135] U. Tariq, K. H. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. S. Huang, X. Lv et T. X. Han : Emotion recognition from an ensemble of features. *In Face and Gesture 2011*, pages 872–877, 2011.

[136] Usman Tariq, Jianchao Yang et Thomas S. Huang : Multi-view Facial Expression Recognition Analysis with Generic Sparse Coding Feature. *In Computer Vision – ECCV 2012. Workshops and Demonstrations*, Lecture Notes in Computer Science, pages 578–588. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-33884-7 978-3-642-33885-4.

[137] Robert Tibshirani : Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246.

[138] Y. Tong, W. Liao et Q. Ji : Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007. ISSN 0162-8828.

[139] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos et Stefanos Zafeiriou : Mnemonic Descent Method : A Recurrent Process Applied for End-To-End Face Alignment. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[140] G. Tzimiropoulos : Project-Out Cascaded Regression with an application to face alignment. *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3659–3667, 2015.

[141] G. Tzimiropoulos et M. Pantic : Gauss-Newton Deformable Part Models for Face Alignment In-the-Wild. *In 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[142] Georgios Tzimiropoulos, Joan Alabort-i-Medina, Stefanos Zafeiriou et Maja Pantic : Generic Active Appearance Models Revisited. *In Computer Vision – ACCV 2012*, Lecture Notes in Computer Science, pages 650–663. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-37430-2 978-3-642-37431-9.

[143] M. Valstar, M. Pantic et I. Patras : Motion history for facial action detection in video. *In 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, volume 1, pages 635–640 vol.1, 2004.

[144] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic et J. F. Cohn : FERA 2015 - second Facial Expression Recognition and Analysis challenge. *In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–8, 2015.

[145] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic et K. Scherer : The first facial expression recognition and analysis challenge. *In Face and Gesture 2011*, pages 921–926, 2011.

[146] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic et K. Scherer : Meta-Analysis of the First Facial Expression Recognition Challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):966–979, 2012. ISSN 1083-4419.

[147] R. L. Vieriu, S. Tulyakov, S. Semeniuta, E. Sangineto et N. Sebe : Facial expression recognition under a wide range of head poses. *In 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7, 2015.

[148] P. Viola et M. Jones : Rapid object detection using a boosted cascade of simple features. *In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518 vol.1, 2001.

[149] Frank Wallhoff : Database with Facial Expressions and Emotions from Technical University of Munich (FEEDTUM), 2006.

[150] J. Wang, S. Wang et Q. Ji : Early Facial Expression Recognition Using Hidden Markov Models. *In 2014 22nd International Conference on Pattern Recognition*, pages 4594–4599, 2014.

[151] Z. Wang, S. Wang et Q. Ji : Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition. *In 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3422–3429, 2013.

[152] Zhan Wang, Qiuqi Ruan et Gaoyun An : Facial expression recognition using sparse local Fisher discriminant analysis. *Neurocomputing*, 174:756–766, 2016. ISSN 0925-2312.

[153] Kilian Q. Weinberger et Gerald Tesauro : Metric Learning for Kernel Regression. *In Artificial Intelligence and Statistics*, pages 612–619, 2007.

[154] Philipp Werner, Frerk Saxen et Ayoub Al-Hamadi : Handling Data Imbalance in Automatic Facial Action Intensity Estimation. *In Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[155] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett et J. Movellan : Toward Practical Smile Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009. ISSN 0162-8828.

[156] J. Whitehill et C. W. Omlin : Haar features for FACS AU recognition. *In 7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 5 pp.–101, 2006.

[157] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S. Bartlett et J. R. Movellan : Action unit recognition transfer across datasets. *In Face and Gesture 2011*, pages 889–896, 2011.

[158] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai et Qiang Zhou : Look at Boundary : A Boundary-Aware Face Alignment Algorithm. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[159] X. Xiong et F. De la Torre : Supervised Descent Method and Its Applications to Face Alignment. *In 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013.

[160] Heng Yang, Xuhui Jia, Chen Change Loy et Peter Robinson : An Empirical Study of Recent Face Alignment Methods. *arXiv :1511.05049 [cs]*, 2015.

[161] P. Yang, Q. Liu et D. N. Metaxas : Exploring facial expressions with compositional features. *In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2638–2644, 2010.

[162] Peng Yang, Qingshan Liu et Dimitris N. Metaxas : Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30(2):132–139, 2009. ISSN 0167-8655.

[163] L. Yin, X. Chen, Y. Sun, T. Worm et M. Reale : A high-resolution 3D dynamic facial expression database. *In 2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6, 2008.

[164] Aliaa A. A. Youssif et Wesam A. A. Asker : Automatic Facial Expression Recognition System Based on Geometric and Appearance Features. *Computer and Information Science*, 4(2):115, 2011. ISSN 1913-8997.

[165] Zhiding Yu et Cha Zhang : Image Based Static Facial Expression Recognition with Multiple Deep Network Learning. *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 435–442, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3912-4.

[166] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao et I. Kotsia : Aff-Wild : Valence and Arousal In-the-Wild Challenge. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1980–1987, 2017.

[167] Stefanos Zafeiriou, Cha Zhang et Zhengyou Zhang : A survey on face detection in the wild : Past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015. ISSN 1077-3142.

[168] J. Zeng, W. Chu, F. De la Torre, J. F. Cohn et Z. Xiong : Confidence Preserving Machine for Facial Action Unit Detection. *IEEE Transactions on Image Processing*, 25 (10):4753–4767, 2016. ISSN 1057-7149.

[169] Jie Zhang, Shiguang Shan, Meina Kan et Xilin Chen : Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment. *In Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 1–16. Springer, Cham, 2014. ISBN 978-3-319-10604-5 978-3-319-10605-2.

[170] Ligang Zhang, Dian Tjondronegoro et Vinod Chandran : Random Gabor based templates for facial expression recognition in images with facial occlusion. *Neurocomputing*, 145:451–464, 2014. ISSN 0925-2312.

[171] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu et Jeffrey M. Girard : BP4D-Spontaneous : A high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. ISSN 0262-8856. Best of Automatic Face and Gesture Recognition 2013.

[172] Yongmian Zhang et Qiang Ji : Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005. ISSN 0162-8828.

[173] Z. Zhang, P. Luo, C. C. Loy et X. Tang : Learning Deep Representation for Face Alignment with Auxiliary Attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):918–930, 2016. ISSN 0162-8828.

[174] G. Zhao et M. Pietikainen : Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. ISSN 0162-8828.

[175] K. Zhao, W. S. Chu et H. Zhang : Deep Region and Multi-label Learning for Facial Action Unit Detection. *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3391–3399, 2016.

[176] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F. Cohn et Honggang Zhang : Joint Patch and Multi-label Learning for Facial Action Unit and Holistic Expression Recognition. *IEEE Transactions on Image Processing*, 25(8):3931–3946, 2016. ISSN 1057-7149, 1941-0042.

[177] Wenming Zheng, Hao Tang, Zhouchen Lin et Thomas S. Huang : Emotion Recognition from Arbitrary View Facial Images. *In Computer Vision – ECCV 2010*, Lecture Notes in Computer Science, pages 490–503. Springer, Berlin, Heidelberg, 2010. ISBN 978-3-642-15566-6 978-3-642-15567-3.

[178] Xiangyu Zhu, Z. Lei, Junjie Yan, D. Yi et S. Z. Li : High-fidelity Pose and Expression Normalization for face recognition in the wild. *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, 2015.

[179] Y. Zhu, L. C. De Silva et C. C. Ko : Using moment invariants and HMM in facial expression recognition. *Pattern Recognition Letters*, 23(1):83–91, 2002. ISSN 0167-8655.