

Ministère de l'enseignement Supérieur et de la recherche Scientifique

وزارة التعليم العالي والبحث العلمي

Badji Mokhtar Annaba University
Université Badji Mokhtar – Annaba
Faculté de Technologie



جامعة باجي مختار – عنابة

كلية التكنولوجيا

قسم الاعلام الالي

Département Informatique

Thèse

Présentée pour obtenir le diplôme de

Doctorat Troisième Cycle

Filière : Informatique

Spécialité : Systèmes Informatiques (SI)

Par :

CHEBAH Ouafa

Thème :

Identification des expressions faciales des visages des apprenants dans un milieu d'apprentissage en ligne

Thèse soutenue le 13 Avril 2023 devant le jury composé de :

N°	Nom et prénom	Grade	Etablissement	Qualité
01	Mohamed Benali Yamina	Prof.	Université Badji Mokhtar -Annaba	Président
02	Laskri Mohamed Tayeb	Prof.	Université Badji Mokhtar -Annaba	Rapporteur
03	Rezeg Khaled	Prof.	Université Mohamed Khider -Biskra	Examineur
04	Kermi Adel	MCA.	Ecole Nationale Supérieure d'Informatique-Alger	Examineur

*A mon défunt Père,
A ma mère, mes deux Sœurs et mon frère,
A mon Mari et mes deux Enfants.*

Remerciements

Je remercie tout d'abord le bon dieu de m'avoir donné le courage et la santé pour accomplir ce travail.

Je souhaite remercier en premier lieu mon directeur de thèse, Pr **Laskri Mohamed Tayeb**, Professeur à l'université Badji Mokhtar, qui m'a encadré tout au long de ces années et qui m'a fait partager ses brillantes intuitions. Qu'il soit aussi remercié pour sa gentillesse, sa disponibilité et pour les nombreux encouragements qu'il m'a prodigués.

Mes sincères remerciements et ma gratitude vont aussi à Pr **Mohamed Benali Yamina** pour avoir accepté de juger ce travail et d'en présider le jury de soutenance. Que vous soyez assuré de mon entière reconnaissance.

Je remercie Pr **Rezeg Khaked**, Dr **Kermi Adel** d'avoir accepté de juger ce travail. Soyez assurés, chers professeurs, de mon estime et de ma profonde gratitude.

Je remercie tous ceux qui m'ont aidé de proche et de loin, ma mère, mon mari et ma soeur pour leur soutien moral et financier.

Identification des expressions faciales des visages des apprenants dans un milieu d'apprentissage en ligne

Résumé

La reconnaissance d'émotions est l'un des aspects les plus importants de l'informatique affective dont l'un des objectifs est l'étude et le développement d'interactions comportementales et émotionnelles entre humains et machine. Des recherches récentes ont montré la nécessité de prendre en compte le facteur émotionnel dans les systèmes d'Interaction Homme Machine, et plus particulièrement dans les systèmes d'enseignement à distance 'e-learning'. En effet, la réussite du tout processus d'apprentissage dépend grandement de l'état émotionnel de l'apprenant, car les émotions jouent un rôle fondamental dans tout processus cognitif et se révèlent primordiales lors de l'apprentissage. Parmi les différentes méthodes utilisées pour la détection de la variation de l'état émotionnel de l'apprenant, les expressions faciales occupent une place de premier plan dans la communication non-verbale et constituent le moyen le plus direct pour l'expression des émotions.

Au cours de cette thèse, nous avons développé un nouveau système pour détecter l'état émotionnel de l'apprenant pendant sa navigation dans un environnement d'apprentissage en ligne à travers ces expressions faciales. Ce système s'appuie sur l'utilisation de trois approches différentes. La première approche est basée sur l'extraction des points caractéristiques géométriques du visage permettant de décrire les déformations ou les mouvements faciaux. Dans la deuxième approche, un nouveau descripteur d'apparence appelé 'Local Directional Double Ternary Coding Patterns (LDDTCP)' est proposé afin de trouver les informations les plus discriminantes qui décrivent les expressions faciales. Enfin, la troisième approche combine quatre modèles CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et CNN_{LDP}) basée sur le calcul de quatre descripteurs de texture LBP, LGC, GDP et LDP pour extraire les caractéristiques faciales. A partir de l'observation et de l'identification de l'état émotionnel de l'apprenant, le tuteur peut entreprendre des actions qui vont influencer la qualité de l'apprentissage.

Des études comparatives ont été menées entre les trois approches proposées et les travaux existants dans la littérature pour montrer leurs efficacités et vérifier leurs faisabilités. Les résultats expérimentaux sont très prometteurs pour l'approche basée sur le nouveau descripteur LDDTCP et l'approche basée sur le fusionnement des quatre modèles CNN car elles ont montré leurs supériorités par rapport aux méthodes existantes en termes de taux de

reconnaissance, néanmoins les résultats restent insuffisantes pour l'approche basée sur les descripteurs géométriques.

Mots clés : Reconnaissance des expressions faciales, émotions, apprenant, systèmes tutoriels intelligents, descripteurs géométriques, descripteurs d'apparences, apprentissage profond, LBP, LGC, GDP, LDP, CNN.

Identifying facial expressions of learners' faces in an online learning environment

Abstract

Emotion recognition is one of the most important aspects of affective computing, one of whose objectives is the study and development of behavioral and emotional interactions between humans and machines. Recent research has shown the need to take into account the emotional factor in Human Machine Interaction systems, and more particularly in distance education systems 'e-learning'. Indeed, the success of any learning process depends greatly on the emotional state of the learner, because emotions play a fundamental role in any cognitive process and are essential during learning. Among the different methods used for the detection of the variation of the emotional state of the learner, facial expressions occupy a prominent place in non-verbal communication and constitute the most direct means for the expression of emotions.

During this thesis, we developed a new system to detect the emotional state of the learner during his navigation in an online learning environment through these facial expressions. This system relies on the use of three different approaches. The first approach is based on the extraction of the geometric characteristic points of the face allowing to describe the facial movements. In the second approach, a new appearance descriptor called 'Local Directional Double Ternary Coding Patterns (LDDTCP)' is proposed in order to find the most discriminating information that describes facial expressions. Finally, the third approach combines four CNN models (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} and CNN_{LDP}) based on the calculation of four texture descriptors LBP, LGC, GDP and LDP to extract facial features. From the observation and identification of the emotional state of the learner, the tutor can take actions that will influence the quality of learning.

Comparative studies were conducted between the three proposed approaches and existing works in the literature to show their effectiveness and verify their feasibility. The experimental results are very promising for the approach based on the new descriptor LDDTCP and the approach based on the fusion of the four CNN models because they have shown their superiority compared to existing methods in terms of recognition rate, but remain insufficient for the approach based on geometric descriptors.

Keywords: Facial expressions recognition, emotions, learner, intelligent tutorial systems, geometric descriptors, appearance descriptors, deep learning, LBP, LGC, GDP, LDP, CNN.

تحديد تعبيرات وجوه المتعلمين في بيئة التعلم عبر الإنترنت

ملخص

يعتبر التعرف على المشاعر من أهم جوانب الحوسبة العاطفية ، ومن أهدافها دراسة وتطوير التفاعلات السلوكية والعاطفية بين الإنسان والآلات. أظهرت الأبحاث الحديثة الحاجة إلى مراعاة العامل العاطفي في أنظمة التفاعل بين الإنسان والآلة ، وبشكل أكثر تحديدًا في أنظمة التعليم عن بعد. في الواقع ، يعتمد نجاح أي عملية تعلم بشكل كبير على الحالة العاطفية للمتعلم ، لأن العواطف تلعب دورًا أساسيًا في أي عملية معرفية وهي ضرورية أثناء التعلم. من بين الأساليب المختلفة المستخدمة للكشف عن تباين الحالة العاطفية للمتعلم ، تحتل تعبيرات الوجه مكانًا بارزًا في التواصل غير اللفظي وتشكل أكثر الوسائل المباشرة للتعبير عن المشاعر.

من خلال هذه الرسالة ، طورنا نظامًا جديدًا لاكتشاف الحالة العاطفية للمتعلم أثناء تنقله في بيئة التعلم عبر الإنترنت من خلال تعابير الوجه. يعتمد هذا النظام على استخدام ثلاث طرق مختلفة. الأسلوب الأول يعتمد على استخراج النقاط المميزة الهندسية للوجه للسماح بوصف التشوهات أو حركات الوجه. في النهج الثاني ، تم اقتراح واصف مظهر جديد يسمى "أنماط الترميز الثنائية الاتجاهية المزدوجة (LDDTCP)" من أجل العثور على المعلومات الأكثر تمييزًا التي تصف تعابير الوجه. أخيرًا ، يجمع النهج الثالث بين أربعة نماذج CNN (CNN_{LBP} ، CNN_{LGC} ، CNN_{GDP} و CNN_{LDP}) بناءً على حساب أربعة واصفات نسيج LBP و LGC و GDP و LDP لاستخراج ملامح الوجه. من خلال ملاحظة الحالة العاطفية للمتعلم وتحديدتها ، يمكن للمدرس اتخاذ إجراءات من شأنها أن تؤثر على جودة التعلم.

أجريت دراسات مقارنة بين الأساليب الثلاثة المقترحة والأعمال الموجودة في الأدبيات لإظهار فعاليتها والتحقق من جدواها. النتائج التجريبية واعدة للغاية بالنسبة للنهج القائم على الواصف الجديد LDDTCP والنهج القائم على اندماج نماذج CNN الأربعة لأنها أظهرت تفوقها مقارنة بالطرق الحالية من حيث معدل التعرف ، لكنها تظل غير كافية للنهج القائم على الواصفات الهندسية.

الكلمات المفتاحية : التعرف على تعابير الوجه ، العواطف ، المتعلم ، أنظمة التعليم الذكية ، الواصفات الهندسية ، واصفات المظهر ، التعلم العميق ، LBP ، LGC ، GDP ، LDP ، CNN.

Table des matières

Introduction générale	23
I. Partie 1 : Etat de l’art	26
1. Les émotions dans les Systèmes Tutoriels Intelligents	27
1.1 Introduction	28
1.2 Notion sur les émotions	28
1.2.1 Définition	28
1.2.2 Type des émotions	30
1.2.2.1 Les émotions primaires	30
1.2.2.2 Les émotions secondaires	30
1.2.3 Représentation des émotions	30
1.2.3.1 Approche catégorielle (discrète)	31
1.2.3.2 Approche dimensionnelle (continue)	32
1.3 Les Systèmes Tutoriels Intelligents et les émotions	34
1.3.1 Notions sur les STI	34
1.3.1.1 Historique des STI	34
1.3.1.2 Définition des STI	35
1.3.1.3 Architecture générale des STI	35
1.3.2 Liens entre émotions et processus cognitifs	41
1.3.3 L’intelligence émotionnelle	42
1.3.4 L’informatique affective	43
1.3.5 Les Systèmes Tutoriels Emotionnellement Intelligents (STEI)	44
1.3.6 Modélisation des émotions dans les STI	44
1.3.7 Méthodes de reconnaissance des émotions de l’apprenant dans les STI	49
1.3.7.1 L’auto-évaluation (self report)	50
1.3.7.2 Les variables d’interaction	52
1.3.7.3 Les mesures comportementales	52
1.3.7.4 Les mesures physiologiques	55
1.3.8 Stratégies d’adaptation émotionnelles dans les STI	58
1.4 Conclusion	62
2. Système d’analyse des expressions faciales	64
2.1 Introduction	65

2.2	Qu'est ce qu'une expression faciale	65
2.3	Expression faciale versus Emotion	67
2.4	Formalismes de représentation des expressions faciales	68
2.4.1	Représentation par le système FACS	68
2.4.2	Représentation par la norme MPEG-4	70
2.5	Expression des émotions primaires	72
2.6	Système d'analyse des expressions faciales	73
2.6.1	Acquisition du visage	74
2.6.2	Détection du visage	75
2.6.2.1	Méthodes basées sur les connaissances acquises	75
2.6.2.2	Méthodes basées sur les caractéristiques invariantes	76
2.6.2.3	Méthodes basées sur la mise en correspondance	77
2.6.2.4	Méthodes basées sur l'apparence globale	77
2.6.3	Prétraitement	78
2.6.4	Extraction des caractéristiques faciales	78
2.6.4.1	Les méthodes basées sur l'apparence	79
2.6.4.1.1	L'Analyse par Composantes Principales (ACP)	79
2.6.4.1.2	L'Analyse Discriminante Linéaire (ADL)	80
2.6.4.1.3	Les filtre de Gabor	81
2.6.4.1.4	Les motifs binaires locaux (LBP)	82
2.6.4.1.5	Local Gradient Code (LGC)	84
2.6.4.1.6	Gradient Directional Pattern (GDP)	85
2.6.4.1.7	Local Ternary Pattern (LTP)	87
2.6.4.1.8	Local Directional Pattern (LDP)	87
2.6.4.2	Les méthodes basées sur les caractéristiques géométriques	89
2.6.4.2.1	Caractéristiques géométriques basées sur la forme des composantes faciales	89
2.6.4.2.2	Caractéristiques géométriques basées sur les points caractéristiques	91
2.6.4.2.3	Caractéristiques géométriques basées sur les modèles	92
2.6.4.3	Les méthodes hybrides	93
2.6.5	Classification des expressions faciales	95
2.6.5.1	Les réseaux de neurones artificiels (ANN)	95
2.6.5.2	Les machines à vecteurs de support (SVM)	98

2.6.5.3 Les K-plus proche voisin (KNN)	100
2.6.5.4 Les modèles de Markov cachés (HMM)	101
2.7 Les systèmes de fusion	102
2.7.1 Fusion de données	103
2.7.2 Fusion d'attributs (caractéristiques)	104
2.7.3 Fusion des décisions	105
2.7.3.1 Les systèmes multi-classifieurs (MCS)	106
2.7.3.2 Architectures de combinaison de classifieurs	107
2.7.3.2.1 Architecture en cascade	107
2.7.3.2.2 Architecture en parallèle	108
2.7.3.2.3 Architecture hybride	108
2.8. Conclusion.....	109
3. Reconnaissance des expressions faciales via l'apprentissage profond	110
3.1 Introduction	111
3.2 Relation entre Intelligence Artificielle, Machine Learning et Deep Learning	111
3.3 Qu'est-ce qu'un apprentissage profond (Deep learning)	114
3.4 Les différents modèles de l'apprentissage profond	114
3.4.1 Les réseaux de neurones convolutionnels (CNN)	115
3.4.1.1 Inspiration des CNNs	115
3.4.1.2 Architecture globale d'un CNN	115
3.4.1.3 Les différentes couches de CNN	116
3.4.2.4 Apprentissage d'un CNN	120
3.4.2.5 Application des CNN pour la reconnaissance des expressions faciales	122
3.4.2 Les réseaux de neurones récurrents (RNN)	123
3.4.2.1 Apprentissage d'un RNN	124
3.4.2.2 Mémoire Court et Long Terme (LSTM)	125
3.4.2.3 Unité de porte récurrente (GRU)	126
3.4.2.4 Application des RNN pour la reconnaissance des expressions faciales.....	127
3.4.3 Les modèles génératives profonds	128
3.4.3.1 Les machines de Boltzmann restreintes (RBM)	128
3.4.3.1.1 Apprentissage des RBM	130

3.4.3.1.2. Application des RBM pour l'analyse des expressions faciales	130
3.4.3.2 Les réseaux de croyance profonds (DBN)	130
3.4.3.2.1 Entraînement des DBN	131
3.4.3.2.2 Application des DBN pour la reconnaissance des expressions faciales	132
3.4.3.3 Les réseaux antagonistes génératifs (GAN)	132
3.4.3.3.1 Apprentissage des GAN	133
3.4.3.3.2 Applications des GAN pour la reconnaissance des expressions faciales	134
3.4.3.4 Les Auto-Encodeurs (AE : Auto-Encoder)	135
3.4.3.4.1 Apprentissage des auto-encodeurs	136
3.4.3.4.2 Application des auto-encodeurs pour la reconnaissance des expressions faciales	136
3.4 Conclusion	137
II. Partie 2 : Contributions	138
4. Approches proposées	139
4.1 Introduction	140
4.2 Architecture générale du système proposé	140
4.3 Système de Reconnaissance des Expressions Faciales (SREF)	141
4.3.1 Approche basée sur les caractéristiques géométriques pour la reconnaissance des expressions faciales	141
4.3.1.1 Détection de visage	142
4.3.1.2 Extraction des caractéristiques	143
4.3.1.3 Classification	147
4.3.2 Approche basée sur les caractéristiques d'apparences pour la reconnaissance des expressions faciales	148
4.3.2.1 Détection et extraction de visage	148
4.3.2.2 Extraction des caractéristiques faciales	149
4.3.2.3 Classification	155
4.3.3 Approche basée sur l'apprentissage profond pour la reconnaissance des expressions faciales	156
4.3.3.1 Détection et extraction de visage	157

4.3.3.2 Calcul des descripteurs de texture	157
4.3.3.3 Architecture et conception du CNN	158
4.3.3.4 Classification et fusion	160
4.4 Conclusion	161
5. Expérimentations, tests et résultats	162
5.1 Introduction	163
5.2 Base de données d'expressions faciales	163
5.3 Critères d'évaluation	165
5.4 Tests et résultats	165
5.4.1 Tests et résultats de l'approche basée sur les caractéristiques géométriques pour la reconnaissance des expressions faciales	166
5.4.2 Tests et résultats de l'approche basée sur les caractéristiques d'apparences pour la reconnaissance des expressions faciales	168
5.4.2.1 Expérimentation 1 : Test et résultats en utilisant le visage entier	168
5.4.2.2 Expérimentation 2 : Test et résultats en utilisant les composantes faciales	171
5.4.2.3. Comparaison des résultats du descripteur LDDTCP avec les autres travaux de l'état de l'art	174
5.4.3 Tests et résultats de l'approche basée sur l'apprentissage profond pour la reconnaissance des expressions faciales	176
5.4.3.1 Résultats d'apprentissage	176
5.4.3.2 Résultats de test	178
5.4.3.3 Résultats de fusionnement en parallèle (agrégation par somme / produit).....	181
5.4.3.4 Discussion	183
5.4.3.5 Comparaison des résultats de l'approche de fusionnement avec les autres travaux de l'état de l'art	184
5.5 Implémentation du système	184
5.5.1 L'application réalisée	185
5.5.1.1 Fenêtre Accueil	185
5.5.1.2 Fenêtre de l'approche basée sur les caractéristiques géométriques pour la reconnaissance des expressions faciales	185
5.5.1.3 Fenêtre de l'approche basée sur les caractéristiques d'apparences (descripteur LDDTCP) pour la reconnaissance des expressions faciales	186

5.5.1.4 Fenêtre de l'approche basée sur l'apprentissage profond pour la reconnaissance des expressions faciales	187
5.6 Conclusion	187
Conclusion générale	189
Références bibliographiques	193

Liste des figures

Figure 1.1 : Représentation des émotions mixte	32
Figure 1.2 : Modèle multidimensionnel de Plutchik	33
Figure 1.3 : Architecture de base d'un STI	36
Figure 1.4 : Diagramme comparatif des méthodes de modélisation de l'expert	38
Figure 1.5 : Modèle overlay simple	39
Figure 1.6 : Modèle overlay avec erreurs (buggy)	39
Figure 1.7 : Sous-réseau d'évaluation des buts	45
Figure 1.8 : Modèle 'Circumplex' de Russell	46
Figure 1.9 : Modèle à quatre quadrants reliant les émotions aux phases d'apprentissage ...	47
Figure 1.10 : Les régions stratégiques des émotions	48
Figure 1.11 : Canaux et mécanismes de détection de l'état émotionnel de l'apprenant	49
Figure 1.12 : L'outil d'auto-évaluation grille d'affect «Affect grid » de Russel	50
Figure 1.13 : L'outil d'auto-évaluation SAM (Self-Assessment Manikin scale)	51
Figure 1.14 : Exemple de chaise utilisée pour la détection émotionnelle	52
Figure 1.15 : Canaux de communication physiologiques et capteurs associés	55
Figure 1.16 : Exemple de capteurs de la conductivité de la peau (a) et (b) représentent respectivement un bracelet et un gant utilisés pour mesurer la réponse galvanique de la peau.....	56
Figure 1.17 : Exemples d'agents émotionnels utilisés dans les STI. De gauche à droite : l'agent COSMO, l'agent STEVE et l'agent EMILIE dans leurs environnement virtuel	59
Figure 1.18 : Interface du système de Prendinger et Ishizuka	60
Figure 1.19 : Interface de l'agent animé développé par [KAP 07]	61
Figure 1.20 : Système proposé par [KAP 07]. De gauche à droite : caméra vidéo, caméra yeux blues, souris à pression, capteur de la conductivité de la peau, siège à pression	61
Figure 1.21 : Exemple d'interface pour induire un état de joie	62
Figure 2.1 : Les muscles faciaux	66
Figure 2.2 : Photographies des expériences de stimulation électrique des muscles de Duchenne	66
Figure 2.3 : Générateur des expressions faciales et des émotions	68

Figure 2.4 : Modèle du visage par le standard MPEG-4 : (a) Représentation des points caractéristiques FPs. (b) Modèle des paramètres unitaires FAPUs	71
Figure 2.5 : Architecture générale d'un système de reconnaissance des expressions faciales	74
Figure 2.6 : Le modèle type exploité par la méthode de Yang et Huang	76
Figure 2.7 : Localisation de visage : (a) image originale, (b) segmentation de couleur, (c) groupement des composants connexes, (d) détermination des régions elliptiques.....	76
Figure 2.8 : L'algorithme de base pour la détection de visage basé sur les réseaux de neurone	78
Figure 2.9 : Exemple de calcul d'Eigenfaces: La première ligne présente des images de visage avec quelques expressions faciale et la deuxième visualise les eigenfaces correspondants	79
Figure 2.10 : Les 5 espaces de Fisher correspondant aux 5 axes du sous espace généré par ACP et ADL	80
Figure 2.11 : Exemple de convolution d'une image avec un ensemble de noyaux de Gabor (5fréquences et 8 orientations)	81
Figure 2.12 : Exemples de réponses d'un banc de filtres de Gabor 2D (avec trois valeurs d'orientations et longueurs d'ondes différentes), appliqué à deux images d'expressions faciales	82
Figure 2.13 : Exemple de calcul du code LBP original pour un pixel d'une image	83
Figure 2.14 : Extension de l'opérateur LBP basique	83
Figure 2.15 : Modèle de voisinage 3*3 de l'opérateur LGC	84
Figure 2.16 : Exemple de calcul de l'opérateur LGC	85
Figure 2.17 : Exemple de calcul du masque de Sobel	86
Figure 2.18 : Exemple de calcul de l'opérateur GDP : (a) Image en niveau de gris, (b) Angles de direction du gradient dans voisinage 3* 3, (c) code binaire, (d) code GDP, (e) Image GDP	86
Figure 2.19 : Exemple de calcul de l'opérateur LTP	87
Figure 2.20 : Les huit masques de Kirsch	88
Figure 2.21 : Exemple de calcul du code LDP avec k= 3	88
Figure 2.22 : Exemples des résultats d'extraction des traits permanents	89
Figure 2.23 : Exemples de la reconstruction finale des traits permanents proposés par Bottino	
Figure 2.24 : Modèles paramétriques pour la représentation des traits du visage : (a) Modèle paramétrique des yeux et des sourcils, (b) Modèle paramétrique de la bouche.	90

Figure 2.25 : Résultats de segmentation des yeux, sourcils et bouche	90
Figure 2.26 : Modèle géométrique des points caractéristiques proposés dans [PAN 04].....	91
Figure 2.27 : Approche proposée par [ABD 11] : (a) Modèle anthropométrique, (b) Points statiques, (c) Points dynamiques, (d) Distances utilisées pour le codage des expressions faciales	92
Figure 2.28 : Modèle des paramètres d'action APs	92
Figure 2.29 : Modèle de détection des déformations faciales HIGEM	93
Figure 2.30 : Le modèle à forme active proposé par [CHA 06]	94
Figure 2.31 : Exemple des résultats d'application du modèle MAA	95
Figure 2.32 : Neurone biologique	96
Figure 2.33 : Neurone artificiel	96
Figure 2.34 : Différents types de fonctions d'activation pour le neurone artificiel	97
Figure 2.35 : Architecture d'un perceptron multicouche	97
Figure 2.36 : Architecture de réseaux de neurones proposée par [ZHA 98]	98
Figure 2.37 : Schéma général d'un classificateur SVM	99
Figure 2.38 : Architecture multi-niveaux de HMMs pour la reconnaissance d'expressions faciales d'après [COH 03]	102
Figure 2.39 : Les différents niveaux de fusion	103
Figure 2.40 : Combinaison séquentielle de classifieurs	107
Figure 2.41 : Combinaison parallèle de classifieurs	108
Figure 2.42 : Combinaison hybride de classifieurs	108
Figure 3.1 : Relation entre Intelligence Artificiel, Machine Learning et Deep Learning ...	112
Figure 3.2 : La différence de performance entre le Deep Learning et la plupart des algorithmes de ML en fonction de la quantité de données	113
Figure 3.3 : Comparaison entre la machine learning et le deep learning	113
Figure 3.4 : Les différents modèles de deep learning	114
Figure 3.5 : Architecture générale d'un CNN	116
Figure 3.6 : Exemple de calcul d'une convolution	117
Figure 3.7 : Exemple de padding et stride dans la couche de convolution	118
Figure 3.8 : Fonction ReLU	118
Figure 3.9 : Exemples de calcul des différents types de pooling avec un filtre 2x2 et un pas de 2.	119
Figure 3.10 : Illustration d'un RNN : (a) version récurrente et (b) version dépliée	123
Figure 3.11 : Unité de base RNN	124

Figure 3.12 : Illustration d'un LSTM	125
Figure 3.13 : Fonctionnement d'une cellule LSTM	126
Figure 3.14 : Unité de base GRU	127
Figure 3.15 : Principe de fonctionnement : (a) Machine de Boltzmann (BM), (b) Machine de Boltzmann Restreintes (RBM)	129
Figure 3.16 : Architecture générale des DBN	131
Figure 3.17 : Principe de fonctionnement d'un réseau GAN	133
Figure 3.18 : Représentation schématique du principe d'un auto-encodeur : la sortie x' est la reconstruction de l'entrée x à partir de la couche cachée z	135
Figure 3.19 : Exemple d'un auto-encodeur	136
Figure 4.1 : Schéma général du système proposé	140
Figure 4.2 : Architecture globale du SREF basé sur les caractéristiques géométriques	141
Figure 4.3 : Détection de visage	143
Figure 4.4 : Localisation de l'axe des yeux et l'axe médian	144
Figure 4.5 : Localisation de l'axe des narines et de la bouche	144
Figure 4.6 : Le modèle géométrique	145
Figure 4.7 : Localisation des composantes faciales	145
Figure 4.8 : Extraction des points caractéristiques	146
Figure 4.9 : Sélection des points caractéristiques faciaux	147
Figure 4.10 : Architecture globale du SREF basé sur les caractéristiques d'apparences ...	148
Figure 4.11 : Procédure d'extraction de visage: (a) Image d'entrée, (b) Maximum de la projection horizontale, (c) Localisation de l'axe des yeux, (d) Maximum de la projection verticale, (e) Localisation de l'axe médian, (f) Détection de visage, (g) Extraction de visage.....	149
Figure 4.12 : Les huit masques de Frei-Chen divisé en deux groupes : (a) Le premier groupe, (b) Le deuxième groupe	150
Figure 4.13 : Résultats de convolution des images avec le masque de Frei-chen: (a) Convolution des images avec le premier groupe du masque (FC_0 , FC_1 , FC_2 , FC_3), (b) Convolution des images avec le deuxième groupe du masque (FC_4 , FC_5 , FC_6 , FC_7)	151
Figure 4.14 : Principe du codage de LDDTCP-1 et LDDTCP-2	152
Figure 4.15 : Principe de calcul de l'opérateur LDDTCP	153
Figure 4.16 : Extraction des caractéristiques de l'image entière à partir du code LDDTCP	154

Figure 4.17 : Extraction des caractéristiques de l'image divisée en plusieurs blocs à partir du code LDDTCP	155
Figure 4.18 : Architecture globale du SREF via l'apprentissage profond	157
Figure 4.19 : Phase d'apprentissage	159
Figure 4.20 : Phase de test	160
Figure 5.1 : Base de données d'expressions faciales utilisée : (a) CK+, (b) JAFFE, (c) FABO, (d) DaFEx	164
Figure 5.2 : Impact des différentes valeurs de K sur la classification	166
Figure 5.3 : Division de l'image de visage entier en régions de taille 3x3	168
Figure 5.4 : Taux de reconnaissance versus variation de la taille de région pour la base CK+	169
Figure 5.5 : Taux de reconnaissance versus variation de la taille de région pour la base JAFFE	169
Figure 5.6 : Procédure d'extraction des composantes faciales : (a) Détection des différents axes de visage, (b) Application du modèle géométriques, (c) Localisation des composantes faciales, (d) Extraction des composantes faciales, (e) Segmentation de chacune des composantes faciales en blocs de taille 3 x 4	171
Figure 5.7 : Variation du nombre d'époque versus le taux de reconnaissance des quatre modèles CNN pour la base CK+	177
Figure 5.8 : Variation du nombre d'époque versus le taux d'erreur des quatre modèles CNN pour la base CK+	178
Figure 5.9 : Variation du nombre d'époque versus le taux de reconnaissance pour la base CK+ en utilisant les fonctions d'activation ReLU et Identity	183
Figure 5.10 : Variation du nombre d'époque versus le taux de reconnaissance pour la base JAFFE en utilisant les fonctions d'activation ReLU et Identity	183
Figure 5.11 : Fenêtre accueil	185
Figure 5.12 : Fenêtre approche basée descripteurs géométriques	186
Figure 5.13 : Fenêtre d'un exemple de calcul du descripteur $P_{LDDTCP-1}$	186
Figure 5.14 : Fenêtre de l'architecture des quatre modèles CNN	187

Liste des tableaux

Tableau 1.1 : Principales définitions des émotions	29
Tableau 1.2 : Liste des émotions basiques selon différents auteurs	31
Tableau 2.1 : Les AUs fréquentes décodées par le manuel FACS	69
Tableau 2.2 : Combinaison des AUs correspondant a chaque émotion	70
Tableau 2.3 : Description des six expressions d'après [LAL 05] et [ESS 97]	73
Tableau 3.1 : Résumé des principaux travaux de la reconnaissance des expressions faciales à partir des modèles CNN	122
Tableau 3.2 : Synthèse des différents travaux de la reconnaissance des expressions faciales basé sur les modèles RNN	128
Tableau 3.3 : Les différents travaux utilisant les modèles RBM pour la reconnaissance des expressions faciales	130
Tableau 3.4 : Les différents travaux adoptant les modèles DBN pour la reconnaissance des expressions faciales	132
Tableau 3.5 : Synthèse des travaux utilisant les modèles GAN pour la reconnaissance des expressions faciales	134
Tableau 3.6 : Les différents travaux adoptant les modèles des auto-encodeurs pour l'analyse des expressions faciales	136
Tableau 4.1 : Entrées de vecteur de caractéristiques	147
Tableau 4.2 : Configuration de CNN	159
Tableau 5.1 : Matrice de confusion pour 6-classes d'émotion de la base FABO (K=3) ...	167
Tableau 5.2 : Matrice de confusion pour 6-classes d'émotion de la base DaFEx (K=3) ...	167
Tableau 5.3 : Résultats obtenus pour le classificateur KNN (K = 3)	167
Tableau 5.4 : Matrice de confusion (%) de la base CK+ en utilisant 7×6 régions et 6 classes d'expression	170
Tableau 5.5 : Matrice de confusion (%) de la base JAFFE en utilisant 7×6 régions et 6 classes d'expression	170
Tableau 5.6 : Matrice de confusion (%) de la base CK+ en utilisant 7×6 régions et 7 classes d'expression	170
Tableau 5.7 : Matrice de confusion (%) de la base JAFFE utilisant 7×6 régions et 7 classes d'expression	171

Tableau 5.8 : Taux de reconnaissance (%) pour les bases de données CK+ et JAFFE en utilisant quatre composantes faciales et 6 - 7 classes d'expressions	172
Tableau 5.9 : Matrice de confusion (%) de la base CK+ en utilisant quatre composantes faciales avec 3 x 4 blocs et 6 classes d'expressions	172
Tableau 5.10 : Matrice de confusion (%) de la base JAFFE en utilisant quatre composantes faciales avec 3 x 4 blocs et 6 classes d'expressions	173
Tableau 5.11 : Matrice de confusion (%) de la base CK+ en utilisant quatre composantes faciales avec 3 x 4 blocs et 7 classes d'expressions	173
Tableau 5.12 : Matrice de confusion (%) de la base JAFFE en utilisant quatre composantes faciales avec 3 x 4 blocs et 7 classes d'expressions	173
Tableau 5.13 : Taux de classification des différents descripteurs de texture en utilisant la base CK+ et 6 classes d'expressions	174
Tableau 5.14 : Taux de classification des différents descripteurs de texture en utilisant la base CK+ et 7 classes d'expressions	175
Tableau 5.15 : Taux de classification des différents descripteurs de texture en utilisant la base JAFFE et 6 classes d'expressions	175
Tableau 5.16 : Taux de classification des différents descripteurs de texture en utilisant la base JAFFE et 7 classes d'expressions	175
Tableau 5.17 : Résultats de test pour CNN _{LBP} et la base de données CK+	179
Tableau 5.18 : Résultats de test pour CNN _{LGC} et la base de données CK+	179
Tableau 5.19 : Résultats de test pour CNN _{GDP} et la base de données CK+	179
Tableau 5.20 : Résultats de test pour CNN _{LDP} et la base de données CK+	179
Tableau 5.21 : Résultats de test pour CNN _{LBP} et la base de données JAFFE	180
Tableau 5.22 : Résultats de test pour CNN _{LGC} et la base de données JAFFE	180
Tableau 5.23 : Résultats de test pour CNN _{GDP} et la base de données JAFFE	180
Tableau 5.24 : Résultats de test pour CNN _{LDP} et la base de données JAFFE	180
Tableau 5.25 : Résultats de test des quatre modèles CNN pour les deux bases de données CK+ et JAFFE	180
Tableau 5.26 : Résultats de fusionnement en utilisant la règle somme et la base CK+	181
Tableau 5.27 : Résultats de fusionnement en utilisant la règle produit et la base CK+	181
Tableau 5.28 : Résultats de fusionnement en utilisant la règle somme et la base JAFFE...	182
Tableau 5.29 : Résultats de fusionnement en utilisant la règle produit et la base JAFFE...	182
Tableau 5.30 : Résultats de fusionnement en utilisant la règle somme et produit pour les deux bases de données CK+ et JAFFE	182

Tableau 5.31 : Etude comparative entre l'approche proposée et les autres systèmes de reconnaissance des expressions faciales	184
---	-----

Introduction générale

Les émotions sont indispensables dans notre vie quotidienne. Elles interviennent dans nos comportements, nos perceptions et nos relations avec les autres. Elles permettent ainsi, d'améliorer la communication entre les individus, d'assurer une meilleure compréhension du message véhiculé et de s'adapter à une situation donnée.

Depuis la fin des années 90, la modélisation des émotions constitue une thématique de plus en plus reconnue dans le domaine de l'Intelligence Artificielle (IA). A cet effet, plusieurs travaux de recherche ont été entrepris sur la nécessité de prendre en compte le facteur émotionnel dans les systèmes d'Interactions Homme-Machine (IHM) et plus particulièrement dans les Systèmes Tutoriels Intelligents (STI). Les récents avancements en neurosciences, psychologie et sciences cognitives, ont démontré que les émotions sont inévitablement sollicitées dans tous les processus cognitifs, et qu'elles ont un impact direct sur l'attention, la mémorisation, le raisonnement, la résolution des problèmes et la prise de décision [ISE 00]. En effet, l'état émotionnel de l'apprenant affecte ses performances. Les émotions positives améliorent la créativité et la flexibilité dans la résolution des problèmes [ISE 00]. Par contre, les émotions négatives peuvent bloquer le fonctionnement des processus cognitifs, rendre l'attention restreinte et gêner en conséquence le processus d'apprentissage.

L'intégration de la dimension émotionnelle dans les STI a donné lieu à une nouvelle génération des systèmes d'apprentissage connus sous le nom de *Systèmes Tutoriels Émotionnellement Intelligents (STEI)* [OCH 04b]. Ces systèmes sont capables de simuler le comportement d'un tuteur humain dans sa capacité à détecter, à analyser et à prendre en compte intuitivement l'état émotionnel de l'apprenant lors de son enseignement.

Un des principaux objectifs d'un STEI est de pouvoir reconnaître et évaluer, à tout instant, les émotions courantes de l'apprenant. De ce fait, les chercheurs ont recours à un large éventail de moyens de détection allant des simples *questionnaires d'auto-évaluation* qui demandent directement à l'apprenant de décrire son état émotionnel ; à la détection d'émotions à partir d'*indices physiologiques* (la fréquence cardiaque, le volume de la pression sanguine, la fréquence de la respiration, etc.) ; en passant par la reconnaissance d'émotions en utilisant des *indices comportementaux* (les expressions faciales, la tonalité de la voix, la gestuelle ou la posture corporelle, etc.).

Parmi les différentes méthodes permettant de détecter la variation des émotions de l'apprenant, les expressions faciales occupent une place irremplaçable dans la communication non-verbale et constituent le moyen le plus direct pour l'expression des émotions. Une

expression faciale est une manifestation visible de l'état d'esprit, de l'activité cognitive, de l'intention, de la personnalité et de la psychopathologie d'une personne. Des travaux de recherche en psychologie ont démontré que les expressions faciales jouent un rôle prépondérant dans la coordination de la conversation humaine, et ont un impact plus important sur l'auditeur que le contenu textuel du message exprimé. Mehrabian [MEH 96] a indiqué le fait que 55 % du message émotionnel est communiqué par les expressions faciales alors que la partie verbale constitue seulement 7 % des effets du message et la partie vocale contribue à 38%.

L'analyse automatique des expressions faciales reste actuellement un challenge. La principale difficulté est liée au fait que les expressions faciales sont le résultat de la déformation du visage par les muscles. Ces déformations sont détectables par le changement de l'aspect visuel qui se résume à des changements de la forme des composantes faciales et des mouvements de la peau pour former des rides et des sillons. Cependant, l'aspect visuel est très lié aux sujets. Les différences inter-sujets et intra-sujets présentent ainsi des difficultés principales de la reconnaissance des expressions faciales. Plusieurs contraintes, liées aux environnements spécifiques à chaque application, existent également, parmi lesquelles nous citons le changement d'illumination, la variation de pose, le changement de résolution et les occultations. L'enjeu majeur pour caractériser une expression faciale consiste alors à identifier de façon unique une expression à partir des informations visibles, qui mélangent l'information utile (expression) à d'autres informations (identité, éclairage, ...).

Objectifs et contributions de la thèse

L'objectif ultime de cette thèse est de concevoir, implémenter et évaluer un nouveau système pour détecter l'état émotionnel de l'apprenant durant sa navigation dans un environnement d'apprentissage en ligne à travers ces expressions faciales. La réalisation du système de reconnaissance des expressions faciales est basée sur la proposition des trois contributions suivantes :

Dans la première contribution, une nouvelle approche basée sur l'extraction des points caractéristiques géométriques du visage est proposée. Cette approche permet de décrire les déformations ou les mouvements faciaux. Ces déformations sont codées en utilisant les distances entre les points. La classification des expressions faciales se fait par la méthode du plus proche voisin (KNN).

Dans la deuxième contribution, nous proposons un nouveau descripteur d'apparence nommé 'Local Directional Double Ternary Coding Patterns (LDDTCP)'. Ce descripteur intègre les concepts des descripteurs Local Directional Pattern (LDP) et Local Ternary Pattern (LTP) pour

coder à la fois les informations de direction et de contraste, ce qui permet d'augmenter la robustesse aux changements d'illumination et de bruit aléatoire.

Enfin, dans la troisième contribution, nous proposons une nouvelle approche qui combine quatre modèles CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et CNN_{LDP}) basée sur le calcul de quatre descripteurs de texture à savoir : Local Binary Pattern (LBP), Local Gradient Code (LGC), Gradient Directional Pattern (GDP) et Local Directional Pattern (LDP). L'approche proposée est capable d'apprendre et de classer avec précision les images des expressions faciales.

Plan de la thèse

Cette thèse est organisée en cinq chapitres. Les trois premiers chapitres sont plus théoriques et consacrés à une revue de la littérature liée à nos recherches, tandis que les chapitres quatre et cinq présentent nos contributions. Chaque chapitre est décrit comme suit :

Dans le *premier chapitre* nous présentons en premier lieu les notions de bases sur les émotions, puis nous décrivons l'impact des émotions dans les STI en présentant le concept de l'intelligence émotionnelle, l'informatique affective et les Systèmes Tutoriels Émotionnellement Intelligents (STEI). Nous exposons également les principales approches utilisées par les STI pour modéliser, reconnaître et réguler l'état émotionnel de l'apprenant.

Le *deuxième chapitre* fournit une étude détaillée sur les différentes étapes de traitement d'un système d'analyse des expressions faciales et décrit les techniques les plus populaires utilisées pour chaque étape. Il traite ensuite l'approche multimodale et présente les différentes stratégies qui ont été utilisés dans la littérature.

Le *troisième chapitre* est consacré à la reconnaissance des expressions faciales via l'apprentissage profond. Nous présentons les différents modèles de l'apprentissage profond tout en donnant une taxonomie complète pour chaque modèle présenté.

Le *quatrième chapitre* est dédié à la conception et l'implémentation de notre système de détection de l'état émotionnel de l'apprenant à travers les expressions faciales.

Finalement, nous exposons dans le *cinquième chapitre* les résultats obtenus et nous dressons un état comparatif qui permet d'analyser les performances de notre système par rapport aux travaux existants dans la littérature en termes de taux de reconnaissance.

Partie I : Etat de l'art

CHAPITRE 1

Les émotions dans les Systèmes Tutoriels Intelligents

Sommaire

1.1 Introduction	28
1.2 Notion sur les émotions	28
1.2.1 Définition	28
1.2.2 Type des émotions	30
1.2.3 Représentation des émotions	30
1.3 Les Systèmes Tutoriels Intelligents et les émotions	34
1.3.1 Notions sur les STI	34
1.3.2 Liens entre émotions et processus cognitifs	41
1.3.3 L'intelligence émotionnelle	42
1.3.4 L'informatique affective	43
1.3.5 Les Systèmes Tutoriels Emotionnellement Intelligents (STEI)	44
1.3.6 Modélisation des émotions dans les STI	44
1.3.7 Méthodes de reconnaissance des émotions dans les STI	49
1.3.8 Stratégies d'adaptation émotionnelles dans les STI	58
1.4 Conclusion	62

1.1 Introduction

Les émotions sont les interfaces de l'organisme avec le monde extérieur. Elles influencent nos réactions et nos perceptions par rapport à l'environnement. Des récents avancements ont montré que les émotions et les processus cognitifs sont intimement reliés, et que les émotions jouent un rôle pivot dans les systèmes d'Interactions Homme-Machine (IHM), et plus particulièrement dans les Systèmes Tutoriels Intelligents (STI). En effet, la réussite du tout processus d'apprentissage dépend grandement de l'état émotionnel de l'apprenant, car les émotions agissent sur divers tâches d'apprentissage, telle que l'attention, la mémorisation à long terme, la prise de décision, etc.

Ce chapitre fournit un état de l'art sur les émotions ainsi que leurs intégrations dans les STI. Cet état de l'art se divise alors en deux parties principales. Dans la première partie nous allons présenter quelques notions concernant les émotions telles que leurs définitions et leurs différents types. Puis, nous décrirons les principales approches de représentation des émotions prédominantes dans la littérature.

La deuxième partie passe en revue l'impact des émotions sur les STI, nous commençons par présenter l'historique de ces systèmes, leurs définitions et leurs évolutions chronologiques. Puis, nous introduirons l'architecture générale de ces systèmes et nous décrivons chacune de ses composantes. Ensuite, nous aborderons les principaux axes de recherches entrepris pour faire évoluer ces systèmes. Ceci va nous amener vers la description de la relation entre les émotions et la cognition, la présentation du concept de l'intelligence émotionnelle, l'informatique affective et les Systèmes Tutoriels Émotionnellement Intelligents (STEI). Nous nous attarderons ensuite, sur les approches de modélisation et de reconnaissance des émotions des apprenants dans les STI. Finalement, nous terminerons ce chapitre par l'étude des différentes stratégies émotionnelles utilisées dans les STI pour réguler l'état émotionnel de l'apprenant.

1.2 Notion sur les émotions

1.2.1 Définition

L'émotion est un terme familier et couramment utilisé dans la vie quotidienne. Dans la littérature, on ne trouve pas de consensus sur la définition du concept d'émotion. Un grand nombre de travaux a été réalisé dans le but de définir et de classifier les émotions humaines. Ainsi, les psychologues, physiologistes et philosophes ont étudié les émotions depuis de nombreuses années. Chacun d'entre eux en a proposé une définition différente. Le tableau 1.1

présente une synthèse de quelques travaux philosophiques et psychologiques qui ont proposé une définition des émotions, sur lesquelles se basent les travaux de recherche actuels.

Auteur	Définition
[JAM 84]	Les émotions sont des sentiments causés par le changement de conditions physiologiques concernant les fonctions autonomes et motrices. "Nous nous sentons tristes parce que nous pleurons, en colère parce que nous frappons quelqu'un et effrayés parce que nous tremblons".
[GRE 88]	L'émotion est un processus mental conscient entraînant un composant important du corps ; elle a aussi une influence importante sur la pensée et l'action de l'individu, notamment pour former des stratégies dans l'interaction sociale.
[STU 92]	L'émotion est un terme large qui normalement a plusieurs connotations cognitives qui affecte l'humeur et qui inclut l'expérience tout en ayant conscience de la perception de la sensation interne.
[MAN 92]	L'émotion est un événement conscient créé par la combinaison d'évaluations cognitives (bon, mauvais, plaisant, déplaisant, nocif, désirable, etc.) et de l'activation physiologique.
[DAM 94]	C'est la combinaison d'un processus mental évaluatif simple ou complexe, avec des réponses dispositionnelles à ce processus pour la plupart vers le corps, ce qui produit comme résultat un état corporel émotionnel, mais aussi vers le cerveau (noyau neurotransmetteur du tronc cérébral), ce qui produit comme résultat des changements mentaux supplémentaires.
[SCH 05]	En général, l'émotion peut être vue comme un processus impliquant cinq composants différents, comprenant le sentiment subjectif, la cognition, l'expression physique, la tendance à l'action ou le désir, et les processus neurologiques.

Tableau 1.1 Principales définitions des émotions.

Comme on peut le voir dans ce tableau, la notion d'émotion n'est pas unanime et apparaît comme polysémique. Il est donc difficile de donner une définition claire et univoque de l'émotion.

Selon les définitions précédentes, nous pouvons retenir deux approches pour définir le concept d'émotion : les *approches cognitives* et les *approches physiques*.

Dans les **approches cognitives** (appelé aussi théories de l'évaluation cognitives), une émotion est considérée comme le résultat d'une évaluation cognitive d'un événement. La perception et l'évaluation de cet événement détermineront le type et l'intensité de l'émotion ressentie [SCH 00], comme l'indique [GRE 88], [STU 92].

Dans les **approches physiques**, c'est plutôt l'aspect corporel de l'émotion qui prédomine. Une émotion est définie par la réaction physiologique qui se manifeste avec l'occurrence de cette émotion. [JAM 84] était le précurseur de cette approche.

Aujourd'hui, aucune des deux approches n'est favorisée : on considère que le corps et le cerveau interagissent dans la génération et dans l'expression des émotions.

1.2.2 Type des émotions

Les émotions sont des séquences courantes et vives qui influencent notre vie quotidienne. Elles se subdivisent en deux catégories : les émotions primaires et les émotions secondaires

1.2.2.1 Les émotions primaires

Les émotions primaires, simples ou basiques sont des réponses naturelles à des stimuli donnés ou bien elles sont déclenchées par des événements spécifiques. Elles se manifestent de façon identique indépendamment de la culture des individus [EKM 93], suivant des expressions faciales ou gestuelles particulières. Elles sont considérées comme une matière première, à partir de laquelle on peut construire toutes les autres émotions. Leur décodage hors contexte est automatique et permet des réactions rapides d'intensité plus ou moins forte [DUV 11].

1.2.2.2 Les émotions secondaires

Les émotions secondaires, complexes ou sociales résultent d'un processus d'évaluation mentale lié à l'expérience et le souvenir. Elles sont considérées comme issues de la combinaison d'émotions primaires c-à-d ces émotions sont apprises et constituées à partir des émotions primaires. Elles regroupent les notions de sympathie, de fierté, de jalousie, de honte, d'orgueil, d'admiration, de gratitude, d'indignation, d'embarras, de culpabilité, de timidité, d'humiliation, de mépris, etc [NUG 09]. L'éducation et la culture sont fortement impliquées dans l'acquisition de ces émotions complexes et par conséquent, elles ne s'avèrent pas universelles [HER 08]. Leur décodage nécessite des procédures de raisonnement qui ne sont applicables qu'en contexte [DUV 11]. Elles ne passent pas si vite et elles ont tendance à rester longtemps. Ce sont des émotions lentes.

1.2.3 Représentation des émotions

La reconnaissance automatique des émotions soulève de nombreuses problématiques. D'abord, au niveau de leurs représentations, il est nécessaire de trouver un formalisme cohérent avec les résultats psychologiques existants, tout en permettant une manipulation aisée. Ensuite, pour un événement donné, il faut pouvoir déterminer l'état émotionnel qui lui est associé. Dans ce contexte, Scherer [SCH 00] propose une description des différents

modèles de représentation de l'émotion. Nous décrivons ici les deux types de représentations les plus utilisés en reconnaissance d'émotion : Les *modèles catégoriels discrets* représentés par une étiquette verbale (par exemple "Peur") et les *modèles dimensionnels continus* représentées par une position, ou un ensemble de positions, dans un espace défini par des dimensions continues.

1.2.3.1 Approche catégorielle (discrète)

Les approches discrètes considèrent les émotions comme des caractéristiques épisodiques et universelles. Le caractère universel des émotions entraîne la définition d'un nombre fini et restreint d'émotions basiques (la peur, la colère, la joie, la tristesse, la surprise, etc.), qui peuvent être observées chez tous les individus, quelque soit leur ethnie ou leur culture.

Le Tableau 1.2 présente une liste de quelques approches discrètes existantes dans la littérature. Par exemple, Izard propose dix types d'émotions basiques à savoir : Colère, mépris, dégoût, angoisse, peur, culpabilité, intérêt, joie, honte, surprise. [TOM 84] propose un ensemble différent de huit émotions : la peur, la colère, l'angoisse, la joie, le dégoût, la surprise, l'intérêt, le mépris et la honte.

Auteur	Liste des émotions
[EKM 82]	Joie, colère, surprise, peur, dégoût, tristesse
[FRI 86]	Peur, plaisir, confiance, intérêt, dégoût, indifférence, colère, choc, surprise
[PLU 80]	Peur, colère, tristesse, joie, dégoût, surprise, acceptation, anticipation
[IZA 71]	Colère, mépris, dégoût, angoisse, peur, culpabilité, intérêt, joie, honte, surprise
[TOM 84]	Peur, colère, angoisse, joie, dégoût, surprise, intérêt, mépris, honte
[JAM 84]	Peur, chagrin, amour, colère

Tableau 1.2 Liste des émotions basiques selon différents auteurs

L'approche catégorielle présente un avantage certain dans le domaine de l'informatique émotionnelle [SCH 00]. En effet, une fois que les émotions à traiter sont clairement identifiées, elles deviennent simples à manipuler et elles se prêtent particulièrement bien à la reconnaissance automatique.

Malgré ces avantages, les modèles discrets comportent des défauts. En effet, la labellisation des émotions sous-entend qu'à chaque label correspond un état émotionnel défini (les étiquettes sont discrètes et ne peuvent représenter pleinement certains aspects émotionnels). Or, une émotion peut être vue comme un processus émotionnel dynamique plutôt que comme

un état émotionnel statique. C'est pourquoi plusieurs travaux se sont concentrés sur la recherche d'espaces continus pour mieux représenter les émotions.

1.2.3.2 Approche dimensionnelle (continue)

Contrairement aux approches discrètes, les approches continues considèrent qu'il existe un nombre infini d'émotions, représentées par une position, ou un ensemble de positions dans un espace de dimensions continues. Les dimensions obtenues les plus utilisées sont bidimensionnelles (plan) ou tridimensionnelles (volume).

Le modèle bidimensionnel le plus connu est l'espace 'Valence / Activation' proposé par Russell [RUS 80]. D'une part, la valence permet de distinguer les émotions positives (ou agréables) des émotions négatives (ou désagréables). D'autre part, l'activation correspond au degré de l'expression corporelle ou gestuelle qui se traduit par des réactions physiologiques (transpiration, accélération du rythme cardiaque, etc.).

Plutchik en 1980 place les émotions primaires sur les différents secteurs d'un cercle. Les émotions secondaires résultent de la combinaison de deux émotions primaires représentées par des secteurs adjacents sur le cercle [PLU 80]. Par exemple, l'amour résulte de la joie et de l'acceptation (voir figure 1.1).

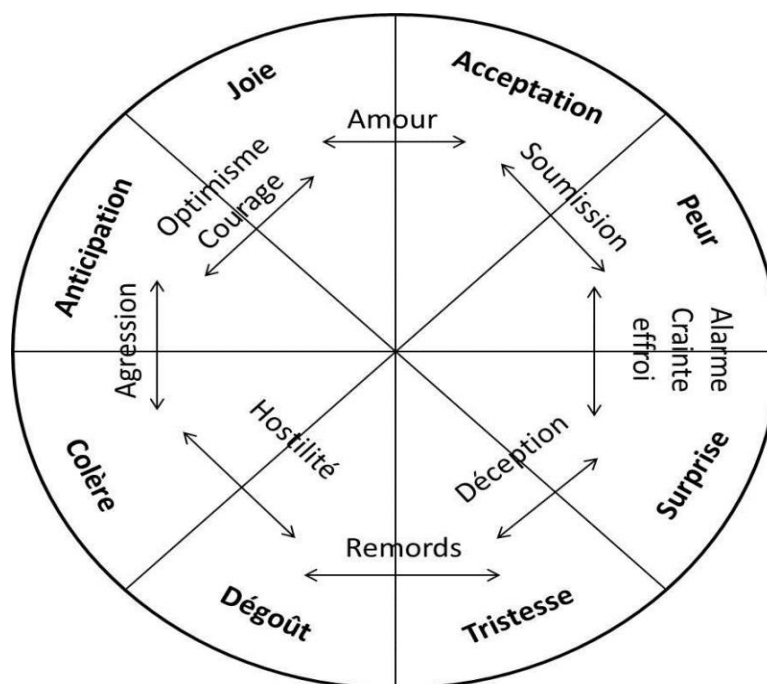


Figure 1.1 Représentation des émotions mixte [PLU 80]

Le modèle bidimensionnel (Valence / Activation) présente plusieurs avantages. Premièrement, il permet de représenter les émotions sans l'utilisation d'étiquettes, mais à l'aide d'un système de coordonnées incluant une signification émotionnelle. Deuxièmement,

l'espace crée à partir de l'analyse des expressions émotionnelles permet d'associer certaines zones à des étiquettes émotionnelles.

Malgré ces avantages, le model bidimensionnel a été critiqué sur les ambiguïtés qu'il induit dans les positions de certaines émotions sur les axes. En particulier, la colère et la peur sont deux émotions à valence fortement négative et à forte activation : elles se retrouvent donc dans la même zone de l'espace.

Le modèle tridimensionnel a été proposé pour remédier à ce problème. Une troisième dimension est ajoutée, à savoir la *dimension de contrôle* (appelée aussi dominance), pour représenter la capacité de contrôle de la situation ayant provoqué l'émotion. Ainsi, la colère est une émotion à forte dominance, alors que la peur est caractérisée par une forte soumission.

La figure 1.2 montre le modèle multidimensionnel de Plutchik où les 8 émotions basiques évoluent par pair et chaque émotion peut être représentée à des degrés ou des niveaux d'intensités différentes (par exemple : appréhension, peur et terreur).

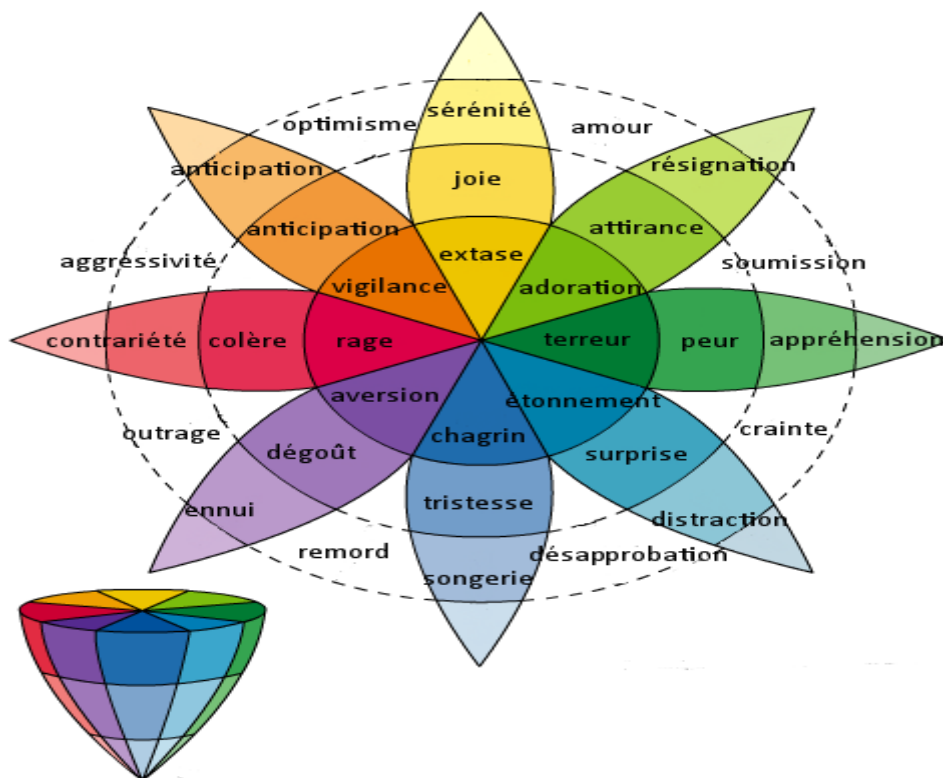


Figure 1.2 Modèle multidimensionnel de Plutchik [PLU 80]

Bien qu'elles offrent une plus grande liberté de représentation des émotions nuancées et permet d'associer des transitions entre différents états émotionnels, les approches continues présentent cependant un inconvénient majeur. En effet, contrairement aux approches discrètes, les approches continues ne permettent pas de donner un sens clair aux

émotions, et la distinction entre certaines émotions peut devenir ambiguë. De ce fait, les deux approches, catégorielle et dimensionnelle sont souvent combinées pour l'étude des émotions.

1.3 Les Systèmes Tutoriels Intelligents et les émotions

Depuis une vingtaine d'années, la modélisation automatique des émotions constitue une thématique de plus en plus centrale dans le domaine de l'Intelligence Artificielle (IA). A cet effet, plusieurs travaux de recherche ont été entrepris sur la nécessité de prendre en compte le facteur émotionnel dans les systèmes d'Interactions Homme-Machine (IHM) et plus particulièrement dans les STI [GRA 04]. Par ailleurs, l'intégration de la composante émotionnelle dans les STI a suscité plusieurs interrogations restent soulevés : Quel est le rôle des émotions dans les tâches d'apprentissage ? Pourquoi intégrer la dimension émotionnelle dans les STI ? Quelles émotions faut-il modéliser lors de l'interaction de l'apprenant avec les STI ? Comment détecter ces émotions ? Et de quelle manière faut-il intervenir pour gérer ou réguler les émotions de l'apprenant ?

1.3.1 Notions sur les STI

Dans cette section nous définissons les STI après un aperçu historique, puis nous présentons leurs différentes composantes et leurs rôles afin de mettre en évidence le potentiel des STIs pour l'apprentissage.

1.3.1.1 Historique des STI

L'idée des systèmes d'apprentissage automatisés remonte à la fin des années 60 – début des années 70, avec les espoirs suscités par le domaine de l'intelligence artificielle (IA), qui a donné lieu à la première génération des systèmes d'*Enseignement Assisté par Ordinateur* (EAO).

Les EAO sont définies comme 'toutes les applications utilisant l'ordinateur à des fins didactiques, qu'elles soient conventionnelles par l'intermédiaire d'un terminal, ou réalisées en temps différé' [BLO 84]. Les premiers systèmes d'EAO sont présentés sous forme de didacticiels exposés sur un écran d'ordinateur, qui se chargeaient d'étaler des informations déterminées sur la matière suivie par des QCM.

Les systèmes EAO semblent alors apporter une amélioration aux premières machines à enseigner car ils tiennent compte des réponses des étudiants et ils leur offrent des outils de progression plus ou moins ajustés à leurs rythmes. Cependant, ces systèmes étaient limités dans la mesure où ils présentaient à l'apprenant une succession de contenu figé et des notions cumulatives soumises à une évaluation plus ou moins arbitraire par le recours à des questions pour tester si le contenu présenté était bien assimilé. Les faiblesses de ces systèmes se

manifestent alors, d'une part, par l'absence de diagnostic à la suite des réponses de l'apprenant et d'autre part par le manque d'interactivité avec ce dernier. Cela représentait un obstacle pour comprendre les intentions de l'apprenant et identifier les difficultés qu'il rencontrait.

Avec les avancées réalisées en Intelligence Artificielle (IA) et en éducation cognitive, une nouvelle génération d'enseignement est apparue vers la fin des années 70 – début des années 80, à savoir, l'*Enseignement Intelligemment Assisté par Ordinateur* (EIAO), appelé aussi *Système Tuteur Intelligent* (STI) [SLE 82]. Ces systèmes se caractérisent par leurs capacités à fournir un enseignement individualisé selon le degré de connaissance de l'apprenant et à accomplir des tâches non prévues d'avance, en utilisant des techniques d'intelligence artificielle afin de ressembler à un tuteur humain.

1.3.1.2 Définition des STI

Les STI sont défini comme étant : « des systèmes informatiques capable de fournir un enseignement et un feedback personnalisé sans intervention humaine » [PSO 88]. Ils visent à simuler le comportement d'un enseignant humain dans ses capacités d'expert pédagogue et d'expert du domaine. Les STI ont ainsi l'intention de prendre en considération l'état de l'apprenant en adaptant les décisions pédagogiques aux besoins particuliers de chaque apprenant. De tels systèmes sont capables de déterminer et d'interpréter les erreurs de l'apprenant, et d'offrir une plus grande flexibilité dans la présentation du matériel à enseigner.

Plusieurs STI ont vu le jour et ont fait leurs preuves, notamment le système *SCHOLAR*, [CAR 79] qui est considéré comme étant le premier STI qui permet d'enseigner la géographie d'Amérique du Sud mettant en œuvre des dialogues à initiative mixte. On trouve aussi le système *GUIDON* [CLA 92] qui permet de diagnostiquer des maladies infectieuses, le système *Andes* [VAN 05] pour l'apprentissage de la physique et le système *Roman Tutor* [BEL 12] pour l'apprentissage de la manipulation d'un bras robot..., etc.

1.3.1.3 Architecture générale des STI

L'architecture de base d'un STI est constituée de quatre modules principaux: le module expert, le module tuteur, le module apprenant et l'interface utilisateur. Chaque module représente une entité abstraite qui traite des fonctionnalités spécifiques du STI. Le module expert du domaine représente les connaissances du domaine à enseigner (le 'quoi' enseigner), le module tuteur regroupe les stratégies et les moyens pédagogiques à incorporer dans le processus d'apprentissage (le 'comment' enseigner), le module apprenant incluant l'ensemble des connaissances sur l'apprenant et représente l'utilisateur du STI (le 'à qui' enseigner), et enfin le

module interface englobe les moyens de communication indispensable pour faciliter l'utilisation du système en le rendant plus convivial [WEN 87]. La figure 1.3 montre les différents composants d'un STI.

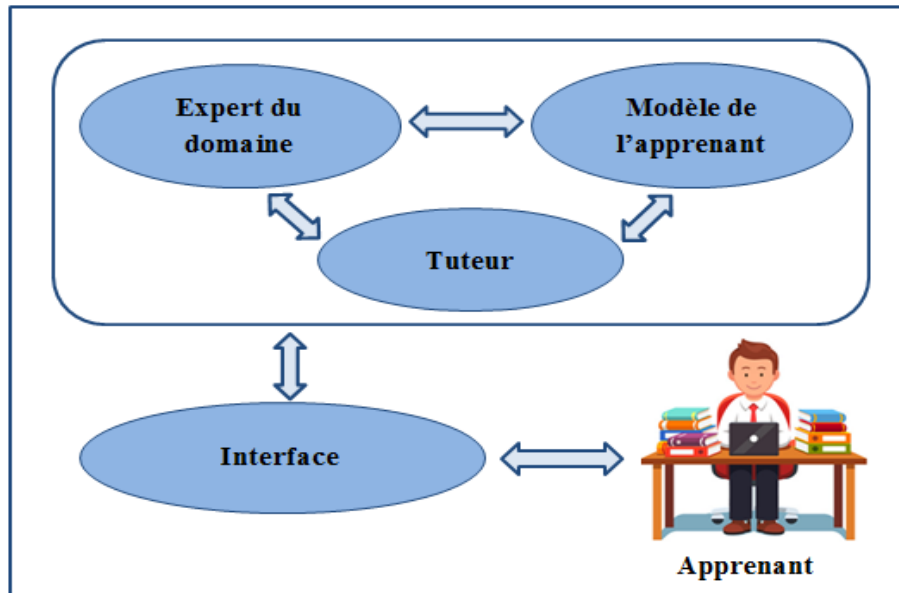


Figure 1.3 Architecture de base d'un STI [BUR 88]

Pour la suite de cette section, nous fournirons une description de chacune des quatre composantes présentées dans la figure précédente. Nous détaillerons ainsi le rôle attribué à chacune de ses composantes.

A. Module interface

Le module interface constitue un moyen de communication entre l'apprenant et le système tutoriel. Il est chargé de contrôler les interactions dans les deux sens : Collection des informations sur l'apprenant, affichage du matériel à enseigner, transfert des interventions, décisions pédagogiques prises par le tuteur, dialogue, etc.

L'interface utilisateur peut se présenter sous différentes formes selon le type d'environnement d'apprentissage (simulation, résolution de problèmes, etc.) et l'objectif visé par l'apprentissage (partage d'expérience, transfert de compétences, etc.). Elle peut prendre la forme d'une interface graphique (boutons, icônes, menus), d'un environnement à réalité virtuelle ou augmentée, d'une interface tangible, d'un langage de commandes ou même du langage naturel (GUIDON [CLA 92]), etc.

De façon générale, l'interface d'un STI doit être conviviale et modélisée de manière à ne pas poser des contraintes ou des obstacles de compréhension supplémentaires à l'apprenant, car elle influe sur la manière dont l'apprenant perçoit les informations présentées. Une interface simple et bien organisée aide énormément l'apprenant à acquérir les notions du

cours, par contre une interface compliquée, incorrectement modélisée ou mal structurée influe négativement sur ses capacités cognitives et peut remettre en cause tout le processus d'enseignement.

B. Module Expert

Le module expert, appelé aussi le *modèle du domaine* ou encore le *curriculum*, contient des connaissances sur la matière à enseigner, et des mécanismes permettant de les manipuler. L'expert du domaine doit être capable de généraliser le contenu didactique et de résoudre les problèmes liés à la matière présentée, en proposant des solutions ou des explications associées à chaque problème. Il doit aussi fournir une norme pour évaluer la performance de l'apprenant afin de détecter ses erreurs ou ses idées fausses [NKA 10], en d'autre terme, il doit interpréter les actions et les réponses de l'apprenant, et de les comparer avec les solutions générées par le système.

Plusieurs formalismes ont été proposés pour représenter les connaissances du domaine dans le module expert telles que les représentations à base de règles (règles de production, logique de premier ordre), les représentations en réseaux (réseaux sémantiques [QUI 68], les graphes conceptuels multicouches [NKA 97], réseaux connexionnistes neuronaux, réseaux de cadres, etc.), les représentations procédurales et la construction de scripts ou de frames [PAQ 11]. Actuellement, les ontologies [GRU 93] deviennent de plus en plus un mécanisme fréquent de représentation des connaissances d'un domaine.

Ces différentes méthodes de représentation amènent à déduire trois stratégies de modélisation de l'expert du domaine [NKA 10] :

- **Stratégie 'Boîte Noire' :** Appliquer une quelconque méthode de raisonnement sur le domaine. L'expert fournit des mécanismes pour résoudre des problèmes et des normes pour évaluer l'apprenant, mais il ne fournit aucune explication sur les démarches à suivre pour résoudre ces problèmes. C-à-d il n'y a aucun moyen d'accéder à la structure de raisonnement de l'expert. Par exemple, un système boîte noire ne serait pas capable de justifier pourquoi il attribue une réponse positive ou négative à une question qui lui est posée.
- **Stratégie 'Boîte Vitree' (système expert) :** L'expert peut expliquer ses raisonnements et les démarches à suivre pour résoudre un problème donné, en utilisant des méthodes de raisonnement basées sur des règles de production. Cette approche offre une représentation articulée des connaissances à la base de l'expertise.

- **Stratégie ‘Modèle Cognitif’** : L’expert simule la manière dont l’humain utilise ses connaissances pour résoudre des problèmes. Les connaissances sont décomposées et organisées de façon à être communicable par le tuteur.

La figure 1.4 présente un diagramme comparatif des méthodes de modélisations exposées précédemment.

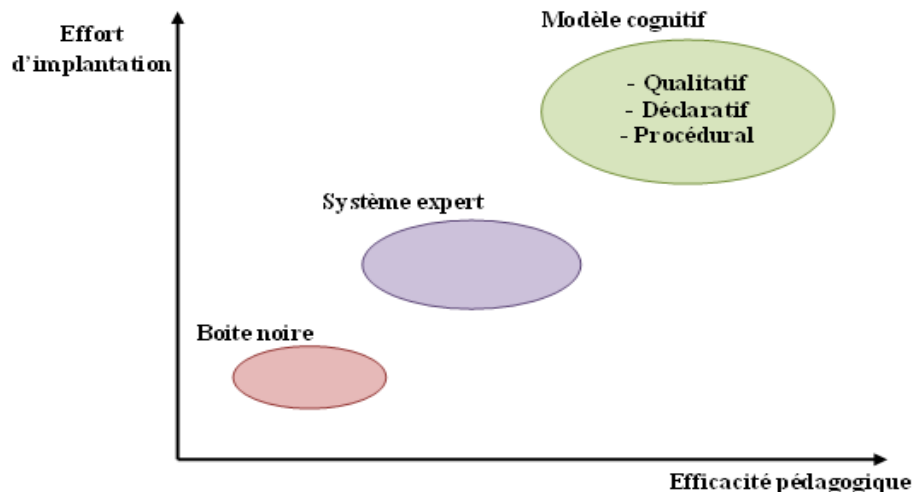


Figure 1.4 Diagramme comparatif des méthodes de modélisation de l’expert.

Bien qu’il demande un grand effort d’implantation, le modèle cognitif assure une efficacité pédagogique plus importante par rapport aux autres approches. C’est également le modèle qui se rapproche le plus de l’intelligence humaine du fait qu’il s’appuie sur le modèle de tutorat humain.

C. Module apprenant

Appelé aussi *module de diagnostic* ou *modèle de l’apprenant*, ce module a pour objectif de représenter à un instant donné, les états courants de l’apprenant (cognitifs, affectifs, psychologiques, émotionnels, etc..), et d’évaluer son niveau de progression par rapport au curriculum, en effectuant des diagnostics et des mises à jour régulières des connaissances acquises par l’apprenant [WEN 87].

Ce modèle est constitué de trois sous-modèles : modèle affectif, modèle inférentiel et modèle cognitif.

- Le **modèle affectif** contient, selon la complexité du modèle de représentation utilisé, des données permettant de décrire l’état émotionnel ou motivationnel de l’apprenant, ces préférences pédagogiques ainsi que son profil psychologique. Ces données sont essentielles à l’adaptation du système d’apprentissage aux besoins des apprenants (sé-

lection du contenu à présenter, choix du mode de communication, choix de l'approche pédagogique, et c.).

- Le **modèle inférentiel** permet au système de faire une inférence ou un diagnostic sur l'état de connaissance (compréhension) de l'apprenant pour déterminer les causes de ses erreurs ou de ses idées fausses. Le diagnostic de l'état de connaissance de l'apprenant peut se faire suivant deux approches :
 1. *Approche « Modèle tracing »* Anderson [AND 83] propose de créer et d'analyser la trace des activités de l'apprenant. Cette approche nécessite une bonne modélisation du processus de résolution de problèmes.
 2. *Approche « Knowledge tracing »* qui propose d'analyser un fragment (épisode) d'apprentissage pour déterminer les connaissances qui ont été utilisées par l'apprenant. Cette approche ne nécessite pas une modélisation complexe du processus de résolution.
- Le **modèle cognitif** considère les connaissances de l'apprenant comme étant un sous-ensemble des connaissances de l'expert, c-à-d les connaissances acquises par l'apprenant sont superposées par rapport à celle de l'expert (overlay). La différence entre les deux structures donne alors une indication sur les connaissances manquantes à acquérir. L'un des exemples typiques de STI dans ce cas de figure est le célèbre STI Guidon [CLA 83]. Il existe deux variantes de modèle de recouvrement 'overlay' :
 1. *Variante simple* dans laquelle, les connaissances de l'apprenant sont entièrement incluses dans celles de l'expert.
 2. *Variante avec erreurs (Buggy)* dans laquelle, sont stockées les connaissances de l'apprenant qui n'appartient pas au domaine de connaissance de l'expert, autrement dit, ses conceptions erronées. Cette variante permet une meilleure planification des interventions du système dans le processus d'apprentissage. Ce type de modèle a été implanté dans le tuteur BUGGY [BRO 78].

Les figures 1.5 et 1.6 présentent les deux variantes du modèle de recouvrement (Overlay) :



Figure 1.5 Modèle overlay simple

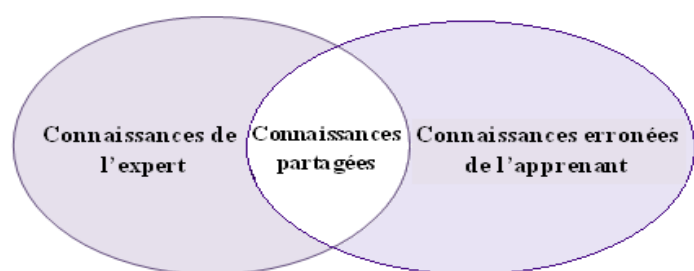


Figure 1.6 Modèle overlay avec erreurs (buggy)

D. Module Tuteur

Le module tuteur, appelé également le *modèle pédagogique*, est la composante responsable de toutes les interventions pédagogiques. Comme un tuteur humain, il a pour objectif d'appliquer les différentes stratégies d'enseignement lors de son interaction avec un apprenant. Son rôle est de planifier et sélectionner les activités pédagogiques à exposer lors de l'apprentissage du cours en les adaptant aux besoins de l'apprenant, tout en communiquant avec le module expert du domaine et le module apprenant. Il permet aussi de répondre de façon appropriée aux questions de l'apprenant, de corriger ses erreurs et de déterminer quand et comment intervenir pour l'aider à compléter sa tâche d'apprentissage.

Pour accomplir ce rôle, le tuteur s'appuie sur des approches éducatives et des stratégies pédagogiques bien déterminées permettant de guider et d'accompagner l'apprenant afin d'améliorer et d'assurer l'efficacité d'apprentissage. Nous présentons ci-après quelques exemples de ces stratégies :

- **Stratégie d'entraînement (coaching)** : le tuteur offre des explications ou des conseils permettant de guider l'apprenant lorsqu'il s'éloigne de la solution.
- **Stratégie du compagnon d'apprentissage** : Cette stratégie a été introduite par [CHA 90], dans le but de stimuler l'apprentissage de l'apprenant par la collaboration et la compétition. Dans cette stratégie, un ou plusieurs co-apprenants (compagnons), ayant le même niveau de performance et le même état de connaissances que l'apprenant, sont simulés par le STI. Ces compagnons peuvent échanger les idées avec l'apprenant pour résoudre des problèmes donnés par le tuteur. Ils peuvent aussi poser des questions sur la matière enseignée, ou donner des conseils à l'apprenant (et vice versa).
- **Stratégie du co-apprenant** : le tuteur est remplacé par un co-apprenant ayant un niveau de connaissances équivalent à celui de l'apprenant ou légèrement supérieur. L'absence du tuteur a pour but d'entraîner l'apprenant à apprendre 'par la découverte'. L'apprentissage par la découverte permet à l'apprenant de construire tout seul des questionnements et à y fournir des réponses à partir de ses propres connaissances. Le co-apprenant peut ainsi expliquer et aider l'apprenant dans la construction de ces connaissances.
- **Stratégie du perturbateur** : Dans cette stratégie, un co-apprenant simulé par le STI va essayer de perturber l'apprenant et lui proposer des solutions erronées, afin de tester sa confiance en ses propres connaissances. L'objectif donc est de pousser l'apprenant à défendre ses points de vue, réorganiser et consolider ses connaissances.

- **Stratégie d'apprentissage par la pratique :** Le tuteur fait quelques démonstrations de résolution de problème étape par étape, avant de demander à l'apprenant d'effectuer la procédure par lui-même.

1.3.2 Liens entre émotions et processus cognitifs

L'émotion a été longtemps considérée comme opposée à la cognition. Ainsi, plusieurs philosophes dont Platon, Descartes et Kant considèrent que les émotions et la cognition sont deux concepts antinomiques, appartenant le plus souvent à des domaines disjoints [ORE 03]. Ils sont concentrés sur le développement des processus cognitifs pour l'apprentissage en délaissant la dimension émotionnelle. Pour eux, la rationalité et la raison ne devraient pas laisser place aux émotions.

Même les théories d'apprentissage reconnaissant l'importance de l'émotion n'ont pas précisé le lien avec la cognition. Dans cette perspective, on trouve les travaux de Bloom et ces collègues [BLO 75][GAR 83][GOL 95] qui ont prévu deux taxonomies: une pour le domaine cognitif et l'autre pour le domaine affectif, mais sans établir un lien entre les deux ou sans montrer l'interdépendance des deux domaines.

Des recherches récentes en neurosciences et en psychologie ont confirmé le lien entre l'émotion et la cognition. Damasio [DAM 95] fut l'un des premiers chercheurs qui a montré que les émotions ont des influences sur plusieurs processus cognitifs tels que l'attention, la mémorisation, la perception et la prise de décision. Il a même affirmé que sans émotions, nous ne pouvons pas décider [DAM 95]. Ce point de vue a été ensuite corroboré par plusieurs chercheurs [WEI 00] [CLO 00] qui sont tous entendus sur le fait que les émotions sont un auxiliaire indispensable à la raison. A cet égard, Isen [ISE 00] a montré que les émotions positives jouent un rôle important pour améliorer la créativité et la flexibilité dans la résolution des problèmes [ISE 00]. Cependant, les émotions négatives peuvent bloquer le fonctionnement des processus cognitifs. Par Exemple, les gens anxieux ont un déficit dans le raisonnement inductif, une attention restreinte et une capacité de mémoire réduite [ISE 00]. Par ailleurs, des chercheurs en informatique s'intéressant à l'avantage des émotions sur l'apprentissage, entre autres, ont pu confirmer cette opinion. Leurs études empiriques menées ont également démontré que les émotions sont fortement liées aux processus d'apprentissage et que certaines émotions pouvaient favoriser le succès de l'apprentissage.

Parallèlement à ces études mettant en évidence le lien étroit entre la cognition et l'émotion, Picard [PIC 97] a développé un nouveau concept, à savoir *l'informatique affective* « affective computing ». Ce courant de recherche vise à doter les systèmes d'Interactions Homme-

Machine (IHM) d'une nouvelle forme d'intelligence, à savoir *l'intelligence émotionnelle* [GOL 95]. S'inspirant de ce courant de recherche, les STI ont progressivement évolué vers des *Systèmes Tutoriels Émotionnellement Intelligents* [OCH 04a] [OCH 04b] : une nouvelle génération de systèmes tutoriels dont les mécanismes d'adaptations sont d'avantage enrichis avec des heuristiques et des considérations émotionnelles.

1.3.3 L'intelligence émotionnelle

Savoir si une machine pourra un jour ressentir des émotions est une question depuis longtemps posée et discutée, sans qu'une réponse définitive ait été apportée. Sloman et Croucher [SLO 87] pensent que les machines intelligentes devraient essentiellement éprouver des émotions. A cet effet, Marvin Minsky [MIN 86] indique qu'on ne doit pas se demander si des machines intelligentes peuvent avoir des émotions, mais plutôt si ces machines peuvent être intelligentes sans avoir des émotions [MIN 86]. Selon Howard Gardner [GAR 83], l'intelligence n'est pas réduite à la simple mesure du quotient intellectuel (QI), mais comporte plusieurs formes, parmi lesquelles l'intelligence interpersonnelle et l'intelligence intrapersonnelle. Reposant sur les travaux de Grander [GAR 83], Mayer et Salovey [MAY 97] ont démontré que les individus n'ont pas tous, les mêmes capacités d'identifier et de gérer leurs propres émotions et celles des autres [SAL 90] ; ces capacités constituent une forme d'intelligence appelée *l'intelligence émotionnelle*.

L'intelligence émotionnelle est définie comme « l'habilité de *percevoir et d'exprimer* des émotions permettant d'améliorer le raisonnement, d'*identifier* les liens entre les émotions et la pensée, de *comprendre et de raisonner* sur les émotions et les connaissances émotionnelles, et de *réguler* ses émotions et celles d'autrui » [MAY 97].

- Les émotions peuvent être perçues et exprimées par différents moyens comme les expressions faciales, la voix, les gestes, etc.
- L'identification des liens entre les émotions et la pensée consiste à déterminer l'influence des émotions sur les processus cognitifs tels que l'attention, la motivation, l'apprentissage ou le raisonnement.
- La compréhension et le raisonnement sur les émotions consistent à distinguer les différents types d'émotions, à les étiqueter, interpréter leurs significations et identifier leurs causes et leurs effets et comprendre les processus de transition d'une émotion à l'autre.
- Finalement, la régulation des émotions consiste à gérer ses propres émotions, à répondre aux émotions d'autrui et les influencer de manière appropriée.

Selon Daniel Golmen, l'intelligence émotionnelle est l'un des facteurs majeurs contribuant à la réussite d'un individu sur le plan personnel et professionnelle. Elle est caractérisée par :

- La conscience de soi et la capacité à comprendre ses propres émotions : pouvoir identifier ses émotions,
- L'autorégulation ou la maîtrise de son propre comportement et de ses pensées : la capacité d'adapter ses émotions à chaque situation,
- l'auto motivation ou la motivation interne,
- La perception et l'influence des émotions sur l'autrui : l'empathie,
- Les aptitudes sociales ou la maîtrise des relations humaines : savoir entretenir des relations courtoises avec les autres.

1.3.4 L'informatique affective :

L'informatique affective est un domaine de recherche interdisciplinaire dérivée de l'Interaction Homme-Machine (IHM), qui vise à introduire les fondements requis pour faire valoir la dimension émotionnelle dans les systèmes interactifs actuels.

Définit comme étant : « *l'informatique qui se réfère à, en résulte, ou influences délibérément les émotions* » [PIC 97], ce courant de recherche permet de concevoir et de développer des systèmes capables de reconnaître et de gérer les émotions de l'utilisateur, et de comprendre et d'exprimer leurs propres émotions [PIC 97]. Trois principaux thèmes de recherche peuvent être distingués en informatique affective [PIC 97] :

- **La reconnaissance des émotions** est un aspect important de l'informatique affective dont l'un des objectifs est l'étude et le développement des techniques et moyens matériels permettant aux machines d'interpréter et modéliser des émotions humaines, il s'agit donc d'améliorer et de renforcer les capacités des machines afin d'identifier et évaluer l'état émotionnel de l'utilisateur et ainsi créer un canal de communication émotionnelle de *l'utilisateur vers la machine*.
- **L'expression des émotions** permet de donner aux machines la capacité d'afficher les émotions d'une façon naturelle, agréable et efficace au cours de leurs interactions avec les utilisateurs. Ces émotions sont généralement transmises aux humains par des agents animés, des robots sociaux ou des avatars graphiques conçus pour avoir un comportement émotionnel crédible et cohérent ce qui permet d'établir un canal de communication émotionnelle de la *machine vers l'utilisateur*.
- **La prise en compte des émotions** : il s'agit de trouver les meilleures stratégies permettant aux machines d'interagir avec pertinence aux émotions des utilisateurs,

donc l'objectif est d'intégrer la dimension émotionnelle dans le fonctionnement des machines afin d'améliorer la satisfaction, l'engagement et les performances des utilisateurs.

1.3.5 Les Systèmes Tutoriels Emotionnellement Intelligents (STEI) :

L'intégration des principes de l'informatique affective a permis l'apparition d'une nouvelle génération des systèmes d'apprentissage connus sous le nom de *Systèmes Tutoriels Émotionnellement Intelligents (STEI)* [OCH 04b].

Les *STEI* sont des systèmes tutoriels intelligents qui possèdent en plus des capacités de l'intelligence émotionnelle. Ils sont capables de déterminer et d'analyser l'état émotionnel des apprenants en adaptant leurs stratégies d'interventions ou d'enseignement en conséquence. L'objectif est d'utiliser la dimension émotionnelle pour améliorer l'efficacité de l'enseignement en optimisant les capacités cognitives de l'apprenant, maximiser ses performances et maintenir son expérience d'apprentissage. Un *STEI* doit principalement assurer les trois tâches suivantes [OCH 04b] :

- Reconnaître l'état émotionnel courant de l'apprenant.
- Déterminer l'influence de chaque stratégie d'intervention sur les émotions de l'apprenant.
- Choisir la stratégie d'intervention optimale à utiliser pour réguler l'état émotionnel de l'apprenant et optimiser ses performances.

Les stratégies de régulation jouent un rôle fondamental dans la logique d'interaction des *STEI* car elles permettent d'amener l'apprenant vers un état émotionnel favorable à l'apprentissage. En effet, ces stratégies cherchent à augmenter les performances d'apprentissage en tenant compte de l'impact qui peuvent exercer les émotions sur les capacités cognitives d'un individu [OCH 04b].

1.3.6 Modélisation des émotions dans les STI

Après avoir décrit les différentes approches de représentation des émotions en psychologie (voir partie 1.2.3), nous introduisons dans cette section les principales approches de modélisation des émotions des apprenants utilisés dans les STI.

Les recherches dans les STI ont été largement influencées par les théories psychologiques dans la conception et la modélisation des émotions. En effet, les approches discrètes existantes en psychologie ont été remises en question par plusieurs chercheurs [DME 07] [KOR 01a]. Ces approches permettent de définir directement des catégories spécifiques des émotions (sans passer par les représentations dimensionnelles). Par exemple, Conati et

Maclare [CON 04] ont développé un modèle probabiliste prenant en considération des émotions spécifiques. Le modèle créé est fondé sur le modèle OCC (Ortony, Clore and Collins) et il a été conçu pour évaluer les émotions des apprenants dans un environnement de jeu éducatif (voir figure 1.7).

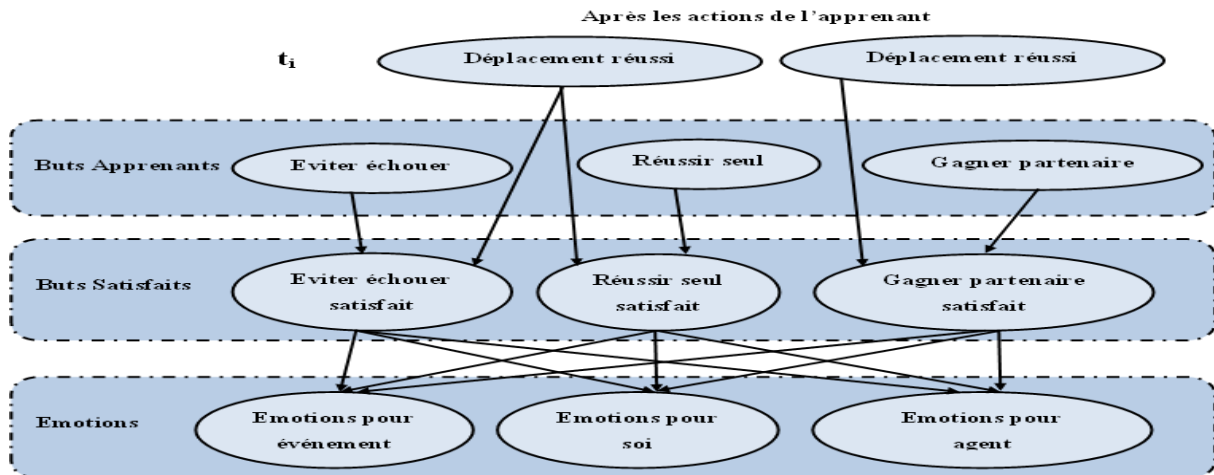


Figure 1.7 Sous-réseau d'évaluation des buts, traduit de [CON 04]

Comme il est indiqué sur la figure ci-dessus, le modèle représente des émotions vis-à-vis d'un événement (joie, tristesse), ou vis-à-vis de soi-même (fierté, honte) ou d'un agent (admiration, reproche). Ce modèle représente alors seulement 6 émotions parmi les 22 émotions définies dans le modèle OCC. De plus, ce modèle ne prend en compte que trois facteurs ou causes reliés à la génération des émotions (les buts, la situation et la personnalité).

Les approches continues introduites précédemment (Partie 1.2.3.2), et particulièrement les modèles bidimensionnels, sont aussi utilisées pour représenter les émotions des apprenants dans les STI et ce à partir de deux dimension : une dimension de valence (négative à positive) et une dimension d'activation ou intensité (faible à élevée) [KAI 06a] [PRE 05]. Ces modèles sont divisés pour reconnaître différents types d'émotions. Par exemple, le modèle circumplex de Russell [RUS 05] permet de représenter les émotions dans une structure circulaire bidimensionnelle dans laquelle se croisent les dimensions de valence et d'intensité de manière à former quatre quadrants (voir figure 1.8) : les émotions à valence négative et faible activation (par exemple l'ennui), les émotions à valence négative et forte activation (par exemple la colère et la peur), les émotions à valence positive et faible activation (par exemple le calme) et les émotions à valence positive et forte activation (par exemple l'excitation et la joie).

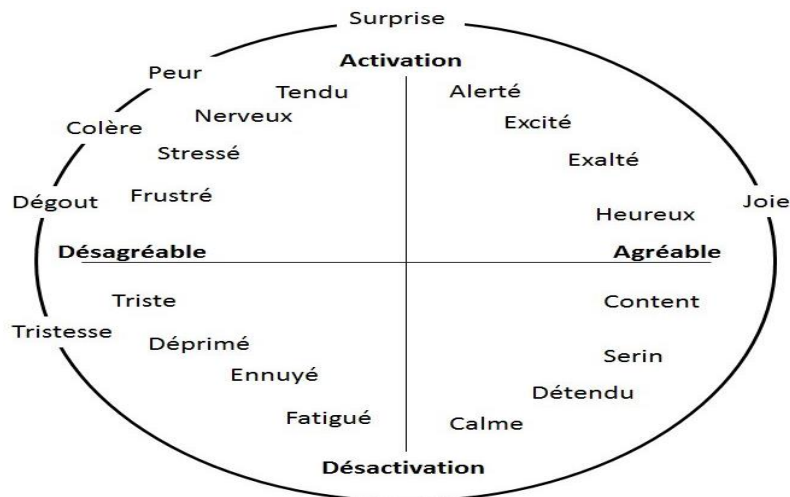


Figure 1.8 Modèle 'Circumplex' de Russell [RUS 05].

La modélisation des émotions de l'apprenant durant une session d'apprentissage est un élément essentiel pour les STI. En effet, le but est non seulement de représenter les émotions qui apparaissent lors de l'interaction des apprenants avec les STI, mais il s'agit aussi de définir de quelle manière ces émotions influencent le comportement de l'apprenant et son expérience d'apprentissage et quelles émotions faut-il considérer comme favorables ou défavorables pour l'apprentissage [DME 07].

Dans les deux dernières décennies, certaines recherches tentent de modéliser la relation entre les émotions et l'apprentissage [KOR 01a] [PIC 97]. Parmi les modèles qui explorent le lien entre les émotions et l'apprentissage on peut citer :

- **Le modèle de Stein et Levine**

Stein et Levine [STE 91] ont identifié le lien entre les objectifs d'un apprenant et ses émotions. Ils tiennent compte de certaines théories sur les émotions qui indiquent que les apprenants préfèrent être dans certains états (la joie) et préfèrent éviter d'autres (la tristesse) [DME 05]. Leurs modèles adoptent une approche de résolution de problèmes axée sur les objectifs, et suppose que les apprenants tentent d'assimiler de nouvelles connaissances en se basant sur des schémas existants (stéréotypes, scripts, croyances, etc...). Les auteurs supposent également que l'expérience émotionnelle est associée à l'interprétation des informations entrantes [DME 05]. Si ces informations sont nouvelles, elles provoquent une inadéquation avec les schémas existants et entraînent une excitation du système nerveux autonome (SNA) qui produit une émotion. Ce modèle théorique prédit que l'apprentissage se déroule normalement lors d'un épisode émotionnel [STE 91] [DME 05].

- **Le modèle de Kort, Reily et Picard**

Ce modèle est connu aussi sous le nom de modèle à quatre quadrants. Il a pour objectif de classer les émotions de l'apprenant et de déterminer celles qui sont impliquées dans le processus d'apprentissage [KOR 01a] [KOR 01b]. Le modèle est présenté par deux axes perpendiculaires (voir figure 1.9). L'axe horizontal représente la valence de l'état émotionnel de l'apprenant ou les émotions peuvent varier de l'état le plus négatif à l'état le plus positif. L'axe vertical représente la répartition des performances de l'apprentissage, qui varie de la situation non optimale (non apprentissage) dans le processus d'interaction à la situation optimale (apprentissage constructif).

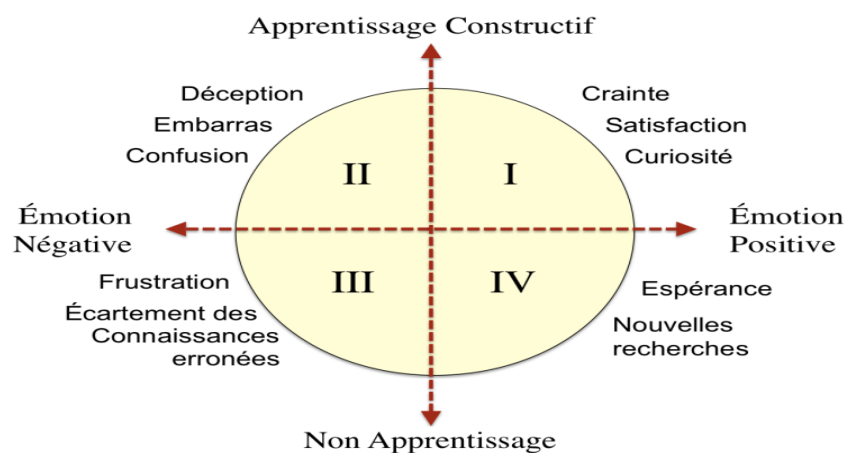


Figure 1.9 Modèle à quatre quadrants de [KOR 01a] reliant les émotions aux phases d'apprentissage.

L'état émotionnel de l'apprenant varie donc d'un quadrant à un autre [KOR 01a]. Le quadrant I (apprentissage et émotions positives), correspondrait à un état où l'apprenant serait curieux de découvrir et d'acquérir de nouvelles connaissances sans aucune lacune; Dans le quadrant II (apprentissage et émotions négatives), des émotions négatives telles que la confusion ou la déception pourraient apparaître car l'apprenant commencerait à avoir des difficultés dans la résolution des activités d'apprentissage. Si l'apprenant n'arrive pas à accomplir ces tâches alors l'apprentissage est basculé vers le quadrant III (non apprentissage et émotions négatives). Dans le quadrant IV (non apprentissage et émotions positives), les connaissances erronées sont éloignées et écartées. L'apprenant se retourne une autre fois vers le quadrant I, si de nouvelles idées sont développées.

- **Le modèle du déséquilibre cognitif**

Un bon nombre de modèles cognitifs démontrent le rôle important du déséquilibre cognitif dans les processus de compréhension et d'apprentissage [GRA 03]. La compréhension profonde se produit lorsque les apprenants sont confrontés à des

contradictions, des anomalies, des obstacles, des perturbations, des surprises et d'autres stimuli inattendus.

Le déséquilibre cognitif permet d'activer la conscience, l'effort, la délibération cognitive et les questions qui visent à rétablir l'équilibre cognitif. Les émotions de confusion et de frustration sont souvent survenues lors d'un déséquilibre cognitif [DME 05]. En effet, des recherches récentes ont montré que la confusion est une émotion importante pour une recherche scientifique [ROZ 03]. Elle indique une incertitude à propos de la prochaine étape à exécuter, et elle accompagne souvent le déséquilibre cognitif. De même, la perturbation et l'hésitation indiquent souvent le besoin de clarification et de plus d'information [ROZ 03].

- **Modèle des émotions de Kaiser et Oertel**

Kaiser et ces collègues utilisent une version étendue du modèle valence/activation de Russel [RUS 80] pour reconnaître les émotions favorables vs défavorables des apprenants durant une session d'apprentissage. Les auteurs mettent en évidence deux régions stratégiques (voir figure 1.10) pour définir ces émotions. Dans la première région des émotions défavorables telles que la colère, la frustration ou l'ennui (régions négatives I et II) sont compris. La deuxième région dite «cible» définie par une valence positive et une activation neutre [KAI 06a] [KAI 06b] [OER 07]. Elle représente l'état favorable et optimal durant laquelle l'apprenant maximise ses productivités, maintient son expérience d'apprentissage et optimise ses performances.

Les émotions de l'apprenant seront classées comme étant soit un état négatif défavorable (valence négative) à éviter lors d'une session d'apprentissage ou un état positif favorable (valence positive) à l'apprentissage.

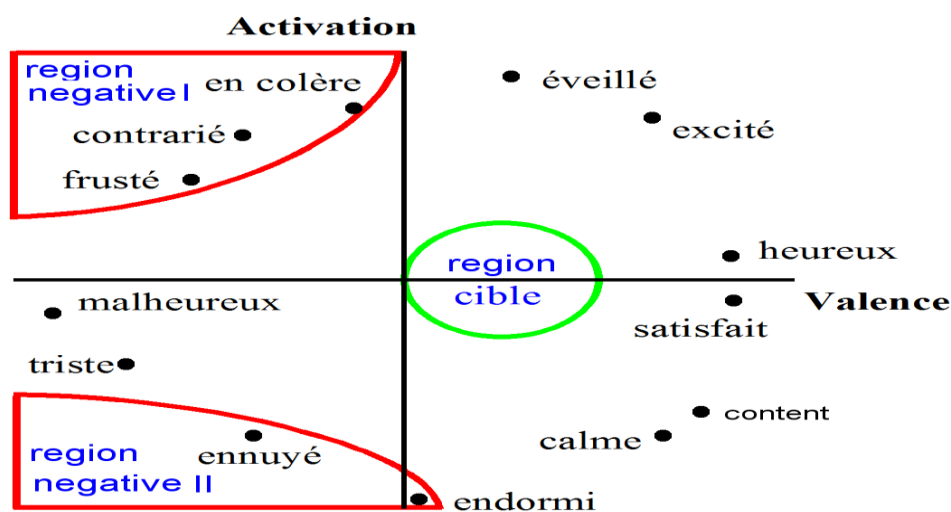


Figure 1.10 Les régions stratégiques des émotions [OER 07].

1.3.7 Méthodes de reconnaissance de l'état émotionnel de l'apprenant dans les STI

La reconnaissance des émotions consiste à identifier l'état émotionnel courant de l'apprenant, à partir d'un ensemble de comportement et d'observations produits suite à un changement émotionnel [PIC 97]. On distingue principalement quatre méthodes de reconnaissance des émotions dans les STI, à savoir : *l'auto-évaluation* (échelle verbale ou imagée, questionnaires, etc), *les variables d'interaction* (le nombre de clics et le mouvement de la souris, le temps passé sur les activités, etc), *les mesures comportementales* (les expressions faciales, la tonalité de la voix, la gestuelle ou la posture corporelle etc.) et *les mesures physiologiques* (la fréquence cardiaque, le volume de la pression sanguine, la fréquence de la respiration, etc.) [PIC 97]. La figure 1.11 résume les différents canaux et mécanismes que nous pouvons utiliser pour la détection des émotions des apprenants dans les STI.

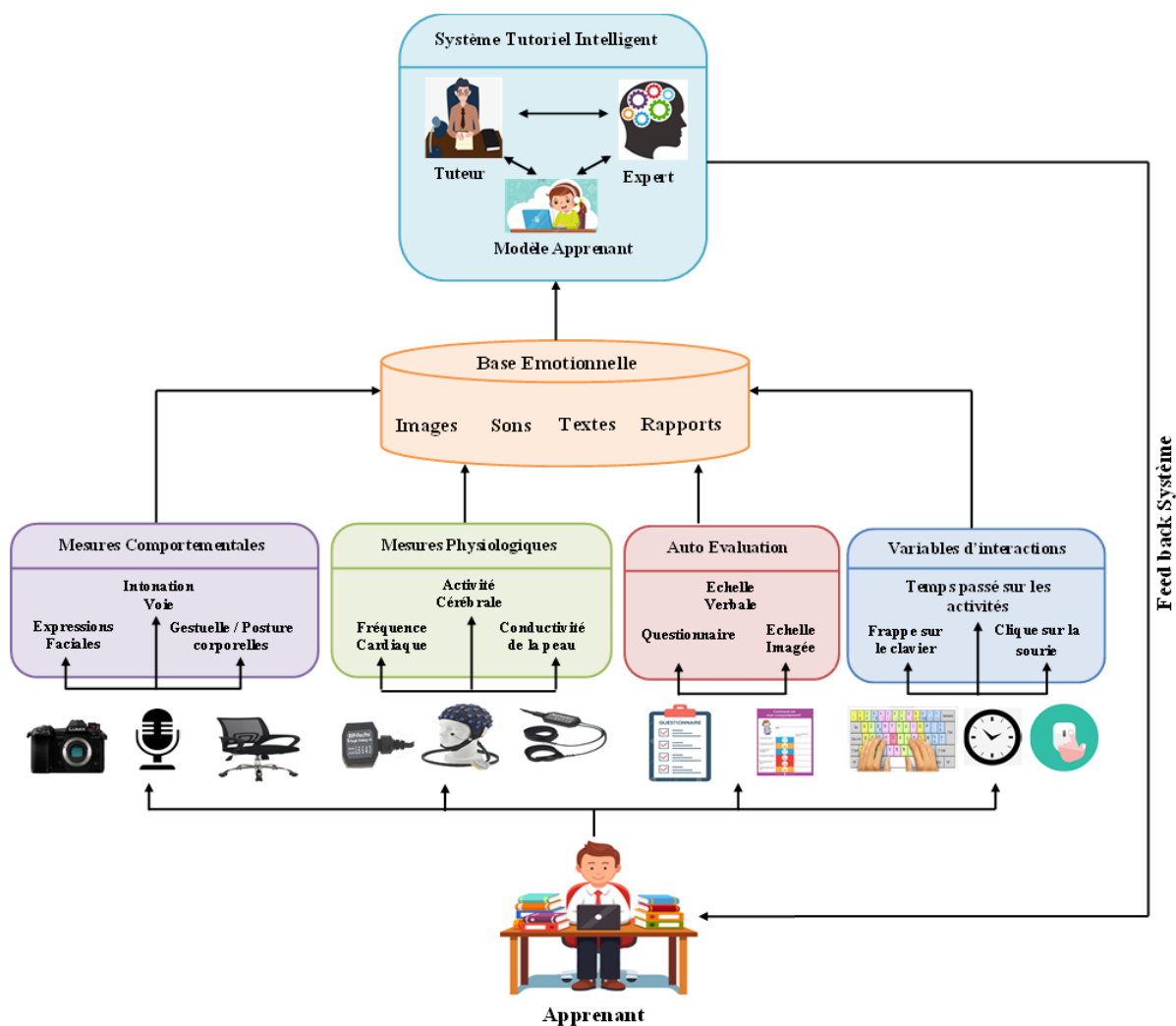


Figure 1.11 Canaux et mécanismes de détection de l'état émotionnel de l'apprenant.

1.3.7.1 L'auto-évaluation (self report)

Dans cette catégorie de mesure, l'apprenant est mené à décrire lui-même ses émotions pendant son interaction avec le système. L'auto-évaluation repose sur l'utilisation des outils comme les questionnaires à choix multiples et les échelles de mesure pour restreindre le choix de réponses de l'apprenant. Ces outils sont conçus pour modéliser les émotions selon deux types d'approches :

- **Outils d'auto évaluation selon l'approche catégorielle :** On trouve un ensemble d'émotions cibles qui sont visées par un questionnaire et en fonction des réponses apportées par l'apprenant, une émotion finale sera déterminée. Parmi les moyens d'auto-évaluation discrète, on peut citer l'échelle des émotions différentielles « Differential emotions scale » [IZA 93] qui permet de différencier entre dix émotions, on trouve aussi le questionnaire d'émotions académiques « Academic Emotions Questionnaire » [PEK 11] qui permet de discriminer entre huit émotions.
- **Outils d'auto évaluation selon l'approche dimensionnelle :** L'évaluation se fait en fonction des dimensions émotionnelles (principalement la valence et l'activation) et ne vise pas une émotion particulière. Parmi les moyens d'auto-évaluation dimensionnelle les plus utilisés, on peut citer la grille d'affect « affect grid » de Russel [RUS 89] qui permet à l'apprenant de décider si ses émotions sont de nature positive ou négative (pour la valence) et si elles ont une forte ou faible intensité (pour l'activation). Pour cela, Russell utilise une grille 9x9 représentant l'espace affectif bidimensionnel de la valence et activation (voir figure 1.12). L'apprenant est alors amené à mettre une croix dans la zone qui traduit au mieux ses émotions et plus cette croix se rapproche du centre de la grille plus les émotions de l'apprenant sont neutres.

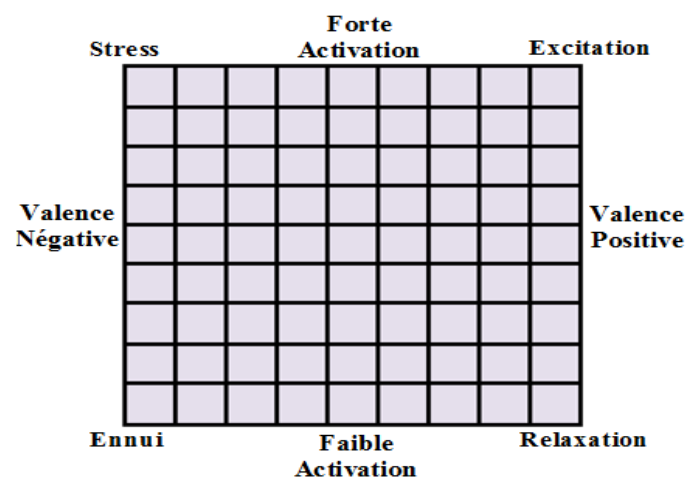


Figure 1.12 L'outil d'auto-évaluation grille d'affect «Affect grid » de Russel [RUS 89].

On peut également citer l'outil Self Assessment Manikin (SAM) [GIL 09] qui permet de représenter toute réponse émotionnelle de l'apprenant dans un espace à trois dimensions indépendantes et bipolaires : Valence, Activation et Dominance. Cette échelle utilise cinq figures représentant les différents niveaux d'émotions pour chaque dimension (voir figure 1.13). L'apprenant devra choisir pour chaque dimension la figure qui représente le mieux son état émotionnel. La combinaison des trois choix représente alors l'état émotionnel final de l'apprenant.

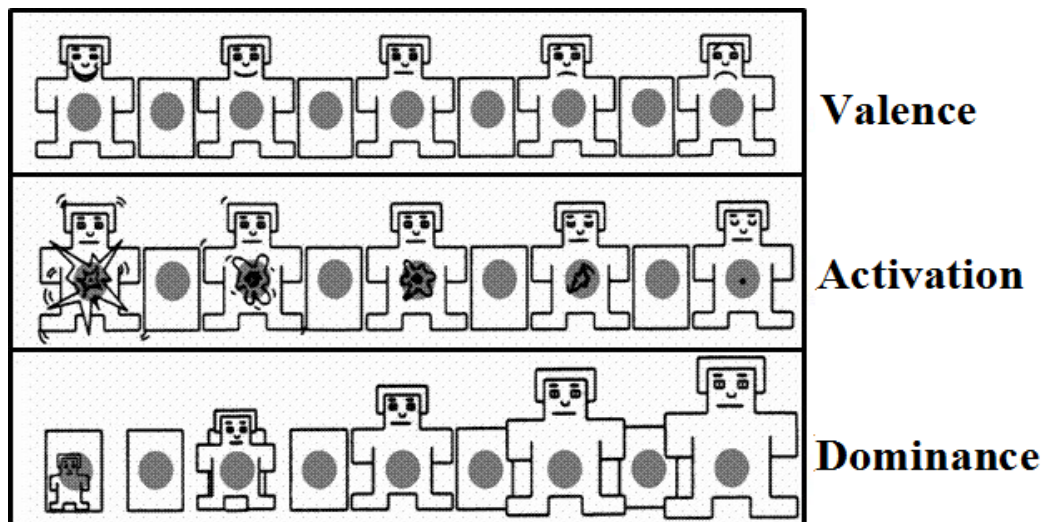


Figure 1.13 L'outil d'auto-évaluation SAM (Self-Assessment Manikin scale) [GIL 09].

L'auto-évaluation est une méthode rapide à administrer, facile à implanter et très fréquemment utilisés dans les STI [DEV 02], mais elle présente également des inconvénients. En effet, la procédure d'auto-évaluation peut perturber la dynamique de la session d'apprentissage, car elle nécessite d'interrompre l'apprenant pour l'interroger sur ses émotions. Si le système procède à une évaluation fréquente des émotions, cela peut devenir fastidieux pour l'apprenant et affecte sa concentration. En revanche, si cette évaluation est rare, le système risque de ne pas avoir une mise à jour régulière de l'état émotionnel de l'apprenant. L'auto-évaluation peut donc détériorer la qualité de l'interaction et le rendement de l'apprenant. De plus, certaines personnes peuvent être gênées à l'idée de révéler directement leurs émotions. D'autres éprouvent certaines difficultés à comprendre ce qu'ils ressentent exactement, ou encore, se préoccupent du fait que leurs réponses vont être enregistrées et que le système percevra leurs états, ce qui pourrait biaiser leurs évaluations. Les méthodes d'auto évaluation ne devraient pas donc être utilisées comme principal moyen de détection des émotions dans les STI [DEV 02]. Dans la littérature, l'auto-évaluation est généralement utilisée comme un moyen permettant de fournir des valeurs de références «

ground truth » pour la reconnaissance des émotions à partir d'autres sources d'informations comme les moyens comportementaux ou physiologiques.

1.3.7.2 Les variables d'interaction

Différentes variables d'interaction enregistrées par les STI sont utilisées pour la détection des émotions des apprenants. Ces variables incluent des indicateurs tels que : le temps émis pour répondre, le temps passé sur les activités, les fréquences d'utilisation des boutons d'aide fournie par le système ou les performances de l'apprenant (capacité de mémorisation, scores obtenus dans des tests ou des exercices, temps de réponse, etc.), la séquence de bonnes réponses ou de mauvaises réponses, le nombre de clics et le mouvement de la souris, etc

Par exemple, De Vicente et al. [DEV 02], analysent le nombre d'erreurs commises pour évaluer le niveau d'attention de l'apprenant. Baker et al. [BAK 04] [BAK 06] utilisent la fréquence de clics sur le bouton aide et le temps passé sur les activités pour déterminer le degré d'intérêt de l'apprenant. Les auteurs vérifient si l'apprenant utilise le système sérieusement ou il tente de le faire déjouer (« gaming the system ») en essayant de passer rapidement au travers le curriculum sans vraiment prêter attention aux activités d'apprentissage et faire le minimum d'effort possible.

Bien que les variables d'interaction permettent de détecter certains aspects émotionnels, leur utilisation est surtout dédiée à l'évaluation et au suivi des performances de l'apprenant dans l'exécution des activités d'apprentissage.

Les émotions peuvent aussi être communiquées à travers le dialogue échangé entre l'apprenant et le STI [DME 06]. En effet, le choix des mots et la ponctuation dans une phrase permettent d'exprimer et de reconnaître des émotions à partir des indicateurs lexicaux ou sémantiques extraits du dialogue.

L'analyse des dialogues de l'apprenant peut cependant être insuffisante, voire même imprécise, car les dialogues ne reflètent pas toujours l'état émotionnel de l'apprenant.

1.3.7.3 Les mesures comportementales

Cette catégorie de mesures permet d'identifier les émotions à partir d'une séquence de caractéristiques apparentes et observables tels que les expressions faciales, la tonalité de la voix, la gestuelle ou la posture corporelle, etc. Le recours à des capteurs physiques pour évaluer ces caractéristiques externe est une pratique très courante dans cette catégorie. En effet, l'utilisation des équipements et des capteurs non intrusifs et à faibles coûts est devenue de plus en plus courante dans les IHM, et plus particulièrement dans les STI [PIC 97] [PIC

01]. Ces capteurs peuvent être facilement intégrés dans les environnements d'apprentissage sans risquer de gêner l'apprenant.

Les **expressions faciales** occupent une place de premier plan dans la communication non-verbale et constituent le moyen le plus direct pour l'expression des émotions [EKM 82]. Ekman et Friesen [EKM 78] ont développé le système d'encodage FACS (Facial Action Coding System) pour coder les expressions faciales en utilisant des mouvements faciaux ou des contractions musculaires du visage. Les mouvements faciaux sont décrits par un ensemble de mouvements élémentaires indépendants appelés les unités d'action (Action Units AUs). Chaque AU est reliée à un muscle et chaque expression faciale peut être décrite par une combinaison d'AUs. Ce système repose sur la description de 46 unités d'actions (par exemple, AU5 correspond à une remontée de la partie supérieure de la paupière), le système ajoute également une lettre parmi 6 lettres possibles (de A jusqu'à E) pour décrire l'intensité de l'action musculaire (la lettre B par exemple correspond à un mouvement léger de l'unité d'action). Le système FACS permet de reconnaître 6 classes d'émotions basiques à savoir : la joie, la surprise, la colère, la peur, la tristesse et le dégoût, à partir de l'analyse des expressions faciales. Par exemple, l'expression de dégoût sera reconnue par un abaissement de la partie intérieure des sourcils et une fermeture de la bouche avec un haussement de la lèvre supérieure.

Des caméras de détection sont utilisées pour analyser les mouvements faciaux et extraire des caractéristiques à partir des différents composants du visage (lèvres, yeux, sourcils, nez, etc.). Ces caractéristiques sont ensuite utilisées pour la reconnaissance des émotions [NKA 04]. Par exemple, Nkambou et ses collègues [NKA 04] ont proposé une architecture d'un STI affectif intégrant plusieurs agents intelligents conçus pour la gestion des émotions. Un de ces agents (Emilie2) est responsable de la capture, l'extraction et la reconnaissance des émotions à partir des expressions faciales exprimées par l'apprenant lors d'une session d'apprentissage.

Tout comme les expressions faciales, l'analyse **automatique de la parole** constitue un vecteur important pour la reconnaissance de l'état émotionnel de l'apprenant. Les émotions du locuteur peuvent être véhiculées à travers deux niveaux : linguistiques et paralinguistiques. Le niveau linguistique est conforme aux règles de la langue, alors que le niveau paralinguistique est lié à la façon dont les mots sont prononcés. Il est généralement difficile d'obtenir d'une manière précise les émotions d'un individu uniquement par l'analyse linguistique de son message [GUN 10], car le choix des mots ainsi que leurs teneurs affectives peuvent énormément varier d'une personne à une autre et d'un contexte à un autre. Pour cela,

l'extraction des émotions à partir de la voix se fait le plus souvent à travers une analyse paralinguistique.

Différents indices prosodiques et acoustiques tels que l'intonation, l'articulation, l'intensité et la qualité du signal peuvent être utilisées pour identifier l'état émotionnel de l'apprenant [BOS 03] [LIS 05]. Par exemple, Liscombe et ces collègues [LIS 05] utilisent des indices acoustiques et prosodiques de la parole extraite à partir d'un corpus collecté du système tutoriel intelligent ITSPOKE, pour identifier quatre états émotionnels de l'apprenant : l'état de certitude, l'état d'incertitude, l'état neutre et l'état mixte. Bosch [BOS 03] utilise des caractéristiques acoustiques du signal vocal pour déterminer l'état de stress chez l'apprenant.

La **posture et les mouvements corporels** constituent également un moyen de détection émotionnelle très utilisé dans les STI. Par exemple, un apprenant intéressé vis-à-vis de l'activité en cours aura tendance à spontanément se rapprocher de l'écran ; en revanche, si l'apprenant est désintéressé ou commence à s'ennuyer, il aura tendance à s'affaisser sur sa chaise ou à avoir une posture qui l'éloigne de l'écran [DME 07]. Si l'apprenant commence à faire des gestes brusques et saccadés, cela peut signifier qu'il est en colère. S'il fait trembler ses genoux, cela peut indiquer qu'il est stressé. Un mouvement soudain en arrière peut exprimer une surprise. Se gratter la tête après une question peut exprimer une certaine réflexion et de la concentration. Un abaissement des épaules peut signifier une tristesse alors qu'un haussement spontané des épaules peut signifier une incertitude.

Une chaise équipée avec des détecteurs de posture placés sur le siège de l'apprenant (assise et dossier) sont utilisés pour mesurer l'amplitude et la direction des mouvements corporels (penchement du corps à droite, penchement en avant, jambes croisées, etc.), ainsi que la pression exercée sur le siège [DME 07] (voir figure 1.14).

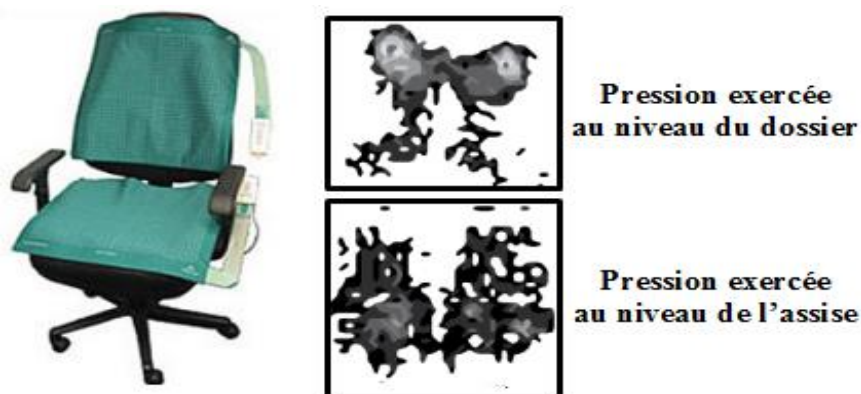


Figure 1.14 Exemple de chaise utilisée pour la détection émotionnelle [DME 07].

1.3.7.4 Les mesures physiologiques

L'activité physiologique d'un individu est étroitement liée à ses états émotionnels. En effet les signaux physiologiques permettent de mesurer des réactions qui reflètent la dynamique interne du corps humain. Ils sont difficiles à contrôler, consciemment, par l'être humain et se produisent spontanément avec le déclenchement d'une émotion. Pour ces raisons, l'approche physiologique représente une solution alternative, complémentaire et prometteuse en informatique affective.

Les signes les plus observés de la provocation émotionnelle concernent les changements dans l'activité du *Système Nerveux Autonome* (SNA). Le SNA prépare l'organisme à utiliser la totalité de ses ressources automatiquement dans des situations de stress, de danger, de défi, etc. Selon les émotions, on peut observer des variations de température et la conductivité électrique de la peau, des augmentations ou des diminutions de la fréquence cardiaque et la tension musculaire, du volume respiratoire, de l'activité cérébrale etc.

Plusieurs indices physiologiques sont actuellement utilisés pour caractériser les émotions. Ces indices sont mesurés à partir des capteurs physiologiques tel que : EEG, BVP, EKG, EMG, etc. (voir figure 1.15).



Figure 1.15 Canaux de communication physiologiques et capteurs associés [HAM 12]

Parmi les indices physiologiques les plus connus pour la détection émotionnelle, on peut citer :

- **La réponse galvanique de la peau (« Galvanic Skin Response », GSR) :** est une mesure de la réponse électrodermale ou du réflexe psycho-galvanique extraite à partir du capteur de la conductivité de la peau. Cet indice traduit l'activité des glandes sudori-

pares particulièrement présentes sur les mains et les pieds, responsables de la sécrétion de la sueur [GIL 09].

La réponse galvanique de la peau est généralement mesurée en positionnant sur les deux doigts de la main deux capteurs dédiés qui mesurent la différence de potentiel entre les deux zones cutanées (voir figure 1.15). Afin de ne pas mobiliser toute la main de l'individu pour mesurer la conductivité de la peau, de nouveaux capteurs existent sous la forme d'un bracelet ou d'un gant, et sont souvent utilisés dans les STI (voir figure 1.16 [a] et [b]).



Figure 1.16 Exemple de capteurs de la conductivité de la peau (a) et (b) représentent respectivement un bracelet et un gant utilisés pour mesurer la réponse galvanique de la peau [JRA 13].

La conductivité de la peau est fortement corrélée avec le niveau d'activation ou l'intensité émotionnelle [LAN 95b] : plus l'intensité de l'émotion est forte, plus la réponse de la conductivité de la peau est élevée et vice versa.

- **La fréquence cardiaque (« Heart Rate », HR) :** Cet indice mesure les variations issues des contractions des muscles du cœur. On peut déduire le rythme cardiaque directement à partir d'un électrocardiogramme ou à partir des données d'un capteur pression du volume sanguin (« Blood Volume Pressure », BVP) (voir figure 1.15) qui permet de calculer l'impulsion cardio-vasculaire à travers le corps humain. Ces capteurs de fréquences cardiaques permettent également d'extraire d'autres indices physiologiques tels que la pression sanguine, l'intervalle inter-battement (l'intervalle qui existe entre chaque deux battements successifs) ou la variabilité de la fréquence cardiaque dont l'augmentation est souvent associée à la présence de stress et de frustration [GUN 10].

La fréquence cardiaque est directement reliée avec la valence émotionnelle. En effet, il existe une corrélation linéaire significative positive entre le pic du rythme cardiaque

et la valence émotionnelle, et il est établi qu'une émotion positive provoque une phase d'accélération cardiaque plus importante qu'une émotion négative [LAN 95b].

- **La fréquence de respiration (RSP)** permet également de renseigner sur l'état émotionnel en indiquant la rapidité et l'intensité (profondeur) de la respiration d'une personne. [GUN 10]. Cet indicateur est mesuré à partir d'un capteur sous forme d'une ceinture placée sur la cage thoracique permettant de calculer le nombre de cycles respiratoires (inspirations et expirations) par une minute. Selon quelques chercheurs la respiration est influencée par l'activation émotionnelle [PRE 03].

Une respiration profonde et lente correspond généralement à un état émotionnel avec faible activation comme l'état de relaxation, alors qu'une respiration profonde et rapide peut indiquer l'excitation (forte activation) comme la colère ou la joie. Une respiration rapide et peu profonde pourrait manifester une anticipation tendue comme la concentration, alors qu'une respiration ralentie et peu profonde pourrait indiquer un état passif comme la dépression.

- **L'activité cérébrale** est considérée comme étant la principale source des émotions. Elle est mesurée à l'aide d'un électroencéphalogramme « Electro Encephalo Gram » (EEG), qui consiste à positionner un ensemble de capteurs (électrodes) sur la surface du cuir chevelu (voir figure 1.15). L'enregistrement de l'activité cérébrale, à travers l'EEG, permet de refléter de subtils changements neuronaux associés au niveau d'attention, de vigilance ou encore de charge de travail mental, et qui peuvent être mesurés avec une très haute précision et résolution temporelle.

L'activité cérébrale est souvent reliée à la détection d'émotions discrètes en utilisant des algorithmes d'apprentissage machine. Par exemple, [HER 09] ont développé un système qui permet de prédire les émotions de l'apprenant en se basant sur des algorithmes d'apprentissage machine, qui utilisent comme prédicteurs des bandes de fréquences extraites de l'activité cérébrale [HER 09] [CHA 09] [JRA 09].

- **La tension musculaire** : Cet indicateur est mesuré à partir d'une électromyographie (EMG). Le signal EMG enregistre l'activité électrique globale au niveau des muscles avec des électrodes placées à la surface de la peau. Selon des recherches existantes, le signal EMG est corrélé avec la valence émotionnelle et il dépend du muscle où il était mesuré. Par exemple, le signal mesuré du muscle zygomatique a une corrélation quadratique positive avec la valence [LAN 93]. Tandis que, le signal EMG mesuré du muscle corrugateur varie linéairement et négativement avec la valence [LAN 93].

1.3.8 Stratégies d'adaptation émotionnelles dans les STI

Dans cette partie, nous allons déterminer les principales stratégies émotionnelles utilisées dans les STI. Ces stratégies ont pour buts d'agir intelligemment sur les émotions des apprenants en définissant d'autres émotions ou en minimisant l'intensité de certaines émotions, cela permet d'augmenter les performances cognitives de l'apprenant, et maintenir son expérience d'apprentissage.

Les stratégies d'adaptation dans les STI définissent le moment et la nature des interventions établies par le système. On distingue principalement deux types de stratégies à savoir : les *stratégies orientées problèmes* et les *stratégies orientées émotions* [ROB 09].

Les *stratégies d'intervention orientées problèmes (ou tâches)* visent à répondre aux émotions négatives de l'apprenant en intervenant sur les éléments de la tâche d'apprentissage qui serait à l'origine de ces émotions [ROB 09]. Il s'agit donc d'agir sur l'activité d'apprentissage que l'apprenant est en train d'exécuter en vue d'améliorer ou de corriger une situation émotionnelle défavorable pour l'apprentissage. Ces stratégies se basent sur différentes interventions que le tuteur peut entreprendre, selon les émotions détectées, pour modifier le contenu tutoriel. Ceci peut être sous forme d'explication ou d'indication pour aider l'apprenant à surmonter certaines difficultés si un état de frustration est détecté, ou en proposant des activités plus stimulantes si un état d'ennui est détecté afin de susciter l'intérêt de l'apprenant. Par exemple, dans [DME 07] [DME 06], si un état de frustration est détecté, le système fournit des définitions associées aux concepts présentés, ou corrige les idées mal assimilées par l'apprenant. Dans [ARR 07] [ARR 09], si le STI détecte que l'apprenant s'ennuie, il lui propose des problèmes plus faciles à résoudre pour le motiver, ou il lui propose d'augmenter le niveau de difficulté si le problème est très facile.

Les *stratégies d'intervention orientées émotions* consistent à agir directement sur les émotions des apprenants sans changer ou remédier à la situation d'apprentissage qui serait à l'origine de ces émotions. Ceci peut être sous forme de messages pour motiver ou encourager l'apprenant [CHA 10].

Certains travaux ont recours à l'intégration des agents émotionnels virtuels au sein des STI pour réagir à l'état émotionnel de l'apprenant en exprimant diverses émotions à travers du contenu textuel, vocal ou corporel [LES 99] [ELL 99] [NKA 04] [PRE 05] [ARR 07] [ARR 09]. Ces agents jouent le rôle d'un partenaire humain qui, en plus de fournir le contenu du cours aux apprenants, peut reconnaître et gérer ses émotions négatives (comme la frustration, l'ennui ou la fatigue) [KAP 07]. Ils permettent aussi de donner des conseils et d'échanger des

discussions pour aider l'apprenant à faire face aux problèmes qu'ils sont en train de résoudre. La figure 1.17 présente quelques exemples d'agents émotionnels existants dans la littérature.

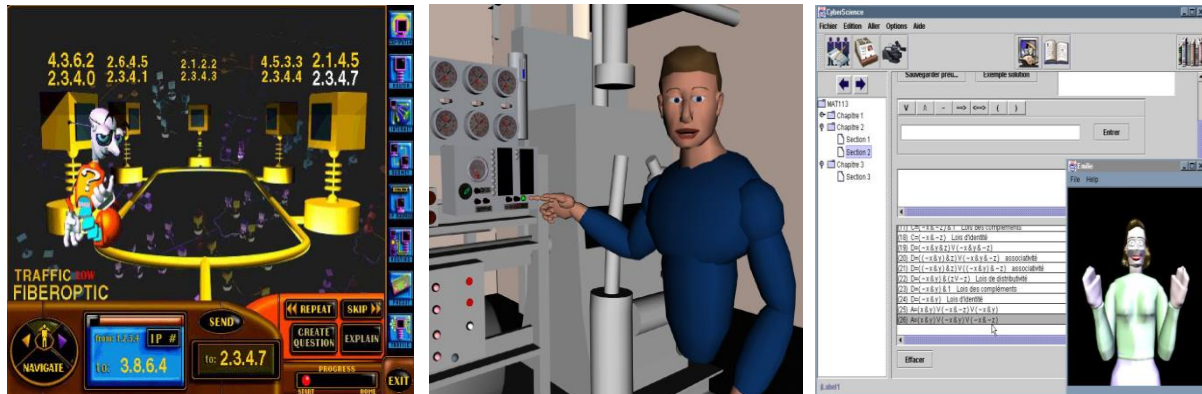


Figure 1.17 Exemples d'agents émotionnels utilisés dans les STI. De gauche à droite : l'agent COSMO [LES 01], l'agent STEVE [ELL 99], et l'agent EMILIE [FAI 02] dans leurs environnements virtuels.

Par exemple, [LES 01] ont développé l'un des premiers STI intégrant un agent virtuel doté d'émotion : **COSMO**. COSMO est un agent pédagogique animé destiné à apprendre les fondements de la topologie du réseau et des mécanismes de routage. Il explique notamment aux apprenants comment les ordinateurs sont connectés, comment gérer l'adressage et le routage des données, comment optimiser le trafic réseau et comment distinguer les différents types de réseaux selon leurs caractéristiques physiques [LES 99]. L'objectif de cet agent ne se limite pas à fournir aux apprenants des connaissances théoriques du cours, mais plutôt de les encourager et d'augmenter leur motivation à apprendre le cours. Il permet d'exprimer les émotions à travers son corps entièrement animé ; son visage (les yeux, les sourcils et la bouche), ses gestes (inclinaison de la tête, mouvement des bras et des mains, déplacement du corps, applaudissement, etc.) et sa voix (voir figure 1.17).

Elliot et ces collègues [ELL 99] ont aussi intégré l'agent virtuel pédagogique émotif **STEVE** (SOAR Training Expert for Virtual Environments) qui évolue dans un monde virtuel semblable à l'environnement de travail de l'apprenant (figure 1.17). L'agent a pour principal objectif d'aider l'apprenant à apprendre à exécuter des tâches, répondre à des questions et fournir de l'aide pour résoudre certains problèmes lors d'une formation sur la maintenance des équipements navals complexes. L'agent STEVE est appelé à contrôler et à réagir sur l'état émotionnel de l'apprenant. Pour ce faire, il dispose de 26 types d'émotions et de 70 expressions faciales.

Un autre exemple d'un STI basé sur une architecture multi-agents est **ÉMILIE** [NKA 04]. C'est un STI destiné à l'enseignement des sciences en ligne. Il intègre notamment deux agents

émotionnels adaptatifs, Émilie-1 et Émilie-2. Émilie-1 est un agent tuteur émotionnel 3D dédié à l'expression des émotions tandis qu'Émilie-2 permet de capturer et de gérer les émotions courantes de l'apprenant.

Les agents émotionnels virtuels intégrés dans les STI peuvent aussi adopter des réponses empathique indiquant l'état émotionnel courant de l'apprenant [PRE 05] [KAP 07]. Par exemple, Prendinger et Ishizuka [PRE 05] ont développé le système nommé « The Empathic Compagnon », qui utilisent une stratégie d'intervention basé sur l'empathie pour s'adapter aux émotions des apprenants (voir figure 1.18).

Ce système intègre un compagnon virtuel empathique, qui a pour but de préparer l'apprenant à la recherche d'un emploi en lui faisant passer un entretien d'embauche. Cet agent peut mesurer les émotions de l'apprenant en analysant son activité physiologique (réponses galvaniques de la peau et rythme cardiaque). Selon l'état émotionnel détecté (calme, joie ou frustration), l'agent fournit des réponses empathiques adaptées. Par exemple, si l'apprenant est frustré suite à une question, l'agent intervient afin de montrer à l'apprenant qu'il est inquiet de son état et il lui exprimera de la sollicitude et l'aidera à surmonter sa frustration à travers une indication particulière. Par contre, si l'apprenant est calme ou joyeux, le compagnon s'intervient avec des messages de félicitation ou d'encouragement.



Figure 1.18 Interface du système de Prendinger et Ishizuka [PRE 05]

L'utilisation des compagnons virtuels ainsi que le concept d'empathie dans les STI ont été également repris par Kapoor et ces collègues [KAP 07]. L'idée était de créer un agent virtuel animé qui simule le comportement de l'apprenant en *miroitant* les mêmes expressions faciales

et les mêmes mouvements de l'apprenant pendant que celui-ci est en train de résoudre un problème de tours de Hanoï (voir figure 1.19). L'agent peut par exemple sourire s'il voit l'apprenant sourire, ou avoir l'air ennuyé s'il détecte que l'apprenant est ennuyé.



Figure 1.19 Interface de l'agent animé développé par [KAP 07]

Cet agent permet de détecter les émotions des apprenants en utilisant une approche multimodale comportant des mesures physiologiques (réponse galvanique de la peau) ainsi que des mesures comportementales (expressions faciales, posture et mouvements corporelles, position des pupilles, nombre de clics et le mouvement de la souris) (voir figure 1.20)

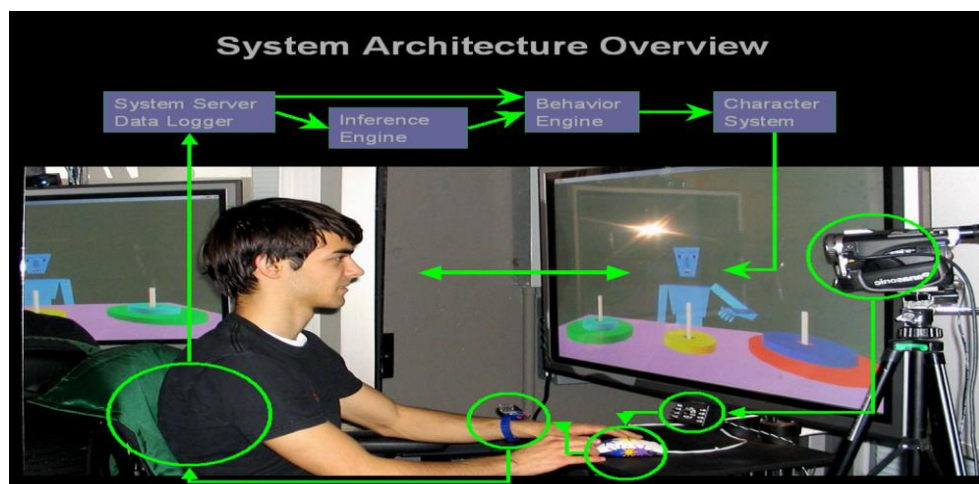


Figure 1.20 Système proposé par [KAP 07]. De gauche à droite : caméra vidéo, caméra yeux bleus, souris à pression, capteur de la conductivité de la peau, siège à pression

D'autres travaux utilisent des techniques d'induction émotionnelle afin d'agir sur l'état émotionnel de l'apprenant [CHA 04] [MAY 95] [PIC 01]. Ces techniques incluent notamment *l'imagination guidée* qui consiste à proposer à l'apprenant d'imaginer des situations plaisantes dans le but de réduire l'impact des émotions négatives [CHA 04]. Ils incluent aussi de la musique, les vidéos ou les images comportant des connotations émotionnelles pour induire certaines émotions.

Par exemple chaffar et ces collègues [CHA 04] ont développé le système émotionnellement intelligent ESTEL, qui utilise des interfaces incluant des vignettes d'imagination guidée, combinée avec de la musique et des images pour induire un état de joie chez l'apprenant. La figure 1.21 montre un exemple d'interface pour induire un état de joie dans le système ESTEL.

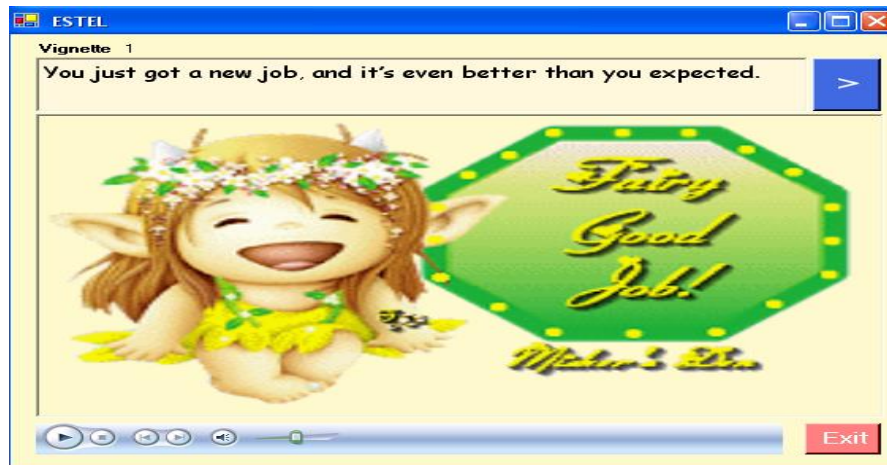


Figure 1.21 Exemple d'interface pour induire un état de joie [CHA 04].

1.4 Conclusion

Ce chapitre nous a permis d'exposer deux parties importantes : dans la première partie, nous avons présenté les émotions humaines avec ses définitions issues du domaine psychologiques, ainsi ses approches de représentations. Nous avons pu constater qu'il n'existe pas de consensus sur la définition du concept d'émotion, ni pour la présentation et la modélisation de celle-ci. Nous pouvons aussi constater que les définitions multiples de l'émotion, ainsi que le grand nombre de représentations associées à ce terme sont autant de difficultés auxquelles doit faire face la recherche en informatique affective.

Dans la deuxième partie, nous avons abordé l'impact des émotions sur les STI. Nous avons présenté dans un premier temps l'historique des STI, leurs définitions ainsi que leurs architectures globales tout en distinguant ses différentes composantes et le rôle de chacune de ces composantes. Nous avons mis l'accent sur ces systèmes puisqu'ils sont des technologies en pleine expansions et ils sont actuellement les tutoriels les plus utilisés dans l'enseignement de diverses disciplines. Dans un second temps, nous avons présenté les principaux concepts liés à la l'intégration de la dimension émotionnelle dans les STI à savoir : l'intelligence émotionnelle, l'informatique affective et les STEI. Ainsi, dans la même section, nous avons passé, en revue les méthodes développées pour la modélisation et la reconnaissance des émotions de

l'apprenant. Enfin, nous avons étudié les différentes stratégies émotionnelles utilisées par les STI pour adapter l'état émotionnel de l'apprenant.

Dans le chapitre suivant, nous établissons une étude détaillée sur la reconnaissance d'émotions de l'apprenant via les expressions faciales, ainsi nous focalisons sur les méthodes et techniques d'évaluation des émotions qui vont répondre aux objectifs de ce travail.

CHAPITRE 2

Système d'analyse des expressions faciales

Sommaire

2.1 Introduction	65
2.2 Qu'est ce qu'une expression faciale	65
2.3 Expression faciale versus Emotion	67
2.4 Formalismes de représentation des expressions faciales	68
2.4.1 Représentation par le système FACS (Facial Action Coding System) ..	68
2.4.2 Représentation par la norme MPEG-4	70
2.5 Expression des émotions primaires	72
2.6 Système d'analyse des expressions faciales	73
2.6.1 Acquisition du visage	74
2.6.2 Détection du visage	75
2.6.3 Prétraitement	78
2.6.4 Extraction des caractéristiques faciales	78
2.6.5 Classification des expressions faciales	95
2.7 Les systèmes de fusion	102
2.7.1 Fusion de données	103
2.7.2 Fusion d'attributs (caractéristiques)	104
2.7.3 Fusion des décisions	105
2.8 Conclusion	109

2.1. Introduction

Les expressions faciales sont une forme de communication non verbale permettant à l'individu de transmettre rapidement à ceux qui l'entourent des informations concernant son propre état émotionnel. Il s'avère que l'être humain est particulièrement compétent pour reconnaître les expressions faciales qu'exprime autrui sur son visage. Néanmoins, la tâche de détection et d'extraction des différents composants qui caractérisent l'expression faciale est une activité extrêmement complexe pour un ordinateur. Ainsi, le paradigme de reconnaissance des expressions faciales a suscité l'intérêt de nombreux chercheurs dans la communauté du traitement d'images et de la vision par ordinateur. Différentes solutions ont alors été proposées permettant ainsi de classer les expressions en quelques grandes catégories émotionnelles, telles que la joie, la tristesse, la colère, la surprise, la peur et le dégoût.

Dans ce chapitre, nous présenterons un état de l'art des différentes méthodes appliquées à la reconnaissance automatique des expressions faciales. Dans un premier temps, nous allons définir le concept des expressions faciales et décrire les différents formalismes qu'il le représente. Ensuite, nous décrirons les étapes de traitement du processus d'analyse des expressions faciales à savoir la détection du visage, l'extraction des caractéristiques et la classification et nous citerons les approches les plus populaires pour chacune de ces étapes. Dans un deuxième temps, nous nous focalisons sur la description des systèmes multimodaux en présentant les différentes stratégies de fusion qui ont été utilisés dans la littérature.

2.2. Qu'est ce qu'une expression faciale

Le visage humain est le premier site de l'expression émotionnelle, il se caractérise par une ressource d'informations très riche dont la forme et le comportement de chaque membre change son apparence. Dans la science anatomique, le visage comporte deux os (le crâne et la mâchoire) et 44 muscles qui peuvent être activés volontairement ou non pour engendrer un nombre total égal à 250 000 expressions faciales [LAL 05].

Une expression faciale résulte alors d'un changement de position et de forme d'un ou de plusieurs traits de visage (Voir Figure 2.1). Ce changement est causé par la contraction de nombreux muscles dont l'interaction avec les tissus du visage produit des déformations visibles.

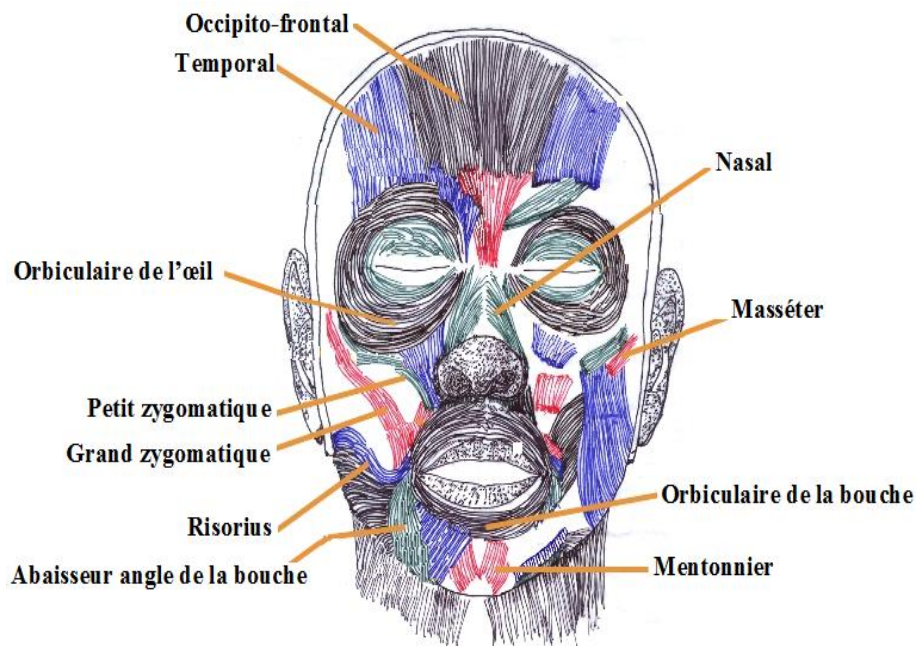


Figure 2.1 Les muscles faciaux

Les expressions faciales émotionnelles sont le principal moyen non verbal d'expression des états émotionnels internes, des intentions ou des communications sociales d'une personne. Au XIXe siècle, le neurologue Duchenne de Boulogne fut le premier à localiser individuellement les différents muscles faciaux en utilisant la stimulation électrique. Il est aussi un des premiers à livrer à la communauté scientifique un ensemble de photographies montrant les mouvements associés à l'expression des émotions. Il remarque ainsi que le sourire sincère reflétant une émotion positive se différencie du sourire volontaire par la contraction du muscle orbiculaire oculi qui se situe autour des yeux (Voir Figure 2.2).



Figure 2.2 Photographies des expériences de stimulation électrique des muscles de Duchenne

En 1872, Charles Darwin, a publié dans le cadre de sa théorie de l'évolution, son ouvrage « L'expression des émotions chez l'homme et les animaux » [DAR 72]. Il a tenté d'expliquer l'origine des principaux comportements expressifs dans différentes espèces. Darwin a conclu que les expressions faciales produites par les émotions étaient comparables chez tous les peuples, quelque soient leur origine raciale ou leur héritage culturel. Pour lui les états d'esprit sont reflétés de la même façon chez tous les hommes et chez les animaux par des mimiques ou des expressions faciales.

Ces travaux sont repris au XXe siècle, notamment par le psychologue Paul Ekman [EKM 69] qui a montré qu'il existe un nombre limité d'expressions du visage à savoir : la surprise, la peur, le dégoût, la colère, la joie, la tristesse et la neutralité, qui sont reconnues universellement de la même manière, sans qu'il n'y ait besoin d'apprentissage. Les six premières expressions sont dites des expressions *primaires* ou de *base*. Le langage universel des expressions faciales semble toutefois s'arrêter aux émotions primaires, toutes les compositions d'émotions plus complexes dépendent fortement du contexte ethnique et à l'intérieur même d'une ethnie, du contexte social, de la localisation, etc.

2.3. Expression faciale versus Emotion

Les émotions et les expressions faciales sont fortement liées et parfois confondue, donc il est important de faire une distinction entre les deux concepts. En effet, l'analyse des expressions faciales cherche à décrire les déformations des traits faciaux et les mouvements du visage issus de l'activation des muscles à partir d'informations purement visuelles. Les expressions faciales peuvent parvenir de l'activité physiologique (douleur, fatigue, etc), de l'état d'esprit (la réflexion) et de la communication non verbale (émotion simulée, clin d'œil, froncement des sourcils, etc).

L'analyse des émotions, quant à elle, est une tentative d'interprétation qui requiert souvent la compréhension de la situation ainsi que des informations contextuelles complètes (Voir figure 2.3). Les émotions résultent de nombreux facteurs et peuvent être révélées sur plusieurs canaux par exemple : la voix, la posture, les gestes, la direction du regard, les expressions faciales, etc.

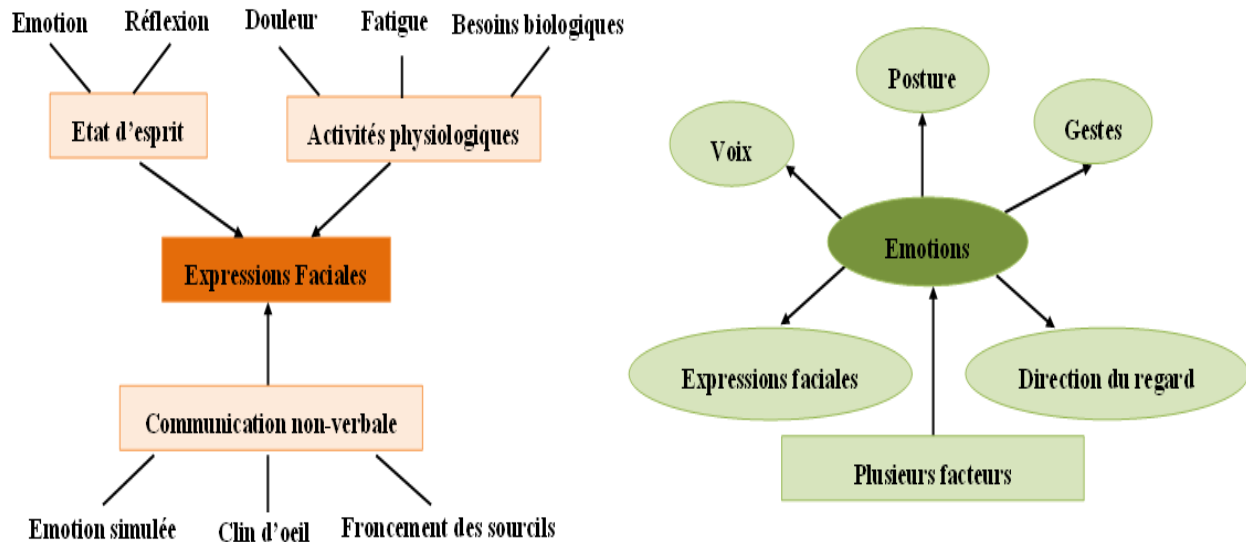


Figure 2.3 Générateur des expressions faciales et des émotions

2.4. Formalismes de représentation des expressions faciales

Les représentations faciales relient les mouvements des muscles faciaux aux émotions exprimées. Nous exposons dans cette section les différents formalismes proposés pour décrire les expressions faciales émotionnelles.

2.4.1. Représentation par le système FACS : Facial Action Coding System

En 1978, Ekman et Friesen ont créé un système destiné à la codification *manuelle* des expressions du visage connu sous le nom 'FACS : Facial Action Coding System' [EKM 78]. De par l'exhaustivité descriptive des mouvements faciaux qu'il propose, ce système est devenu l'outil principal des psychologues, animateurs, et de tout chercheur s'intéressant aux mécanismes de productions des expressions faciales.

Le système FACS est un guide technique qui explique la façon de classer les comportements faciaux, fondés sur les muscles, et comment l'action musculaire est liée à l'apparence faciale. Il permet de décomposer tous les mouvements *visibles* du visage en *unités d'action* (AU - Actions Units) qui décrivent les mouvements *élémentaires* des muscles, par exemple AU1, relèvement de la partie intérieure des sourcils, consiste en une position actuelle des sourcils plus haute que sa position précédente [EKM 03]. AU4 correspond à l'abaissement et le rapprochement des sourcils, AU7 indique l'existence d'une tension au niveau des paupières, etc. Le tableau 2.1 donne quelques exemples des actions unitaires du système FACS.

AU	Nom	Exemple
1	Relève de la partie intérieure des sourcils	
2	Relève de la partie extérieure des sourcils	
4	Abaissement et rapprochement des sourcils	
5	Relève de la paupière supérieure	
6	Relève les joues	
7	Relève et tend des paupières inférieures	
9	Plissement de la peau du nez vers le haut	
10	Relève de la lèvre supérieure	
11	Creusement des rides nasales	
12	Étirement des coins des lèvres	
13	Bomber les joues	
15	Abaissement des coins externes des lèvres	
16	Ouverture de la lèvre inférieure	
17	Relève du menton	
20	Étirement externs des lèvres	
23	Serrer les lèvres	
24	Presser les lèvres	
25	Ouverture de la bouche	
26	Ouverture de la mâchoire	

Tableau 2.1 Les AUs fréquentes décodées par le manuel FACS

Une expression faciale observée peut donc être représentée sous la forme d'une combinaison d'AUs. Les émotions de base définies par Ekman ont été aussi décrites par une combinaison de plusieurs AUs spécifiques à chaque expression d'émotion. A titre d'exemple, une expression de *surprise* est définie par l'association des AUs suivant : AU1 (haussement

du sourcil intérieur) + AU2 (haussement du sourcil extérieur) + AU5 (haussement de la lèvre supérieure) + AU25 (écartement des lèvres). Le tableau 2.2 présente la description des combinaisons des AUs correspondant à chaque émotion de base définie par Ekman.

Emotion	Action Unit
Colère	4+5+17+23+24
Dégoût	9+10+17+25
Joie	6+10+11+12+13+25+37
Peur	1+2+4+5+20+25
Surprise	1+2+5+25
Tristesse	1+4+11+15+17

Tableau 2.2 Combinaison des AUs correspondant a chaque émotion

Le répertoire FACS compte 46 AUs classées par zones du visage (visage haut, visage bas) et environ 7000 combinaisons différentes d'AUs lors de la production d'expressions spontanées. A chaque AU est associé un score qui varie en fonction de l'intensité (1 à 5 niveaux de magnitude), de la durée ou encore de l'asymétrie. L'association des scores de chaque AU pour un visage produit un code d'identification unique pour l'expression de ce visage.

Il existe plusieurs variantes du système de codage *FACS*, parmi lesquelles on peut citer : *FAST* (Facial Action Scoring Technique), *EMFACS* (Emotional Facial Action Coding System), *MAX* (Maximally Discriminative Facial Movement Coding System), *EMG* (Facial Electromyography), *AFFEX* (Affect Expressions by holistic Judgment), *FACSAID* (FACS Affect Interpretation Database) et *FACS* pour nourrissons/bébés.

Bien que FACS s'est imposé comme un outil puissant de description des mimiques du visage bénéficiant d'une grande maturité (environ vingt années de développement), il souffre cependant de quelques inconvénients comme la complexité, la difficulté de manipulation par une machine et la manque de précision. De plus, dans le système FACS, la combinaison de plusieurs mouvements faciaux ne peut pas être décrite facilement par la combinaison visuelle de chacun des mouvements isolés. Le système n'est pas du tout annoté, il s'agit juste d'une description des visages sans y associer des étiquettes émotionnelles.

2.4.2. Représentation par la norme MPEG-4

Un deuxième codage très utilisé pour la description des expressions faciales est la norme de compression MPEG-4 [STA 01]. Cette norme est utilisée principalement pour l'animation et la synthèse du visage. Elle dispose d'un modèle du visage humain développé par le groupe d'intérêt Face and Body AdHoc Group. C'est un modèle 3D articulé, capable de s'adapter à

une morphologie particulière et à des expressions particulières. Ce modèle est construit sur un ensemble d'attributs faciaux appelés paramètres d'animation faciale (*FAPs : Facial Animation Parameters*) nécessaire au codage de la déformation d'un visage à l'état neutre. Les *FAPs* sont définis en fonction de deux unités de mesures : Les points caractéristiques (*FP : Feature Points*) et les unités de paramètres d'animation faciale (*FAPU : Face Animation Parameter Units*).

Les *FPs* représentent un ensemble de points caractéristiques clés permettant de définir l'amplitude de la déformation faciale. La norme MPEG-4 définit 84 *FPs* sur un visage neutre [MEP 98] (Voir Figure 2.4 (a)). Les *FAPUs* (équivalentes aux *AUs* du système *FACS*) permettent de définir des mouvements élémentaires du visage de manière transposable. Elles représentent les rapports des distances entre les principales caractéristiques faciales. La figure 2.4 (b) montre le modèle de visage dans son état neutre avec les *FAPUs* correspondants. Comme exemples de *FAPU*, on peut citer *la largeur de la bouche*, *la distance de séparation entre la bouche et le nez*, *la distance de séparation entre les yeux et le nez*, etc. Par exemple, l'étirement du coin de la lèvre gauche (Facial Animation Parameter 6 stretch_l_cornerlip) est défini comme le déplacement vers la droite du coin de la lèvre gauche d'une distance égale à la longueur de la bouche.

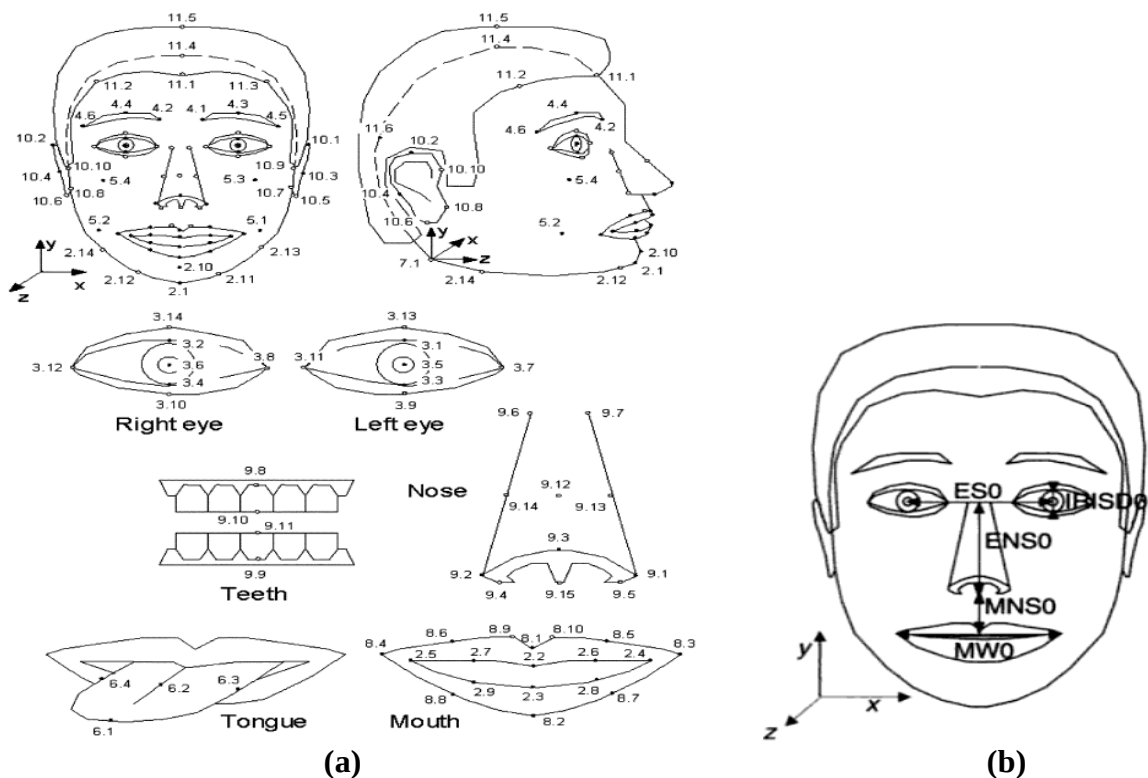


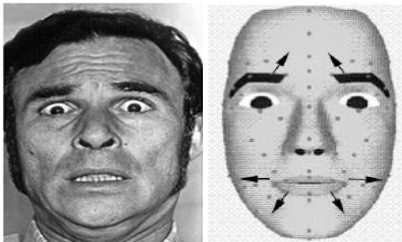
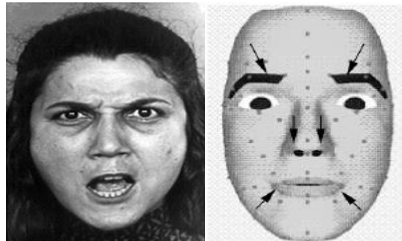
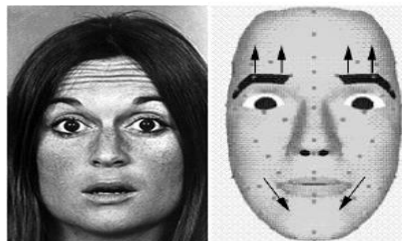
Figure 2.4 Modèle du visage par le standard MPEG-4 : (a) Représentation des points caractéristiques *FPs*. (b) Modèle des paramètres unitaires *FAPUs*

Comme le codage FACS, le codage par FAPs est très lié aux changements des muscles faciaux. Il définit également six expressions faciales émotionnelles qui coïncident avec les émotions de base définies par le système FACS.

2.5. Expression des émotions primaires

Les expressions faciales se déclinent en d'innombrables mimiques rapportant l'état émotionnel interne d'un individu. Le dénominateur commun de toutes ces expressions est l'utilisation universelle de six expressions faciales liées aux six émotions primaires : la peur, le dégoût, la surprise, la joie, la colère et la tristesse. Ces expressions sont inscrites dans notre cerveau et ont été reconnues depuis de nombreuses années comme universelles et indépendantes du contexte culturel ou social.

Dans ce travail de thèse, nous étudierons la reconnaissance des expressions faciales en utilisant ces émotions de base. Afin de classer correctement une expression faciale requête, nous allons donner une brève description de chacune des expressions mises au point par Ekman pour fixer leurs caractéristiques pertinentes.

<i>Expression faciale</i>	<i>Description</i>
Peur 	Emotion négative induite par des situations dangereuses ou stressantes. Elle est caractérisée par les <i>sourcils qui se tirent et s'élèvent</i>, les <i>lèvres légèrement tirées en arrière</i> et la <i>bouche ouverte horizontalement</i>, les <i>paupières supérieures mises en avant</i> et les <i>rides au milieu du front</i>.
Colère 	Emotion négative indiquant l'état d'une personne dans une réaction violente et agressive lors d'une menace. Elle est caractérisée par les <i>sourcils qui se baissent et se rapprochent</i>, les <i>lèvres restent rapprochées</i>, les <i>yeux sont quasiment fermés</i>.
Surprise 	Emotion soudaine indiquant l'état d'une personne étonnée par quelque chose d'inattendu. Elle est caractérisé par la <i>bouche ouverte</i>, les <i>sourcils sont courbés et élevés</i>, les <i>yeux sont grands ouverts</i> et <i>l'apparition des rides horizontales traversant le front</i>.

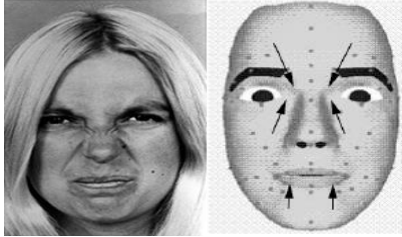
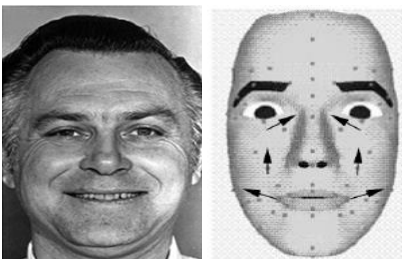
Dégout 	Emotion négative généralement évoquée par l'odorat, le goût ou la vision. Il se caractérise par le <i>nez fronce</i>, les <i>pommettes élevées</i>, les <i>coins intérieurs des sourcils</i> sont <i>légèrement baissés</i>, les <i>yeux</i> sont <i>semi-ouverts</i> et la <i>bouche</i> est <i>fermée et légèrement déplacée vers le haut</i>.
Tristesse 	Emotion négative qui apparaît quand une personne souffre moralement suite à une insatisfaction et des soucis. Elle est caractérisée par les <i>coins intérieurs des sourcils</i> sont <i>tirés vers le bas</i> et la <i>peau en dessous des paupières</i> devient <i>triangulaire</i> et <i>s'élève</i>, les <i>coins de la bouche</i> sont <i>abaissés</i>.
Joie 	Emotion positive qui apparaît dans une condition de satisfaction intense. Elle est caractérisée par les <i>coins des yeux élevés</i>, les <i>coins des lèvres tirés vers le haut</i>, la <i>bouche ouverte légèrement</i> et l'<i>apparition des rides sur le bas du nez</i>, les <i>cotés des lèvres</i> et l'<i>extérieur des coins des yeux</i>.

Tableau 2.3 Description des six expressions d'après [LAL 05] et [ESS 97].

2.6. Système d'analyse des expressions faciales

La reconnaissance des expressions faciales consiste à classer des déformations faciales en classes abstraites, basées sur l'information visuelle.

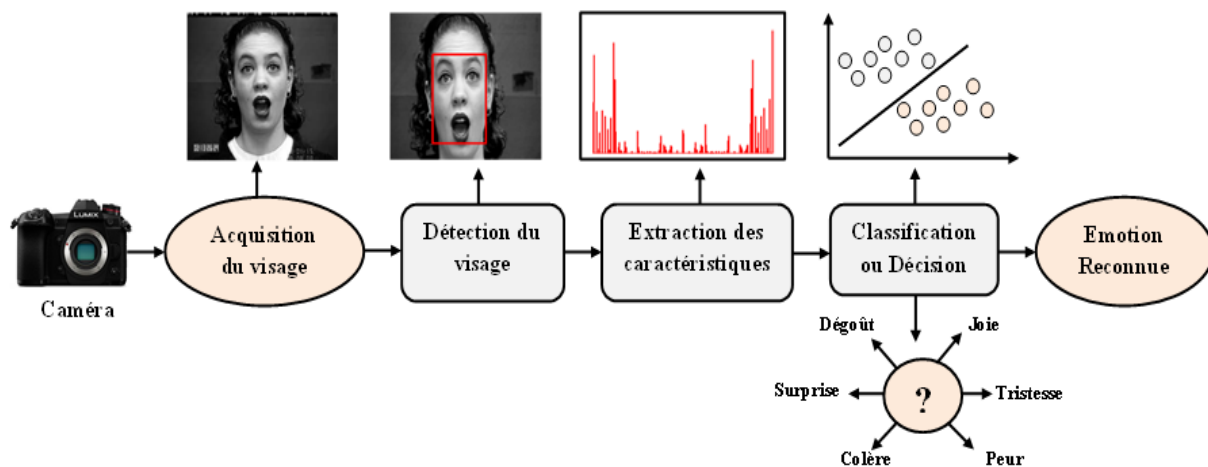


Figure 2.5 Architecture générale d'un système de reconnaissance des expressions faciales

Un système d'analyse automatique des expressions faciales s'effectue généralement selon les trois étapes suivantes : La première étape consiste à *détecter la présence d'un visage* dans l'image et d'en *localiser sa position* ce qui permet de limiter la zone de recherche. La deuxième étape consiste à *extraire les informations* nécessaires à la description de l'expression faciale. A la fin, en se basant sur ces informations, l'image sera *affectée à une catégorie d'expressions* à l'aide d'un classifieur. La figure 2.5 ci-dessus présente les différentes étapes d'un système de reconnaissance des expressions faciales.

2.6.1. Acquisition du visage

La première étape de tout système d'analyse d'expression faciale concerne l'acquisition du visage. Le système d'acquisition est généralement équipé d'un *appareil photo* ou d'une *caméra* qui permet d'acquérir une image 2D du visage. L'image acquise sera traitée pour détecter et localiser la région du visage et ses limites. La précision de l'acquisition et l'enregistrement du visage est particulièrement importante dans des conditions réalistes (non contrôlés), où la présence du visage dans une scène et sa localisation globale ne sont pas connues a priori. Un système complet d'acquisition et de localisation du visage devrait faire face aux défis décrits ci-dessous :

- **Variation de pose** : Les traits du visage, y compris les yeux, le nez et la bouche, peuvent être partiellement ou complètement invisibles en raison de la position relative du visage ou de la caméra.
- **Occlusion** : Un visage peut apparaître à moitié dans une image ou parfois masqués partiellement par un objet dans la scène, ou par le port d'accessoire tels que des lunettes, une écharpe, etc. De même, le maquillage peut provoquer l'apparition de régions artificielles sur le visage ou cacher les limites faciales normales.
- **Présence ou absence des composants structurels** : La présence des composants structurels telle que la barbe, la moustache, peut modifier énormément les caractéristiques faciales telles que la forme, la couleur, ou la taille du visage. De plus, ces composants peuvent cacher les caractéristiques faciales de base causant ainsi une défaillance du système de reconnaissance.
- **Les conditions d'éclairage et d'illumination** : Les conditions d'éclairage et les changements dans les caractéristiques de la caméra affectent de manière significative la chrominance des régions du visage. Certaines caractéristiques peuvent être masquées ou combinées avec des ombres ou des brillances sur le visage provoquant ainsi une perte d'informations sur les couleurs, d'où la nécessité de faire des prétraitements de l'image de minimiser les effets d'éclairage et d'illumination.

- **Expression faciale** : L'expression faciale peut modifier d'une manière significative la géométrie ou la forme d'un visage et donc influencer la décision d'un système de d'acquisition et de détection de visages.
- **Dans le contexte de la vidéo**, il faut en plus assurer un temps de calcul compatible au temps réel (vitesse de traitement, la complexité algorithmique).

2.6.2. Détection du visage

La détection automatique du visage est une étape indispensable dans les systèmes de reconnaissance des émotions. Elle consiste à déterminer la présence ou l'absence de visages dans une image et en cas de présence à définir sa localisation. Les régions d'intérêts déterminées sont ensuite utilisées pour l'extraction des caractéristiques. Plusieurs méthodes de détection de visages ont été proposées au cours de ces dernières années. On peut les classer selon les quatre catégories décrites ci-dessous [YAN 02] :

2.6.2.1. Méthodes basées sur les connaissances acquises

Ces méthodes se basent sur la connaissance des différents éléments qui constituent un visage et des relations qui existent entre eux. Les relations définies sont traduites par un ensemble de règles. Voici quelques exemples qui peuvent être appliqué sur une image pour tenter de déterminer la présence d'un visage :

- La partie centrale d'un visage a des valeurs d'intensité lumineuse uniforme ;
- Un visage apparaît souvent avec deux yeux qui sont symétriques entre eux, un nez et une bouche ;
- Les rapports existants entre les différentes composantes faciales peuvent être représentés par leurs distances et positions relatives.

Un exemple typique de ces méthodes est l'approche introduite par Yang et Huang [YAN 94], dans laquelle les auteurs ont employé une stratégie hiérarchique (par une simple moyenne et un sous échantillonnage) pour détecter les visages (voir Figure 2.6). Dans [KOT 97], Kotropoulos et Pitas ont proposé une autre approche basée sur la projection horizontale et verticale pour déterminer des visages candidats.

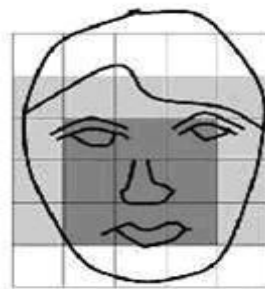


Figure 2.6 Le modèle type exploité par la méthode de Yang et Huang [YAN 94].

Bien qu'elles soient efficaces pour détecter des visages frontaux dans des images avec un arrière-plan simple, ces méthodes présentent cependant un inconvénient majeur. En effet, il est difficile de traduire des connaissances humaines en des règles bien définies. De plus, il est difficile de rendre ces méthodes capables de détecter des visages dans différentes orientations car il ne semble pas réalisable d'énumérer tous les cas possibles.

2.6.2.2. Méthodes basées sur les caractéristiques invariantes

Ces méthodes visent à trouver les caractéristiques structurelles (forme, texture, couleur de la peau, contour) qui existent même lorsque la pose, l'angle de vue, ou les conditions d'illumination varient.

Il existe principalement trois familles des méthodes basées sur les caractéristiques invariantes. La première famille utilise la propriété de la couleur de la peau pour capter des régions contenant des visages. Un exemple de cette famille on peut citer les travaux de D. Saxe et al. [SAX 96] et K. Sandeep et al. [SAN 02]. La deuxième famille vise à détecter les différentes caractéristiques faciales. Par exemple, Leung et al [LEU 95] ont développé une méthode probabiliste pour localiser un visage dans une scène encombrée qui se base sur la détection des caractéristiques locales et la comparaison de graphe des correspondances aléatoires (Random Graph Matching), et la troisième famille combine une multitude de caractéristiques de nature différentes. Comme exemple de cette famille on peut citer les travaux de Sobottka et Pitas [SOB 96] (voir figure 2.7).

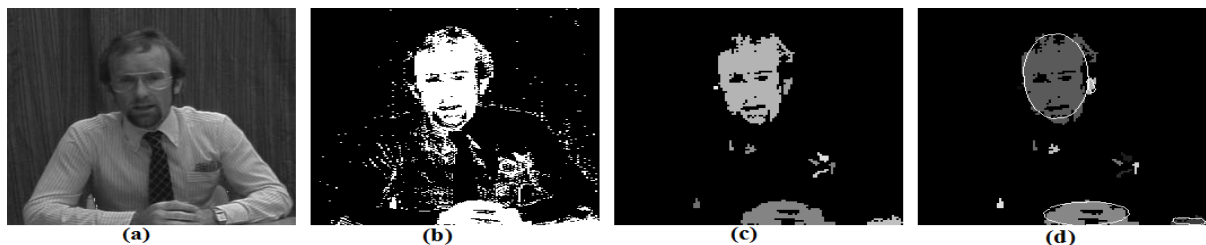


Figure 2.7 Localisation de visage : (a) image originale, (b) segmentation de couleur, (c) groupement des composants connexes, (d) détermination des régions elliptiques [SOB 96].

Le problème de ces méthodes est le manque d'efficacité si les images en entrée sont de faible luminosité ou si certaines parties d'un visage sont masquées. De plus, le fait qu'un visage soit ombragé rend l'extraction des caractéristiques faciales très difficiles.

2.6.2.3. Méthodes basées sur la mise en correspondance

L'idée principale de ces méthodes est de créer des modèles standards capables de décrire un visage ou une partie de visage. Les valeurs de corrélation entre le modèle et l'image d'entrée sont calculées et la présence du visage est déterminée en utilisant ces valeurs de corrélation.

On peut citer comme exemple d'application de ces méthodes les travaux d'A. Lanitis et al. qui utilisent des modèles de visage déformables [LAN 95a] pour détecter la présence d'un visage. Luhong Liang et al. [LIA 01] ont aussi proposé une approche qui utilise non seulement le modèle de visage mais aussi le modèle des yeux, car ils pensent que les yeux jouent un rôle très important dans la forme du visage.

Les méthodes de mise en correspondance sont faciles à implémenter. Toutefois, elles ne pouvaient pas faire face aux changements d'orientation, de forme et de taille des visages. Donc, la construction des modèles de visage s'avère très difficile. C'est pourquoi des méthodes basées sur des modèles de visage multi-résolution, multi-échelles et déformables ont été proposées par la suite.

2.6.2.4. Méthodes basées sur l'apparence globale

Le principe de ces méthodes est de considérer le problème de la détection de visage comme un problème de classification : il s'agit de classer un modèle capturé dans l'un des deux classes : classe de visages et classe de non-visages. Ces méthodes utilisent l'apprentissage automatique et les techniques d'analyse statistiques pour construire des machines capables de trouver les traits du visage pertinents et séparer les visages des non-visages. Les réseaux de neurones, les machines à vecteurs de support (SVM), les classifieurs bayésiens, les modèles de Markov cachés (HMM) sont parmi les techniques d'apprentissage automatique les plus souvent utilisées.

Les méthodes basées sur l'apparence globale présentent l'avantage de s'exécuter très rapidement mais demandent un long temps d'entraînement. Elles ont montré de bons résultats par rapport aux 3 autres types de méthodes citées précédemment. Parmi les approches appartenant à cette catégorie on peut citer la méthode basée sur les réseaux de neurones (RN) de Rowley et al. [ROW 98] (Voir figure 2.8), la méthode de Schneiderman et Kanade [SCH 98] basée sur un classifieur bayésien, ainsi que le fameux algorithme de Viola et Jones [VIO 01] fonctionnant en temps réel.

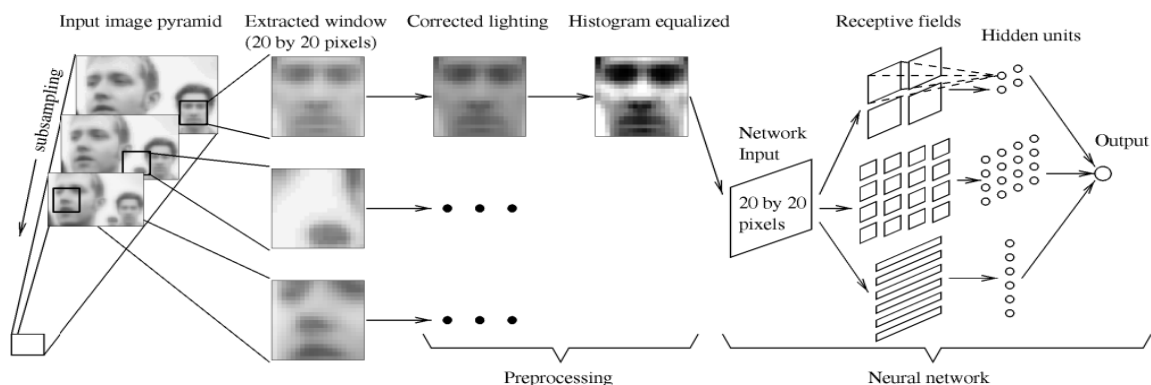


Figure 2.8 L'algorithme de base pour la détection de visage basé sur les RN [ROW 98].

2.6.3. Prétraitement

Cette étape permet d'effectuer des transformations sur les visages détectés afin d'éliminer certaines variations (telle que les rotations du visage dans le plan d'image, les variations du facteur d'échelle, les variations d'illumination, le bruit, etc) causés par la qualité des dispositifs optiques ou électroniques lors de l'acquisition d'image en entrée. Cela permet de retirer les parties inutiles et de conserver que les informations nécessaires pour préparer l'image à l'étape d'extraction de caractéristique. Il existe plusieurs types de traitement et d'amélioration de la qualité de l'image, telle que : la normalisation, l'égalisation d'histogramme, le filtre médian...etc.

2.6.4. Extraction des caractéristiques faciales

Cette étape consiste à extraire les informations pertinentes qui décrivent les expressions faciales. Ces informations doivent être suffisamment discriminantes et robustes pour que le visage d'une personne ne ressemble pas à celui d'une autre personne (variations inter-individu) et en même temps qu'il ressemble à lui-même dans d'autres conditions d'acquisition (variations intra-individu).

Un algorithme d'extraction de caractéristiques permet d'effectuer le traitement du visage dans un sous espace de travail plus simple que l'espace originale des caractéristiques. Cet espace assure une meilleure exploitation de données, et donc permettre l'utilisation, seulement, des informations utiles et non redondantes. L'efficacité de cette étape à une influence directe sur la performance du système de reconnaissance des expressions faciales.

Il existe plusieurs méthodes d'extraction des structures faciales que l'on peut regrouper en trois grandes catégories : les méthodes basées sur l'*apparence*, les méthodes basées sur les *caractéristiques géométriques* et les *méthodes hybrides* qui combinent les méthodes d'apparences et les méthodes géométriques.

2.6.4.1. Les méthodes basées sur l'apparence

Les caractéristiques d'apparence représentent les changements transitoires de la texture de la peau du visage tels que les rides sur le front, autour de la bouche, sur les côtés des yeux et bien d'autres zones qui dépendent en général de l'élasticité de la peau et de l'expression. Ces caractéristiques peuvent être extraites sur tout le visage ou sur des régions spécifiques du visage.

Plusieurs méthodes d'extraction de texture ont été utilisées dans la littérature pour la reconnaissance des émotions. Parmi celles-ci nous pouvons citer l'Analyse par Composantes Principales (ACP) [KAP 03], l'Analyse Discriminante Linéaire (LDA) [DEN 05] [SHA 20],

les filtres de Gabor [BAS 08] [DEN 05], la méthode des motifs binaires locaux (LBP) [SHA 09] [ZHA 07] et ces variantes : Local Gradient Code (LGC) [TON 14], Gradient Directionnel Pattern (GDP) [AHM 12], Local Directionnel Pattern (LDP) [JAB 10a] [JAB 10a], Local Ternary Pattern (LTP) [TAN 10],...etc. Dans ce qui suit, nous décrivons brièvement les méthodes d'extraction des caractéristiques mentionnées ci-dessus.

2.6.4.1.1. L'Analyse par Composantes Principales (ACP)

L'ACP est une méthode de réduction de dimension linéaire introduite par Pearson [PEA 01]. Elle permet de définir, à partir d'un jeu de données d'apprentissage, un sous espace permettant de simultanément conserver l'information discriminante et supprimer les informations secondaires. L'idée principale consiste à trouver une nouvelle base de l'espace des données dont tous les vecteurs sont orthogonaux entre eux.

En 1991, Turk et Pentland [TUR 91] ont proposé la méthode des visages propres ou 'Eigenfaces' qui adapte l'ACP à la problématique de la reconnaissance des visages. Cette méthode consiste à extraire les caractéristiques les plus appropriées d'une image de visage, pour l'encoder aussi efficacement que possible afin de la comparer à une base de données de modèles encodés de manière similaire. Les images résultats de cette méthode dites images fantômes en raison de leurs apparences (voir figure 2.9).



Figure 2.9 Exemple de calcul d'Eigenfaces: La première ligne présente des images de visage avec quelques expressions faciales et la deuxième visualise les eigenfaces correspondants [MAA 12].

Padgett et Cottrell [PAD 96] utilisent des gabarits d'œil et de bouche, calculés par l'ACP d'un ensemble d'apprentissage, en conjonction avec des réseaux de neurones pour reconnaître les sept expressions faciales (six expressions plus le neutre). La moyenne de taux de reconnaissance réalisée était de 86%.

Huang et Huang [HUA 97] remplacent les 44 unités d'actions (AUs) du système FACS par 10 paramètres d'actions (APs). Ils enlèvent les 10 APs en se basant sur la différence entre un visage neutre et un autre expressif et appliquent une ACP sur ces derniers pour

réduire la dimension à deux et simplifier le processus de reconnaissance. Les auteurs calculent le profil de corrélations des APs en utilisant un classificateur de distance minimale pour le traitement des six expressions universelles. Le système présente un taux de reconnaissance égale à 84.5%.

2.6.4.1.2. L'Analyse Discriminante Linéaire (ADL)

L'ADL est une méthode supervisée linéaire introduite en 1936 par Robert Fisher [FIS 36]. À la différence de l'ACP, l'ADL analyse les vecteurs propres de la matrice de dispersion tout en cherchant à minimiser les variations au sein d'une même classe (intra classes) et maximiser les variations entre des classes différentes (inter classes).

Belhumeur et al. [BEL 97] ont proposé une méthode appelée visages de Fisher ou 'Fisherfaces' qui consiste à appliquer une ACP sur l'ensemble des images de visage afin de réduire l'espace d'origine, puis une ADL afin de calculer les Fisherfaces. La réduction d'espace par l'ACP proposée par les auteurs résout le problème de singularité de la matrice calculée par l'ADL, du fait que les données sont souvent sous-représentées.

Dubuisson et al. [DUB 02] utilisent l'ACP pour la réduction de la dimension du visage et l'ADL pour générer l'espace de Fisher ou le nouvel échantillon est classé. Le classifieur de l'arbre de décision est adopté comme une mesure entre la projection des échantillons de test et chaque vecteur moyen d'estimation dans l'espace de Fisher. La Figure 2.10 montre 5 espaces de Fisher générés et appliqués au corpus d'apprentissage contenant les 6 classes des expressions faciales.



Figure 2.10 Les 5 espaces de Fisher correspondant aux 5 axes du sous espace généré par ACP et ADL [DUB 02]

2.6.4.1.3. Les filtres de Gabor

L'une des méthodes les plus utilisées pour l'extraction des textures est la méthode de Gabor. Ceci est motivé par le fait que les filtres de Gabor présentent deux propriétés intéressantes : la localisation *fréquentielle* et la sélectivité en *orientation*. Les filtres de Gabor en deux dimensions ont été introduits en biométrie par [DAU 93] pour la reconnaissance d'iris. Ils ont été depuis utilisés dans plusieurs applications telles que la reconnaissance du visage

[CHE 08], la reconnaissance des empreintes [CHI 99]...etc. Sa robustesse dans ces domaines a favorisé son application dans la reconnaissance des expressions faciales. A noter que les travaux de Hubel et Wiesel [HUB 62] ont démontré l'existence d'une ressemblance dans le fonctionnement des filtres de Gabor et celui du système visuel humain [HUB 62].

Un noyau de filtre de Gabor est le produit d'une sinusoïde de fréquence et d'orientations particulières avec une enveloppe gaussienne. Une ondelette de Gabor est définie comme suit

$$\psi_{u,v}(z) = \frac{\|k_{u,v}\|^2}{\sigma^2} e^{\left(\frac{\|k_{u,v}\|^2 \|z\|^2}{2\sigma^2}\right)} [e^{ik_{u,v}z} - e^{-\frac{\sigma^2}{2}}] \quad (2.1)$$

Où $z = (x, y)$ est le point de coordonnées (x, y) dans le plan image. Les paramètres u et v définissent l'orientation et la fréquence des noyaux de Gabor. $\|\cdot\|$ est l'opérateur norme, et σ l'écart-type de l'enveloppe gaussienne. k est le vecteur d'onde dont $k_{u,v} = k_{u,v} e^{i\phi u}$.

La représentation d'une image $I(x, y) = I(z)$ en filtre de Gabor résulte du produit de convolution de cette image avec un ensemble de noyaux de Gabor de fréquence et d'orientation différente (voir figure 2.11). La convolution de l'image I et d'un noyau de Gabor $\psi_{u,v}(z)$ est définie par :

$$G_{u,v}(z) = I(z) * \psi_{u,v}(z) \quad (2.2)$$

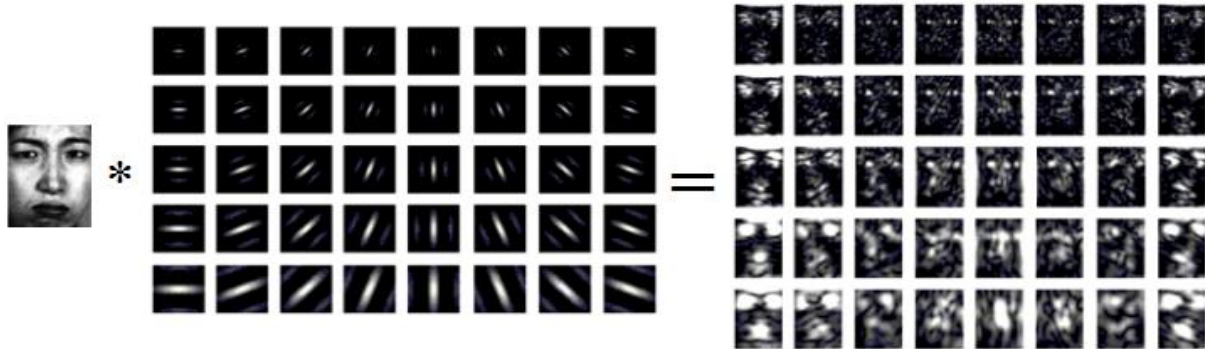


Figure 2.11 Exemple de convolution d'une image avec un ensemble de noyaux de Gabor (5fréquences et 8 orientations) [DEN 05]

Appliqués à la reconnaissance des expressions faciales, les filtres de Gabor permettent de capturer les informations de visage dans des orientations et des résolutions différentes. De plus, ils sont robustes aux changements de pose et d'illumination, aux distorsions et aux variations d'échelle.

Lyons et al [LYO 98] se basent sur l'utilisation d'un ensemble de filtres de Gabor multi-orientés et multi-résolus pour la classification des 6 expressions universelles plus l'expression neutre (figure 2.12). Un modèle de 34 points des caractéristiques faciales est manuellement initialisé dans le visage. Les coefficients de l'ondelette de Gabor sont calculés pour chaque point du maillage. Ces points sont soumis à une ACP puis analysé à l'aide

d'une ADL afin de séparer les vecteurs dans des classes ayant des attributs faciaux différents. Enfin, la classification est effectuée par la projection du vecteur d'entrée de l'image test le long du vecteur de distinction afin de l'affecter au groupe le plus proche. Les taux de reconnaissances étaient de 92% pour les sujets familiers et de 75% pour les sujets inconnus.

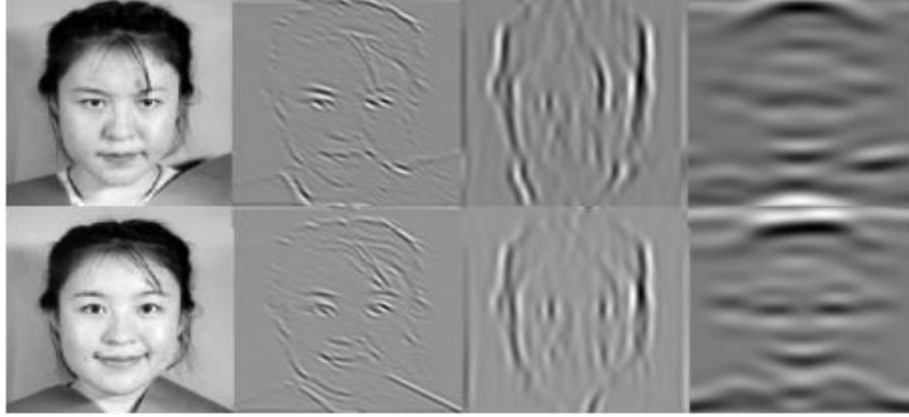


Figure 2.12 Exemples de réponses d'un banc de filtres de Gabor 2D (avec trois valeurs d'orientations et longueurs d'ondes différentes), appliqué à deux images d'expressions faciales [LYO 98].

Bashyal et al [BAS 08] appliquent 18 filtres de Gabor sur un ensemble de 34 points positionnés sur les yeux, la bouche, les sourcils et sur le contour du visage. Les points sont localisés par une interface utilisateurs. Une ACP est ensuite appliquée pour réduire la taille des descripteurs obtenus. La quantification vectorielle est adoptée pour la classification des 7 expressions faciales. Le taux de reconnaissance obtenu est de 90.2% pour la base JAFEE.

2.6.4.1.4. Les motifs binaires locaux (Local Binary Pattern : LBP)

L'opérateur LBP a été initialement proposé par Ojala et al. [OJA 96] afin de caractériser les textures présentes dans des images. Cet opérateur propose de décrire chaque pixel par un code binaire calculé à partir des valeurs relatives des niveaux de gris des 8 pixels voisins. Le niveau de gris du pixel central (G_c) est comparé à ceux de ses voisins (G_p). Si la valeur du niveau de gris du pixel voisin est supérieure ou égale à celle du pixel central, on lui attribue la valeur de 1, sinon 0. La valeur du pixel central LBP est définie par la formule suivante :

$$LBP_{(x_c, y_c)} = \sum_{p=0}^{P-1} S(G_p - G_c) 2^p \quad (2.3)$$

$$S(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{autrement} \end{cases} \quad (2.4)$$

Le code LBP du pixel courant est alors produit en concaténant ces 8 valeurs pour former un code binaire. Ce code est ainsi converti en code décimal. La figure 2.13 donne un exemple de traitement de l'opérateur LBP. On obtient donc, comme pour une image en

niveaux de gris, une matrice des valeurs LBP contenant des valeurs d'intensité comprises entre 0 et 255.

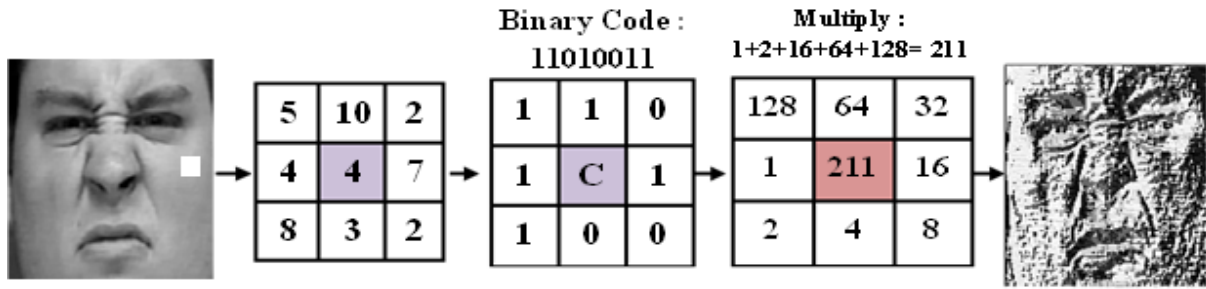


Figure 2.13 Exemple de calcul du code LBP original pour un pixel d'une image

Le descripteur LBP original a été étendu ultérieurement en utilisant différentes tailles de voisinage pour intégrer des changements d'échelle [OJA 02]. Dans ce cas, un cercle de rayon R autour du pixel central est considéré. Les valeurs des P points sur le bord de ce cercle sont prises et comparées avec la valeur du pixel central. Comme ces P points ne tombent pas nécessairement au centre d'un pixel de d'image, leurs valeurs sont obtenues par interpolation bilinéaire selon l'équation ci-dessous.

$$(x_{pi}, y_{pi}) = \left(x_p + R \cos\left(\frac{i}{v} \times 2\pi\right), y_p - R \sin\left(\frac{i}{v} \times 2\pi\right) \right) \quad (2.5)$$

La notation LBP (P, R) est adoptée pour définir les LBP construits pour un rayon R et un voisinage de taille P. La figure 2.14 suivante montre quelques exemples de différents points de voisinage P et de rayon R.

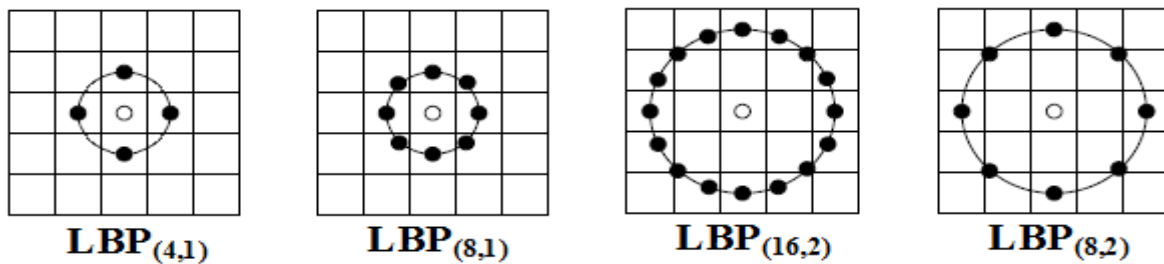


Figure 2.14 Extension de l'opérateur LBP basique.

En raison de sa simplicité de calcul, de sa tolérance aux changements d'illumination et de son pouvoir discriminant, le LBP a été utilisé pour résoudre différents problèmes d'analyse de textures [LIU 12] et de visages comme la détection de visages et la reconnaissance des expressions faciales [SHA 09]...etc. Par exemple, Shan et al [SHA 09] utilisent les motifs binaires locaux (LBP) pour extraire la texture de visage et représenter les expressions faciales. Adaboost a été utilisé pour sélectionner les descripteurs les plus discriminants. Cette approche a montré sa robustesse face aux changements de résolutions dans la base

PETS. Les auteurs ont montré aussi que les caractéristiques LBP sont plus discriminantes et efficaces que celles de Gabor.

Liao et al [LIA 06] ont quant à eux comparé les motifs binaires et les filtres de Gabor sur la base JAFEE. Ils montrent que les motifs binaires locaux (LBP) se sont avérés plus performants que les filtres de Gabor dans le cas de visages à faibles résolutions.

2.6.4.1.5. Local Gradient Code (LGC)

Bien que l'opérateur LBP considère uniquement la relation entre le pixel central et ses pixels voisins, l'opérateur LGC compare les valeurs de niveaux de gris des pixels voisins (autour du pixel central g_c) entre eux, dans les directions horizontales, verticale et diagonale suivant le modèle 3×3 présenté dans la figure 2.15. Le code binaire ainsi obtenu est converti en code décimal [TON 14].

g_1	g_2	g_3
g_4	g_c	g_5
g_6	g_7	g_8

Figure 2.15 Modèle de voisinage 3*3 de l'opérateur LGC

L'opérateur LGC peut être exprimé par la relation suivante:

$$LGC_{(x_c, y_c)} = S(g_1 - g_3)2^7 + S(g_4 - g_5)2^6 + S(g_6 - g_8)2^5 + S(g_1 - g_6)2^4 + \\ S(g_2 - g_7)2^3 + S(g_3 - g_8)2^2 + S(g_1 - g_8)2^1 + S(g_3 - g_6)2^0 \quad (2.6)$$

$$S(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{autrement} \end{cases} \quad (2.7)$$

La figure 2.16 ci-dessous présente un exemple de calcul de l'opérateur LGC :

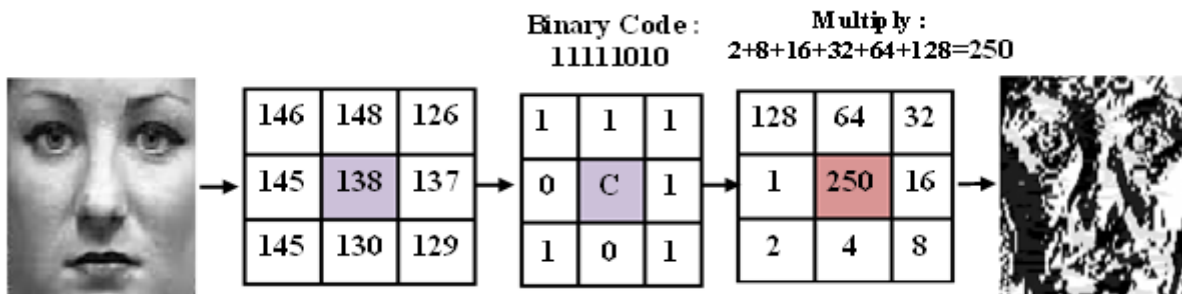


Figure 2.16 Exemple de calcul de l'opérateur LGC

Tong et al. [TON 14] proposent une variante de l'opérateur LGC, appelée LGC-HD (LGC-based Horizontal and Diagonal Gradient Prior Principal). Cet opérateur est capable de

réduire le temps d'exécution et d'augmenter le taux de reconnaissance par rapport au LGC original. L'opérateur LGC-HD est défini comme suit :

$$LGC_{(x_c, y_c)} = S(g_1 - g_3)2^4 + S(g_4 - g_5)2^3 + S(g_6 - g_8)2^2 + S(g_1 - g_8)2^1 + S(g_3 - g_6)2^0 \quad (2.8)$$

Al-Sumaidae et al [SUM 15] ont proposé un système de reconnaissance des expressions faciales basé sur nouveau descripteur connu sous le nom LGGC-HD (Local Gabor Gradient Code-Horizontal Diagonal). Ce nouveau descripteur est le résultat de fusionnement de l'opérateur LGC-HD et les filtres de Gabor. Les auteurs utilisent la mesure de similarité Chi-square pour classifier les 7 expressions faciales. Le taux de reconnaissance est estimé à 93,33%.

2.6.4.1.6. Gradient Directional Pattern (GDP)

L'opérateur GDP encode la texture d'une région locale d'une image en calculant les angles de direction de gradient (Gradient Direction) des pixels à l'aide de l'équation suivante [AHM 12] :

$$GD(x, y) = \tan^{-1}(G_y/G_x) \quad (2.9)$$

$GD(x, y)$ représente l'angle de direction du gradient du pixel (x, y) ; G_x et G_y sont les deux éléments du vecteur de gradient que l'on peut obtenir en appliquant l'opérateur de Sobel sur l'image source $I(x, y)$. L'opérateur de Sobel convolue l'image avec un noyau horizontal et vertical pour obtenir les valeurs de G_x et G_y respectivement, comme le montre la figure 2.17 ci-dessous.

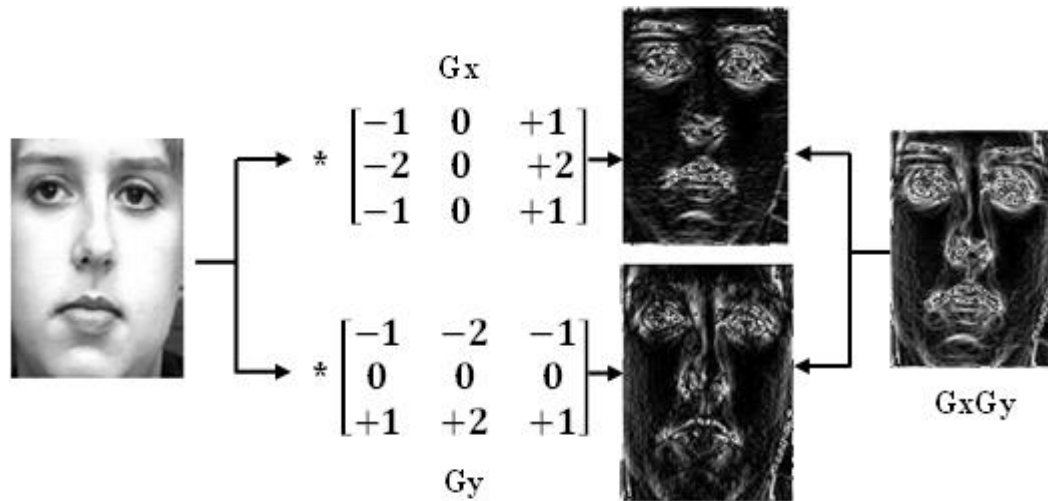


Figure 2.17 Exemple de calcul du masque de Sobel

Après cela, l'opérateur GDP sélectionne un voisinage 3*3 autour de chaque pixel de l'image et quantifie les angles de direction du gradient voisin par rapport à l'angle de direction du pixel central en utilisant un seuil t . Les voisins avec un angle de gradient $\pm t$ autour de l'angle central sont quantifiés à 1, et les autres sont quantifiés à 0. Le code

binaire ainsi obtenu est converti en valeur décimal. Cette valeur sera affectée au pixel central.

$$GDP(x_c, y_c) = \sum_{p=0}^{P-1} S(GD_p - GD_c) 2^p \quad (2.10)$$

$$S(GD_p, GD_c) = \begin{cases} 1 & GD_c - t \leq GD_p \leq GD_c + t \\ 0 & \text{otherwise} \end{cases} \quad (2.11)$$

GD_c représente l'angle de direction du gradient du pixel central (x_c, y_c), GD_p représente les angles de direction de ses voisins et t est un seuil spécifié par l'utilisateur. La figure 2.18 illustre un exemple de calcul de l'opérateur GDP.

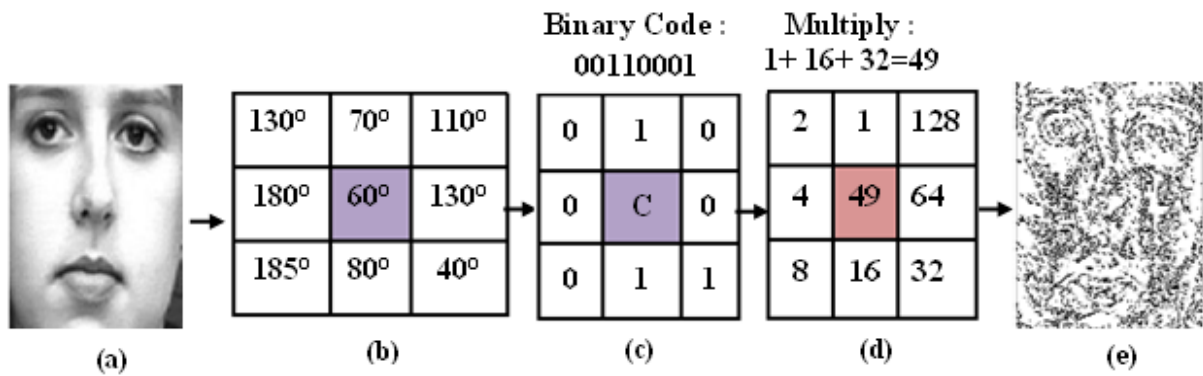


Figure 2.18 Exemple de calcul de l'opérateur GDP : (a) Image en niveau de gris, (b) Angles de direction du gradient dans voisinage 3*3, (c) code binaire, (d) code GDP, (e) Image GDP.

2.6.4.1.7. Local Ternary Pattern (LTP)

L'opérateur LBP est très sensible au bruit aléatoire. Pour surmonter ce problème, Tan et Triggs [TAN 10] ont proposé un nouveau descripteur appelé Local Ternary Pattern (LTP) qui étend le code binaire LBP (0,1) à un code ternaire à 3 valeurs (-1, 0, 1) en utilisant un seuil t . Les valeurs de niveau de gris dans une zone de largeur $\pm t$ autour du pixel central I_c sont quantifiées à 0, celles inférieures à $I_c - t$ sont quantifiées à -1 et celles supérieures à $I_c + t$ sont quantifiées à +1.

$$S(I_p, I_c, t) = \begin{cases} +1 & I_p \geq I_c + t \\ 0 & |I_p - I_c| < t \\ -1 & I_p \leq I_c - t \end{cases} \quad (2.12)$$

Où I_c est la valeur de niveau de gris du pixel central, I_p est la valeur de niveau de gris des pixels voisins et t est un seuil spécifié par l'utilisateur (ici $t=10$).

Tan et Triggs [TAN 10] divisent le code ternaire LTP en deux codes binaires LBP traités séparément et rassemblés ensuite lors de la phase de comparaison. Tout d'abord, les auteurs ont créé le motif supérieur (Upper pattern) en convertissant les codes négatifs en zéro (convertir les -1 en 0). Ensuite, ils ont créé le motif inférieur (Lower Pattern) en réglant les

valeurs positives à zéro (rendre les 1 des 0) tout en convertissant les valeurs négatives en valeurs positives (convertir les -1 en 1). La figure 2.19 représente le principe de calcul de l'opérateur LTP.

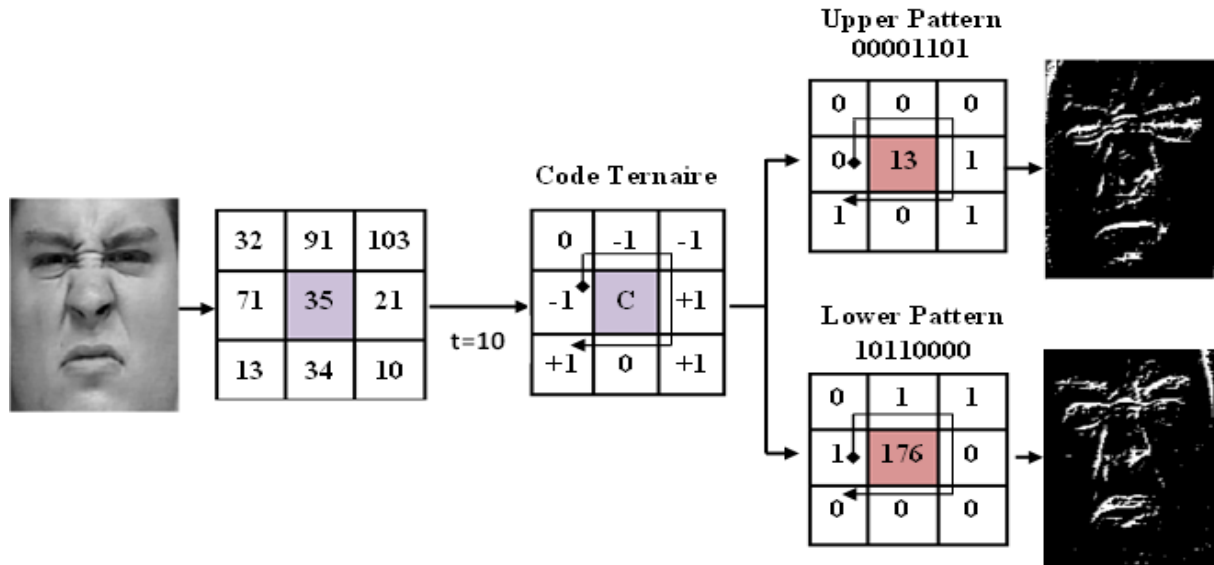


Figure 2.19 Exemple de calcul de l'opérateur LTP

2.6.4.1.8. Local Directional Pattern (LDP)

Afin d'obtenir de meilleures performances en présence de variation d'éclairage et de bruit, le descripteur Local Directional Pattern (LDP) a été développé par Jabid al [JAB 10a] [JAB 10b]. Dans cet opérateur, huit directions du masque de Kirsch (M0-M7) de taille 3x3 sont convolués avec des régions d'image de taille 3x3 pour obtenir un ensemble de 8 valeurs de masque. Les masques de Kirsch (M_i , $i = 0, \dots, 7$) sont illustrés dans la figure 2.20.

$$\begin{array}{cccc}
 \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} & \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} & \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} & \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \\
 \text{Est } M_0 & \text{Nord Est } M_1 & \text{Nord } M_2 & \text{Nord Ouest } M_3 \\
 \\
 \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} & \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} & \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} & \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix} \\
 \text{Ouest } M_4 & \text{Sud Ouest } M_5 & \text{Sud } M_6 & \text{Sud Est } M_7
 \end{array}$$

Figure 2.20 Les huit masques de Kirsch

Appliquant les huit masques, les auteurs ont obtenu huit valeurs de réponse de contour (Edge Response) $ER_0, ER_1, \dots, ER_7$, chacune représentant la signification de contour dans sa direction respective. Les valeurs de réponse ne sont pas tous de la même importance dans

toutes les directions. La présence d'un coin ou d'un bord provoque des valeurs de réponse élevées dans certaines directions.

Afin de générer le code LDP, les k directions les plus importantes doivent être calculés. Ensuite, les k valeurs supérieures (top k bits) de $|ER_j|$ sont mis à 1, les autres $(8 - k)$ bits restant de $|ER_j|$ sont définis par 0. Le code LDP pour chaque pixel est calculé à l'aide des formules ci-dessous :

$$LDP_k = \sum_{i=0}^7 b_i (ER_i - ER_k) 2^i \quad (2.13)$$

$$b_i(a) = \begin{cases} 1 & \text{if } a \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.14)$$

La figure 2.21 représente un exemple de calcul du code LDP avec $k=3$.

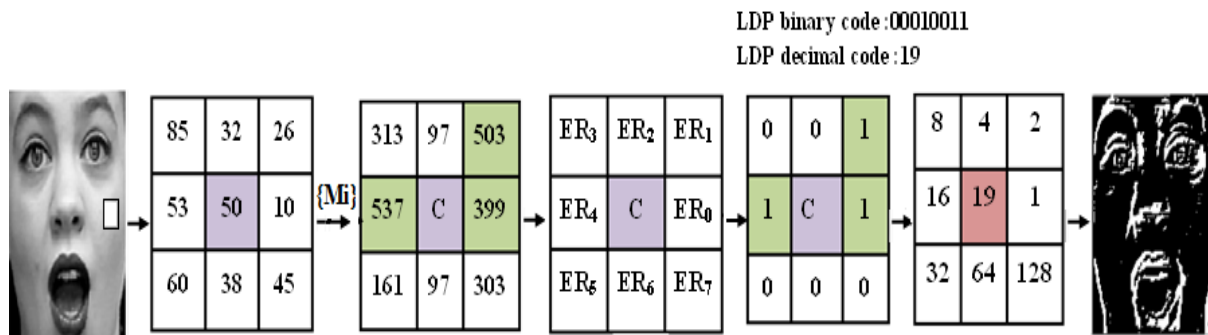


Figure 2.21 Exemple de calcul du code LDP avec $k=3$.

2.6.4.2. Les méthodes basées sur les caractéristiques géométriques

Par opposition aux caractéristiques d'apparence, qui se basent sur l'analyse des textures des visages pour décrire les expressions faciales, les caractéristiques géométriques représentent des changements de la forme des traits permanents du visage comme la bouche, le nez, les yeux, les sourcils et décrivent les relations spatiales et/ou spatio-temporelles qui existent entre eux. L'intérêt des caractéristiques géométriques vient du fait que les positions relatives de ces parties du visage, ainsi que leurs mouvements, sont utilisés pour former un vecteur de caractéristiques représentant la géométrie du visage.

Les méthodes géométriques présentent le mouvement des muscles soit par la détection de la *forme des composantes faciales*, soit par le mouvement des *points caractéristiques*, soit par l'utilisation des *modèles* qui détectent directement les mouvements des muscles.

2.6.4.2.1. Caractéristiques géométriques basées sur la forme des composantes faciales

Plusieurs techniques ont été présentées pour détecter la forme, la position et l'état des composants faciaux. Tian et al [TIA 01] ont proposé un système d'analyse automatique des expressions faciales par reconnaissance d'unité d'action (AU) (6 unités d'actions de la partie supérieure du visage et 10 AU de la partie inférieure). Ce système est basé sur la détection des

positions et des formes des traits permanents et transitoires du visage. Les sourcils et le haut des joues sont modélisés par trois points avec deux segments connectés, la bouche et les yeux sont représentés par un modèle multi états : état ouvert, état semi-ouvert et état fermé. Le modèle de l'œil ouvert correspond à deux paraboles et un cercle pour l'iris, par contre l'œil fermé est modélisé par un segment liant les deux coins de l'œil (voir figure 2.22). La classification est effectuée par les réseaux de neurones. La reconnaissance de l'expression faciale finale est réalisée par la combinaison des unités d'actions des deux parties supérieure et inférieure du visage.

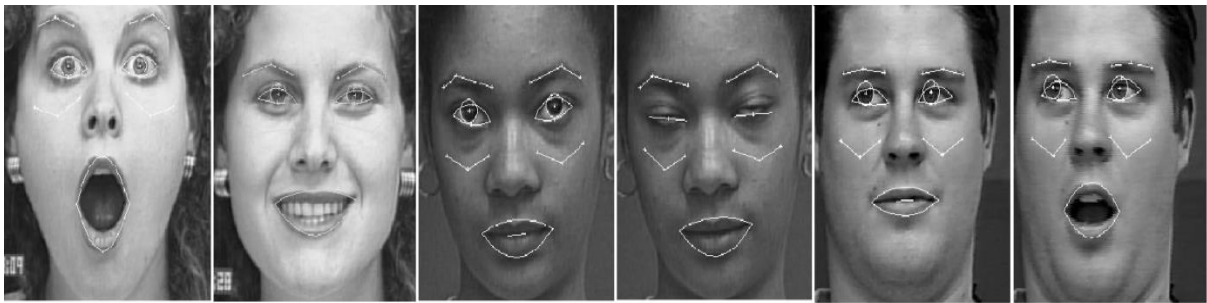


Figure 2.22 Exemples des résultats d'extraction des traits permanents proposés par [TIA 01]

Bottino et al [BOT 02] ont adopté une approche similaire à l'approche proposée par Tian et al. pour détecter les traits permanents du visage. Ils utilisent deux segments de droite pour les sourcils, et des courbes paramétriques pour la bouche et les yeux (voir figure 2.23). La détection est basée sur l'étiquetage manuel des points caractéristiques dans la première image. Dans le reste de la séquence, la localisation devient automatique en utilisant l'algorithme de Kanade [LUC 81].



Figure 2.23 Exemples de la reconstruction finale des traits permanents proposés par Bottino

Hammal et al [HAM 05] ont proposé une approche basée sur l'utilisation des modèles paramétriques pour la segmentation et la détection des traits permanents faciaux. Les contours supérieurs des yeux et les sourcils sont représentés par des courbes de Bézier contrôlées par trois points, le contour du bas des yeux est défini par un contour parabolique contrôlé également par trois points. La bouche est modélisée par quatre courbes cubiques (voir figure 2.24).

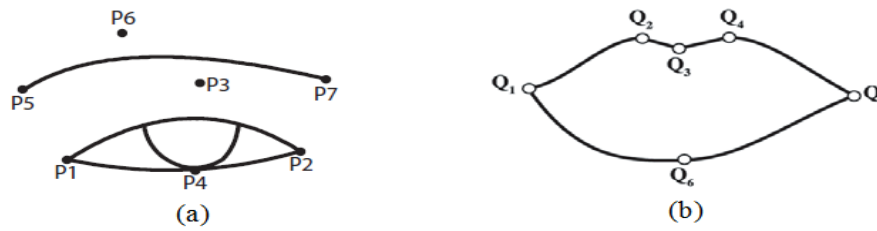


Figure 2.24 Modèles paramétriques pour la représentation des traits du visage : **(a)** Modèle paramétrique des yeux et des sourcils, **(b)** Modèle paramétrique de la bouche.

La théorie des croyances transférables est utilisée pour classer les changements faciaux en une des quatre expressions (l'expression neutre, la joie, le dégoût et la surprise) ou une union de deux expressions. Le cas de doute entre les expressions étant considéré, il permet l'ajout de caractéristiques transitoires telles que les rides qui apparaissent au-dessus du nez dans le système (voir figure 2.25). L'approche proposée est flexible et robuste aux conditions d'éclairage, à l'origine ethnique, au port de lunettes et aux déformations des traits faciaux qui peuvent survenir sur le visage en cas d'expression faciale.



Figure 2.25 Résultats de segmentation des yeux, sourcils et bouche proposés par [HAM 06]

2.6.4.2.2. Caractéristiques géométriques basées sur les points caractéristiques

Les points caractéristiques décrivent les expressions faciales soit par leurs mouvements, soit par la variation des distances entre eux. Le mouvement des points caractéristiques nécessite deux étapes à savoir l'extraction de leurs positions et leur suivi. Plusieurs techniques basées sur le déplacement d'un ensemble de points caractéristiques pour le codage des expressions faciales sont proposées ces dernières années. Pantic et al [PAN 04] proposent un modèle géométrique basé sur 32 points caractéristiques définis sur les deux vues frontale et latérale (voir figure 2.26). Les modèles géométriques sont extraits après application d'une méthode de calcul de contours sur les traits permanents du visage. La différence est ensuite calculée entre le modèle détecté et le modèle correspondant du visage à l'état neutre. Basé sur les connaissances acquises de FACS, les règles de production classifient le modèle détecté dans la classe AUs appropriée. Le taux de reconnaissance réalisé était de 86%.

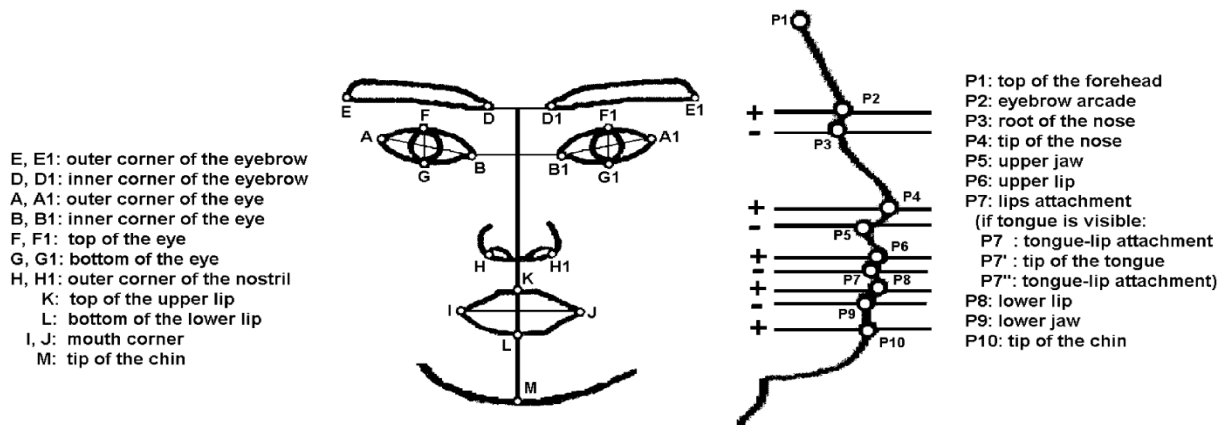


Figure 2.26 Modèle géométrique des points caractéristiques proposés dans [PAN 04]

Abdet et al [ABD 11] ont proposé un modèle anthropométrique composé de 38 points caractéristiques qui sont divisés en 18 points statiques (fixes) et 20 points dynamiques (mobiles) (voir figure 2.27 (a)). Les points statiques sont des points qui ne bougent pas pendant la production d'une expression faciale, ils sont localisés sur le contour du visage et sur les zones invariantes par l'expression (voir figure 2.27 (b)), alors que les points dynamiques subissent un mouvement selon le type de l'expression. Ils sont localisés sur les zones variables liées à l'expression notamment les traits du visage (voir figure 2.27 (c)). Les distances entre les points mobiles et les points fixes sont calculées et normalisées par les distances de l'image neutre. Chaque distance représente le mouvement d'un muscle, et chaque muscle est représenté par deux points : un point statique et un point dynamique (voir figure 2.27 (d)). Le classifieur Séparateur à Vaste Marge (SVM) est adopté pour classifier les visages dans l'une des 7 expressions faciales. Le taux de reconnaissance est de 97.5% pour la base FEEDTUM et 95.8% pour la base Cohn-Kanade.

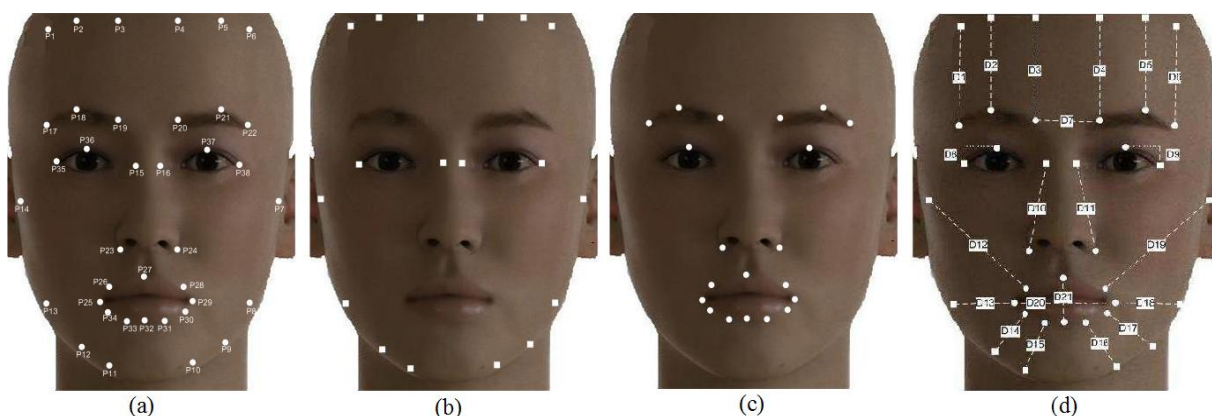


Figure 2.27 Approche proposée par [ABD 11] : (a) Modèle anthropométrique, (b) Points statiques, (c) Points dynamiques, (d) Distances utilisées pour le codage des expressions faciales

2.6.4.2.3. Caractéristiques géométriques basées sur les modèles

Plusieurs systèmes de reconnaissance des expressions faciales ont employé des techniques basées sur des modèles. Ces modèles codent souvent les déformations des muscles et les changements des états des traits faciaux soit par le mouvement des nœuds placés sur les caractéristiques du visage [KOS 07], soit en le transformant en une information 3D, qui renseigne sur l'intensité du mouvement [BEN 14]. A noter que la définition et l'initialisation de ces modèles se fait généralement sur un visage neutre.

Huang et Huang [HUA 97] calculent 10 paramètres d'actions (APs) en se basant sur la différence entre les paramètres du modèle utilisé dans un visage neutre et les paramètres du modèle utilisé dans un visage expressif étudié pour les mêmes personnes (voir figure 2.28). Les deux premières composantes de l'ACP sont utilisées pour représenter les variations des (APs). Le classificateur de la distance minimale est adopté pour classifier ces composantes dans l'une des six classes des expressions faciales. Le taux de reconnaissance réalisé est de 84,5%.

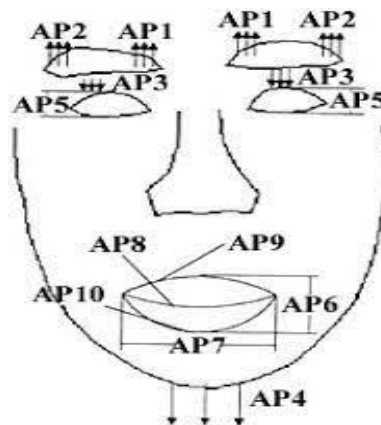


Figure 2.28 Modèle des paramètres d'action APs [HUA 97].

Benli et al [BEN 14] utilisent un modèle facial appelé HIGEM (High polygon generic face model), qui comporte 612 nœuds et 1128 polygones. Ce modèle inclut 18 muscles faciaux et chaque muscle est représenté par un point d'insertion (sur la peau) et un point d'attache (sur le crâne) (voir la figure 2.29). En se basant sur ces muscles, les régions d'intérêt sont définies, permettant ainsi l'initialisation des points caractéristiques dans l'image du visage neutre. Ces points sont suivis par les flux optiques et leur déplacement au cours de la séquence permet l'estimation de la pose de tête ainsi que le déplacement des sommets du polygone. L'estimation de la déformation du visage est également résolue par un système basé sur les contraintes de l'anatomie faciale permettant d'obtenir des niveaux d'activation musculaire. Une sélection séquentielle de l'ensemble des muscles les plus descriptifs est appliquée. Le taux de reconnaissance est estimé à 77,6% sur une base d'expressions posées.

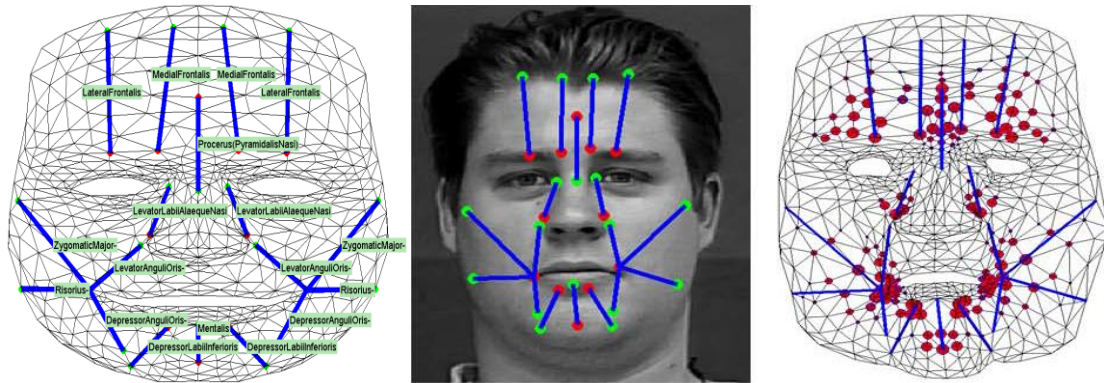


Figure 2.29 Modèle de détection des déformations faciales HIGEM [BEN 14]

2.6.4.3. Les méthodes hybrides

Les méthodes hybrides combinent le changement de la forme des caractéristiques géométriques avec la variation de la texture des caractéristiques d'apparence. En effet, la forme et la texture du visage sont fortement liées et la présence des deux informations est indispensable pour une bonne reconnaissance des expressions faciales.

Les méthodes représentant l'apparence seulement sont souvent influencées par un manque de représentation des changements des traits du visage, tandis que, les méthodes représentant uniquement la forme et les mouvements géométriques ignorent des informations telles que les rides et les sillons qui peuvent être des caractéristiques indispensables à la différenciation entre les expressions faciales. Ainsi, les méthodes hybrides peuvent constituer une approche efficace pour représenter à la fois l'information d'apparence et de forme ce qui permet de réduire la complexité des classifieurs et améliorer leur capacité de généralisation.

Le fusionnement des approches géométriques et d'apparences peut être effectué par l'utilisation des modèles de visage comme les modèles à forme active (*Active Shape Model (ASM)*) [COO 92] et les modèles à apparence active (*Active Appearance Model (AAM)*) [COO 01].

Les modèles à forme active (ASM) ont été introduits par Cootes et Taylor [COO 92] afin de retrouver automatiquement les points caractéristiques du visage en utilisant conjointement les contraintes de forme et les contraintes d'apparence. Le principe de l'ASM est de faire correspondre le modèle de forme à une nouvelle image, en trouvant la transformation permettant d'optimiser la correspondance entre ce modèle et cette nouvelle image. Shbib et Zhou [SHB 15] ont développé un système d'analyse des expressions faciales basé sur l'utilisation des ASM. Après la détection des visages par le classificateur Adaboost, un ensemble des points caractéristiques du visage ont été extraits en appliquant la technique

d'ajustement ASM. Le déplacement géométrique entre les coordonnées des points caractéristiques de l'ASM projetées et la forme moyenne de l'ASM ont été utilisés pour évaluer l'expression faciale. Le classificateur Séparateur à Vaste Marge (SVM) est adopté pour classifier les visages dans l'une des 6 expressions faciales. Le taux de reconnaissance est estimé à 93 %. Chang et al [CHA 06] ont aussi proposé une approche similaire à l'approche proposée par Shbib et Zhou [SHB 15]. Un ASM défini par 58 points faciaux est utilisé pour reconnaître les six expressions faciales (voir figure 2.30).

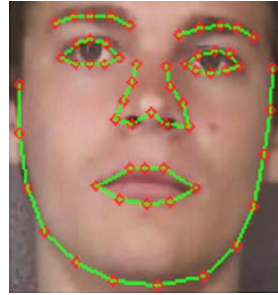


Figure 2.30 Le modèle à forme active proposé par [CHA 06]

Les modèles d'apparence active (AAM) sont des modèles performants, qui considèrent à la fois les variations de forme et de texture dans l'alignement et l'extraction des caractéristiques faciales. Le principe des AAMs est de générer, par apprentissage à partir d'exemples, un modèle statistique caractérisant l'apparence et la forme des visages, et d'effectuer une mise en correspondance entre ce modèle et l'image à traiter afin de produire un masque de point sur le visage (Voir figure 2.31). Davoine et al [DAV 04] proposent une solution pour reconnaître les six expressions faciales d'un visage par discrimination linéaire. La solution se résume en un modèle MAA hiérarchique basé sur l'utilisation du modèle MAA initialement proposé par Cootes et al [COO 01] et des filtres de Gabor. Pour la classification, les auteurs utilisent la distance de Mahalanobis comme un critère de similarité entre le vecteur MAA de l'image de test et le vecteur moyen dans l'espace de Fisher, par la suite le visage testé sera assigné à la classe qui a la moyenne la plus proche.



Figure 2.31 Exemple des résultats d'application du modèle MAA proposé par [DAV 04]

2.6.5. Classification des expressions faciales

Après avoir détecté le visage et extrait les informations pertinentes, l'étape suivante consiste à identifier l'expression faciale correspondante. Ce traitement est généralement effectué à l'aide d'un classifieur. Un classifieur est un système qui regroupe les données similaires dans une même classe. Il est capable de faire la correspondance entre les paramètres calculés et les expressions faciales. Les classifieurs les plus couramment utilisés sont les réseaux de neurones artificiels (ANN : Artificiel Neural Network), les machines à vecteurs de support (Support Vector Machine, SVM), les K-plus proche voisin (K Nearest Neighbor, KNN), les modèles de Markov cachés (Hidden Markov Model, HMM),... etc.

2.6.5.1. Les réseaux de neurones artificiels (ANN)

Les réseaux de neurones artificiels sont des systèmes informatiques s'inspirant du fonctionnement du cerveau humain pour apprendre. Ils sont constitués d'un ensemble de neurones (noeuds) connectés entre eux par des liens qui permettent de propager les signaux de neurone à neurone.

Un réseau neuronal, neuro mimétique ou connexionniste est un graphe orienté et pondéré composé d'objets élémentaires appelés les *neurones formels*. Grâce à leur capacité d'apprentissage, les réseaux neuronaux permettent de découvrir des relations complexes non-linéaires entre un grand nombre de variables, sans intervention externe.

De façon très réductrice, un neurone biologique est le composant cellulaire de base du cerveau qui se caractérise par (voir figure 2.32):

- Les *dendrites* ou entrée de neurone, qui sont les récepteurs principaux du neurone, captant les signaux qui lui parviennent ;
- Les *axones* ou sortie de neurone, qui permet de transmettre les signaux émis par le corps cellulaire aux autres neurones ;
- Le *corps cellulaire* ou le noyau, qui active les sorties en fonction des stimulations en entrée;
- Les *synapses* ou les points de connexion, qui permettent aux neurones de communiquer avec les autres neurones via les axones et les dendrites.

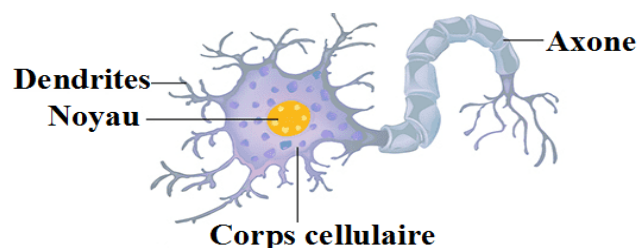


Figure 2.32 Neurone biologique.

Le neurone formel est un modèle mathématique simplifié du neurone biologique, il présente un certain nombre d'entrées, un corps traitant les entrées, et un axone véhiculant la réponse du neurone.

La Figure 2.33 montre la structure d'un neurone artificiel. Chaque neurone reçoit un nombre variable d'entrées en provenance des neurones amont. A chacune de ces entrées est associé un poids W_i représentatif de l'entrée X_i .

Chaque neurone élémentaire est doté d'une fonction de transfert (fonction d'activation) qui donne une sortie unique Y , et se ramifie ensuite pour alimenter un nombre variable de neurones avals :

$$Z = \sum_{i=1}^n w_i \cdot x_i \quad \text{Et} \quad Y = f(Z) = \begin{cases} 1, & \sum_{i=1}^n w_i \cdot x_i > 0 \\ -1, & \text{sinon} \end{cases} \quad (2.15)$$

Où W_i désigne le poids de la connexion liant le neurone j à l'entrée i

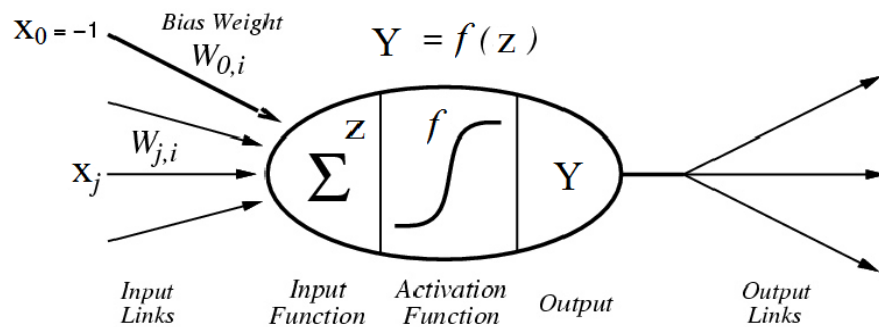


Figure 2.33 Neurone artificiel.

Il existe de nombreuses formes possibles pour la fonction d'activation. Les principaux types sont présentés dans la figure 2.34 ci-dessous :

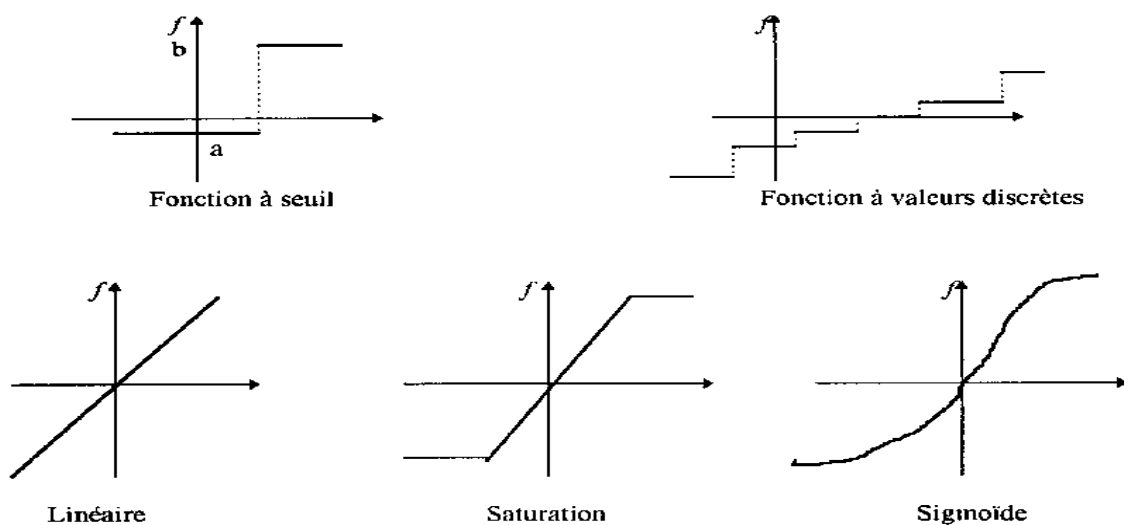


Figure 2.34 Différents types de fonctions d'activation pour le neurone artificiel.

Le perceptron multicouche est un classifieur linéaire de type réseau neuronal formel organisé en plusieurs couches au sein desquelles une information circule de la couche d'entrée vers la couche de sortie uniquement (voir figure 2.35). Chaque couche est constituée d'un nombre variable de neurones, les neurones de la couche de sortie correspondant toujours aux sorties du système.

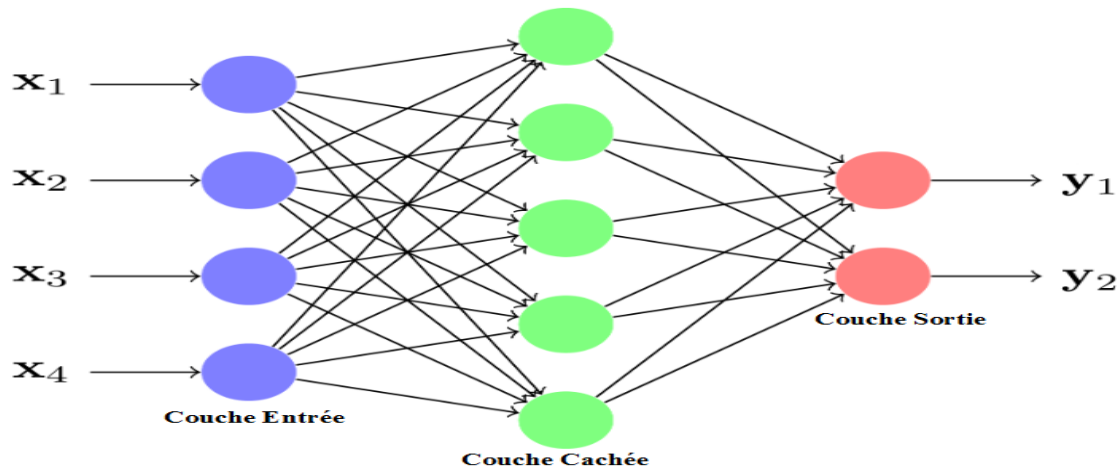


Figure 2.35 Architecture d'un perceptron multicouche

La phase d'apprentissage des MLP consiste à adapter les poids des connexions en fonction des erreurs de prédiction constatées à chaque classification d'une nouvelle instance. Plusieurs méthodes sont adoptées pour entraîner les réseaux MLP parmi lesquelles on peut citer : Rétro-propagation du gradient de l'erreur, règle d'apprentissage de Hebb, apprentissage de Rosenblatt, apprentissage de Widrow-Hoff...etc.

La rétro-propagation du gradient (back-propagation) [RUM 86] est la méthode la plus utilisée pour l'adaptation des poids. Cet algorithme permet de déterminer le gradient de l'erreur pour chaque neurone du réseau en partant de la dernière couche et en arrivant jusqu'à la première couche cachée.

Les ANN ont été largement utilisées dans les domaines de la vision par ordinateur tels que la classification d'images, la détection d'objets, la segmentation, la reconnaissance de visage, la reconnaissance des expressions faciales. Zhang et al [ZHA 98] utilisent les réseaux de neurones pour reconnaître les sept expressions faciales. Deux types de caractéristiques sont extraits : caractéristiques géométriques basé sur la détection de 34 points d'intérêts et caractéristiques d'apparence basé sur le calcul de 18 coefficients d'ondelettes de Gabor à chaque point. Ces caractéristiques représentent des entrées pour le perceptron multicouche. Les auteurs développent un perceptron à deux couches et ils l'entraînent par Rprop (Resilient propagation). La première couche effectue une réduction non linéaire de la dimension des entrées et la seconde met une décision statistique tout en se basant sur un jeu de fonctionnalité

des unités cachées (voir figure 2.36). Le taux de reconnaissance du système proposé atteint 90.1%.

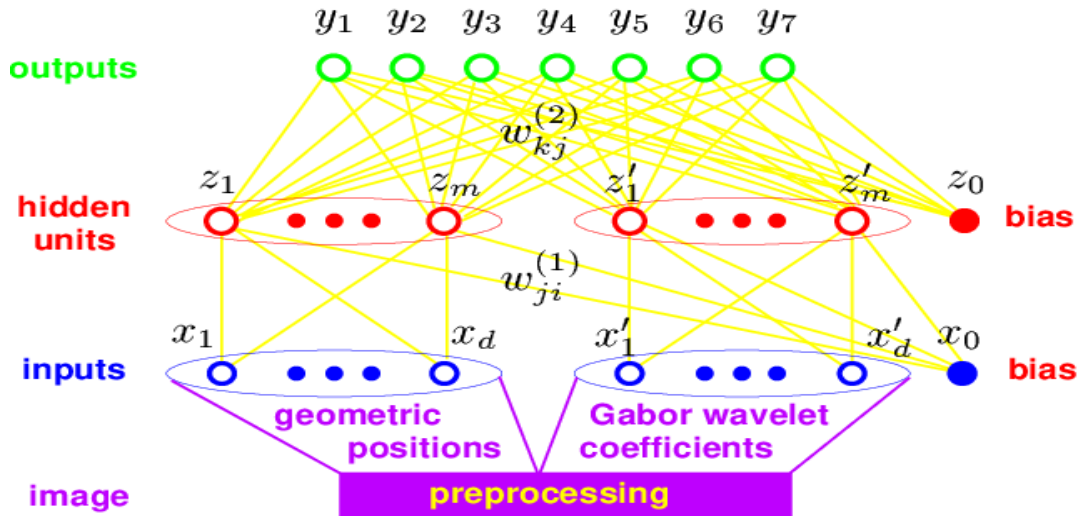


Figure 2.36 Architecture de réseaux de neurones proposée par [ZHA 98].

2.6.5.2. Les machines à vecteurs de support (SVM)

Les machines à vecteurs de support développés par V. Vapnik [VAP 99] sont des techniques discriminante d'apprentissage statistique supervisé qui font partie des algorithmes de classification les plus performants.

L'idée essentielle de cette approche consiste à séparer les données en deux classes, en construisant un hyperplan entre les points des différentes classes. Un tel hyperplan est considéré comme un séparateur optimal qui aura une meilleure capacité à classer les nouveaux exemples inconnus. Les points des données les plus proches de l'hyperplan sont appelés vecteurs de support. Ce processus est illustré sur la Figure 2.37.

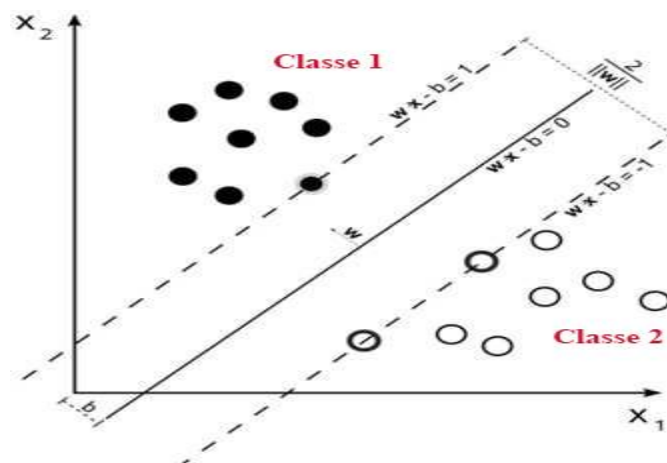


Figure 2.37 Schéma général d'un classificateur SVM

Etant donné les couples :

$$(x_1, c_1), (x_2, c_2), \dots, (x_n, c_n), x_j \in \mathbb{R}^n, c_j \in \{+1, -1\}$$

Où x_i est le point qui vient d'être classé, c_i représente la classe d'appartenance.

La première classe correspond à une réponse positive ($c_j = +1$), la deuxième à une réponse négative ($c_j = -1$). Les SVM linéaires sont la forme la plus simple de cet algorithme. Ils sont applicables dans le cas où les données sont linéairement séparables. Ils essaient de séparer les vecteurs à classe positive des vecteurs à classe négative par un hyperplan défini par l'équation suivante :

$$w \cdot x - b = 0. \quad w \in \mathbb{R}^n, b \in \mathbb{R}$$

Le vecteur des points w est perpendiculaire à l'hyperplan de séparation et b représente une constante de décalage par rapport à l'hyperplan.

Si les données d'apprentissage ne sont pas linéairement séparables, les SVM transforment l'espace de représentation des données d'entrées (appartenant à des classes différentes) non linéairement séparables, dans un espace de plus grande dimension appelé espace de caractéristiques, de façon à ce que les données deviennent linéairement séparables. Ceci est réalisé grâce à une fonction noyau, qui calcul les produits scalaires dans un espace de grande dimension. La tâche principale pour entraîner un SVM non linéairement séparables consiste donc à résoudre le problème d'optimisation quadratique selon la fonction suivante :

$$f(x) = \text{sign}(\sum_{i=1}^l \alpha_i l_i K(x_i, x) + b) \quad (2.16)$$

Où : x_i : Les exemples d'apprentissage, α_i : Les multiplicateurs de Lagrange à déterminer, K : La fonction noyau (Kernal) utilisée et b : Le biais ou le paramètre seuil à déterminer.

La fonction noyau K surmonte le problème de non-linéarité, ce qui permet de traiter des relations non linéaires avec des machines linéaires. En pratique, plusieurs fonctions standard sont considérées parmi lesquelles on peut citer ici :

- La fonction de noyau linéaire:

$$K(x_i, x_j) = x_i^T \cdot x_j \quad (2.17)$$

- La fonction du noyau polynomial:

$$K(x_i, x_j) = (x_i^T \cdot x_j + C)^d \quad (2.18)$$

- La fonction noyau RBF (Radial Basis functions) :

$$K(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right) \quad (2.19)$$

Depuis l'introduction des SVM dans le domaine de reconnaissance de formes principalement en reconnaissance des visages et des expressions faciales, plusieurs travaux ont montré leurs efficacités, Valstar et al proposent en [VAL 07] une approche qui traite le problème de reconnaissance de sourire spontanée par opposition à sourire sournoise. Cette approche fu-

sionne entre différents niveaux d'abstraction de types de caractéristiques faciales et diverses règles de fusion. Ils implémentent trois stratégies de fusion et utilisent un classificateur multi-classe dit GentleSVM qui se base sur un noyau sigmoïde. Le taux de reconnaissance de l'approche proposé est égal à 94% et il est atteint par la troisième modalité de fusion. Rapp et al. [RAP 12] ont combiné deux types de descripteurs (LGBP et AMM) en utilisant un SVM à plusieurs noyaux (MKL) pour reconnaître les micromouvements faciaux d'un visage.

2.6.5.3. Les K-plus proche voisin (KNN)

L'algorithme des k plus proches voisins est un algorithme d'apprentissage supervisé qui peut être utilisé pour la classification et pour la régression. L'hypothèse sur laquelle s'appuie l'algorithme KNN consiste en l'idée qu'une instance est plus proche des instances de la même classe que celles des autres classes. Par conséquent, pour prédire la classe d'une instance inconnue, l'algorithme cherche les K plus proches voisins de cette nouvelle instance et prédit la réponse la plus fréquente de ces K plus proches voisins. La méthode utilise donc deux paramètres : le nombre K et la fonction de similarité pour comparer la nouvelle instance aux instances déjà classés. En effet, la distinction entre deux instances, similaires ou non est faite grâce au calcul de la distance. Plus les deux points sont proches l'un de l'autre, plus ils sont semblables et vice versa. Il existe plusieurs fonctions de calcul de distance, Parmi lesquelles on peut citer :

- La distance Euclidienne : $D_e(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
- La distance de Manhattan : $D_m(x, y) = \sum_{i=1}^n |x_i - y_i|$
- La distance de Hamming : $D_h(x, y) = \max |x_i - y_i|$

Le choix de la valeur de K dépend de la nature, le type et le jeu de données utilisées. Cette valeur a un effet sur la performance du classifieur KNN. Si le K est petit (moins qu'on utilise de voisins), plus on aura un sous-apprentissage. Par contre si le nombre K est grand (plus on utilise de voisins), on risque d'avoir un sur-apprentissage (over fitting) et par conséquent on aura un modèle qui se généralise d'une façon incorrecte sur les instances inconnues.

Les KNNs ont été utilisés dans les applications de reconnaissances des formes et particulièrement dans la reconnaissance de visage et d'expression faciale [SOH 07] [KUM 14]. Sohail et al [SOH 07] ont proposé une méthode pour reconnaître les six expressions faciales de base à partir d'images de visage frontales. Après la localisation des composants faciaux, onze points caractéristiques ont été extraite pour représenter les principaux changements dans les mouvements musculaires. Les auteurs utilisent une approche multi détecteur pour extraire ces points d'intérêt. Le vecteur de caractéristiques est ensuite obtenu en calculant la distance euclidienne entre les onze points d'intérêt. Le classifieur KNN est utilisé pour classifier les vi-

sages dans l'une des six expressions faciales. Le taux de reconnaissance moyen a été atteint 90,76%.

2.6.5.4. Les modèles de Markov cachés (HMM)

Les HMM [EDD 96] sont des algorithmes de classification génératifs qui définissent une loi de probabilité, pour chaque classe, selon laquelle les séquences en entrée sont générées. Un HMM est utilisé pour caractériser les propriétés stochastiques d'une séquence de données, où des probabilités d'observation sont associées à chaque état et des probabilités de transition sont associées à chaque arc.

Les HMMs ont été utilisés dans les applications de reconnaissances de paroles et de caractères, ensuite ils sont appliqués à la reconnaissance de visage et d'expression faciale. Cohen et al. [COH 03] proposent une méthode hiérarchique des HMMS pour la segmentation et la reconnaissance des expressions faciales à partir d'une séquence vidéo (voir figure 2.38). Le premier niveau de l'architecture est composé de six HMMs indépendants correspondants aux six expressions universelles. Les caractéristiques des mouvements du visage par rapport à l'état neutre sont utilisées comme des entrées des HMMs du bas niveau, et la séquence des sorties de ces derniers est codée puis utilisée comme vecteur d'observation du modèle du niveau haut de l'architecture qui est constitué de 7 états (6 pour les expressions universelles et un pour l'état neutre). Les auteurs montrent que la performance de leur approche atteint un taux de reconnaissance environ 82.46% pour une personne familière et un taux environ 58.63% pour une personne non familière.

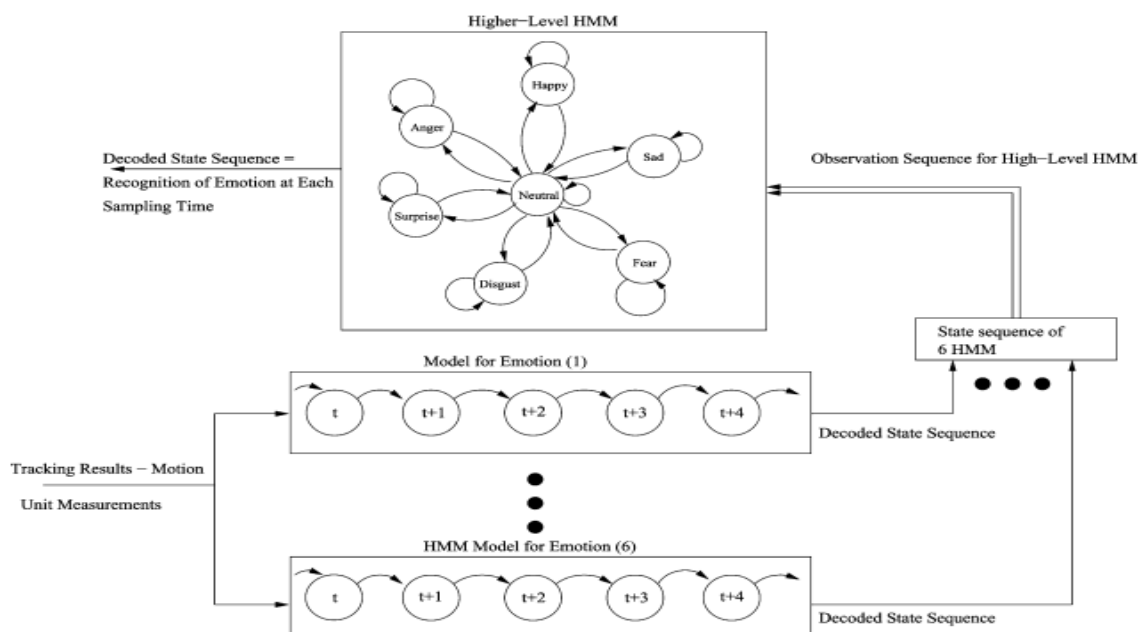


Figure 2.38 Architecture multi-niveaux de HMMs pour la reconnaissance d'expressions faciales d'après [COH 03].

2.7. Les systèmes de fusion

Afin d'augmenter la précision et la robustesse des systèmes de reconnaissance des expressions faciales, de nouvelles méthodes, dites de fusion, ont été développées [BUS 04] [SEB 05]. En effet la reconnaissance des émotions basée sur une seule modalité fournit de l'information discriminante limitée et une applicabilité locale, et ne permet pas d'atteindre des performances satisfaisantes.

Les méthodes de fusionnement consistent à combiner les multiples sources d'information afin de tirer parti des avantages et de la complémentarité de chaque information présentée. Leurs intérêts résident dans la possibilité d'améliorer et de fiabiliser la reconnaissance des d'émotions et de lever des ambiguïtés qui peuvent survenir lors de l'utilisation d'une seule modalité.

La fusion d'informations peut prendre différentes formes selon le niveau ou elle est effectuée. La Figure 2.39 illustre les trois possibilités de fusion à différents stades de la reconnaissance. D'une façon générale, il est possible de fusionner (i) les données directement après l'extraction des signaux (niveau signal), (ii) fusionner les attributs provenant des différentes modalités (niveau caractéristique), ou (iii) fusionner les informations durant la phase décision (niveau décision).

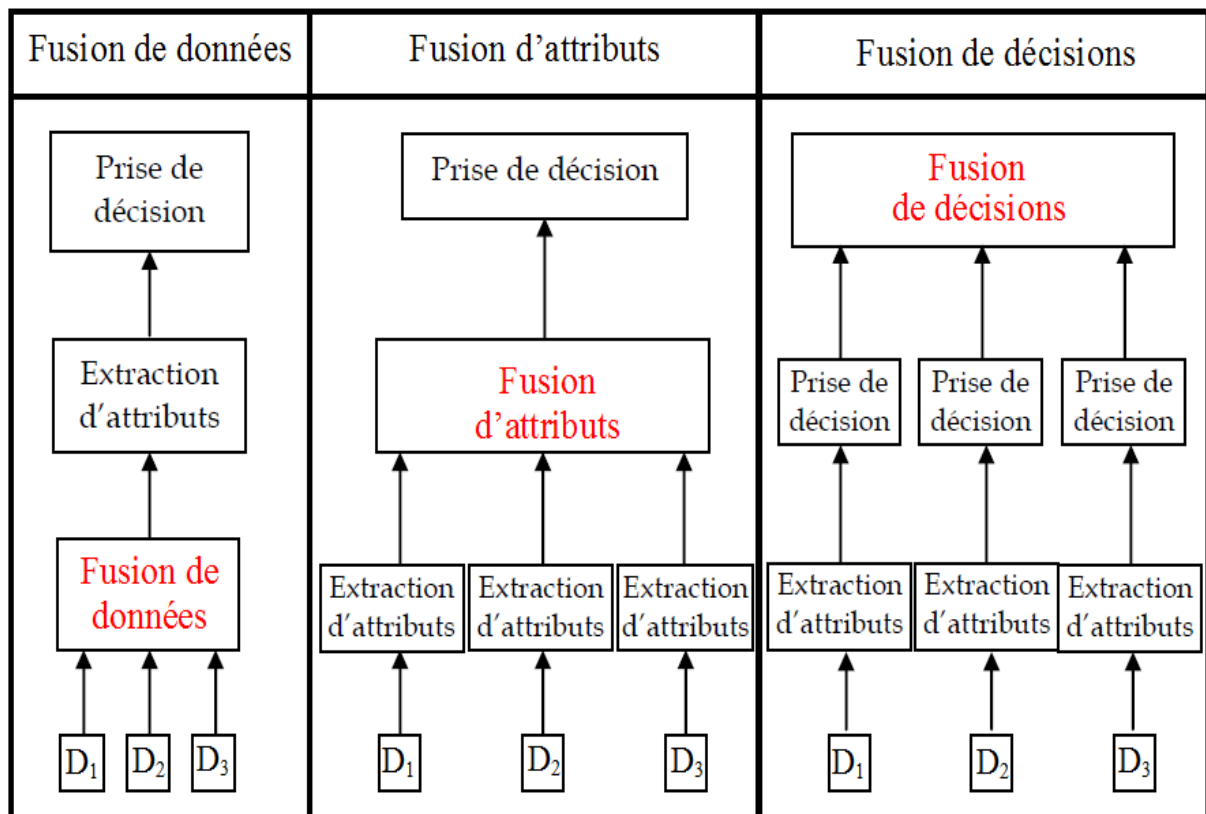


Figure 2.39 Les différents niveaux de fusion

2.7.1. Fusion de données

La fusion au niveau des échantillons consiste à combiner les données brutes ou les signaux provenant directement des capteurs afin de construire de nouvelles données ou des signaux de meilleure qualité. Cette stratégie ne peut être appliquée que lorsque les signaux sont compatibles et ont la même nature et la même résolution temporelle. Ainsi, le couplage et la synchronisation de plusieurs signaux n'est généralement pas possible car les différentes modalités ont souvent des capteurs possédant différentes caractéristiques de signal. Cette technique est donc rarement utilisée en raison de la difficulté d'intégration des signaux et de la sensibilité au bruit, produite par un mauvais fonctionnement des capteurs utilisés. Par exemple, il est peut-être difficile de fusionner des images de visages provenant de caméras ayant des résolutions différentes.

2.7.2. Fusion d'attributs (caractéristiques)

Cette stratégie de fusion est effectuée sur l'ensemble des caractéristiques extraites à partir des données de chaque modalité impliquée. Les vecteurs de caractéristiques calculés sont ensuite fusionnés afin d'en construire un seul qui va être utilisé lors de l'apprentissage et de la mise en correspondance. Le processus de fusion peut être effectué par une simple concaténation ou par l'utilisation de la méthode de corrélation entre les différents vecteurs de caractéristiques. A noter qu'une étape de prétraitement (normalisation, réduction d'espace...) est nécessairement appliquée avant la combinaison des descripteurs afin de normaliser et réduire la dimension du vecteur caractéristique final.

Plusieurs travaux dans la littérature ont préféré l'utilisation de l'approche de fusion de descripteurs pour la reconnaissance des expressions faciales. Huang and al [HUA 98] ont proposé une approche simple qui permet de concaténer des caractéristiques audio-visuel. Cette approche est basée d'une part sur l'analyse vidéo et d'autre part sur la parole. Les auteurs ont comparé les résultats de reconnaissance des émotions en utilisant trois tests : vidéo seul, audio seul, audio et vidéo combinés. Ils ont constaté en utilisant la méthode des plus proches voisins (KNN), que l'utilisation des caractéristiques provenant de la vidéo seule présente une confusion entre la colère, la surprise et la tristesse, et l'utilisation de la parole seule présente une confusion entre la tristesse et l'aversion, alors que l'utilisation de l'approche multimodale était plus performante. Le taux de reconnaissances était 75% pour l'audio et 69.4% pour la vidéo contre 97.2% pour l'approche multimodale.

Zhengyou et al [ZHE 98] fusionnent les coefficients obtenus par l'application de 18 filtres de Gabor et les coordonnées de 34 points caractéristiques détectés manuellement. La

combinaison est appliquée au niveau de perceptrons à deux couches. La première couche réduit de façon non linéaire les descripteurs de chaque type d'information. La deuxième calcule les taux de reconnaissance pour chaque expression émotionnelle. Le taux de reconnaissance atteint 90% en utilisant la base JAFFE.

La stratégie de fusion au niveau des caractéristiques a l'avantage de nécessiter une seule phase d'apprentissage pour l'ensemble des modalités. Cependant, son principal inconvénient réside dans la malédiction de dimension ainsi que la non-homogénéité de données fusionnées. En outre, Cette méthode de fusion suppose que les descripteurs de chaque modalité soient de nature similaire. De plus, pour que cette fusion ait un sens, il est nécessaire que les flux, et donc les descripteurs qui leur sont associés, soient synchronisés. Une alternative à cette approche de fusion est la fusion au niveau des décisions qui est largement étudiée dans la littérature.

2.7.3. Fusion des décisions

Cette stratégie de fusion est connue aussi comme une fusion de haut niveau, vu qu'elle intervient après l'étape de classification. Les classifieurs sont d'abord construits pour chaque modalité et leurs sorties respectives sont ensuite combinées pour obtenir une vision globale de l'état émotionnel. Ainsi, cette approche consiste à fusionner les décisions prises séparément pour chaque classifieur.

La combinaison de deux ou plusieurs classifieurs intègre les avantages de chacun d'eux pour améliorer la qualité du système de reconnaissance, ce qui conduit à des solutions plus robustes. En effet, les méthodes de combinaison des classifieurs exploitent la complémentarité potentielle pour obtenir des performances supérieures à celles d'un seul classifieur.

Suivant le niveau d'information apporté par le classifieur, on peut distinguer trois catégories de sorties:

- **Sortie de type classe :**

C'est le type le plus général mais qui apporte le moins d'informations. Le classifieur ne donne que la proposition du type (la classe) de l'entrée à reconnaître sans aucune autre information.

$$e_j(x) = C_i \ (i \in \{1..N\}).$$

Ou C_i est la classe parmi N classes attribuées par le classifieur e_j à x .

La combinaison de type classe est basée sur le principe du vote qui consiste à interpréter chaque sortie d'un classifieur comme un vote pour l'une des classes possibles. La classe ayant un nombre de votes supérieur à un seuil préfixé est retenue

comme décision finale. Ces méthodes sont les plus simples à mettre en oeuvre. Elles contiennent trois catégories : vote avec seuil, majorité sans conflit et majorité notoire.

- **Sortie de type rangs :**

Ce type de sortie reflète l'ordre de préférence des propositions fournies par le classifieur, cela se traduit par l'attribution d'un rang pour chaque classe ; Plus cette dernière est probable moins le rang est élevé. La liste des propositions peut contenir toutes les classes possibles ou seulement les mieux classées.

$$e_j(x) = [r_j^1, r_j^2, \dots, r_j^N]$$

Où r_j^i est le rang attribué à la classe C_i par le classifieur e_j .

Les méthodes de type rang combinent des listes de propositions de classifieurs ordonnées de manière décroissante selon leurs préférences. Les rangs attribués par chaque classifieur peuvent être considérés comme des votes pour toutes les classes. Ayant un nombre N de classes, chaque classifieur attribue un poids N à la classe placée au début de la liste, un poids $N-1$ à la seconde classe, etc. La classe placée à la fin de la liste reçoit un seul vote. Les méthodes non-paramétriques de type rang les plus reconnues sont le Bordacount et le meilleur rang.

- **Sortie de type mesure :**

Cette sortie est la plus riche en informations puisque le classifieur dans ce cas associe à chaque classe une mesure de confiance qui peut être, par exemple, une distance, une probabilité a posteriori, une valeur de confiance, un score, une mesure floue, ... etc..

$$e_j(x) = [M_j^1, M_j^2, \dots, M_j^N]$$

Où M_j^i est la mesure attribuée à la classe C_i par le classifieur e_j .

Les méthodes de type mesure combinent des mesures qui reflètent le degré de confiance des classifieurs sur l'appartenance de la forme à reconnaître en chacune des classes. Toutefois, comme les sorties des classifieurs ne sont pas toujours comparables, une normalisation est souvent nécessaire. Les méthodes de combinaison non-paramétrique les plus utilisées sont les *règles fixes* et la *moyenne de Bayes*.

Il existe différentes variantes de règles fixes, qui permettent de cumuler les scores produits par les classifieurs en fusion. Ces variantes sont les suivantes :

- **La règle de maximum :** Cette règle trouve la valeur maximale pour chaque classe parmi les classifieurs et assigne l'objet à la classe qui a la valeur maximale.

- **La règle de minimum** : Trouve la valeur minimale de chaque classe parmi les classifieurs et assigne la forme à la classe qui a la valeur maximale.
- **La règle de la somme** : Additionne les résultats des classifieurs et assigne la classe avec la valeur maximale à la forme.
- **La règle de produit** : Multiplie les résultats des classifieurs et assigne la classe avec la valeur maximale à la forme.
- **La règle de la moyenne** : Trouve la moyenne de chaque classe parmi les classifieurs et assigne l'objet à la classe avec la valeur maximale.

2.7.3.1. Les systèmes multi-classifieurs (Multiple Classifier Systems MCS)

Au cours des dix dernières années les systèmes multi-classifieurs sont devenus une méthode établie pour concevoir des systèmes de classification. Plusieurs preuves théoriques et expérimentales ont montré que les systèmes multi-classifieurs peuvent surpasser un classifieur unique dans plusieurs applications réelles, en termes de précision de la classification.

Un système multi-classifieurs est constitué d'un ensemble de classifieurs et d'une fonction de décision pour combiner leurs sorties. La description d'un MCS suit les trois phases suivantes : génération, sélection et intégration. Dans la première phase un ensemble de classifieurs complémentaires est généré. Dans la deuxième phase, un sous-ensemble de ces classifieurs est sélectionné puis combinés pour arriver à une solution optimale, tandis que dans la dernière phase, une fonction de combinaison est définie pour donner une décision finale, qui sera prise en se basant sur les prédictions des classifieurs sélectionnés.

2.7.3.2. Architectures de combinaison de classifieurs

Les principaux schémas de combinaison de classifieurs sont les architectures en cascade, en parallèle et hybride :

2.7.3.2.1. Architecture en cascade :

La fusion en cascade, appelée également, la fusion *séquentielle ou en série*, consiste à positionner les classifieurs les uns à la suite des autres. Elle est donc organisée en niveaux successifs de décision permettant de réduire progressivement le nombre de classes possibles. Dans chaque niveau, il existe un seul classifieur qui prend en compte la réponse fournie par le classifieur placé en amont afin de confirmer la décision prise auparavant ou traiter les rejets (voir figure 2.40).

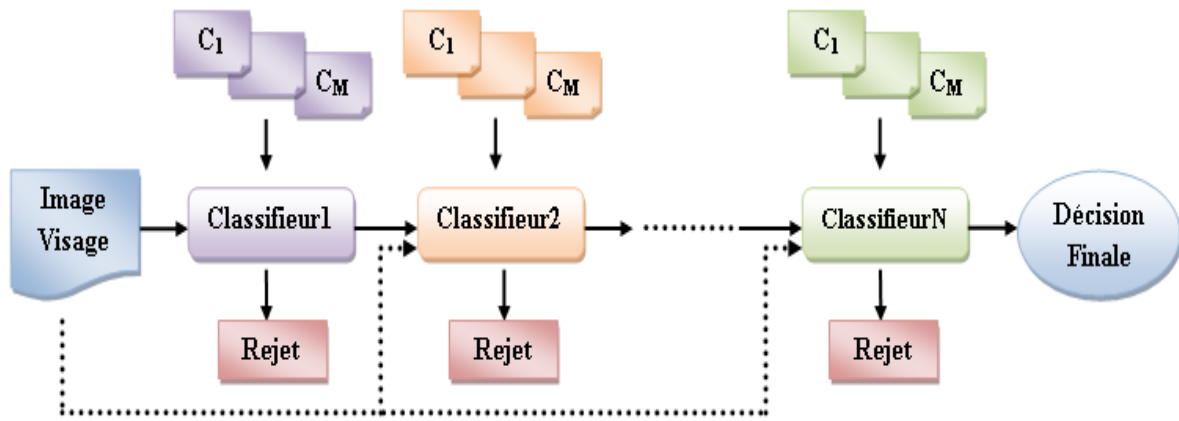


Figure 2.40 Combinaison séquentielle de classifieurs

L'avantage principal de la fusion en cascade réside dans le gain de temps de recherche en filtrant les décisions des classifieurs permettant ainsi de diminuer le taux d'erreur globale de la reconnaissance. Cependant ce type de combinaison demeure sensible à l'ordre dans lequel les classifieurs sont placés. Les premiers classifieurs utilisés doivent être robustes et capables d'identifier la solution réelle, sinon, l'erreur va se propager jusqu'au dernier classifieur.

2.7.3.2.2. Architecture en parallèle

Dans une architecture parallèle, tous les classifieurs opèrent indépendamment l'un de l'autre puis leurs réponses respectives sont combinées (voir figure 2.41). Cette combinaison est faite soit de manière démocratique, dans le sens où elle ne favorise aucun classifieur par rapport à un autre, soit au contraire dirigée et, dans ce cas, on attribue à la réponse de chaque classifieur, un poids en fonction de ses performances. L'ordre d'exécution des classifieurs n'intervient pas dans cette architecture.

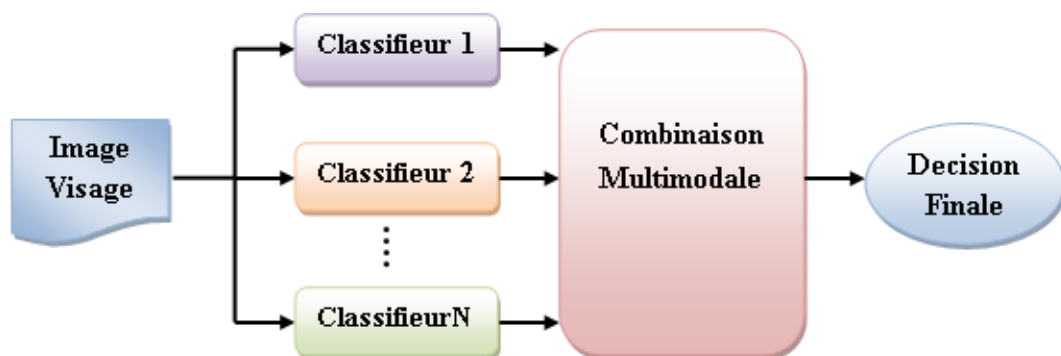


Figure 2.41 Combinaison parallèle de classifieurs

2.7.3.2.3. Architecture hybride

Comme son nom l'indique, l'approche hybride consiste à utiliser un schéma qui reprend simultanément les principes de combinaison séquentielle et parallèle afin d'obtenir une dé-

cision finale plus sûre. La figure 2.42 montre une représentation de la fusion hybride dans laquelle on combine un classifieur série avec deux classifieurs en parallèle.

Ce type de configuration nécessite une bonne connaissance de chacun des classifieurs utilisés de façon à choisir la topologie la plus adaptée.

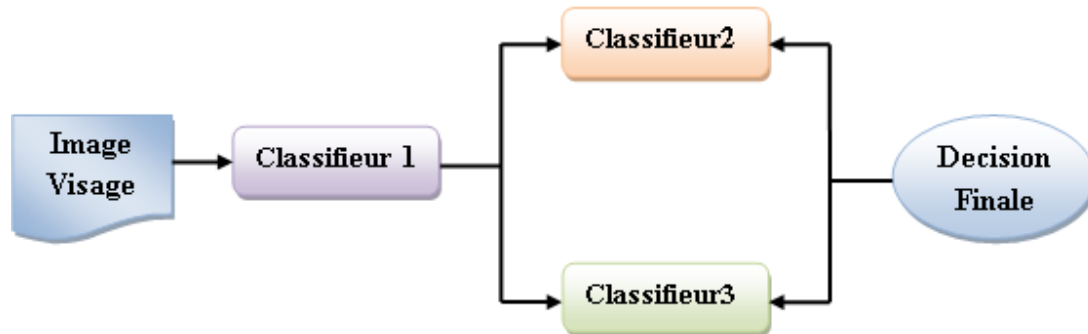


Figure 2.42 Combinaison hybride de classifieurs

Dans la littérature, plusieurs travaux ont adopté l'approche de fusion au niveau décisionnel. Song et al. [SON 10] combinent les décisions issues de la modalité géométrique et la modalité d'apparence pour différencier deux types d'expressions à savoir les expressions symétriques les expressions non symétriques. La modalité géométrique est basée sur l'extraction des paramètres d'animation faciale (FAPs) codés par 27 points, tandis que la modalité d'apparence est basée sur la détection des déformations de la peau en calculant les rapports d'intensité dans 8 patches. L'identification des émotions est ensuite établie en fonction du type de l'expression symétrique ou asymétrique. Les taux de reconnaissance dépassent les 88% en utilisant la base Cohn-Kanade, la base JAFFE et une base préparée par les auteurs.

Zeng et al. [ZEN 07] ont utilisé la méthode de vote pour fusionner les résultats de la classification à partir de deux canaux audio-visuelle. Du côté audio, ils ont utilisé le pitch et l'énergie de la prosodie et du côté vidéo ils ont utilisé des mouvements du front du visage. Cette fusion a permis d'améliorer le taux de reconnaissance de 7.5% par rapport à l'utilisation de chaque modalité toute seule.

2.8. Conclusion

Dans ce chapitre, nous avons présenté les différentes approches utilisées pour l'analyse des expressions faciales. Nous avons abordé tout d'abord la terminologie des expressions faciales ainsi que les différents formalismes proposés pour leurs représentations. Ensuite, nous avons décrit les différentes méthodes élaborées pour la détection du visage, l'extraction des caractéristiques faciales et la classification des émotions. Nous avons focalisé notre analyse sur l'étape d'extraction des caractéristiques que nous l'avons divisée en trois catégories (géométriques, d'apparence et hybrides), tout en exposant les approches les plus populaires

pour chacune de ces catégories. Finalement, Nous avons traité dans la dernière partie l'approche multimodale et nous avons passé en revue les trois stratégies de fusion qui ont été proposées dans la littérature. Nous avons pu constater que la multi-modalité était un concept largement répondu pour l'analyse des expressions faciales et que les méthodes multi modales ont démontré un gain notable de précision par rapport à l'utilisation de chaque modalité séparément.

Dans le chapitre suivant, nous nous focalisons sur la reconnaissance des expressions faciales via l'apprentissage profond (deep learning). Nous détaillons particulièrement les différents modèles de deep learning utilisées pour l'analyse des émotions.

CHAPITRE 3

Analyse des expressions faciales via l'apprentissage profond

Sommaire

3.1 Introduction	111
3.2 Relation entre IA, ML et DL	111
3.3 Qu'est ce qu'un apprentissage profond (Deep learning)	114
3.4 Les différents modèles de l'apprentissage profond	114
3.4.1 Les réseaux de neurones convolutionnels (CNN)	115
3.4.2 Les réseaux de neurones récurrents (RNN)	122
3.4.3 Les modèles génératives profonds	128
3.4.3.1 Les machines de Boltzmann restreintes (RBM)	130
3.4.3.2 Les réseaux de croyance profonds (DBN)	130
3.4.3.3 Les réseaux antagonistes génératifs (GAN).....	132
3.4.3.4 Les Auto-Encodeurs (AE : Auto-Encoder)	135
3.5 Conclusion	137

3.1 Introduction

Au sein de la révolution que connaît l'Intelligence Artificielle aujourd'hui, l'apprentissage profond occupe une place de tout premier plan. En effet, il est impossible d'ignorer la prédominance des modèles profonds dans le domaine de traitement d'image et de vision par ordinateur, et en particulier pour l'analyse des expressions faciales. Le succès remarquable du l'apprentissage profond repose à la fois sur des avancées en terme de modélisation des représentations de l'information, sur l'utilisation des méthodes d'apprentissage exploitant des données massivement annotées, et sur des calculateurs très puissants adaptés à ces architectures.

Dans ce chapitre, nous allons présenter un aperçu détaillé des différents modèles de l'apprentissage profond utilisés dans la reconnaissance des expressions faciales. Nous allons d'abord commencer par donner une comparaison entre l'Intelligence Artificielle, le Machine Learning et le Deep Learning, puis nous allons définir l'apprentissage profond. Nous aborderons ensuite les différents modèles de l'apprentissage profond tout en donnant une taxonomie complète pour chaque modèle présenté.

3.2 Relation entre Intelligence Artificielle (IA), Machine Learning (ML) et Deep Learning (DL)

L'*intelligence artificielle* est en plein développement ces dernières années. De nouveaux lexiques liés à l'émergence de ce domaine sont apparus. En effet, lorsqu'on parle d'intelligence artificielle, on fait très souvent l'allusion aux technologies associées comme *l'apprentissage automatique* ou *l'apprentissage approfondi*.

L'intelligence artificielle, l'apprentissage automatique, et l'apprentissage approfondi sont souvent regroupés ensemble, alors qu'en réalité, ces trois notions ne sont pas exactement similaires. Avant d'entamer la définition d'apprentissage profond, il nécessaire de positionner et comprendre la relation entre ces trois concepts (voir figure 3.1). Donc pour commencer, revenons sur ces trois définitions essentielles :

- **L'intelligence artificielle (IA) (1950 – 1980) :** est une discipline scientifique qui regroupe l'ensemble des théories et des techniques permettant de comprendre et reproduire le fonctionnement d'un cerveau humain. Autrement dit, il s'agit d'un ensemble de méthodes mises en œuvre en vue de fabriquer des machines capables de simuler l'intelligence humaine [KAP 19].
- **L'apprentissage automatique ou le Machine Learning (ML) (1980 - 2010) :** est un domaine de l'IA donnant la capacité aux machines d'apprendre automatiquement un

ensemble de règles, de résoudre des problèmes et de mettre en œuvre des solutions à partir de jeux de données et d'algorithmes d'apprentissage [DUT 96].

- **L'apprentissage profond ou le Deep Learning (DL) (depuis 2010) :** est un sous-domaine de ML qui utilise généralement des réseaux neuronaux profonds avec des dizaines voire des centaines de couches de neurones empilées pour apporter une plus grande quantité d'informations pour la formation ou l'apprentissage [DEE 20].

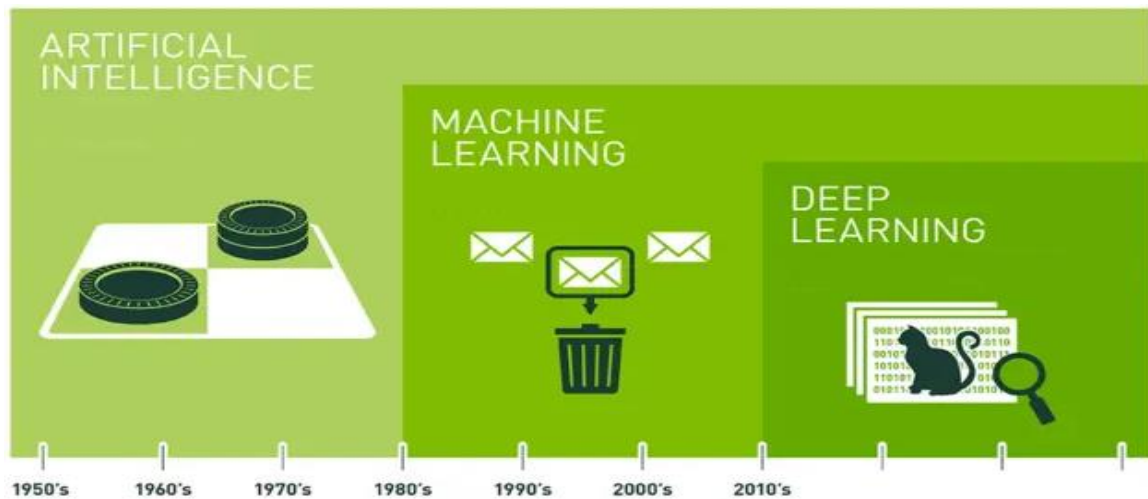


Figure 3.1 Relation entre Intelligence Artificiel, Machine Learning et Deep Learning [COP 16]

Les algorithmes de ML fonctionnent bien pour une grande variété de problèmes, cependant ils ont échoué à résoudre quelques problèmes majeurs de l'IA telle que la reconnaissance vocale et la reconnaissance d'objets. La naissance de l'apprentissage profond fut motivée en partie par l'échec des algorithmes traditionnels dans la résolution de telle tâche de l'IA [WAN 18b].

Une des grandes différences entre la technique d'apprentissage profond et les algorithmes d'apprentissage automatique traditionnelles est l'adaptabilité. En effet, les techniques d'apprentissage profond évoluent beaucoup mieux avec plus de données que les outils traditionnels de ML qui ont des problèmes lorsqu'ils sont confrontés au Big Data [GHE 17] [KUA 15]. Autrement dit, plus la quantité de données fournie est grande plus les performances d'un algorithme d'apprentissage profond sont meilleurs. Contrairement à plusieurs algorithmes de l'apprentissage automatique classiques qui possèdent une borne supérieure à la quantité de données qu'ils peuvent recevoir. La figure 3.2 suivante explique le succès du Deep learning de point de vue performance qui augmente avec l'augmentation de données par rapport à celui des algorithmes de ML classique qui cesse de s'améliorer après un point de saturation et montre ses limites.

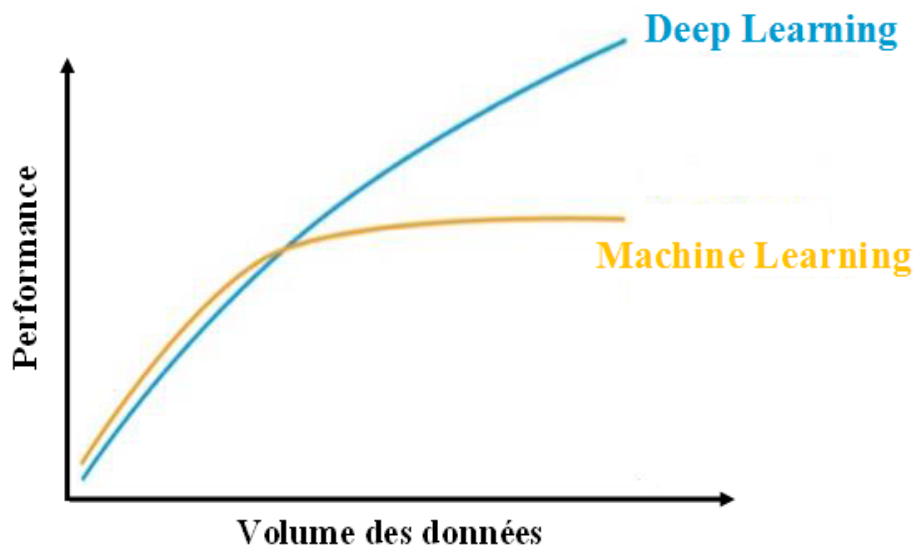


Figure 3.2 La différence de performance entre le Deep Learning et la plupart des algorithmes de ML en fonction de la quantité de données [MER 20a].

Une autre différence entre les algorithmes de ML traditionnelles et les algorithmes de deep learning est dans l'étape de l'extraction de caractéristiques. En effet, les algorithmes de ML traditionnelles requièrent une expertise humaine en matière d'extraction de caractéristiques [WAN 18b], tandis que les méthodes de deep learning traitent et créent leurs propres caractéristiques directement à partir des données.

Les algorithmes de machine learning classique dépendent donc des caractéristiques qui ont été conçues pour créer le modèle de classification, alors que les deep learners sont potentiellement capables d'extraire une meilleure représentation des caractéristiques des données brutes afin de créer des modèles bien meilleurs.

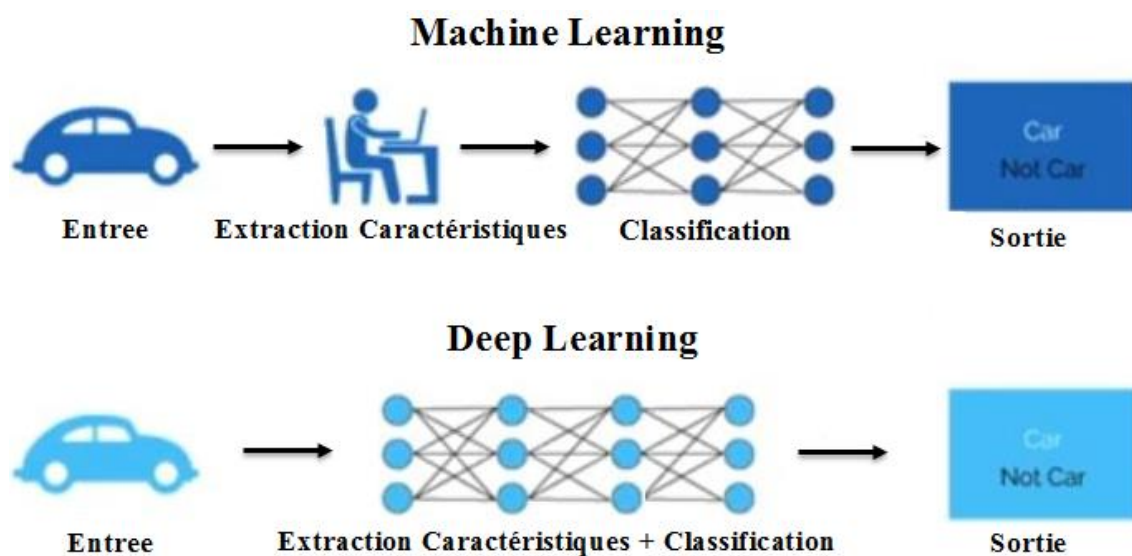


Figure 3.3 Comparaison entre la machine learning et le deep learning [MER 20a].

3.3. Qu'est ce qu'un apprentissage profond (Deep learning)

Le deep learning est un ensemble de méthodes d'apprentissage automatique qui tentent d'apprendre à plusieurs niveaux d'abstraction dans le but de modéliser des relations complexes entre les données. Il est basé sur des algorithmes capables de mimer les actions du cerveau humain grâce à des réseaux de neurones artificiels.

Les réseaux sont composés de plusieurs couches de neurones, chacune recevant et interprétant les informations de la couche précédente. Le DL permet de gérer de larges quantités de données en ajoutant des couches au réseau. Un modèle de DL a la capacité d'extraire des caractéristiques à partir des données brutes grâce aux multiples couches de traitement composé de multiples transformations linéaires et non linéaires, et apprendre sur ces caractéristiques petites à petites à travers chaque couche avec une intervention humaine minime.

3.4. Les différents modèles de l'apprentissage profond

Il existe un grand nombre de variantes d'architectures profondes. La plupart d'entre elles sont dérivées de certaines architectures originales (voir figure 3.4). On peut citer les architectures profondes suivantes :

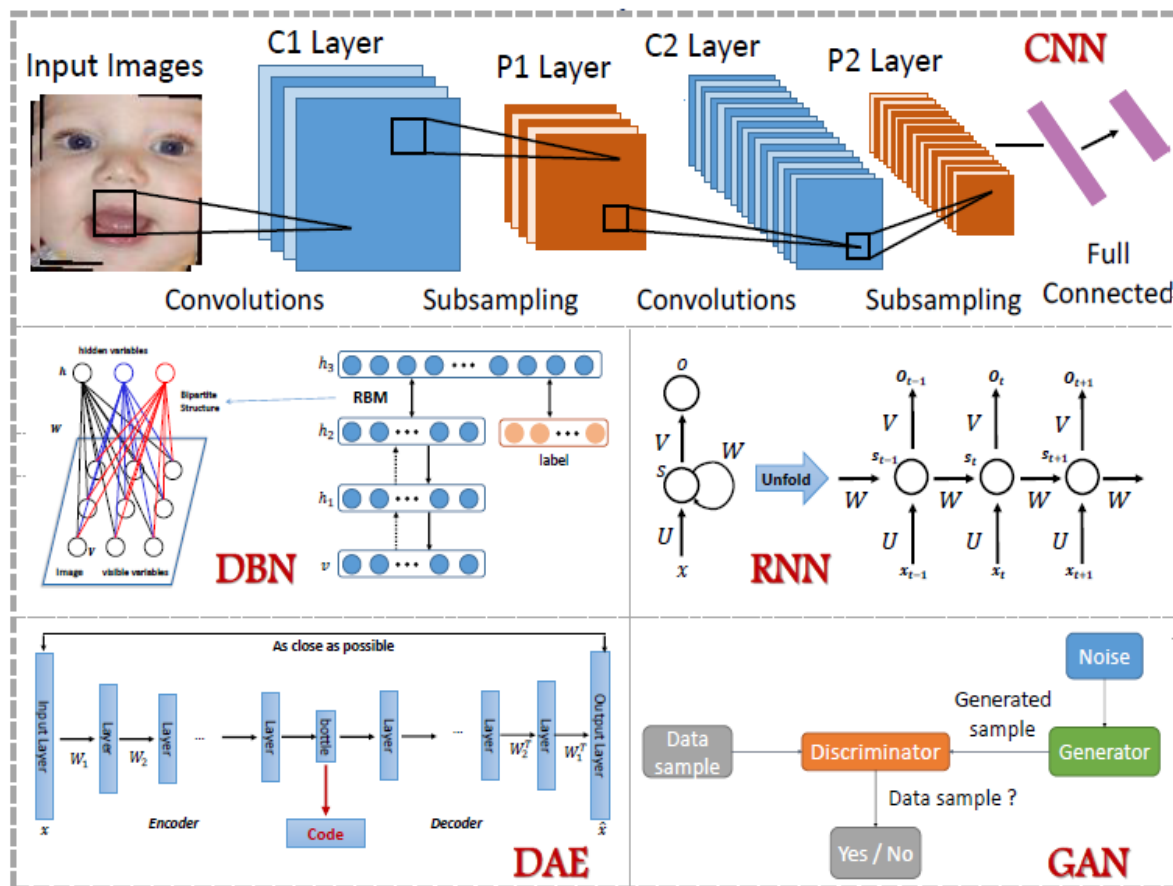


Figure 3.4 Les différents modèles de deep learning [LI 20].

3.4.1 Les réseaux de neurones convolutionnels (CNN : Convolutional Neural Networks)

Les CNN sont un spécialisés pour le traitement de données ayant une topologie semblable à une grille, qui peuvent être considérées comme une grille 1D (Vecteur) ou une grille 2D de pixels (Matrice). Le nom '*réseau de neurones convolutif*' indique que le réseau utilise une opération mathématique qui s'appelle *convolution*. Il s'agit d'une opération linéaire spéciale. Les réseaux convolutifs sont simplement des réseaux de neurones qui utilisent la convolution à la place de la multiplication matricielle dans au moins une de leurs couches [LEC 98].

3.4.1.1 Inspiration des CNNs

Le modèle de connectivité entre les neurones d'un CNN s'inspire de l'organisation du cortex visuel animal [HUB 62]. Les neurones corticaux individuels répondent aux stimuli dans une région restreinte de l'espace connu sous le nom de champ réceptif. Les champs réceptifs des types de réseau de neurones acycliques différents neurones se chevauchent partiellement de sorte qu'ils couvrent le champ visuel. La réponse d'un neurone individuel aux stimuli dans son champ réceptif peut être approchée mathématiquement par une opération de convolution [HUB 62].

3.4.1.2 Architecture globale d'un CNN

L'architecture typique d'un CNN comprend en plus de la couche d'entrée, deux blocs bien distincts (voir figure 3.5) [MER 20b] :

Le *premier bloc* fait la particularité de ce type de réseaux, puisqu'il fonctionne comme un *extracteur de caractéristiques* des images. Une image est passée à travers une succession de filtres ou noyaux de convolution, créant de nouvelles images appelées cartes de caractéristiques. Ces cartes sont ensuite normalisées (avec une fonction d'activation) et/ou redimensionnées (pooling). Ce procédé peut être réitéré plusieurs fois. Les valeurs des dernières cartes de caractéristiques sont concaténées dans un vecteur. Ce vecteur définit la sortie du premier bloc, et l'entrée du second.

Le *second bloc* représente la couche entièrement connectée (perceptron multicouche) d'un CNN. Dans cette couche, les valeurs du vecteur en entrée sont transformées (avec plusieurs combinaisons linéaires et fonctions d'activation) pour renvoyer un nouveau vecteur en sortie [MER 20b]. Ce dernier vecteur contient autant d'éléments qu'il y a de classes : l'élément i représente la probabilité que l'image appartienne à la classe i . Donc la couche entièrement connectée permet de combiner les valeurs de vecteur en sortie de la partie convolutive pour classer les images d'entrée dans différentes classes.

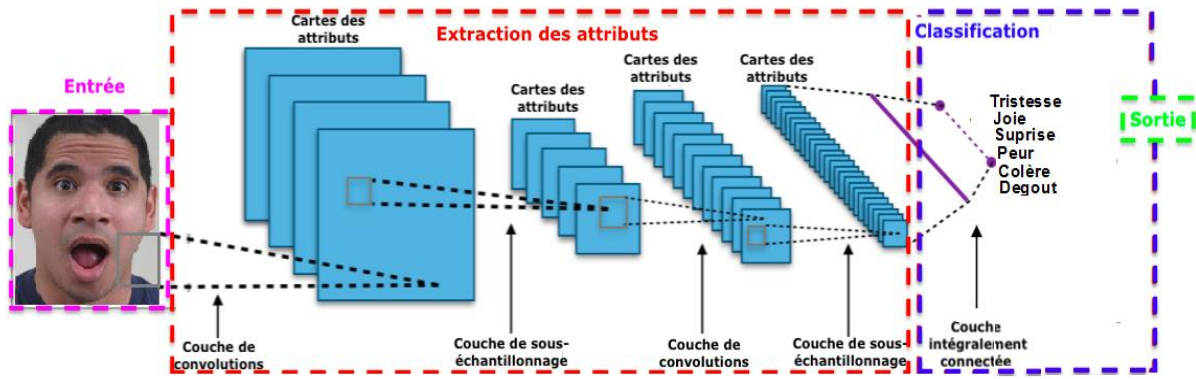


Figure 3.5 Architecture générale d'un CNN.

Il est à noter que la couche d'entrée est constituée de la matrice de données à classifier. Elle est représentée généralement par la matrice de l'image dans le cas des images en niveau de gris, ou les 3 matrices de l'image pour le cas d'images RGB.

3.4.1.3 Les différentes couches de CNN

Un CNN classique est généralement composé de cinq types de couches, tout en ayant un rôle bien défini :

A. Couche de convolution

La couche de convolution est l'élément central des réseaux neuronaux convolutifs. Elle consiste à appliquer un ensemble de filtre ou noyau de convolution à l'image d'entrée pour en extraire des cartes de caractéristiques [KRI 12]. Le terme 'convolution' indique que la couche de convolution utilise une opération mathématique linéaire appelée *convolution* ou *produit de convolution*. Soient f et g deux fonctions définies sur R , la convolution discrète entre f et g est généralement notée $f * g$ et elle est définie par l'équation suivante :

$$S(x) = (f * g)(x) = \sum_t f(t) g(x + t) \quad (3.1)$$

Dans un CNN, le premier argument f est assimilé à l'entrée, le second argument g est assimilé au noyau de convolution (kernel) et la sortie est assimilée à la carte de caractéristiques (feature maps).

Pour les signaux bidimensionnels tels que les images, nous considérons la convolution 2D suivante :

$$S(i, j) = (K * I)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \quad (3.2)$$

K est un noyau de convolution appliqué à un signal 2D (ou image) I .

Le principe de la convolution 2D est de faire glisser un noyau de convolution sur l'image. A chaque position, on obtient la convolution entre le noyau et la partie de l'image qui est actuellement traitée.

La figure 3.6 illustre l'opération de convolution. On considère une image ou une carte de caractéristiques de 4x7 pixels et un filtre de 3x3, la convolution est la somme des produits des éléments qui se trouvent aux mêmes positions des deux matrices.

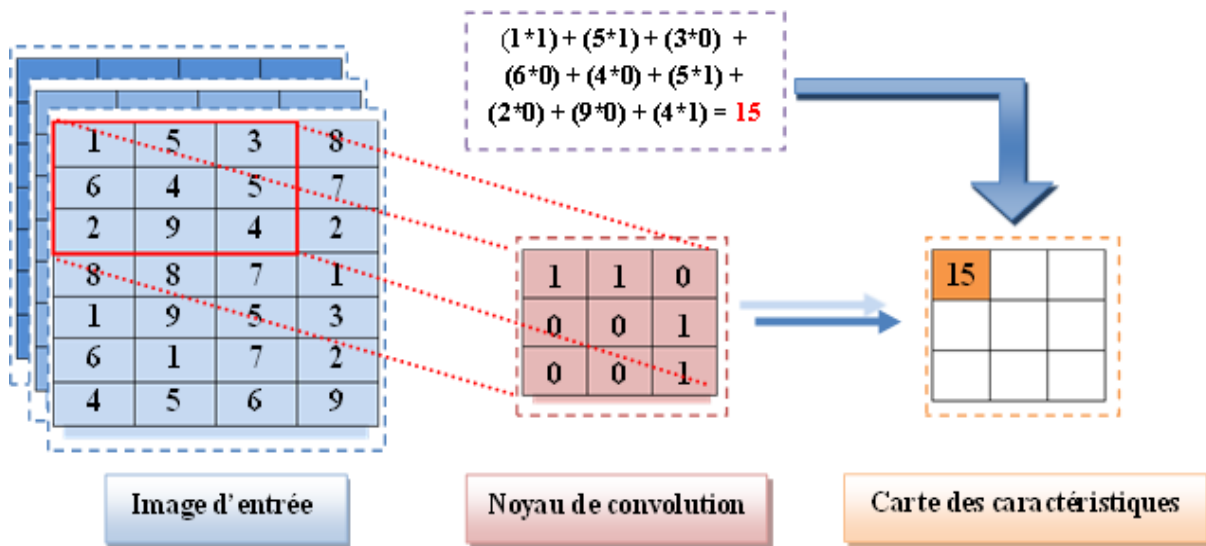
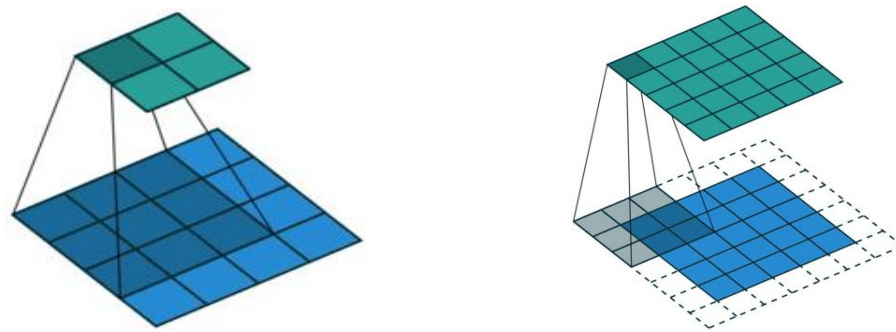


Figure 3.6 Exemple de calcul de l'opération de convolution

Une couche de convolution est caractérisée par quatre hyper-paramètres :

- Les dimensions des noyaux de convolution, généralement une convolution à une dimension égale à 2 avec des noyaux carrés [DAH 20].
- Le nombre des noyaux de convolution ou la profondeur D : c'est le nombre de cartes de caractéristiques, en sortie associée à la convolution (ou nombre de neurones associés à un même champ récepteur) [DAH 20];
- Le pas de convolution ou *Stride* S : C'est le pas de décalage du noyau de convolution à chaque calcul. Par exemple, un pas de 1 signifie qu'on déplace la fenêtre d'un pixel à la fois. Le pas permet de contrôler le chevauchement des champs récepteurs. Plus le pas est petit, plus les champs récepteurs se chevauchent et plus le volume de sortie sera grand [MER 20b] ;
- La marge à zéro ou le *zero-padding* P : On ajoute à l'image en entrée de la couche un contour des zéros d'épaisseur P pixels (voir figure 3.7). Sans ce contour, les dimensions en sortie sont plus petites. Ainsi, plus on empile de couches de convolution avec $P=0$, plus l'image en entrée du réseau rétrécit. On perd donc beaucoup d'informations rapidement, ce qui rend la tâche d'extraction de caractéristiques difficile [MOK 18].



Pas de padding, Stride=1, Filtre 3*3

Zéro padding, Stride=1, Filtre 3*3

Figure 3.7 Exemple de padding et stride dans la couche de convolution [DAH 20].

B. Couche de correction (ReLU)

Souvent, il est possible d'améliorer l'efficacité du traitement en introduisant entre les couches de traitement une couche qui va opérer une fonction mathématique (fonction d'activation) sur les signaux de sortie [MER 20b]. La correction se fait notamment par les fonctions d'activations suivantes :

- La correction par tangente hyperbolique : $f(x) = \tanh(x)$
- La correction par la tangente hyperbolique saturante : $f(x) = |\tanh(x)|$
- La correction par la fonction sigmoïde : $f(x) = \frac{1}{1+e^x}$
- La correction par la fonction ReLU (Rectified Linear Units) : $f(x) = \max(0, x)$. Cette fonction remplace chaque valeur négative d'un neurone par un 0, ce qui permet d'augmenter les propriétés non linéaires de la fonction de décision et de l'ensemble du réseau sans affecter les champs récepteurs de la couche de convolution.

En pratique, la correction Relu est préférable, car elle accélère le processus d'entraînement, sans faire une différence significative à la généralisation de précision. La figure 3.8 présente le principe de fonctionnement de la fonction ReLU.

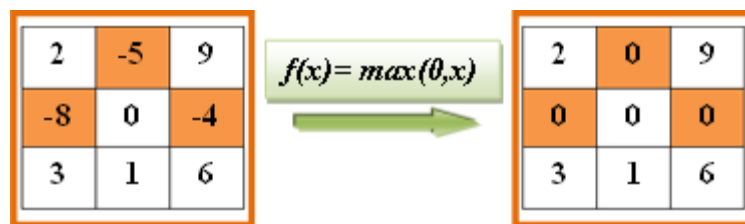


Figure 3.8 Fonction ReLU

La sortie de cette couche est une carte de caractéristiques rectifiées, tout en préservant la taille de la couche précédente.

C. Couche de pooling

La couche de pooling est une opération de sous-échantillonnage typiquement appliquée après une couche convolutive. Elle est chargée de réduire progressivement la taille spatiale

des cartes de caractéristiques. Cela permet de réduire la quantité de paramètres et de calcul dans le réseau ainsi que l'invariance aux petites translations, ce qui à la fois raccourcit le temps d'entraînement et empêche le sur-apprentissage (over fitting).

Les couches Pooling sous-échantillonnent chaque carte de caractéristiques indépendamment, en réduisant la hauteur et la largeur, tout en laissant la dimension de profondeur inchangée [DAH 20]. Le sous-échantillonnage consiste à appliquer un noyau de taille $n \times n$ sur la carte d'activation en le faisant glisser avec un *pas* préalablement défini (le *pas* est généralement égal à la taille du noyau n pour éviter le phénomène de chevauchement). En pratique, on utilise souvent des noyaux de petite taille pour ne pas perdre trop d'informations. Les choix les plus communs sont des cellules adjacentes de taille 2×2 pixels appliqués avec un *pas* de 2 réduisant ainsi la dimension de l'entrée par deux. On obtient en sortie le même nombre des cartes de caractéristiques qu'en entrée, mais celles-ci sont bien plus petites.

Il existe différents types de pooling [DAH 20b] :

- Le *max pooling* : renvoie la valeur maximale de la partie de l'image couverte par le noyau. Il permet de supprimer le bruit et de garder les caractéristiques détectées. C'est le type le plus communément utilisé, car il est rapide à calculer et permet de simplifier efficacement l'image.
- Le *mean pooling* ou *average pooling* : Chaque opération de pooling sélectionne la valeur moyenne de la surface couverte par le noyau. Ce type de pooling effectue simplement une réduction de la dimensionnalité de la carte des caractéristiques.
- Le *sum pooling* : revient à calculer la somme de la sélection couverte par le noyau.

La figure 3.9 présente un exemple de calcul des différents types de pooling avec une fenêtre de taille 2×2 et un pas de 2.

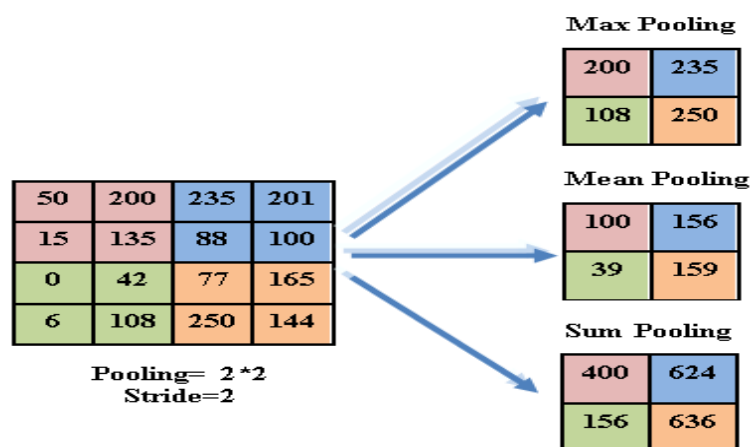


Figure 3.9 Exemples de calcul des différents types de pooling avec un filtre 2×2 et un pas de 2.

D. Couche entièrement connectée (FC : Fully Connected layer)

La couche entièrement connectée est un perceptron multicouche traditionnel qui utilise généralement, la fonction d'activation softmax dans le vecteur de sortie afin d'ajouter la non-linéarité. Le terme '*entièrement connecté*' implique que chaque neurone dans la couche précédente est connecté à chaque neurone de la couche suivante. Leurs fonctions d'activations peuvent donc être calculées avec une multiplication matricielle suivie d'un décalage de polarisation [BUY 16].

La sortie des couches de convolution et de pooling représente des caractéristiques de haut niveau de l'image d'entrée. Le but de la couche entièrement connectée est de pouvoir utiliser ces caractéristiques pour classer l'image d'entrée dans différentes classes en fonction de l'ensemble de données d'apprentissage.

Pour éviter le sur-apprentissage causé par un nombre important de paramètres, les couches entièrement connectées sont généralement suivies d'un **dropout** [SRI 14]. Ce dernier agit sur les poids de ces couches afin de déconnecter un certain nombre de neurones avec une certaine probabilité pendant une itération de l'apprentissage pour réduire le nombre de paramètres.

La somme des probabilités en sortie de la couche entièrement connectée est : $P = \sum_i \hat{y} = 1$. Cela est garanti en utilisant la fonction d'activation Softmax dans la couche en sortie de la couche entièrement connectée.

E. Couche de perte (LOSS)

La couche de perte constitue la dernière couche d'un CNN. Elle spécifie comment l'entraînement du réseau pénalise l'écart entre le signal prévu et réel. Différentes fonctions de perte adaptées à diverses tâches peuvent y être utilisées :

- La *perte par entropie croisée sigmoïde* : utilisée pour prédire N valeurs de probabilité indépendante dans [0,1]
- La *perte euclidienne* : utilisée pour régresser vers des valeurs réelles.
- La *perte Softmax* : utilisée pour prédire une seule classe parmi N classes mutuellement exclusives [MOK 18]. En effet, la fonction Softmax prend un vecteur de scores arbitraires à valeurs réelles et le réduit à un vecteur de valeurs comprises entre zéro et

un. Elle s'écrit sous la forme : $\hat{y}_i = \text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$

3.4.1.4 Apprentissage d'un CNN

L'entraînement consiste à modifier les paramètres des réseaux de manière à ce que la fonction qu'ils modélisent permette de répondre à une tâche désirée. La *descente de gradient* est un algorithme d'optimisation souvent utilisé pour entraîner des méthodes d'apprentissage

profond tel que les CNN. Cet algorithme itératif permet d'ajuster les poids et les coefficients du réseau afin de minimiser l'erreur entre la sortie estimée (prédite) du réseau et la sortie désirée [FOU 19].

Soit θ le vecteur contenant tous les paramètres d'un réseau profond. Soit $J(\theta, g(x), y)$ la fonction de coût représentant l'erreur entre la vérité terrain $g(x)$ (liée à la donnée d'entraînement x) et la valeur prédite y estimée par le réseau en utilisant les paramètres θ . La formule générale de la descente de gradient est donnée par [FOU 19] [MOU 17]:

$$\theta_{t+1} = \theta_t - \eta * \nabla_{\theta} J(\theta, g(x), y) \quad (3.3)$$

η est un hyper-paramètre appelé *pas* ou *taux d'apprentissage*. Il détermine la longueur du *pas* à faire dans la direction du gradient. Le choix d'une valeur de pas d'apprentissage satisfaisante est difficile. Un taux d'apprentissage peu adapté peut empêcher l'algorithme de converger vers le minimum. En revanche, plus le taux d'apprentissage est élevé, plus on suivra loin la direction indiquée par le gradient

En apprentissage profond, la fonction objective que l'on cherche à minimiser est souvent non convexe et non régulière. La convergence de la descente du gradient vers le minimum global n'est donc pas garantie et la convergence même vers un minimum local peut être extrêmement lente. Une solution à ce problème consiste en l'utilisation des variantes de l'algorithme de descente de gradient suivante [FOU 19] [MOU 17] :

- **Descente de gradient par batch** : Cette variante utilise toutes les données de la base d'entraînement pour calculer l'erreur globale. Ainsi, le gradient donne la direction exacte de la pente permettant de minimiser l'erreur globale.
- **Descente de gradient stochastique (SGD)** : Cette variante est applicable lorsque la fonction objective qu'on cherche à minimiser peut-être écrite sous la forme de la somme de fonctions différentiables. Ce processus est alors réalisé de manière itérative sur des exemples de données tirés aléatoirement. Chaque fonction objective minimisée de cette manière est une approximation de la fonction objective globale. Le SGD permet de mettre à jour les poids après chaque évaluation, il est donc beaucoup plus rapide pour faire évoluer les paramètres du réseau.
- **Descente de gradient par mini batch** : cette variante permet de calculer le gradient pour un sous-ensemble de n données, tirées aléatoirement dans la base d'apprentissage. Ce sous-ensemble est appelé mini-batch. L'utilisation d'un mini-batch pour calculer l'erreur permet d'obtenir un gradient plus précis, tout en gardant l'avantage de ne pas évaluer l'ensemble des données pour chaque pas de l'algorithme.

Il existe plusieurs méthodes qui accélèrent l'optimisation d'apprentissage des modèles en deep learning tels que :

- L'optimiseur d'inertie (*momentum*) : qui permet d'estimer une direction globale et d'ajouter une inertie dans le gradient.
- L'optimiseur NAG (*Nesterov accelerated gradient*) [NES 83], qui calcule le gradient après avoir avancé d'un pas dans la direction précédente,
- L'optimiseur Adagrad [DUC 11], qui adapte le pas d'apprentissage η pour chacun des paramètres indépendamment,
- L'optimiseur Adam (*Adaptive Moment Estimation*) [KIN 15], qui permet de calculer non seulement les moyennes m_t des paramètres, mais aussi leurs variances v_t .

3.4.1.5 Application des CNN pour la reconnaissance des expressions faciales :

Le but d'un CNN est d'apprendre des fonctionnalités d'ordre supérieur dans les données via des convolutions. Ils sont bien adaptés à la détection d'objets, le traitement d'image et particulièrement à la reconnaissance des visages et des expressions faciales. Plusieurs travaux utilisant les modèles CNN ont été proposés dans la littérature pour la reconnaissance des expressions faciales. Le Tableau 3.1 présente une synthèse des différents travaux adoptant les modèles CNN pour la reconnaissance des expressions faciales :

Travaux	Type d'image utilisée	Architecture du système	BDD	Taux de reconnaissance
[LOP 17]	Image normalisée	CNN	CK+, JAFFE, BU-3DFE	96,76% (CK+)
[MOH 17]	Niveau de gris	CNN	CK+	97,01%
[CAI 18]	Niveau de gris	SBN-CNN	CK+, JAFFE	95,24%, 96,87%
[JAI 19]	Image normalisée	CNN	CK+, JAFFE	95,23%, 93,24%
[YOL 19]	Image Originale	CNN	RafD	94,44%
[YAN 17]	Niveau de gris + LBP	WMDNN	CK+, JAFFE, Oulu CASIA	97,02%, 92,21%, 92,89%
[ALS 17]	Image normalisée + SIFT	CNN + SIFT + Dense SIFT	CK+, FER-2013	99,10%, 73,4%
[FEN 18]	Image normalisée + LBP TOP	Spatial CNN, Temporel CNN, Two Stream CNN	CK+	93,7% (Two Stream CNN)
[KON 19]	Niveau de gris + Improved LBP	LBP DCNN	ORL, CMU-PIE	95,20%, 95,10%
[YAN 19]	Image original + LBP	Improved CNN	CK+, JAFFE, FER-2013	95%

Tableau 3.1 Principaux travaux de la reconnaissance des expressions faciales à partir des modèles CNN.

3.4.2 Les réseaux de neurones récurrents (RNN : Recurrent Neural Networks)

Les RNNs sont un type spécifique du réseau neuronal artificiel [RUM 86] conçus pour prendre en compte l'information temporelle contenue dans chaque événement en entrée [HOC 97]. L'idée consiste à garder des informations au cours du temps à l'intérieur des couches de neurones afin de donner un contexte aux données analysées. En effet, l'état caché du RNN mémorise toutes les informations spécifiques pour prédire des séquences de données dans le futur proche [HOC 97].

Les RNNs sont constitués d'unités (neurones) interconnectés interagissant non-linéairement et pour lesquels il existe au moins un cycle dans la structure. Les unités sont reliées par des arcs (synapses) qui possèdent un poids. Pour chaque unité, la sortie de l'étape précédente est envoyée comme entrée supplémentaire à l'étape actuelle c-à-d la sortie du RNN, à un instant t , va dépendre non seulement de l'entrée à l'instant t mais également de l'état du RNN calculé à l'instant $t - 1$. La Figure 3.10 présente l'architecture classique d'un RNN.

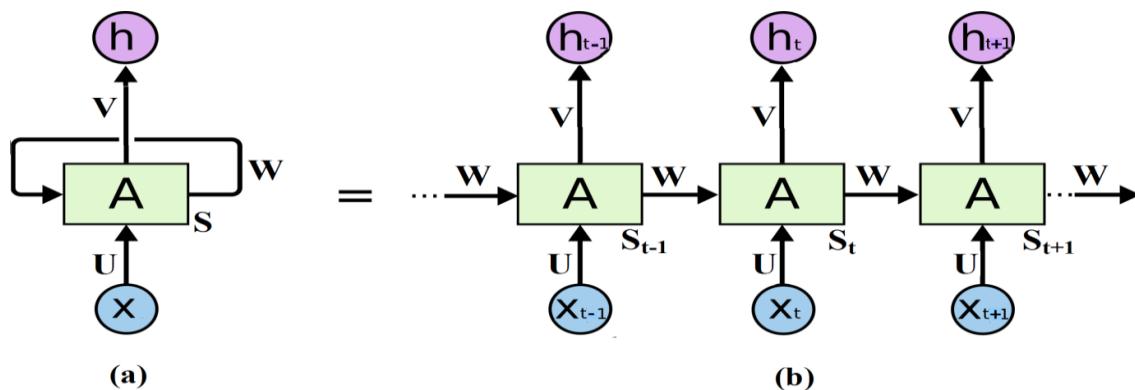


Figure 3.10 Illustration d'un RNN : (a) version récurrente et (b) version dépliée (déroulée)
[HAI 21].

Suivant cette architecture, un RNN prend en entrée une séquence d'événements $x = (x_1, x_2, \dots, x_T)$ et définit la séquence d'états cachés $s = (s_1, s_2, \dots, s_T)$ pour produire la séquence de vecteurs de sortie $h = (h_1, h_2, \dots, h_T)$.

Ou :

- x_t est l'entrée au moment t .
- $U; V$ sont les paramètres que le réseau va apprendre des données de l'apprentissage.
- s_t est l'état caché au moment t .
- h_t est la sortie au moment t .
- T est le nombre total de vecteurs d'entrée et W est la matrice de poids.

3.4.2.1 Apprentissage d'un RNN

L'algorithme de la rétro-propagation dans le temps (Back Propagation Through Time : BPTT) est utilisé pour entraîner les RNNs [RUM 86]. L'idée principale de cet algorithme est de déplier dans le temps le RNN de façon à obtenir un réseau à L couches. Le nombre L est directement lié à la taille de l'ensemble de données. Chaque couche représente une nouvelle itération dans laquelle une nouvelle donnée est présentée. L'algorithme d'apprentissage BPTT comporte trois étapes :

- **Propagation vers l'avant** : l'information circule comme dans un réseau non bouclé normal, du bas vers le haut. À chaque instant t (variant de 1 à T), l'état caché s_t est calculé en fonction de la valeur de l'état caché à l'instant précédent (s_{t-1}) ainsi que le vecteur d'entrée de l'instant t (x_t) :

$$s_i(t) = f_i(\text{net}_i(t-1)) + x_i(t) \quad \text{Avec} \quad \text{net}_i(t-1) = \sum_{j \in \text{pred}(i)} w_{ij} s_j(t-1) \quad (3.4)$$

Où f_i est une fonction non linéaire telle que : ReLu ou Hyperbolic tangent (tanh). La figure 3.11 montre l'unité de base d'un RNN utilisant la fonction Hyperbolic tangent (tanh).

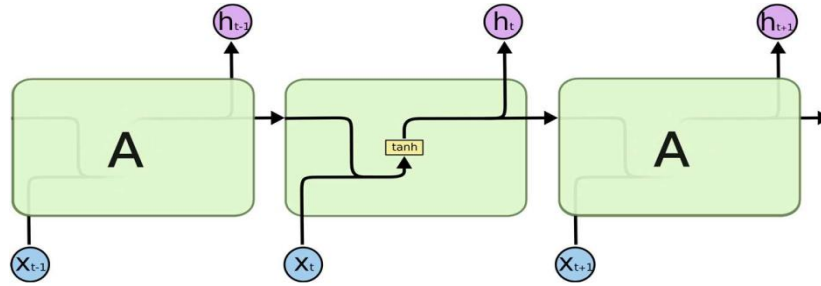


Figure 3.11 Unité de base RNN [HAI 21].

- **Propagation vers l'arrière** : Pour chaque instant $t \in [1..T]$, la variation d'un poids Δw_{ij} se calcule sous forme de la somme de toutes les variations. En appliquant la descente du gradient sur l'erreur totale, on obtient l'équation suivante :

$$\Delta w_{ij}(t_1; t_l - 1) = -\eta \frac{\delta E(t_1; t_l)}{\delta w_{ij}} \quad (3.5)$$

- **Adaptation des paramètres** : les poids sont révisés et la correction à appliquer aux poids des liens peut ensuite être déterminée à l'aide de la règle suivante:

$$\Delta w_{ij}(t_1; t_l - 1) = -\eta \sum_{\tau=t_1}^{t_l-1} \frac{\delta E(t_1; t_l)}{\delta \text{net}_i(\tau)} s_j(\tau) \quad (3.6)$$

Bien que les RNN soient construits spécifiquement pour le traitement des données séquentielles, il s'avère qu'ils sont extrêmement difficiles à entraîner pour gérer la dépendance à

long terme [BEN 94] en raison du problème d'évanescence du gradient (Vanish Gradient) ou de l'explosion du gradient (exploding gradient) lors de la phase d'apprentissage. En effet, la fonction d'activation tanh utilisée dans le RNN prend trop de valeur proche de zéro lors des opérations de dérivés pendant la descente du gradient. De plus, les RNNs simples ne sont capables de mémoriser que le passé dit proche, et commencent à oublier au bout d'une cinquantaine d'itérations environ [BEN 94].

Pour surmonter ces lacunes, de nouvelles variantes RNN telles que les *LSTM* [HOC 97] [GER 00] et les *GRU* [CHO 14] ont été introduites dans la littérature. Ces variantes permettent d'oublier certaines informations contenues dans l'état caché s et d'ainsi éviter que l'information contenue dans celui-ci ne disparaisse lorsqu'on traite des séquences très longues.

3.4.2.2 Mémoire Court et Long Terme (LSTM: Long Short Term Memory)

Les LSTM sont un type particulier de RNN, capable de modéliser et d'apprendre efficacement les dépendances à long terme. Ils ont été introduits initialement par S. Hochreiter et al [HOC 97], puis amélioré par F. Gers et J. Schmidhuber [GER 00].

Chaque unité LSTM est composée d'une mémoire interne appelée cellule et de trois portes : une *porte d'entrée* (Input Gate), une *porte de sortie* (Output Gate) et une *porte d'oubli* (Forget Gate). Une représentation graphique des LSTM est illustrée dans la figure 3.12 ci-dessous.

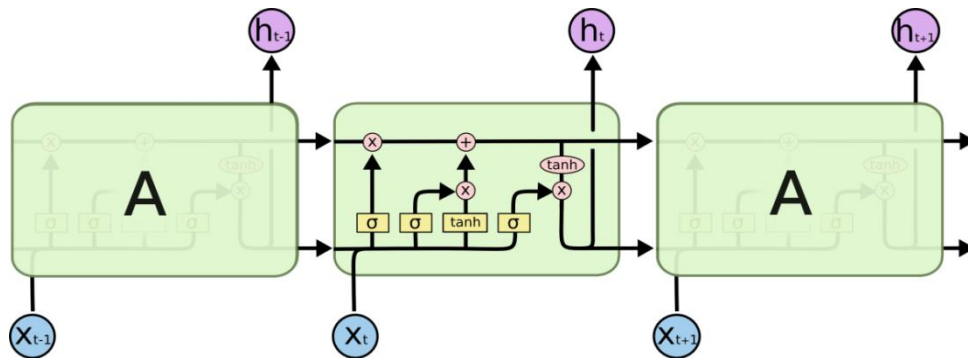


Figure 3.12 Illustration d'un LSTM [HAI 21].

- La *porte d'oubli* f (forget gate) contrôle la partie de la cellule précédente qui était utile au temps $t-1$ mais qui sera oubliée à l'instant t . Elle est définie par l'équation :

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (3.7)$$

- La *porte d'entrée* i (input gate) choisit les informations pertinentes transmises à la mémoire. Elle est définie par l'équation :

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3.8)$$

- La *porte sortie* o (output gate) contrôle l'information qui sera transmise au temps $t+1$ en fonction de la mémoire c et de la fonction d'activation. Elle est définie par les équations ci-dessous.

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (3.9)$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (3.10)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (3.11)$$

$$h_t = o_t * \tanh(c_t) \quad (3.12)$$

Où :

- σ est la fonction sigmoïde,
- f_t est le vecteur d'activation de la porte d'oubli,
- i_t le vecteur d'activation de la porte d'entrée,
- o_t le vecteur d'activation de la porte de sortie,
- $\tilde{c}_t$ est le vecteur d'activation d'entrée de cellule et c_t état de la cellule
- h_t est le vecteur de sortie de l'unité LSTM.
- W et U sont des poids, b est le vecteur biais et le symbole $*$ pour le produit de Hadamard.

La mémoire dynamique c , qui évolue en fonction de la séquence de données temporelle, permet de mémoriser les valeurs sur des intervalles de temps arbitraires, alors que les trois portes régulent le flux d'informations entrant et sortant de la cellule. La figure 3.13 présente le fonctionnement d'une cellule LSTM.

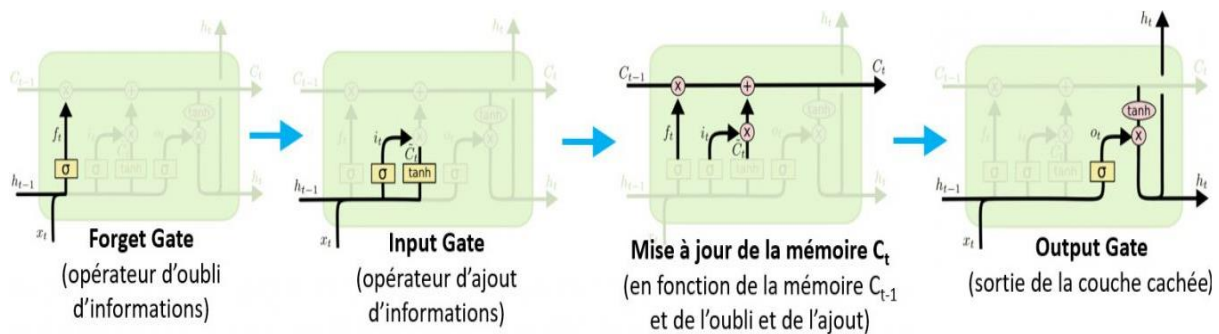


Figure 3.13 Fonctionnement d'une cellule LSTM [HAI 21].

3.4.2.3. Unité de porte récurrente (GRU : Gated Recurrent Unit)

L'unité récurrente fermée GRU a été introduite en 2014 par Cho et Al [CHO 14] pour résoudre le problème de disparition du gradient rencontré par les RNNs classiques mais aussi pour proposer une architecture avec moins de paramètres à entraîner par rapport à une LSTM. L'unité GRU est l'élément de base d'une architecture GRU. Elle dispose de deux portes et un état en sortie (voir figure 3.14) :

- *Porte de reset (Reset gate)* : sert à contrôler combien d'information passée le réseau doit oublier. Elle est définie par l'équation :

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (3.13)$$

- *Porte de mise à jour (Update gate)* : décide des informations à conserver et de celles à oublier. Elle est définie par l'équation :

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (3.14)$$

- *Sortie du réseau GRU* : L'état caché précédent (partiellement effacé par la porte de reset) est combiné avec l'entrée du réseau et normalisé par un tanh entre -1 et 1. Elle est définie par l'équation :

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t * h_{t-1}) + b_h) \quad (3.15)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (3.16)$$

Où :

- σ est la fonction sigmoïde,
- z_t est le vecteur d'activation de la porte de mise à jour,
- r_t le vecteur d'activation de la porte de réinitialisation,
- $\tilde{h}_t$ est le vecteur candidat et h_t est le vecteur output de l'unité GRU.
- W et U sont des poids, b est le vecteur biais et le symbole $*$ pour le produit de Hadamard.

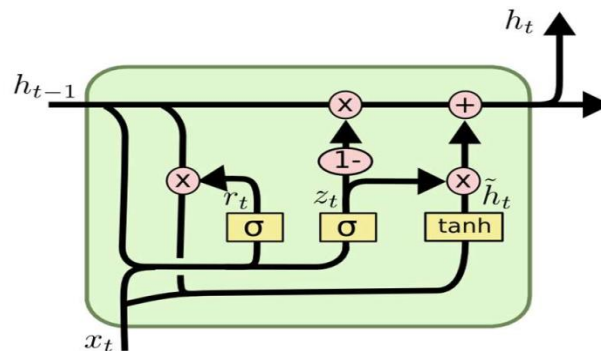


Figure 3.14 Unité de base GRU [HAI 21].

3.4.2.4 Application des RNN dans la reconnaissance des expressions faciales

Les RNN ont connu un grand succès dans de nombreuses tâches. Ils sont très largement utilisés dans les domaines du traitement automatique du langage naturel (Modélisation du langage et génération de texte, traduction automatique, etc...) et de la vision par ordinateur tels que la classification d'images, détection d'objets et la reconnaissance de visages et des expressions faciales. Le tableau 3.2 résume les différents travaux utilisant les modèles RNN pour la reconnaissance des expressions faciales.

Travaux	Architecture du système	BDD	Taux de reconnaissance
[GRA 08]	LSTM Bidirectionnel LSTM unidirectionnel	CK+	/
[YU 18]	STC-NLSTM	CK+, Oulu CASIA, MMI, BP4D	99,8%, 93.45%, 84.53%
[MOS 18]	CNN-LSTM	BioVid Emo	82%
[JAI 18]	CNN-RNN	JAFFE, MMI	94.91%, 92.07%
[KIM 19]	CNN-LSTM	MMI, CASMEII	78.61%, 60.98%
[GUO 19]	RNN hybride (GRNN)	CK+, Oulu CASIA	98.45%, 93.92%
[KAN 19]	CNN-Convolutional GRU	AFEW	/
[LIA 20]	DCBiLSTM	CK+, Oulu CASIA, MMI	99.6%, 91.07%, 80.71%
[SAL 20]	CNN-LSTM	UNBC-MSPEA	73.31%
[ZHU 22]	ResNeXt-50 + Pyramid + cascaded attention block + GRU	CK+, Oulu-CASIA, RAF-DB	99.23%, 89.29%, and 86.80%,

Tableau 3.2 Synthèse des différents travaux de la reconnaissance des expressions faciales basé sur les modèles RNN.

3.4.3 Les modèles génératives profonds :

Bien que les modèles discriminatifs (CNN, RNN, MLP,...) sont utilisés pour prédire les données du label $p(y/x)$ et de l'entrée x , les modèles génératifs décrivent comment générer les données. Ils apprennent pour faire des prédictions en utilisant la loi de Bayes [NG 02]. Voici quelques exemples de modèle génératif :

- Les machines de Boltzmann (BM)
- Les machines de Boltzmann Restreintes (RBM)
- Les réseaux de croyance profonds (DBN)
- Les réseaux antagonistes génératifs (GAN)
- Les auto-encodeurs (AE)

3.4.3.1 Les machines de Boltzmann restreintes (RBM : Restricted Boltzmann Machines)

Les RBM sont des réseaux de neurones stochastiques génératifs capables d'apprendre une distribution de probabilités à partir d'unités d'entrée. Ils ont été initialement inventés sous le nom « Harmonium » par P. Smolensky [SIM 91], puis popularisés plus récemment par les travaux de G. E. Hinton [HIN 02].

Une RBM est composée de deux couches : Une *couche d'unités visibles* et une *couche d'unités cachées*. Les variables *observables* sont représentées dans les unités visibles, tandis

que les variables *latentes* sont représentées par les unités cachées [DEN 14]. En effet, les unités visibles peuvent avoir des valeurs réelles, alors que les unités cachées fournissent des décisions binaires. Les unités sont reliées les unes aux autres par des connexions pondérées bidirectionnelles, dont les paramètres sont calculés durant l'apprentissage. Chaque unité fournit une décision en tenant compte de l'apport des unités voisines. Les poids affectés à ces connexions sont symétriques et peuvent prendre des valeurs positives ou négatives. La somme totale de tous les poids sur toutes les connexions représente l'énergie d'une RBM (voir figure 3.15.b). A noter qu'il n'y a pas de connexion entre les unités visibles, de même entre les unités cachées, les seules interactions possibles sont les connexions entre une unité cachée et une unité visible. Cette restriction est d'ailleurs à l'origine du nom des RBM, par rapport aux machines de Boltzmann standards (BM), pour lesquelles toutes les connexions sont autorisées (voir figure 3.15.a).

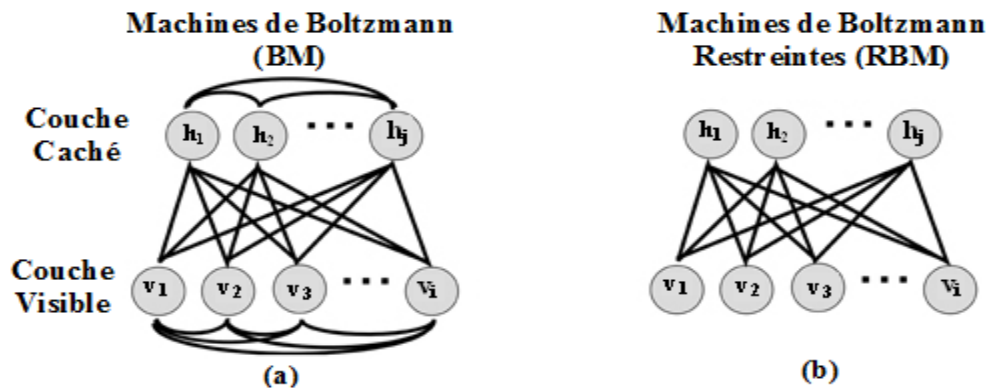


Figure 3.15 Principe de fonctionnement : (a) Machine de Boltzmann (BM), (b) Machine de Boltzmann Restreintes (RBM) [SAL 09]

Comme le montre la figure 15.b, les neurones sont limités pour former un graphe biparti dans une RBM. Supposons que v_i est l'état d'activation de l'unité visible i , et h_j est l'état d'activation de l'unité cachée j , la fonction énergétique de l'état (v_i, h_j) est alors donnée par :

$$E(v_i, h_j) = -(\sum_{i,j} w_{ij} v_i h_j + \sum_i a_i v_i + \sum_j b_j h_j) \quad (3.16)$$

- w_{ij} est le poids entre le neurone j et le neurone i ,
- a_i et b_j sont respectivement les biais des neurones v_i et h_j

Les RBMs assignent une probabilité conjointe pour chaque configuration (v_i, h_j) comme suit :

$$P(v_i, h_j) = \exp(-E(v_i, h_j)) / Z \quad (3.17)$$

Où $E(v_i, h_j)$ est la fonction d'énergie définie dans l'équation, tandis que Z indique la fonction de partition, définie par :

$$Z = \sum_i \sum_j \exp(-E(v_i, h_j)) \quad (3.18)$$

3.4.3.1.1 Apprentissage des RBM

Une RBM peut être considérée comme un cas particulier d'un champ aléatoire de Markov connue sous le nom d'un échantillonneur de Gibbs. Les chaînes de Markov sont une séquence d'états de variables aléatoires qui sont transférés en utilisant des probabilités conditionnelles connues.

L'échantillonneur Gibbs est adopté pour entraîner une RBM, en partant d'un état aléatoire dans une couche et en effectuant un échantillonnage Gibbs, nous pouvons générer des données à partir d'une RBM. Une fois que les états des unités d'une couche sont donnés, toutes les unités des autres couches seront mises à jour. Ce processus de mise à jour se poursuivra jusqu'à ce que la distribution d'équilibre soit atteinte. Il convient de noter que le processus d'entraînement sera plus efficace si l'on utilise l'algorithme de *contrastive divergence (CD)* de Hinton et al. [HIN 02], qui se base sur un schéma d'auto-encodage, et qui vise à minimiser la fonction d'énergie exprimée par l'équation précédente.

3.4.3.1.2 Application des RBM pour l'analyse des expressions faciales

Aujourd'hui, les modèles RBMs jouent un rôle important dans diverses applications telles que la modélisation de sujets, la régression, la réduction de la dimensionnalité, le traitement d'image et particulièrement la reconnaissance de visage et des expressions faciales. Le tableau 3.3 ci-dessus représente quelques travaux utilisant les modèles RBM pour la reconnaissance des expressions faciales :

Travaux	Architecture du système	BDD	Taux de reconnaissance
[WU 13]	RBM	CK+, MMI, ISL, ASL	/
[WAN 14]	Gaussian RBM+ rKCCA	FRGCv2.0	96%
[NIE 15]	LRBM	CK+	99.6%
[ELA 16]	Spatio - temporal RBM	CK+, SFEW, MMI	95.66%, 81.63%, 46.36%
[ELF 18]	RBM	MMI	/
[NAG 19]	msDBM	Radboud, CAFE	85.9%, 56%
[WAN 19]	CNN-CRBM	JAFFE, FER2013, SFEW, RAF-DB.	99.2%, 73.75%, 96.09%, 92.97%

Tableau 3.3 Les différents travaux utilisant les modèles RBM pour la reconnaissance des expressions faciales.

3.4.3.2 Les réseaux de croyance profonds (DBN : Deep Belief Networks) :

Les DBN sont des modèles génératifs qui ont été introduits par G. Hinton [HIN 02]. Ils sont composés de plusieurs couches de variables stochastiques latentes (unités cachées). Les variables des unités cachées sont typiquement binaires.

Les DBN peuvent être considérés comme un ensemble de machines de Boltzmann restreintes (RBM) empilées les unes sur les autres [HIN 02]. Toutes les deux couches adjacentes forment une RBM. Chaque couche RBM utilise les neurones cachés de la couche RBM précédente comme entrée, donc chaque couche cachée du RBM représente la couche visible de la RBM suivante comme la montre la figure 3.16. L'architecture profonde avec plusieurs couches peut alors extraire une représentation hiérarchique profonde des données d'apprentissage.

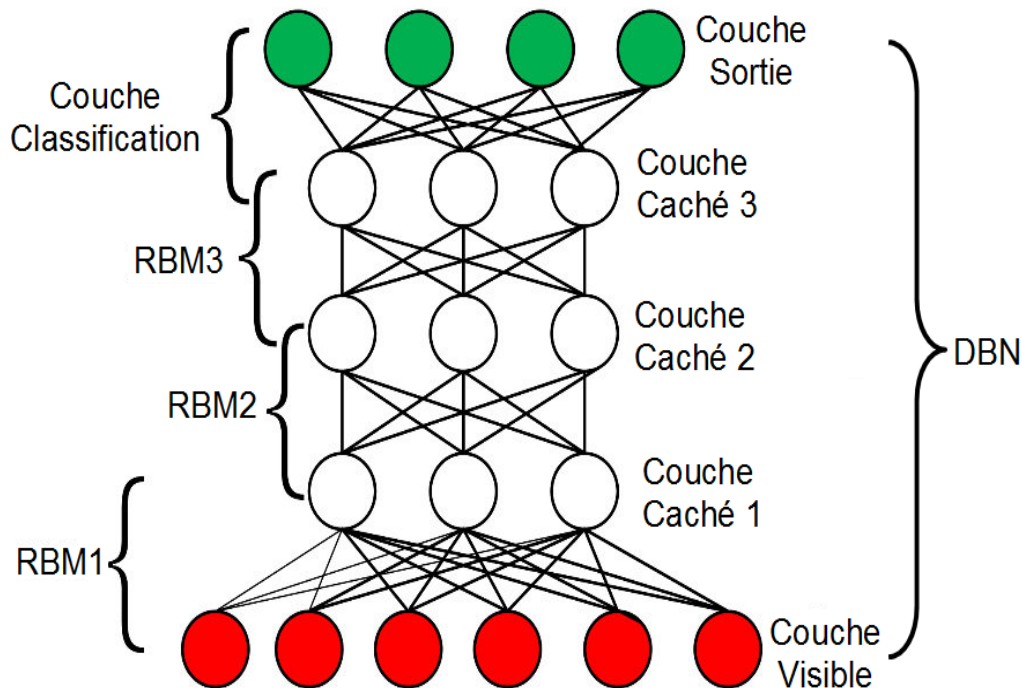


Figure 3.16 Architecture générale des DBN

3.4.3.2.1. Entraînement des DBN

Les différentes couches de RBM dans un DBN sont entraînées de manière séquentielle. En effet, les RBM inférieures sont entraînées en premier, puis les plus élevées. Une fois que les caractéristiques sont extraites par la RBM supérieure, elles sont propagées vers les couches inférieures.

Le processus d'entraînement d'un DBN peut être divisé en deux étapes : la première phase, appelée *pré-entraînement*, effectue un apprentissage non supervisé de chaque couche RBM séparément de bas en haut pour l'extraction des caractéristiques ; la deuxième phase, appelée *ajustement*, utilise un algorithme ascendant basé sur un apprentissage supervisé pour ajuster les paramètres du réseau ultérieurement (couche classification).

L'algorithme *contrastive divergence (CD)* introduit dans la section précédente peut être utilisé pour pré-entraîner un DBN [HIN 06]. Cependant, les performances sont généralement insatisfaisantes. Pour surmonter ce problème, l'algorithme *gourmand couche par couche*

(*greedy layer-by-layer*) a été introduit. Cet algorithme élimine une grande partie des problèmes généralement rencontrés avec l'algorithme de rétro-propagation de gradient et optimise les poids d'un DBN à la taille et à la profondeur du réseau [HIN 06].

Donc pour entrainer le réseau DBN, l'apprentissage *greedy layer-by-layer* est d'abord effectué. Ensuite, l'apprentissage par l'algorithme de rétro-propagation est appliqué à l'ensemble du réseau afin d'apprendre à distinguer les étiquettes de classes.

3.4.3.2.2 Application des DBN pour la reconnaissance des expressions faciales

Les DBN ont été utilisés pour générer des séquences vidéo, des données de capture de mouvement et reconnaître des images tels que les visages, les expressions faciales...etc. Le tableau 3.4 résume quelques travaux adoptant les modèles DBN pour la reconnaissance des expressions faciales :

Travaux	Architecture du système	BDD	Taux de reconnaissance
[LIU 14]	Boosted DBN	CK+, JAFFE	96.7%, 68%
[WU 17]	DBN	JAFFE	87.62%
[HAO 17]	DWLD+ DBN	JAFFE	92.66%
[VAS 18]	DBN	JAFFE	95.28%
[KUR 19]	DBN	CK+, RaFD, MMI	98.57 %, 94.50 %, 98.75 %

Tableau 3.4 Les différents travaux adoptant les modèles DBN pour la reconnaissance des expressions faciales.

3.4.3.3 Les réseaux antagonistes génératifs (GAN : Generative Adversarial Networks)

Les GANs sont des modèles génératifs non supervisés qui permettent de modéliser et d'imiter une distribution de probabilités à partir de données [GOO 14]. Un GAN est basé sur la compétition entre deux réseaux de neurones profonds : un *réseau générateur*, et un *réseau discriminateur*. Le réseau générateur produit des données candidates à partir d'un code tiré d'une distribution aléatoire connue. Il permet donc de créer des contenus synthétiques les plus réalistes possibles. Le réseau discriminateur va quant à lui apprendre à distinguer les données générées des données réelles. Il cherche à déterminer si les données présentées sont issues de la vraie distribution ou bien elles sont le résultat du générateur.

Notons par z le bruit aléatoire, et $G(z)$ l'instance générée par le générateur G à partir de ce bruit, notons aussi x une instance des données (de l'ensemble d'apprentissage, ou générée) et $D(x)$ la probabilité estimée par le discriminateur que x soit une instance générée (créée par le

générateur G et n'appartienne pas à l'ensemble d'apprentissage). La figure 3.17 illustre le fonctionnement d'un réseau GAN.

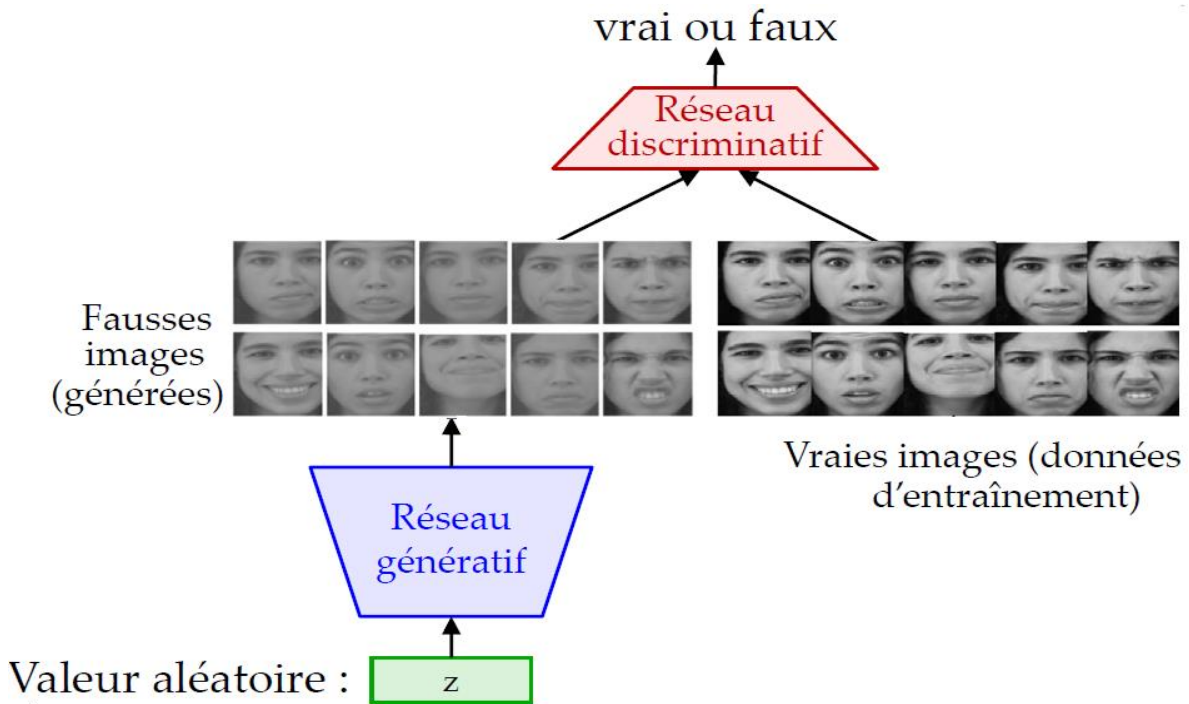


Figure 3.17 Principe de fonctionnement d'un réseau GAN [GIG 14].

3.4.3.3.1 Apprentissage des GAN :

Lors de l'apprentissage, les deux réseaux sont entraînés simultanément. L'entraînement simultané permet d'amener le générateur à améliorer de plus en plus la qualité des contenus générés, alors que le discriminateur doit trouver des détails de plus en plus fins pour détecter les contenus synthétiques [GIG 14]. Le générateur et le discriminateur apprennent l'un contre l'autre et jouent à un jeu minimax à deux joueurs :

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{data(x)}} [\log D(x)] + \mathbb{E}_{z \sim P_z(z)} [\log(1 - D(G(z)))] \quad (3.19)$$

Où : $V(D, G)$ représente la fonction objective d'un GAN.

D'une part, le générateur G cherche à tromper le discriminateur D , afin de minimiser V (trouver les paramètres de G qui produisent un $x' = G(z)$ tel que D lui associe le score 1 "réel").

$$\min_G \max_D V(D, G) = \underbrace{\mathbb{E}_{x \sim p_{data(x)}}}_{\text{Vraies données}} [\log \overbrace{D(x)}^{\downarrow 0}] + \underbrace{\mathbb{E}_{z \sim p_z(z)}}_{\text{Fausses données}} [\log(1 - \overbrace{D(G(z))}^{\uparrow 1})] \quad (3.20)$$

D'autre part, le discriminateur change ses poids afin de maximiser V (trouver les paramètres de D qui associent le score 1 aux données réelles et 0 aux données produites par G).

$$\min_G \max_D V(D, G) = \underbrace{E_{x=p_{data}(x)}}_{\substack{\text{Vraies} \\ \text{données}}} [\log \overbrace{D(x)}^{\uparrow 1}] + \underbrace{E_{z=p_z(z)}}_{\substack{\text{Faussees} \\ \text{données}}} [\log(1 - \overbrace{D(G(z))}^{\downarrow 0})] \quad (3.21)$$

La compétition qui a lieu entre ces deux réseaux adversaires joue un rôle essentiel dans l'entraînement, l'un ne pouvant être entraîné sans l'autre. Le générateur finalement obtenu permettra d'échantillonner des contenus et pourra être utilisé dans diverses applications telles que l'augmentation des données ou la modification de contenus [VIE 19] (super résolution, inpainting, ... etc). Le discriminateur permettra d'obtenir une représentation particulièrement robuste d'un contenu et pourra notamment être affiné pour résoudre de nouvelles tâches relatives à ce type de contenu [VIE 19].

En pratique, il est souvent conseillé d'entraîner chaque réseau d'un GAN contre un adversaire statique, autrement dit, d'entraîner le discriminateur seul, puis le générateur seul. De plus, il est souvent déconseillé d'utiliser un discriminateur trop performant pour entraîner un générateur faible et inversement. L'idéal est que les deux adversaires aient des performances comparables et puissent apprendre ensemble.

3.4.3.3.2 Applications des GAN dans la reconnaissance des expressions faciales

Depuis leurs introduction en 2014 par Ian Goodfellow [GOO 14], les GANs ont été utilisés avec beaucoup de succès dans plusieurs domaines applicatifs tels que la génération automatique des images photo réalistes, la sécurité informatique, les patterns de mouvement dans les vidéos, etc. Pour la reconnaissance des expressions faciales, plusieurs travaux utilisant les modèles GAN ont été proposés comme il est indiqué dans le tableau 3.5.

Travaux	Architecture du système	BDD	Taux de reconnaissance
[LAI 18]	GAN	Multi-PIE, BU-3DFE	87.08%, 73.13%
[WAN 18a]	CNN-GAN	JAFPE, MMI, CK+	59.62%, 61.86%, 76.58%
[LEE 19]	RcGAN	CK+, MMI	97.93%, 82.86%
[POR 20]	GAN + augmentation des données	CK+, ExpW	83.3%, 40.4%,
[CAI 21]	IF-GAN	BU-3DFE, CK+, MMI, RAF-DB	85.25%, 97.52%, 75.48%, 88.33%
[WAN 21]	GAN	JAFPE, CK+, FER2013	96.6%, 95.6%, 72.8%,
[LU 21]	GAN	Multi-PIE	90.25%

Tableau 3.5 Synthèse des travaux utilisant les modèles GAN pour la reconnaissance des expressions faciales.

3.4.3.4 Les Auto-Encodeurs (AE : Auto-Encoder)

Les auto-encodeurs sont des réseaux de neurones artificiels utilisés pour apprendre des représentations efficaces à partir de données brutes de manière non supervisée [HIN 12]. En effet, pour un ensemble donné d'entrées, un auto-encodeur tente de trouver une représentation compressée et distribuée. Par conséquent, les auto-encodeurs peuvent être considérés comme effectuant une réduction de dimension (ou une compression avec perte). La sortie du réseau auto-encodeur est une reconstruction des données d'entrée sous la forme la plus précise [ARN 11].

Les auto-encodeurs consistent essentiellement en deux parties séparées [BAL 11] : un encodeur et un décodeur. *L'encodeur* est un réseau de neurones qui compresse l'information contenue dans l'entrée dans une couche cachée $\mathbf{z}$ (transforme les données d'entrée $\mathbf{x}$ de l'espace R_n dans un nouvel espace R_m avec $n > m$). Inversement, *le décodeur* prend les attributs latents de la sortie de l'encodeur ($\mathbf{z}$) et essaie de reconstruire l'image (retransforme les données $\mathbf{x}$ de l'espace R_m dans l'espace R_n).

Le *goulot d'étranglement* ou le *codage* est le vecteur latent émis par le codeur (le vecteur $\mathbf{z}$ dans l'image) et sur-échantillonné par le décodeur. Il contient les attributs latents produits par le décodeur. Le vecteur $\mathbf{z}$ est de dimension beaucoup plus faible que le signal d'entrée, ce qui implique que seules les informations les plus significatives sont conservées. A noter que le nombre de neurones dans la dernière couche du décodeur est égal à la dimension des données d'entrée du réseau.

Notons $\mathbf{x}$ l'entrée d'un auto-encodeur et $Q_{\phi i}$ la fonction d'encodage de la couche i pour un ensemble de paramètres ϕ donné. Notons également $\mathbf{x}'$ la reconstruction de l'entrée $\mathbf{x}$ par le décodeur et $P_{\theta i}$ la fonction de décodage liée à la couche i . Un exemple d'auto encodeur est montré dans la Figure 3.18.

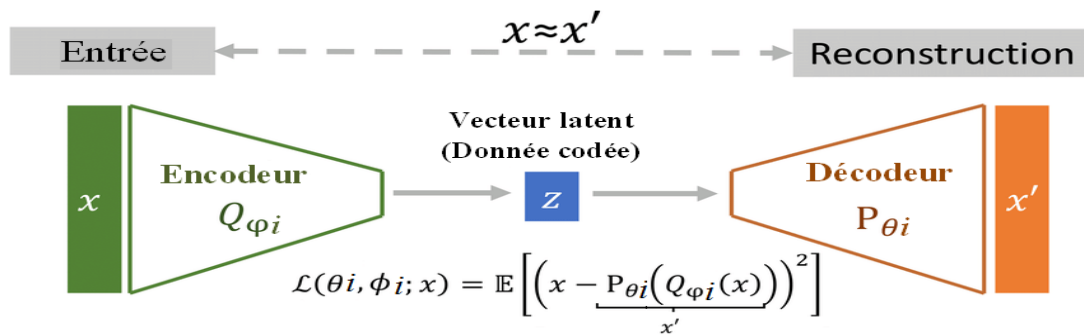


Figure 3.18 Représentation schématique du principe d'un auto-encodeur : la sortie $\mathbf{x}'$ est la reconstruction de l'entrée $\mathbf{x}$ à partir de la couche cachée $\mathbf{z}$ [FER 19].

3.4.3.4.1 Apprentissage des auto-encodeurs

Dans un auto-encodeur, l'apprentissage du codage s'appuie principalement sur la rétro-propagation où le réseau tente de réduire l'erreur de reconstruction entre l'entrée et sa représentation de la couche caché (en utilisant des poids inverses).

De nombreuses variantes des auto-encodeurs existent, parmi lesquelles on peut citer les Auto-Encodeur Variationnel (VAE : Variational Auto-Encoder) [KIN 14] et les Auto-Encodeur Débruitant (DAE : Denoising Auto-Encoder) [VIN 10], etc.

3.4.3.4.2 Application des auto-encodeurs pour la reconnaissance des expressions faciales

Les auto-encodeurs sont utilisés à des fins telles que la découverte de produits pharmaceutiques, la prédiction de popularité, le traitement d'images et particulièrement la reconnaissance des expressions faciales (Voir figure 3.19).

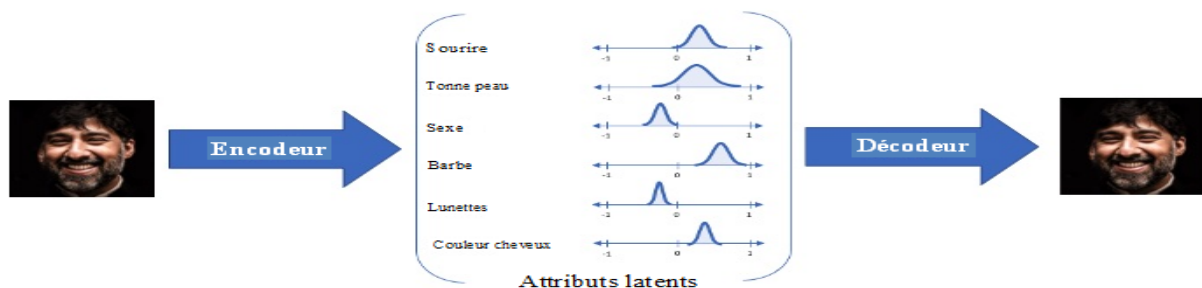


Figure 3.19 Exemple d'un auto-encodeur [GEN 21].

En utilisant les attributs latents (sourire, sexe, barbe, lunettes,...), ce réseau est capable de *générer une image* en respectant les attributs donnés. Les travaux adoptant les modèles des auto-encodeurs pour l'analyse des expressions faciales sont présentés dans le tableau 3.6 ci-dessus :

Travaux	Architecture du système	BDD	Taux de reconnaissance
[LIU 14]	Flux optique+ Spare AE (SAE)	<i>CK+</i>	92.03%
[HUA 15]	Spare Auto-Encoder (SAE)	<i>JAFFE</i>	94,01%
[SUN 17]	Binarized Auto-encoders (BAEs) + Stacked Binarized AE (SBAEs)	<i>SFEW</i>	49.25%
[RUI 17]	CNN+ Stacked Convolutional AE (SCAE)	<i>KDEF</i>	92.52%
[USM 17]	HOG + Autoencoder (AE)	<i>CK+</i>	99.6%
[ZEN 18]	Deep Spare AE (DSAE)	<i>CK+</i>	95.79%
[LAK 21]	HOG+ LBP+ Stacked AE (SAE)	<i>CK+, JAFFE</i>	97.66%, 90.83%

Tableau 3.6 Les différents travaux adoptant les modèles des auto-encodeurs pour l'analyse des expressions faciales

3.5 Conclusion

Dans ce chapitre, nous avons présenté et détaillé le principe général de l'apprentissage profond, ses différentes architectures et comment il se différencie des algorithmes de machine learning traditionnelles. Nous avons également présenté les trois familles majeures de deep learning à savoir les CNN, les RNN ainsi que les modèles génératifs (RBM, DBN, GAN, auto-encodeurs.) tout en montrant leurs importances dans le domaine de reconnaissance des expressions faciales. Dans le chapitre suivant, nous présenterons notre approche et le système proposé pour la reconnaissance des expressions faciales.

Partie II : Contributions

Sommaire

4.1 Introduction	140
4.2 Architecture générale du système proposé	140
4.3 Système de Reconnaissance des Expressions Faciales (SREF)	141
4.3.1 Approche basée sur les caractéristiques géométriques pour la reconnais-	
sance des expressions faciales	141
4.3.2 Approche basée sur les caractéristiques d'apparences pour la reconnais-	
sance des expressions faciales	148
4.3.3 Approche basée sur l'apprentissage profond pour la reconnaissance des	
expressions faciales	156
4.4 Conclusion	160

4.1. Introduction

Dans ce chapitre, nous présenterons notre système de détection de l'état émotionnel de l'apprenant durant sa navigation dans un environnement d'apprentissage en ligne à travers ses expressions faciales. Le système proposé est conçu en utilisant trois sous-système différents : Le premier sous-système extrait les *descripteurs géométriques* par une méthode qui code les mouvements faciaux en se basant sur un ensemble de points. La classification des émotions se fait par la méthode du plus proche voisin (KPPV). Dans le deuxième sous système, un nouveau *descripteur d'apparence* nommé 'Local Directional Double Ternary Coding Patterns (LDDTCP)' est proposé. La classification dans ce cas est assurée par un support de vecteurs machine (SVM). Le troisième sous-système combine quatre réseaux de neurones convolutifs 'CNN' (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et CNN_{LDP}) en se basant sur le calcul de quatre descripteurs de texture LBP, LGC, GDP et LDP. Les trois sous-systèmes comportent les étapes suivantes :

1. Une phase de localisation des visages, d'élimination de l'arrière-plan et de prétraitement.
2. Extraction des caractéristiques faciales.
3. Classification et reconnaissance des émotions.

4.2. Architecture générale du système proposé

L'objectif de notre travail est de proposer un nouveau système pour détecter l'état émotionnel de l'apprenant durant sa navigation dans un environnement d'apprentissage en ligne à travers ces expressions faciales. La figure 4.1 présente le schéma général du système proposé [CHE 19a].

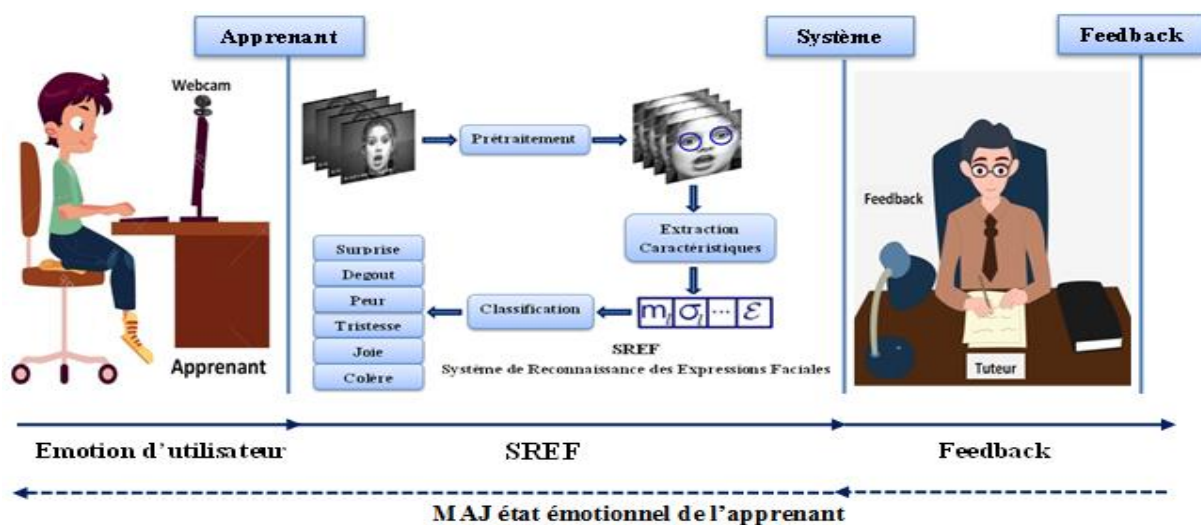


Figure 4.1 Schéma général du système proposé

Les apprenants sont placés face à une webcam et connectés au système d'enseignement à distance (Système Tutoriel Intelligent) pour apprendre un sujet donné. Une fois l'apprenant est connecté, le système de reconnaissance des expressions faciales (SREF) commence à analyser les émotions ressenties par l'apprenant. A partir de l'observation et de l'identification de l'état émotionnel de l'apprenant, le tuteur peut entreprendre des actions qui régularisent la qualité de l'apprentissage, puis les apprenants mettent à jour leurs états émotionnels.

4.3. Système de Reconnaissance des Expressions Faciales (SREF)

Le système d'analyse des expressions faciales consiste à extraire les traits du visage et à classer les déformations faciales dans l'une des six émotions universelles : joie, tristesse, colère, dégoût, surprise, peur, et l'émotion neutre.

Dans le cadre de cette thèse, notre système de reconnaissance d'émotions à partir des expressions faciales est présenté en utilisant trois approches différentes : Approche basée sur les *descripteurs géométriques*, approche basée sur les *caractéristiques d'apparences* et approche basées sur l'*apprentissage profond*.

4.3.1. Approche basée sur les caractéristiques géométriques pour la reconnaissance des expressions faciales

Dans cette approche, on s'intéresse à des points caractéristiques particuliers du visage afin d'en étudier les déformations ou les mouvements faciaux. Ces déformations sont codées soit en utilisant directement les coordonnées des points, soit en utilisant les distances entre les points. La figure 4.2 illustre l'architecture globale du système de reconnaissance des expressions faciales basé sur les descripteurs géométriques [CHE 19b]:

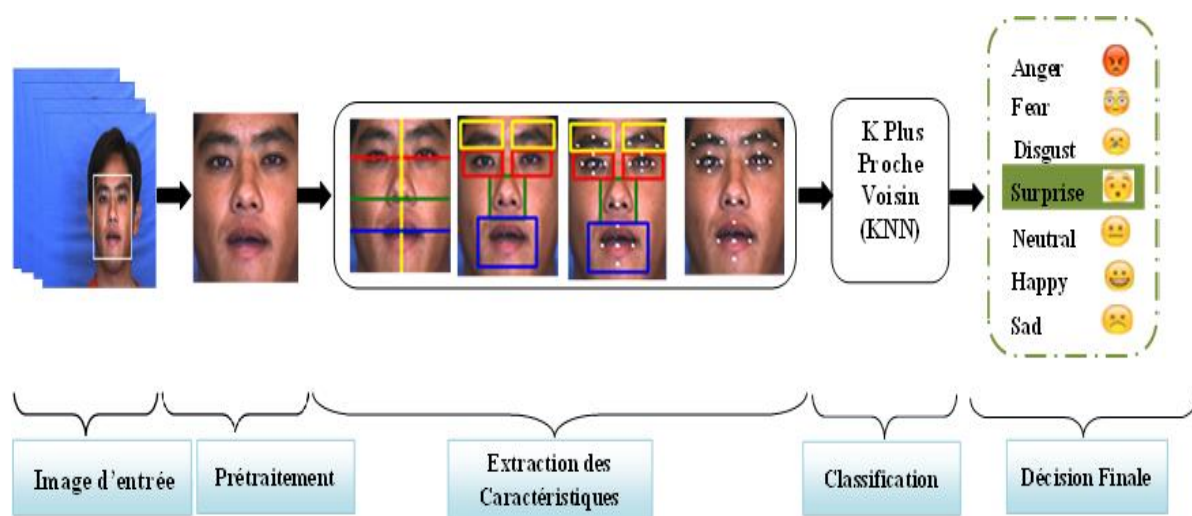


Figure 4.2 Architecture globale du SREF basé sur les caractéristiques géométriques.

Tout d'abord, les images capturées sont prétraitées et les visages sont détectés. Ensuite, vient l'étape de localisation des régions d'intérêts (ROI : Region Of Interest) tels que les yeux, le nez et la bouche. A partir des différentes fenêtres des ROIs localisés, un ensemble de points caractéristiques sont extraits permettant de décrire les mouvements faciaux. Ces caractéristiques représentent des entrées pour le k plus proche voisin (KNN : K Nearest Neighbor) facilitant ainsi l'étape de reconnaissance.

4.3.1.1. Détection du visage

Il existe différentes méthodes de détection de visage décrites dans la littérature. Ces méthodes présentent une différence dans leurs complexités, leurs performances, le type et la nature des images. Pour réaliser cette première étape, nous appliquons un algorithme de détection de visage basé sur la couleur de peau en utilisant l'espace RGB [BEN 07] [GAS 06], car il offre de bonnes performances (Figure 4.3.b).

Un pixel est dit pixel de peau, si les composantes R (Rouge), V (Vert) et B (Bleu) de la couleur vérifient les conditions suivantes (segmentation et binarisation) :

$$\left\{ \begin{array}{l} \text{Si } (R > 95) \text{ et } (V > 40) \text{ and } (B > 20) \text{ et} \\ \quad ((\text{Max } [R, V, B] - \text{Min } [R, V, B]) > 15) \text{ et} \\ \quad (\text{Abs } (R - V) > 15) \text{ and } (R > V) \text{ and } (R > B) \\ \text{Alors} \quad \quad \quad \text{Skin pixel (1)} \\ \text{Sinon} \quad \quad \quad \text{Skin pixel (0)} \end{array} \right. \quad (4.1)$$

Après la sélection des pixels de peau, on obtient une image binaire dont les différentes valeurs de pixel sont égales à 0 ou 1. L'image obtenue est suivie par un prétraitement morphologique (érosion) (Figure 4.3.d) et un filtre médian (Figure 4.3.c) pour éliminer les fausses détections.

Le principe de détermination des coordonnées du visage est de réaliser une projection horizontale (HPI) et une projection verticale (VPI) de l'image binaire.

$$HPI = \sum_{y=1}^m I(x, y) \quad (4.2)$$

$$VPI = \sum_{x=1}^n I(x, y) \quad (4.3)$$

On cherche les positions des deux premières valeurs qui sont différentes du zéro dans les deux côtés des vecteurs de projection HPI et VPI. Ces positions sont les coordonnées du visage dans l'image d'origine (Figure 4.3.f).

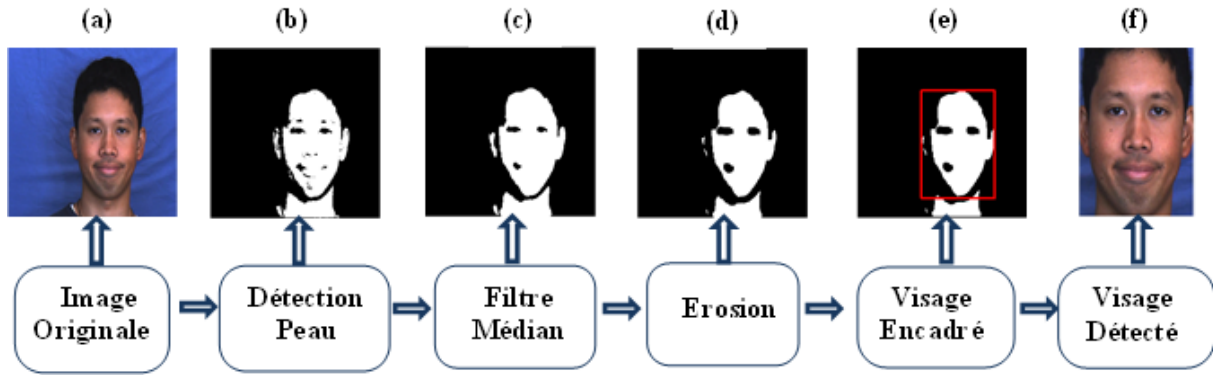


Figure 4.3 Détection du visage [CHE 19a] [CHE 19b].

4.3.1.2. Extraction des caractéristiques

Après la détection de visage, nous allons extraire les points caractéristiques faciaux dans le cadre qui englobe le visage. Il existe de nombreuses techniques consacrées à l'extraction des caractéristiques faciales. Notre choix s'est porté sur la méthode développée dans Abdat et al [ABD 07]. Cette méthode utilise le suivi du gradient pour localiser chaque composante faciale. Elle permet ainsi d'extraire des points caractéristiques nécessaires pour former un modèle géométrique. La méthode de Abdat et al [ABD 07] présente une robustesse face aux petits angles de rotation, aux conditions d'éclairage, à la couleur de peau ou aux accessoires (moustache, lunettes, barbes).

A. Localisation des caractéristiques faciales

Pour localiser les composantes du visage à l'aide du modèle géométrique, nous utilisons les étapes suivantes :

- (a) Calcul du niveau de gris de l'image I :

$$I = R * 0.299 + V * 0.587 + B * 0.114 \quad (4.4)$$

Où I représente le niveau de gris pour un pixel donné, R, V, B sont les composantes de couleur du pixel.

- (b) Calcul du gradient dans la direction horizontale :

$$\nabla I_y = \frac{\partial I}{\partial y} \vec{j} \quad (4.5)$$

Où y et J sont respectivement l'indice et la direction horizontale, I est l'image en niveaux de gris.

- (c) Localisation de l'axe des yeux (N_{eyes}): L'axe des yeux est déterminé par le maximum de la projection horizontale qui a un gradient horizontal élevé. Cet axe correspond à plusieurs transitions : peau vers l'œil, blanc de l'œil vers l'iris, l'iris vers la pupille.

$$HPI_y = \sum_{y=1}^m \nabla I_y(x, y) \quad (4.6)$$

$$N_{eyes} = \max(HPI_y) \quad (4.7)$$

- (d) Localisation de l'axe médian (N_{centre}): l'axe médian permet de couper le visage vertical en deux parties symétriques. Pour déterminer l'axe médian, nous cherchons la position du niveau de gris le plus élevé sur l'axe des yeux.

$$N_{center} = \max(\nabla I_y(N_{eyes}, y) \quad (4.8)$$

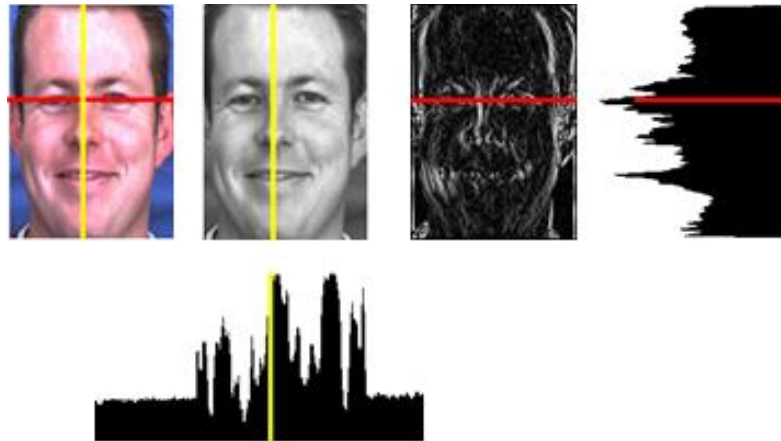


Figure 4.4 Localisation de l'axe des yeux et l'axe médian [CHE 19a] [CHE 19b].

- (e) Localisation de l'axe des narines : L'axe des narines correspond à la ligne qui a le plus fort gradient situé au-dessous de l'axe des yeux. Pour localiser l'axe des narines, on cherche le gradient le plus élevé dans une fenêtre étroite, au-dessous de l'axe des yeux et autour de l'axe médian avec une largeur de Δx (voir figure 4.5).
- (f) Localisation de l'axe de la bouche : En se basant sur le même principe que l'étape précédente, l'axe de la bouche correspond à la ligne qui a le plus fort gradient situé au-dessous de l'axe des narines (voir figure 4.5).

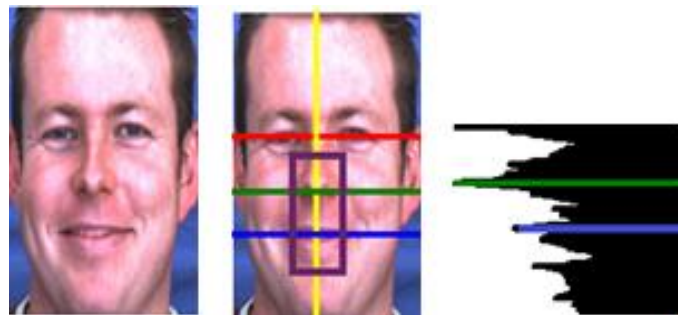


Figure 4.5 Localisation de l'axe des narines et de la bouche.

- (g) Application du modèle géométrique : Une fois que l'axe des yeux et de la bouche sont localisés, nous appliquons le modèle géométrique de visage donné dans [SHI 04] [LAS 02] (voir figure 4.6).

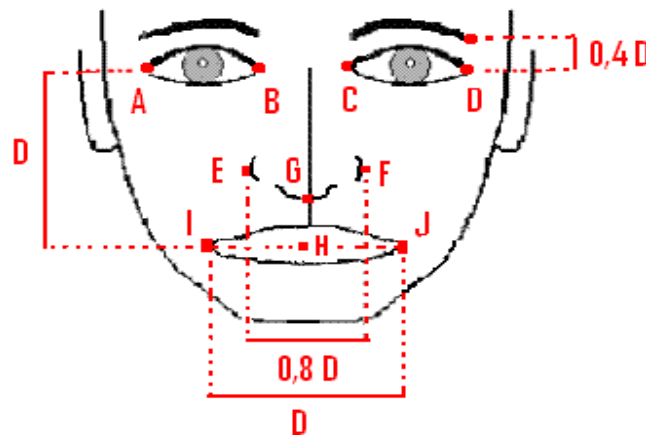


Figure 4.6. Le modèle géométrique

Ce modèle suppose que:

- La distance verticale entre les yeux et le centre de la bouche est D .
- La distance verticale entre les yeux et le centre des narines est $0,6D$.
- La largeur de la bouche est D .
- La largeur du nez est $0,8D$.
- La distance verticale entre les yeux et les sourcils est $0,4D$.

La figure 4.7 montre les résultats obtenus pour l'extraction des caractéristiques faciales

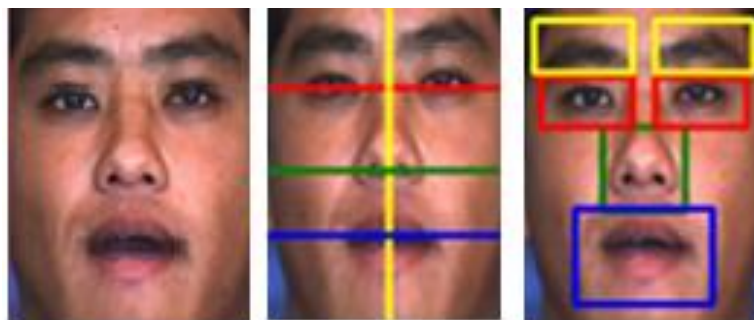


Figure 4.7 Localisation des composantes faciales [CHE 19a] [CHE 19b].

B. Sélection des points caractéristiques faciales

Après avoir localisé les différentes composantes faciales (yeux, bouche, sourcils et nez), nous avons opté pour la détection des coins dans les rectangles englobant chaque composante. A cet effet, le détecteur de point d'angle de Harris [HAR 88] est mis en œuvre.

Le détecteur de Harris [HAR 88] cherche à détecter des points en analysant les directions du gradient de l'image. Il est basé sur le calcul de la matrice d'autocorrélation (M) de son gradient. Cette matrice décrit la distribution du gradient dans un voisinage local d'un point sur l'image. Elle est définie par la formule suivante :

$$M = w(x, y) \otimes \begin{pmatrix} I_x^2 & I_x \cdot I_y \\ I_x \cdot I_y & I_y^2 \end{pmatrix} \quad (4.9)$$

Où $W(x, y)$ est un filtre gaussien, $\otimes$ est le produit de convolution, I_x et I_y sont les dérivées premières de l'image selon x et y . Elles sont obtenues à partir de la convolution de l'image I avec le filtre Sobel :

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

L'algorithme suivant résume les étapes de la détection des coins en utilisant la méthode de Harris : [ELK 17]

- (a) Calculer I_x et I_y sur toute l'image $I(x, y)$
- (b) Pour chaque point de l'image p :
 - a. A partir de la matrice M définie par la formule (4.9) :
 - b. Calculer $\text{Det}(M)$, et $\text{Trace}(M)$, le déterminant et la trace de M ;
 - c. Calculer la réponse $R = \text{Det}(M) - k (\text{Trace}(M))^2$; $k \in [0,04 - 0,06]$;
 - d. Si $R > \text{seuil}$, enregistrer les p points dans une liste C ;
 - Trier la liste C dans l'ordre décroissant de R ;
 - Parcourez la liste triée de haut en bas. Pour chaque point courant p , supprimer tous les points apparaissant plus loin dans la liste qui appartiennent au voisinage de p .

La figure 4.8 montre les résultats obtenus après application du détecteur de Harris [HAR 88] sur les différentes composantes faciales :

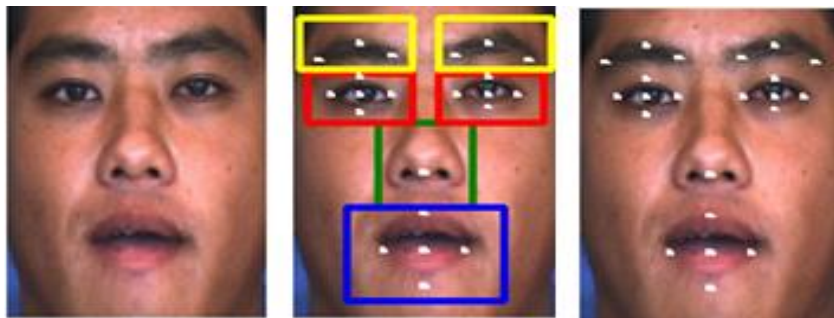


Figure 4.8 Extraction des points caractéristiques

La figure 4.9 suivante présente les points caractéristiques sélectionnés pour représenter les mouvements faciaux.

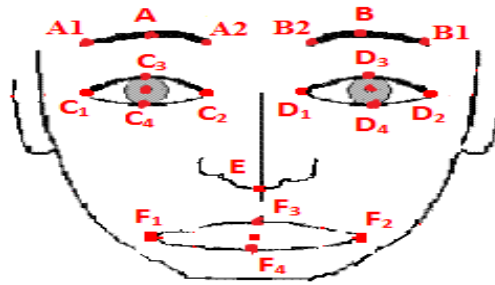


Figure 4.9 Sélection des points caractéristiques faciales [CHE 19a] [CHE 19b].

C. Construction du vecteur de caractéristiques

Une fois que les points caractéristiques sont extraits, l'étape suivante consiste à former le vecteur de caractéristiques qui représente une entrée pour le K plus proche voisin. En se basant sur ce vecteur, le KPPV classe les mouvements faciaux dans l'une des six expressions universelles. Le tableau 4.1 présente les différentes entrées de vecteur de caractéristique.

Formule	Définition
Largeur de l'œil droit	$V1 = D3 - D4$
Distance entre le sourcil droit coin gauche et l'œil droit coin gauche	$V2 = B2 - D1$
Longueur de la bouche	$V3 = F1 - F2$
Largeur de la bouche	$V4 = F3 - F4$
Distance entre l'œil droit coin droit et la bouche coin droit	$V5 = D2 - F2$
Longueur de l'œil droit	$V6 = D1 - D2$
Distance entre le centre du sourcil gauche et le centre du sourcil droit	$V7 = A - B$
Distance entre le sourcil gauche coin droit et l'œil gauche coin droit	$V8 = A2 - C2$
Longueur de l'œil gauche	$V9 = C1 - C2$
Largeur de l'œil gauche	$V10 = C4 - C3$
Distance entre l'œil gauche coin gauche et la bouche coin gauche	$V11 = C1 - F1$
Distance entre le centre du sourcil droit et le centre du l'œil droit	$V12 = B - D5$
Distance entre le centre du sourcil gauche et le centre du l'œil gauche	$V13 = A - C5$
Distance entre le centre de l'œil droit et le centre du nez	$V14 = D5 - E$
Distance entre le centre de l'œil gauche et le centre du nez	$V15 = C5 - E$
Distance entre le centre du nez et le centre de la bouche	$V16 = E - F5$
Distance entre l'œil gauche coin droit et l'œil droit coin droit	$V17 = C2 - D1$
Distance entre le centre de l'œil gauche et le centre de la bouche	$V18 = C5 - F5$
Distance entre le centre de l'œil droit et le centre de la bouche	$V19 = D5 - F5$

Tableau 4.1 Entrées de vecteur de caractéristiques

4.3.1.3. Classification

Pour évaluer la puissance du système de reconnaissance des expressions faciales basé sur l'utilisation des descripteurs géométriques, nous avons utilisé le classificateur KPPV. Le KPPV est un algorithme d'apprentissage non paramétrique fréquemment utilisée dans le domaine de reconnaissance des formes. Dans notre cas, la méthode de KPPV est utilisée pour calculer la *distance euclidienne* entre le vecteur de caractéristiques des exemples à prédire

(exemples de test) avec tous les vecteurs de caractéristiques des exemples connus (exemples d'apprentissage), puis affecter les exemples de test à la classe présentant la distance minimale.

4.3.2. Approche basée sur les caractéristiques d'apparences pour la reconnaissance des expressions faciales

Les caractéristiques d'apparence décrivent les changements transitoires de la texture de la peau tels que les rides, les gonflements et les sillons. Ces changements sont principalement localisés sur le front et autour de la bouche et des yeux. Les caractéristiques d'apparences peuvent être extraites soit sur tout le visage, soit sur des régions spécifiques d'une image de visage. La figure 4.10 schématise notre approche basée sur les descripteurs d'apparences pour la reconnaissance des expressions faciales [CHE 22b].

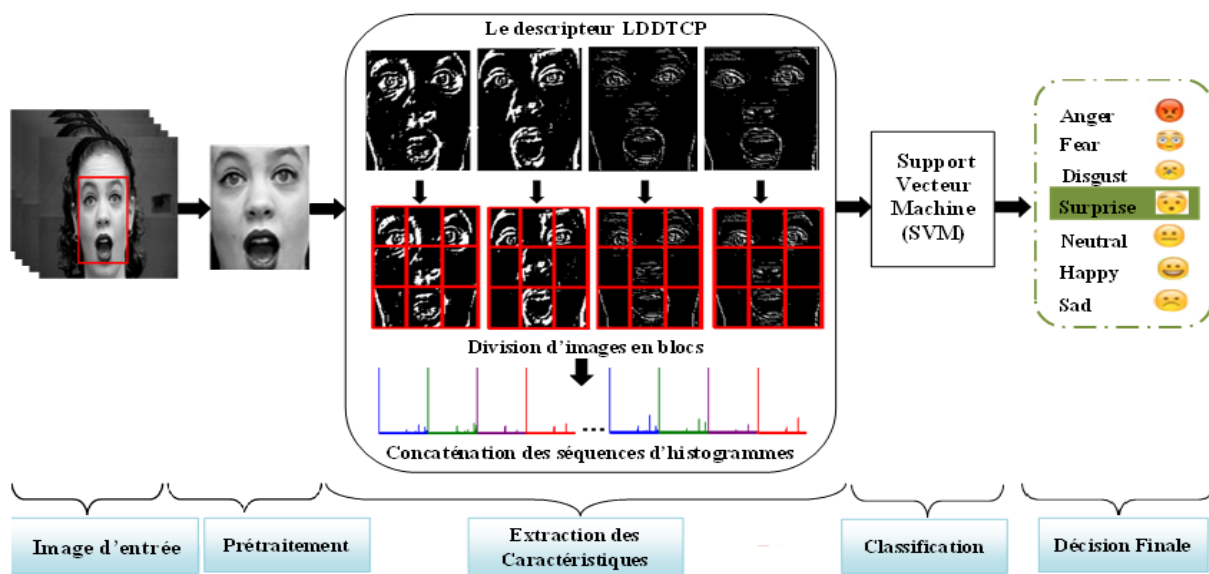


Figure 4.10 Architecture globale du SREF basé sur les caractéristiques d'apparences

La première étape consiste à détecter la présence d'un visage dans l'image et d'en localiser sa position. Une fois extrait, un nouveau descripteur d'apparence appelé *Local Directional Double Ternary Coding Pattern (LDDTCP)* est proposé pour mieux représenter l'image du visage. Des caractéristiques faciales sont ensuite extraites de ces visages traités en utilisant une concaténation des séquences d'histogrammes. Ces histogrammes sont calculés après avoir divisé l'image faciale en un certain nombre de blocs non superposés. L'ensemble des caractéristiques ainsi extraites sont introduites dans le classificateur SVM pour accomplir la tâche de reconnaissance [CHE 22b].

4.3.2.1. Détection et extraction de visage

La partie du visage est détecté en utilisant l'axe des yeux et l'axe médian, qui sont localisés selon les étapes (a), (b) (c) et (d) décrites dans la section A partie 4.3.1.2. Une fois que le vi-

sage est détecté, la zone est extraite du reste de l'image, puis normalisé à une taille fixe de 140 x140 pixels. La figure 4.11 illustre la procédure de détection et d'extraction de visages.

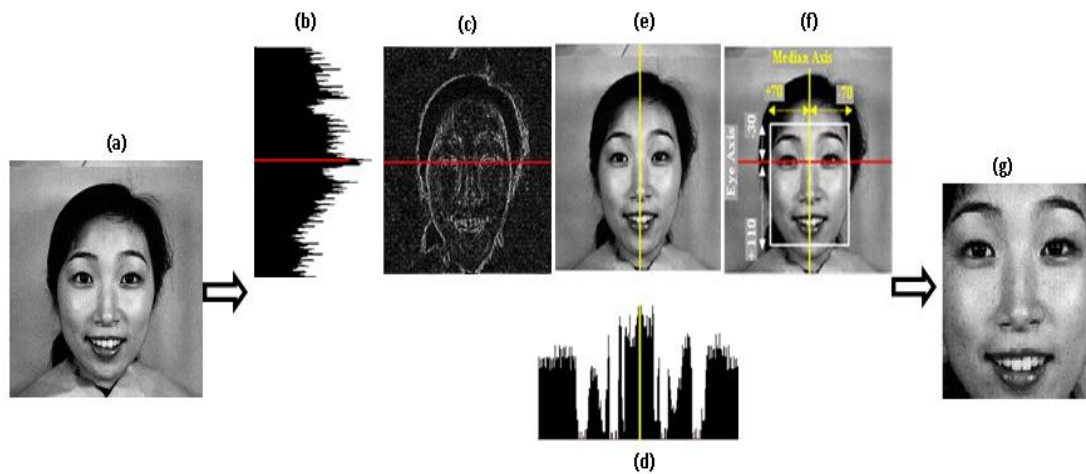


Figure 4.11 Procédure d'extraction de visage: (a) Image d'entrée, (b) Maximum de la projection horizontale, (c) Localisation de l'axe des yeux, (d) Maximum de la projection verticale, (e) Localisation de l'axe médian, (f) Détection de visage, (g) Extraction de visage.

4.3.2.2. Extraction des caractéristiques faciales

Après avoir localisé le visage, nous allons proposer dans la section suivante un nouveau descripteur d'apparence nommé *Local Directional Double Ternary Coding Pattern (LDDTCP)* qui tente d'extraire les informations pertinentes pour former les vecteurs de caractéristiques nécessaires à l'analyse des expressions faciales [CHE 22b].

A. Approche proposée : Local Directional Double Ternary Coding Patterns

Bien que l'opérateur LBP fournit une tolérance aux changements d'illumination monotones et présente une efficacité en termes de calcul, des recherches récentes ont démontré que ce descripteur considère uniquement la relation entre le pixel central et ses pixels voisins, ce qui le rend sensible aux grandes variations d'éclairage et au bruit aléatoire. Pour résoudre cette limitation, des descripteurs tels que Local Ternary Pattern (LTP) [TAN 10] et Local Directionnel Pattern (LDP) [JAB 10a] [JAB 10b] ont été proposés.

D'une part, l'opérateur LTP a été développé afin de réduire l'influence des variations d'illuminations et d'augmenter la robustesse au bruit aléatoire dans les régions uniformes en étendant le code binaire LBP (0,1) à un code ternaire à 3 valeurs (-1, 0, 1) en utilisant un seuil t . Bien que le descripteur LTP présente des avantages par rapport à l'opérateur LBP, il n'est pas exactement invariant aux changements d'échelle de gris et il ignore la plupart des informations du voisinage. De plus, le choix de la valeur de seuil n'est pas simple [ELK 18].

D'autre part, le descripteur LDP a introduit un nouveau concept qui exploite les réponses de contour (bord) en utilisant huit directions au lieu des intensités des pixels [AHM 13]. Alors

que LDP atteint des performances supérieures par rapport à l'opérateur LBP, il produit toujours des modèles inadéquats dans des régions uniformes. De plus dans l'opérateur LDP, toutes les directions sont traitées de la même manière, ainsi, il manque certaines informations directionnelles comme le signe des réponses [ELK 18].

Pour surmonter les problèmes de représentation de texture des méthodes sous mentionné (LBP, LTP et LDP), un nouveau descripteur d'apparence nommé Local Directional Double Ternary Coding Patterns (LDDTCP) est proposé [CHE 22b]. Cet opérateur intègre les concepts des descripteurs LDP et LTP pour coder à la fois les informations de direction et de contraste, ce qui permet d'augmenter la robustesse aux changements d'illumination et de bruit aléatoire.

L'opérateur LDDTCP encode la texture locale d'une image en calculant un code ternaire pour chaque pixel basé sur les valeurs de réponse de contour et de ligne (Edge and line response) dans différentes directions autour du pixel central [CHE 22b].

Pour obtenir les valeurs de réponse de contour et de ligne, huit directions du masque de Frei-Chen (FC₀-FC₇) [APD 17] de taille 3x3 sont convolués avec des régions d'image de taille 3x3. Les masques de Frei-Chen (FC_i, i= 0,....., 7) sont présentés dans la figure 4.12.

$$\begin{array}{cccc}
 \text{(a)} & & & \\
 \frac{1}{2\sqrt{2}} \begin{bmatrix} 1 & \sqrt{2} & 1 \\ 0 & 0 & 0 \\ -1 & -\sqrt{2} & -1 \end{bmatrix} & \frac{1}{2\sqrt{2}} \begin{bmatrix} 1 & 0 & -1 \\ \sqrt{2} & 0 & -\sqrt{2} \\ 1 & 0 & -1 \end{bmatrix} & \frac{1}{2\sqrt{2}} \begin{bmatrix} 0 & -1 & \sqrt{2} \\ 1 & 0 & -1 \\ -\sqrt{2} & 1 & 0 \end{bmatrix} & \frac{1}{2\sqrt{2}} \begin{bmatrix} \sqrt{2} & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -\sqrt{2} \end{bmatrix} \\
 \text{Groupe 1- FC}_0 & \text{Groupe 1- FC}_1 & \text{Groupe 1- FC}_2 & \text{Groupe 1- FC}_3 \\
 \\
 \text{(b)} & & & \\
 \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} & \frac{1}{2} \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{bmatrix} & \frac{1}{6} \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} & \frac{1}{6} \begin{bmatrix} -2 & 1 & -2 \\ 1 & 4 & 1 \\ -2 & 1 & -2 \end{bmatrix} \\
 \text{Groupe 2- FC}_0 & \text{Groupe 2- FC}_1 & \text{Groupe 2- FC}_2 & \text{Groupe 2- FC}_3
 \end{array}$$

Figure 4.12. Les huit masques de Frei-Chen divisé en deux groupes : **(a)** Le premier groupe, **(b)** Le deuxième groupe

Pour atteindre la robustesse, nous partitionnons les huit masques de Frei-chen obtenus en deux groupes selon leurs directions et nous adoptons ces deux groupes pour calculer les valeurs de convolution de chaque pixel. Les quatre premiers masques de Frei-Chen (FC₀ –FC₃) sont utilisés pour la détection des contours ; les quatre masques suivants (FC₄ –FC₇) sont utilisés pour la détection des lignes, la figure 4.13 montre les résultats de convolution des images avec les deux groupes de masques de Frei-Chen.

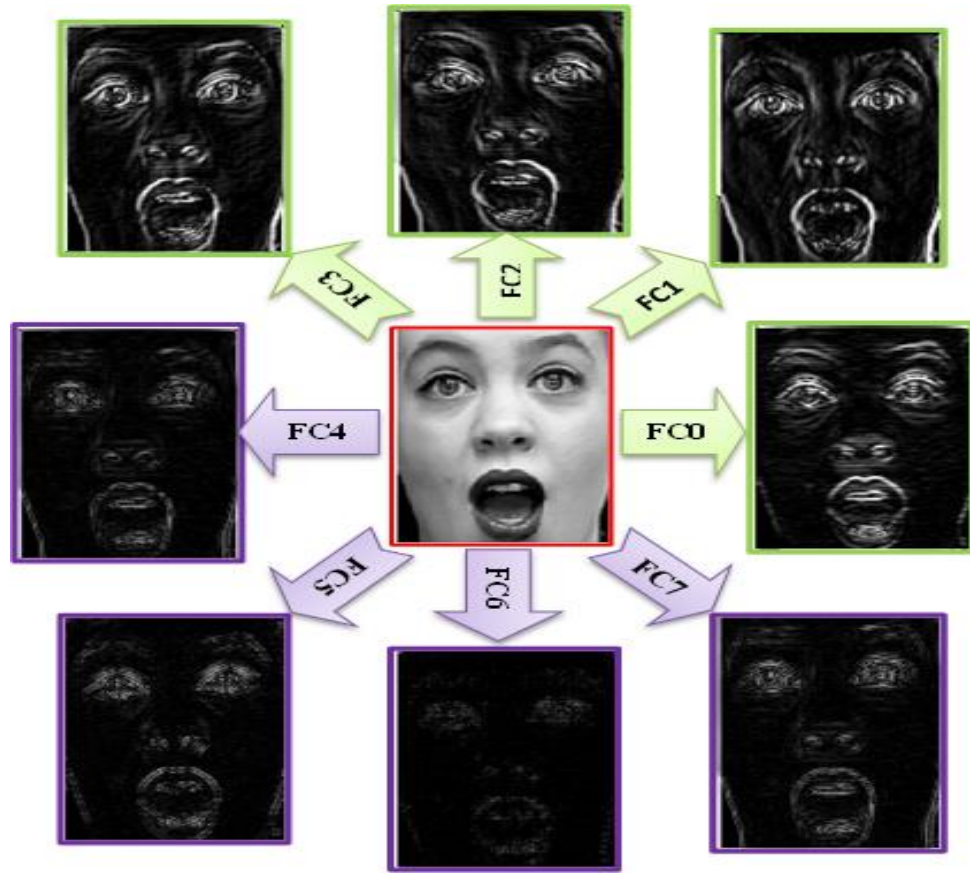


Figure 4.13 Résultats de convolution des images avec le masque de Frei-chen : **(a)** Convolution des images avec le premier groupe du masque (FC₀, FC₁, FC₂, FC₃), **(b)** Convolution des images avec le deuxième groupe du masque (FC₄, FC₅, FC₆, FC₇) [CHE 22b].

En appliquant le premier groupe des quatre masques, nous obtenons quatre valeurs de réponse de contour FC₀, FC₁, FC₂ et FC₃. La présence d'un contour produira des valeurs de réponse de contour élevées dans leurs directions respectives.

L'opérateur LDDTCP-1 (LDDTCP pour le premier groupe) utilise un code ternaire à trois valeurs pour différencier les régions uniformes des régions non uniformes à l'aide d'un seuil t_1 . Après avoir calculé la moyenne du premier groupe $Average_1$, les valeurs de réponse de contour dans une marge de largeur $\pm t_1$ autour de la moyenne $Average_1$ sont quantifiées à 0, celles inférieures à $Average_1 - t_1$ sont quantifiées à -1 et celles supérieures à $Average_1 + t_1$ sont quantifiées à +1 [CHE 22b].

$$S_{LDDTCP-1}(FC_{1i}) = \begin{cases} +1 & FC_{1i} > Average_1 + t_1 \\ 0 & Average_1 - t_1 \leq FC_{1i} \leq Average_1 + t_1 \\ -1 & FC_{1i} < Average_1 - t_1 \end{cases} \quad (4.10)$$

$$Average_1 = \frac{1}{4} \sum_{i=0}^3 FC_{1i} \quad (4.11)$$

Où FC_{1i} est la valeur de réponse de contour dans la $i^{ème}$ direction ($i=0,...,3$), $Average_1$ est la valeur moyenne de réponse de contour et t_1 est le seuil du premier groupe (ici $t_1=15$).

Nous répétons la même opération pour le deuxième groupe des quatre masques restants, nous obtenons quatre valeurs de réponse de ligne FC_4 , FC_5 , FC_6 et FC_7 . La présence d'une ligne produira des valeurs de réponse de ligne élevées dans leurs directions respectives.

Après cela, nous calculons la moyenne du deuxième groupe $Average_2$. Les valeurs de réponses de ligne situant entre $Average_2 + t_2$ et $Average_2 - t_2$ sont quantifiées à 0, tandis que celles inférieures à $Average_2 - t_2$ et supérieures à $Average_2 + t_2$ sont respectivement quantifiées à -1 et $+1$, comme suit :

$$S_{LDDTCP-2}(FC_{2i}) = \begin{cases} +1 & FC_{2i} > Average_2 + t_2 \\ 0 & Average_2 - t_2 \leq FC_{2i} \leq Average_2 + t_2 \\ -1 & FC_{2i} < Average_2 - t_2 \end{cases} \quad (4.12)$$

$$Average_2 = \frac{1}{4} \sum_{i=0}^4 FC_{2i} \quad (4.13)$$

Où FC_{2i} est la valeur de réponse de ligne dans la $i^{\text{ème}}$ direction ($i=4,...,7$), $Average_2$ est la valeur moyenne de réponse de ligne et t_2 est le seuil du deuxième groupe (ici $t_2 = 7$). La figure 4.14 montre le principe de codage des opérateurs LDDTCP-1 et LDDTCP-2.

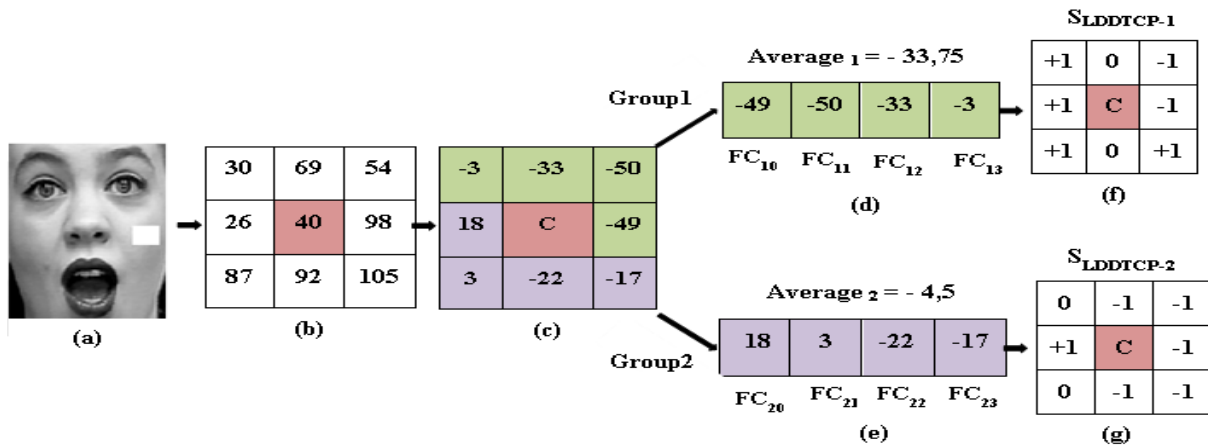


Figure 4.14 Principe du codage de LDDTCP-1 et LDDTCP-2 : (a) Image en niveaux de gris. (b) Exemple d'image de taille 3×3 . (c) Valeurs du pixel centrale après application du masque de Frei-chen (FC). (d) Moyenne du premier groupe des masques FC_{1i} . (e) Moyenne du deuxième groupe des masques suivants FC_{2i} . (f) Code LDDTCP-1 pour $t_1 = 15$. (g) Code LDDTCP-2 pour $t_2 = 7$ [CHE 22b].

LDDTCP est formé par concaténation de LDDTCP-1 et LDDTCP-2, ce qui donne un vecteur de caractéristiques de grande dimension (3^8 pour chacun de LDDTCP-1 et LDDTCP-2). Pour réduire la longueur du vecteur de caractéristiques, chaque code LDDTCP-1 et LDDTCP-2 est divisé en deux codes négative et positive : $P_{LDDTCP-1}$, $N_{LDDTCP-1}$ pour LDDTCP-1 et $P_{LDDTCP-2}$, $N_{LDDTCP-2}$ pour LDDTCP-2 (voir Figure 4.15). Par conséquent, le nombre de caractéristiques passe de 2×3^8 à 2×2^8 pour chacune. Les $P_{LDDTCP-1}$ et $N_{LDDTCP-1}$ sont calculés comme suit [CHE 22b] :

$$P_{LDDTCP-1} = \sum_{i=0}^7 S_P (S_{LDDTCP-1}(FC_{1i})) * 2^i \quad (4.14)$$

$$N_{LDDTCP-1} = \sum_{i=0}^7 S_N (S_{LDDTCP-1}(FC_{1i})) * 2^i \quad (4.15)$$

On répète le même calcul pour les opérateurs $P_{LDDTCP-2}$ and $N_{LDDTCP-2}$:

$$P_{LDDTCP-2} = \sum_{i=0}^7 S_P (S_{LDDTCP-2}(FC_{2i})) * 2^i \quad (4.16)$$

$$N_{LDDTCP-2} = \sum_{i=0}^7 S_N (S_{LDDTCP-2}(FC_{2i})) * 2^i \quad (4.17)$$

Où:

$$S_P(v) = \begin{cases} 1, & \text{si } v = 1 \\ 0, & \text{sinon} \end{cases} \quad S_N(v) = \begin{cases} 1, & \text{si } v = -1 \\ 0, & \text{sinon} \end{cases} \quad (4.18)$$

La figure 4.15 représente le principe de calcul de l'opérateur LDDTCP.

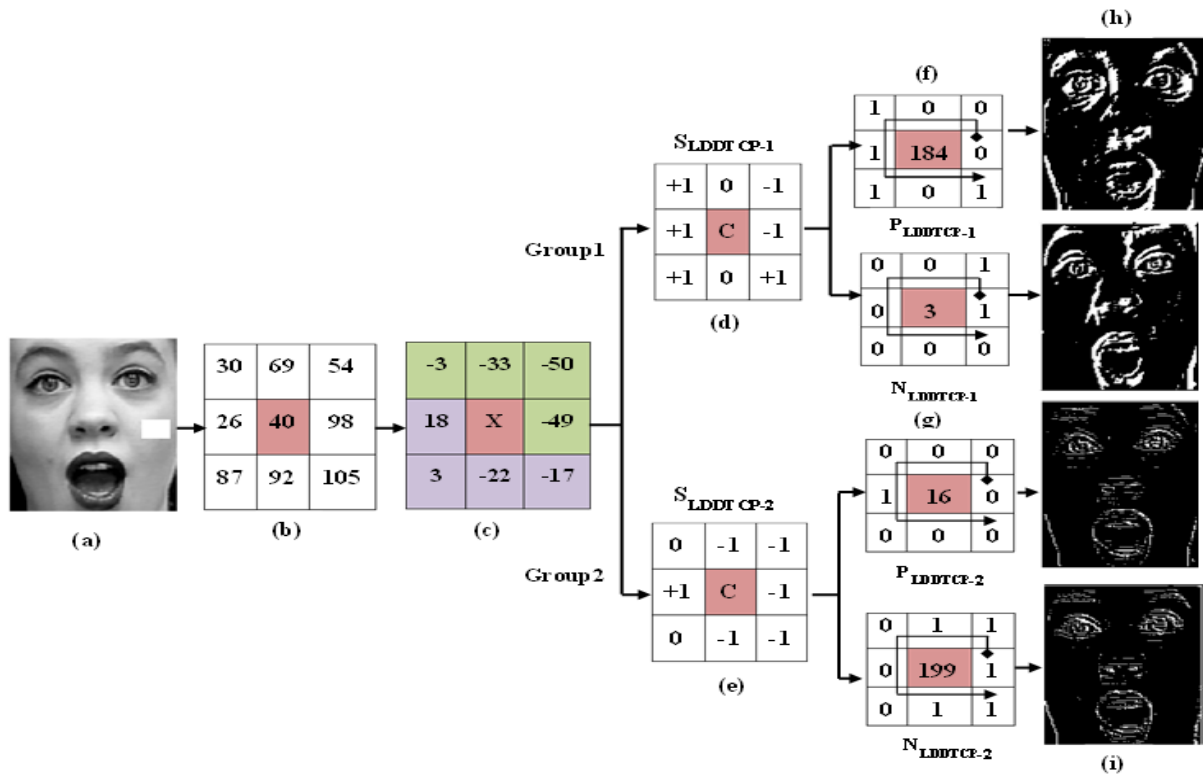


Figure 4.15 Principe de calcul de l'opérateur LDDTCP: (a) Image en niveau de gris. (b) 3 × 3 sample image. (c) Valeurs du pixel centrale après application du masque de Frei-chen (FC). (d) Code de LDDTCP-1 avec $t_1 = 15$. (e) Code de LDDTCP-2 avec $t_2 = 7$. (f) Division du code LDDTCP-1 en deux codes positive and négative pour détecter les contours (g) Division du code LDDTCP-2 en deux codes positive and négative pour détecter les lignes. (h) Image codée positive and négative du code LDDTCP-1. (i) Image codée positive and négative du code LDDTCP-2 [CHE 22b].

B. Représentation des caractéristiques faciales basé sur le code LDDTCP

Une fois le code LDDTCP est calculé pour chaque pixel (t, u), quatre images encodées seront produites : deux images encodées pour le premier groupe $P_{LDDTCP-1}$, $N_{LDDTCP-1}$ et deux autres pour le deuxième groupe $P_{LDDTCP-2}$, $N_{LDDTCP-2}$ [CHE 22b].

L'image d'entrée I de taille $M \times N$ est représentée par un histogramme H_{LDDTCP} . L'histogramme H_{LDDTCP} est le résultat de la concaténation des quatre histogrammes ($H(P_{LDDTCP-1})$, $H(N_{LDDTCP-1})$, $H(P_{LDDTCP-2})$ et $H(N_{LDDTCP-2})$) (voir Figure 4.16) calculés à partir des quatre images encodées à l'aide des équations suivantes :

$$\begin{cases} H_{P_{LDDTCP-1}}(\alpha) = \sum_{t=1}^M \sum_{u=1}^N f(P_{LDDTCP-1}(t, u), \alpha) \\ H_{N_{LDDTCP-1}}(\alpha) = \sum_{t=1}^M \sum_{u=1}^N f(N_{LDDTCP-1}(t, u), \alpha) \end{cases} \quad (4.19)$$

$$\begin{cases} H_{P_{LDDTCP-2}}(\alpha) = \sum_{t=1}^M \sum_{u=1}^N f(P_{LDDTCP-2}(t, u), \alpha) \\ H_{N_{LDDTCP-2}}(\alpha) = \sum_{t=1}^M \sum_{u=1}^N f(N_{LDDTCP-2}(t, u), \alpha) \end{cases} \quad (4.20)$$

$$f(a, \alpha) = \begin{cases} 1 & a = \alpha \\ 0 & \text{otherwise} \end{cases} \quad (4.21)$$

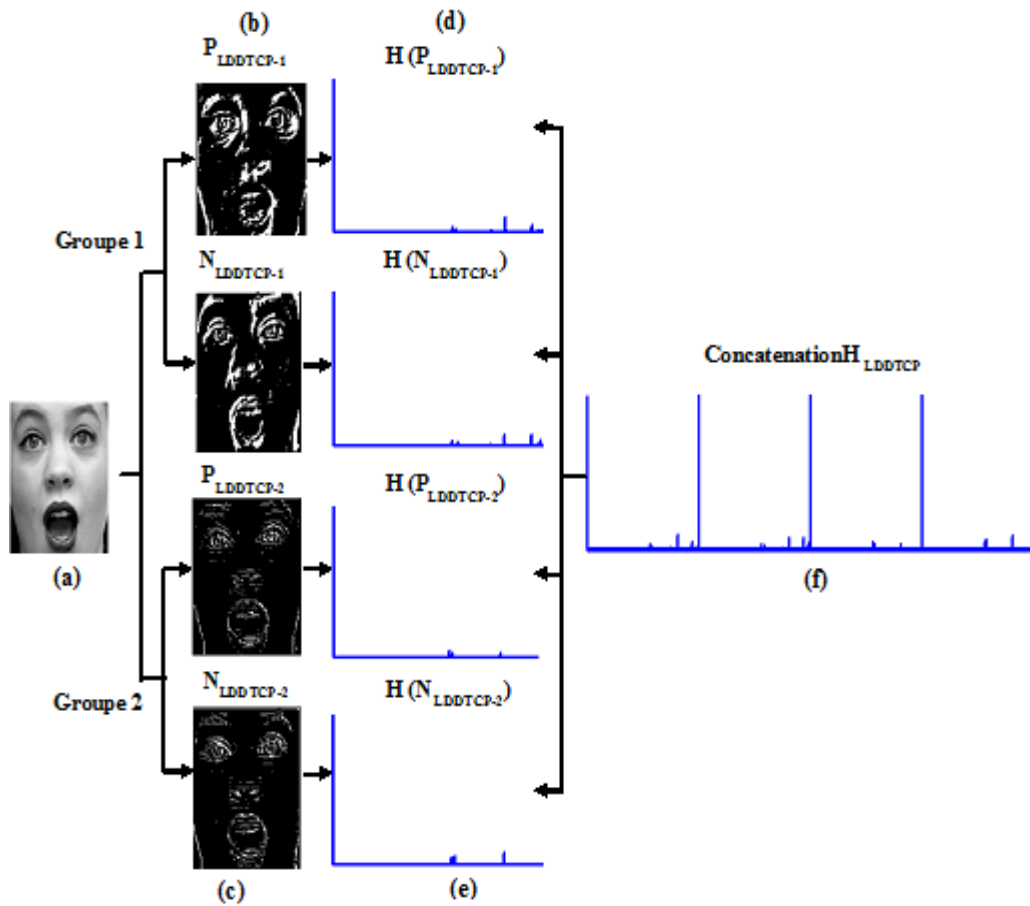


Figure 4.16 Extraction des caractéristiques de l'image entière à partir du code LDDTCP : (a) Image en niveau de gris. (b) Image encodée positive and négative ($P_{LDDTCP-1}$ et $N_{LDDTCP-1}$) du premier groupe LDDTCP-1. (c) Image encodée positive and négative ($P_{LDDTCP-2}$ and $N_{LDDTCP-2}$) du deuxième groupe LDDTCP-2. (d) Histogramme positif et négatif ($H(P_{LDDTCP-1})$ et $H(N_{LDDTCP-1})$) du premier groupe LDDTCP-1. (e) Histogramme positif et négatif ($H(P_{LDDTCP-2})$ et $H(N_{LDDTCP-2})$) du deuxième groupe LDDTCP-2. (f) Concaténation des histogrammes LDDTCP-1 et LDDTCP-2 pour former l'histogramme LDDTCP [CHE 22b].

En réalité, afin d'incorporer plus informations spatiales au vecteur représentant le visage, l'image codée par l'opérateur LDDTCP est divisée en m petites régions $R_0, R_1, \dots, R_m$ et l'histogramme $H(R_i)$ est construit pour chaque région R_i en accumulant dans des bins chacun des codes LDDTCP extraits. Les histogrammes $H_{LDDTCP-1}$ et $H_{LDDTCP-2}$ sont ensuite concaténés afin de former un histogramme final H_{LDDTCP} représentant l'image des caractéristiques faciales (voir figure 4.17).

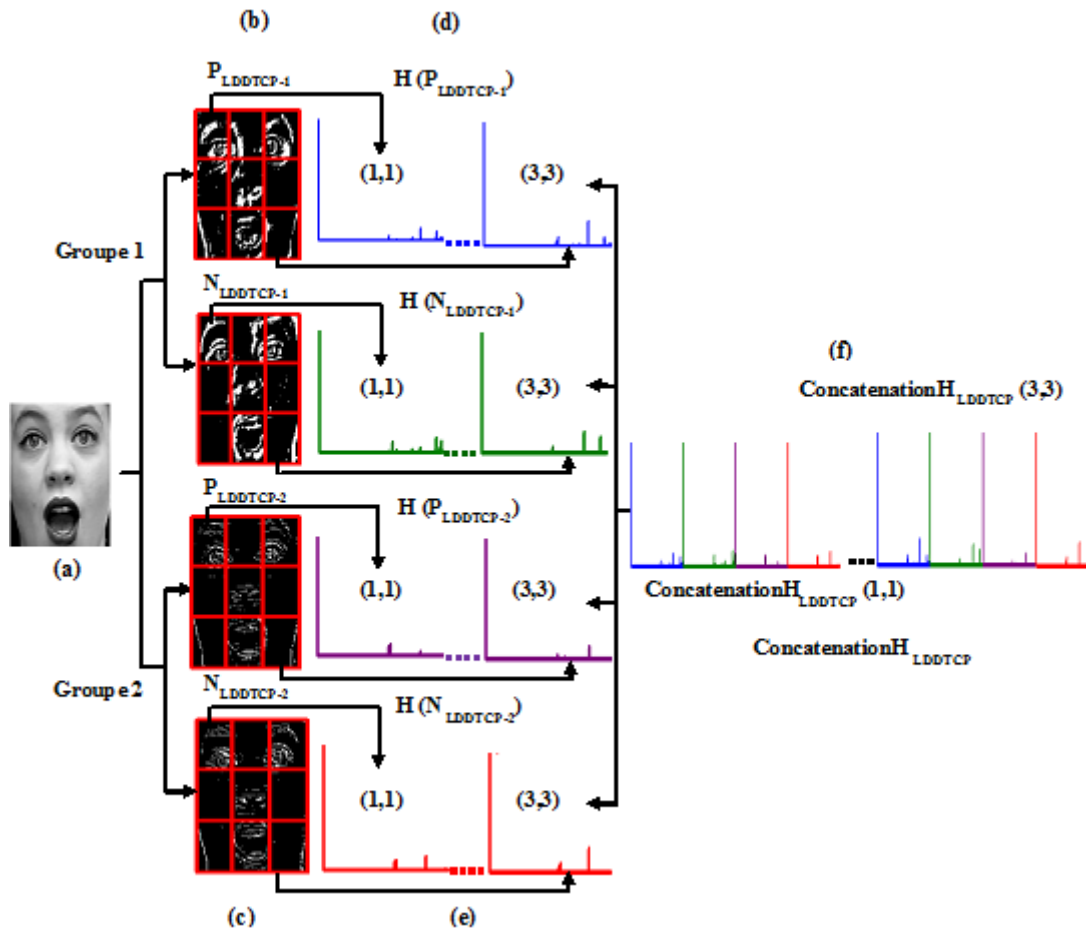


Figure 4.17 Extraction des caractéristiques de l'image divisée en plusieurs blocs à partir du code LDDTCP : (a) Image en niveau de gris. (b) Segmentation de l'image encodée positive and négative ($P_{LDDTCP-1}$ et $N_{LDDTCP-1}$) du premier groupe LDDTCP-1 en bloc de taille 3×3 . (c) Segmentation de l'image encodée positive and négative ($P_{LDDTCP-2}$ et $N_{LDDTCP-2}$) du deuxième groupe LDDTCP-2 en bloc de taille 3×3 . (d) Concaténation des histogrammes positives et négatives du premier groupe LDDTCP-1. (e) Concaténation des histogrammes positives et négatives du deuxième groupe LDDTCP-2. (f) Concaténation des histogrammes LDDTCP-1 et LDDTCP-2 pour former l'histogramme LDDTCP [CHE 22b].

4.3.2.3. Classification

Pour évaluer la validité et l'efficacité de la méthode proposée, nous utilisons le classificateur statistique supervisé SVM qui permet de faire la correspondance entre deux histogrammes (histogrammes des images d'apprentissage avec les histogrammes des images de

test) et classer les vecteurs de caractéristiques dans l'une des six expressions faciales universelles.

SVM prend des décisions binaires en construisant l'hyperplan de séparation entre les exemples positifs et négatifs. Pour un problème de classification multiple ou multi-classe deux approches sont généralement adoptées : **Un contre un** et un **contre tous** :

- Approche **un contre un** consiste à construire un SVM par paire de classe, puis à entraîner $k(k-1)/2$ SVM (k est le nombre de classes) afin de distinguer les exemples d'une classe de ceux d'une autre classe. La classification d'une nouvelle instance test est déterminée par un vote majoritaire sur l'ensemble des SVM. La classe avec un maximum de votes détermine l'étiquette de l'instance.
- Approche un **contre tous** consiste à entraîner un SVM binaire en utilisant les éléments d'une classe contre tous les autres. Il s'agit de résoudre un problème SVM d'ordre C et de taille N .

Dans notre cas, nous avons opté pour la méthode un-contre-un, qui permet de séparer les exemples de différentes classes. Pour implémenter la classification par SVM, nous avons utilisé la machine d'apprentissage SVM^{light} [JOA 99] et la fonction noyau RBF (Radial Basis functions).

4.3.3. Approche basée sur l'apprentissage profond pour la reconnaissance des expressions faciales

Dans cette approche, nous développons un système de reconnaissance des expressions faciales (SREF) basé sur la combinaison des réseaux de neurones convolutionnels 'CNN'[CHE 22a]. Le modèle proposé est capable d'apprendre et de classer avec précision les images des expressions faciales. Le SREF comprend quatre étapes importantes : le prétraitement, l'extraction des traits du visage, la classification des expressions faciales et enfin la fusion des différents classificateurs (fusion des CNN). La figure 4.18 montre l'architecture globale de l'approche proposée et ses différentes étapes [CHE 22a].

Premièrement, une étape préalable d'obtention des images de visage a été effectuée. Ensuite vient l'étape de prétraitement qui comprend l'extraction de l'image du visage et le calcul des descripteurs de texture. Les visages sont extraits en utilisant la méthode développée par Abdat [ABD 07], et les descripteurs de texture sont calculés à partir des opérateurs Local Binary Pattern (LBP), Local Gradient Code (LGC), Gradient Directional Pattern (GDP) et Local Directional Pattern (LDP). Deuxièmement, les données multimodales générées à partir de chaque descripteur représentent une entrée pour un CNN spécifique (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et

CNN_{LDP}) qui sera utilisé pour extraire les traits du visage et classer les images faciales dans l'une des six expressions faciales universelles. Enfin, nous appliquons des classificateurs séparés pour chaque modalité, puis nous combinons les quatre sorties multimodales à l'aide de deux méthodes de fusion (le produit et la somme) pour obtenir la décision finale de notre système [CHE 22a].

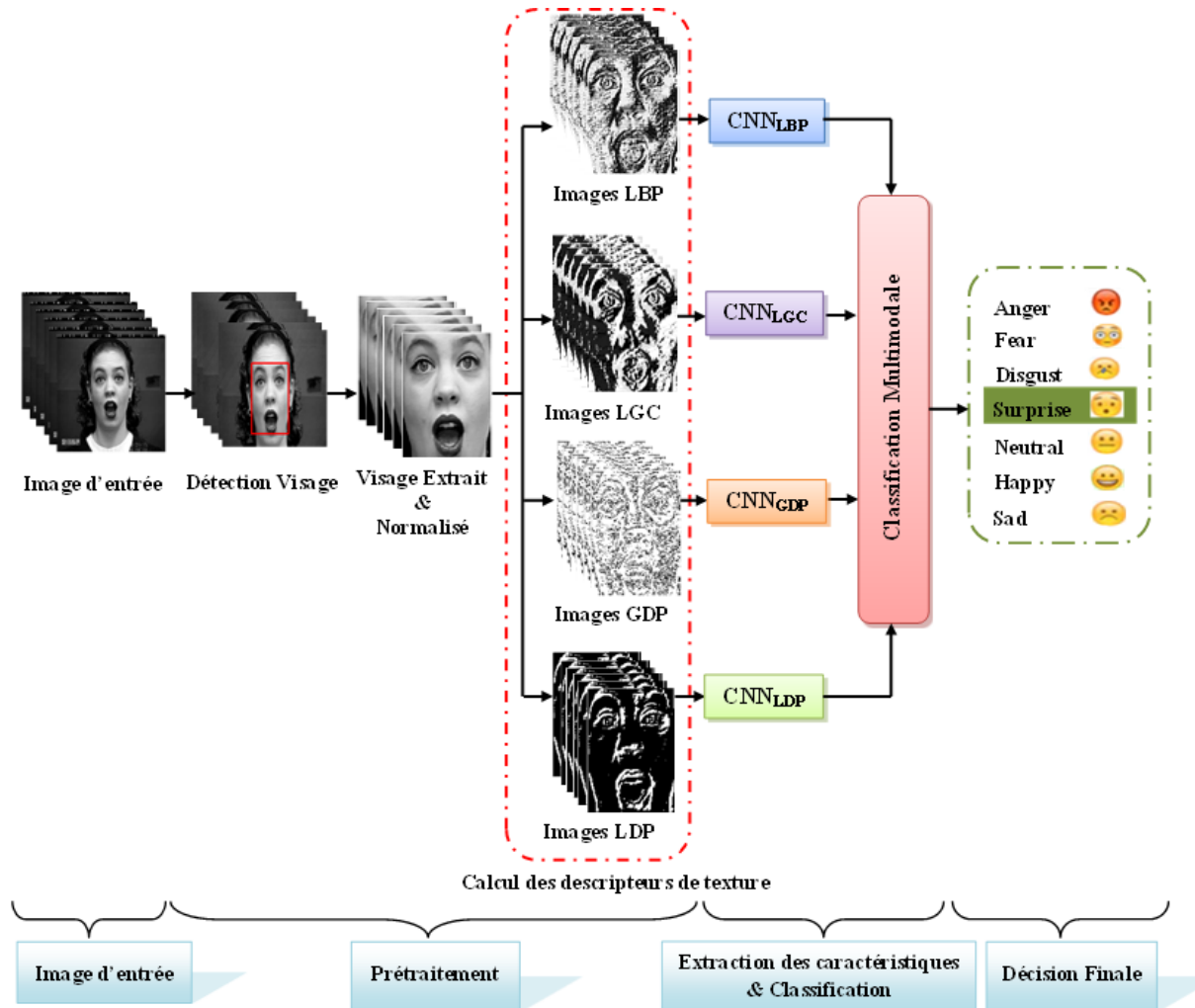


Figure 4.18 Architecture globale du SREF via l'apprentissage profond [CHE 22a].

4.3.3.1. Détection et extraction de visage

Les visages sont détectés, extraits du reste de l'image puis normalisés à une taille fixe de 140 x 140 pixels en utilisant la même procédure décrite ci-dessous (Section 4.3.2.1 Partie 4.3.2)

4.3.3.2. Calcul des descripteurs de texture

Pour atteindre la robustesse et rendre les CNN plus faciles à apprendre, nous avons calculé quatre descripteurs de texture qui sont LBP, LGC, GDP et LDP à partir des images en ni-

veaux de gris. Les données multimodales générées à partir de chaque descripteur sont ensuite envoyées à un CNN spécifique.

Le calcul de chacun des quatre opérateurs d'apparences *LBP*, *LGC*, *GDP* et *LDP* est présenté respectivement dans les sections 2.6.4.1.4, 2.6.4.1.5, 2.6.4.1.6 et 2.6.4.1.8 (partie 2.6.4.1 du chapitre 2).

4.3.3.3. Architecture et conception du CNN

Nous avons développé une architecture CNN qui extrait les caractéristiques et classe les expressions faciales en même temps au sein d'un même réseau de neurones. L'architecture proposée fonctionne en deux phases principales : l'apprentissage et le test.

A. Apprentissage et génération de modèle

Notre réseau de neurones est réalisé après plusieurs tests. Nous avons utilisé onze (11) couches dans les quatre structure CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et CNN_{LDP}) : Couche de convolution C1, Couche de sous-échantillonnage S1, Couche de convolution C2, Couche de sous-échantillonnage S2, Couche de convolution C3, Couche de sous-échantillonnage S3, Couche de convolution C4, Couche de sous-échantillonnage S4, Couche entièrement connectée FC1, Couche entièrement connectée FC2 et Couche de sortie O (voir Figure 4.19) [CHE 22a].

L'entrée du réseau est $140 \times 140 \times 1$, où 1 fait référence au nombre de canaux existant dans les images d'entrée. Dans notre cas, les images LBP, LGC, GDP et LDP sont calculées à partir des images en niveaux de gris, donc le nombre de canaux est égal à 1. Dans les *couches de convolution*, l'image d'entrée est passée à travers une succession de filtres ou noyaux de convolution, créant ainsi de nouvelles images appelées cartes de caractéristiques. Ces cartes sont ensuite normalisées avec une *fonction d'activation*. La fonction d'activation ReLU est utilisée pour les deux descripteurs LBP et LDP et la fonction d'activation identité (Identity) est utilisée pour les descripteurs LGC et GDP. Les couches de convolution sont souvent suivies par des *couches de sous-échantillonnage (pooling)* qui permettent de réduire progressivement la taille spatiale des cartes de caractéristiques. La dernière couche (*couche entièrement connectée*) est similaire à un perceptron multicouche traditionnel qui implique que chaque neurone dans la couche précédente est connecté à chaque neurone de la couche suivante. Cette couche permet d'effectuer la tâche de classification. Une *couche de suppression (dropout)* est placée après les couches entièrement connectées pour éviter le sur-apprentissage (over fitting) causé par un nombre important de paramètres. Enfin, une *couche Soft-Max* à six sorties (6 classes d'expressions faciales) est insérée au dernier étage du réseau. La figure 4.19 présente le modèle d'apprentissage proposé [CHE 22a].

Tous les paramètres utilisés dans l'architecture des quatre CNN sont énumérés en détail dans le tableau 4.2.

Type de couche	Taille / Stride	Sortie
Données d'entrée	140 * 140	-
Couche de Convolution C1	9*9 / 1	132*132*16
Couche de Max-Pooling S1	2*2 / 2	66*66*16
Couche de Convolution C2	7*7 / 1	66*66*32
Couche de Max-Pooling S2	2*2 / 2	33*33*32
Couche de Convolution C3	5*5 / 1	33*33*64
Couche de Max-Pooling S3	2*2 / 2	16*16*64
Couche de Convolution C4	5*5 / 1	16*16*128
Couche de Max-Pooling S4	3*3 / 2	7*7*128
Couche Entièrement Connecté FC1	-	3200 Dropout=0,5
Couche Entièrement Connecté FC2	-	1600 Dropout=0,5
Couche Sortie O		6

Tableau 4.2 Configuration de CNN [CHE 22a].

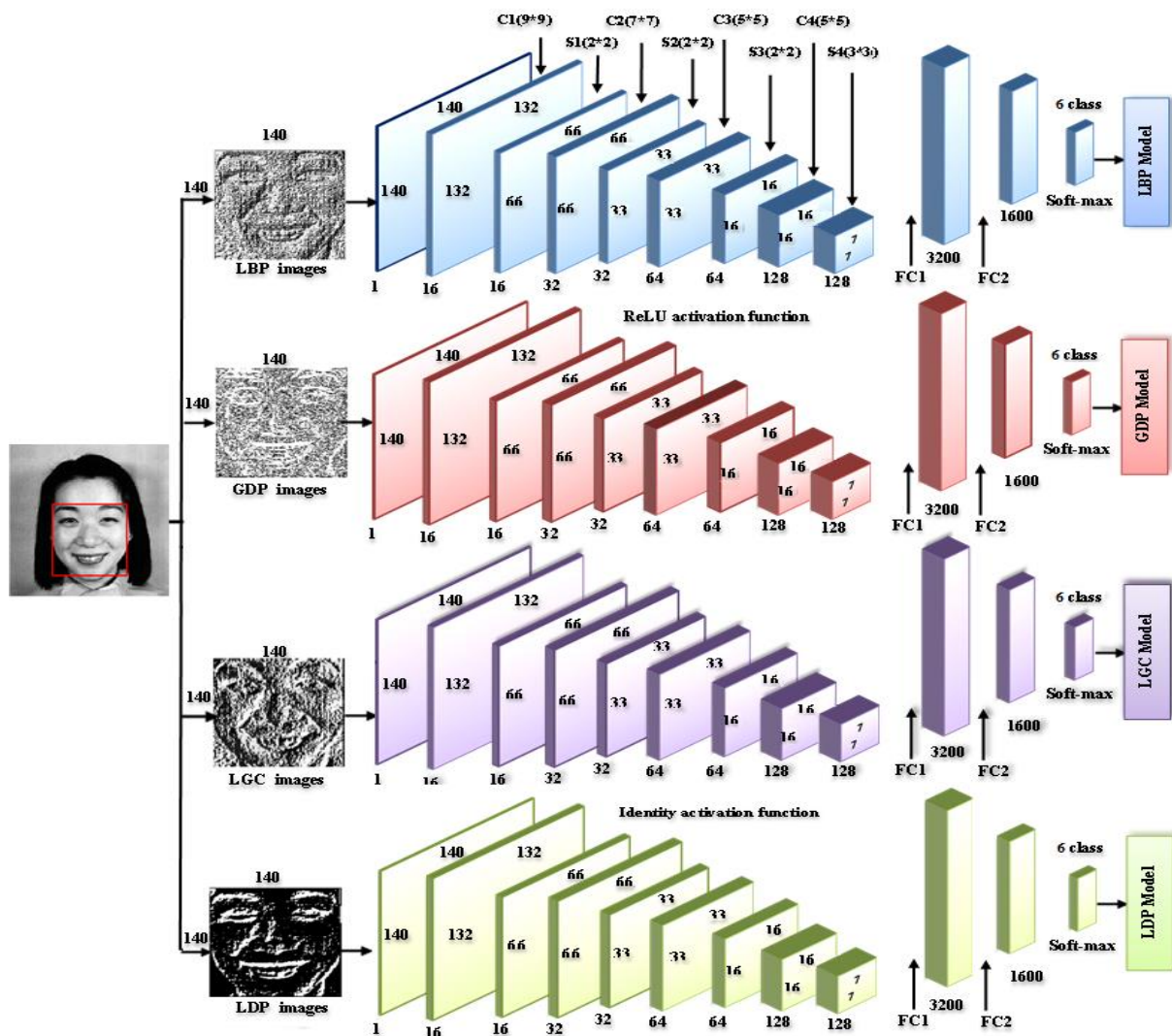


Figure 4.19 Phase d'apprentissage [CHE 22a].

B. Phase de test

Les quatre modèles CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et CNN_{LDP}) générés lors de la phase d'apprentissage ont été testés sur les deux bases de données CK+ et JAFFE. Le système général sera capable de reconnaître toute nouvelle image faciale. Les quatre classificateurs CNN sont appliqués et leurs sorties sont combinées à l'aide de la règle somme et produit. La décision finale est attribuée à la classe avec la probabilité la plus élevée. La figure 4.20 présente le processus de test de notre CNN.

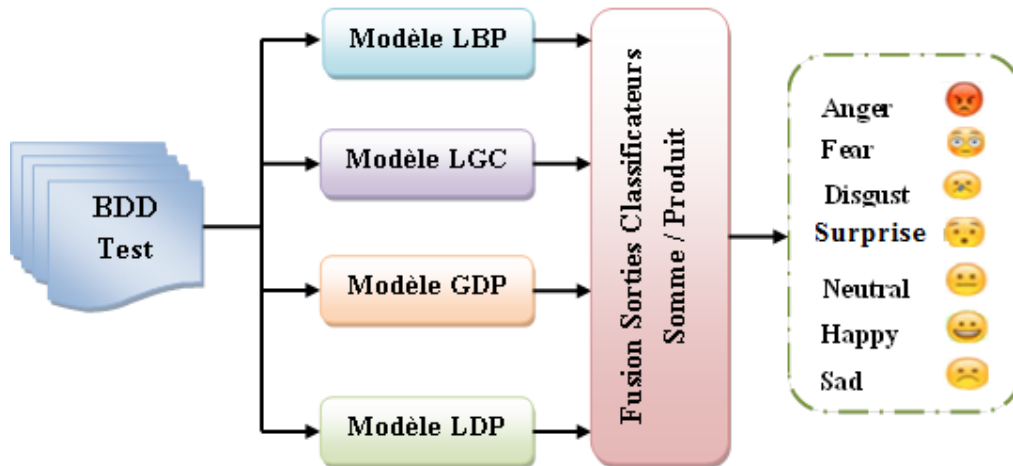


Figure 4.20 Phase de test [CHE 22a].

4.3.3.4. Classification et fusion

Récemment, de nombreuses études ont montré qu'il n'existe pas de classificateur unique capable de traiter une connaissance complète de la distribution des données et pouvant résoudre tous les problèmes de reconnaissance de formes. Par conséquent, l'utilisation d'un ensemble de classificateurs pour résoudre des problèmes de classification est plus efficace que l'utilisation d'un classificateur unique.

Dans ce travail, nous adoptons l'architecture parallèle où les quatre CNN (CNN_{LBP} , CNN_{LGC} , CNN_{LDP} et CNN_{GDP}) sont générés en parallèle. L'avantage principal d'utiliser cette méthode est d'exploiter la complémentarité entre les quatre classificateurs ce qui permet d'obtenir des performances supérieures à celles d'un seul classificateur. Les sorties des multiples classificateurs sont combinées selon la règle somme et produit pour donner la décision finale de notre système multimodal.

Dans la règle de produit, les scores fournis par les quatre classificateurs sont multipliés et l'étiquette de la classe avec le score maximum est attribuée à l'entrée donnée. Dans la règle de la somme, on additionne les résultats des quatre classificateurs, et on assigne la classe avec la valeur maximale à l'entrée donnée.

4.4 Conclusion

Dans ce chapitre, nous avons d'abord présenté la structure générale de notre système de reconnaissance des émotions des apprenants à partir des expressions faciales. Ce système a été conçu en utilisant trois sous-système différents. Chaque sous système est constitué principalement de trois étapes : la première étape est la détection du visage. Elle se base sur la méthode proposée par Abdat. La dernière étape est la classification de l'expression dans l'une des six classes (joie, colère, peur, dégoût, tristesse, surprise). Nous nous sommes intéressés de plus près à la deuxième étape qui décrit l'expression. Trois approches d'extraction de caractéristiques faciales ont alors été étudiées. L'approche *basée sur l'aspect géométrique* code les mouvements faciaux en se basant sur le changement des distances entre des points caractéristiques. L'approche *basée sur l'aspect apparence* extrait les traits du visage en se basant sur un nouveau descripteur appelé 'Local Directional Double Ternary Coding Patterns (LDDTCP)'. La troisième approche *basée sur l'apprentissage profond* combine quatre CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et CNN_{LDP}) pour extraire les caractéristiques faciales. Dans le chapitre subséquent, nous donnons les résultats expérimentaux des trois approches inclus dans notre système de reconnaissance.

CHAPITRE 5

Expérimentations, Tests et Résultats

Sommaire

5.1 Introduction	163
5.2 Base de données d'expressions faciales	163
5.3 Critères d'évaluation	165
5.4 Tests et résultats	165
5.4.1 Tests et résultats de l'approche basée sur les caractéristiques géométriques pour la reconnaissance des expressions faciales	166
5.4.2 Tests et résultats de l'approche basée sur les caractéristiques d'apparences pour la reconnaissance des expressions faciales	168
5.4.3 Tests et résultats de l'approche basée sur l'apprentissage profond pour la reconnaissance des expressions faciales	176
5.5 Implémentation du système	184
5.5.1 L'application réalisée	185
5.6 Conclusion	189

5.1 Introduction

L'objectif de ce chapitre est de présenter les tests et les résultats des trois expérimentations que nous avons menées afin d'évaluer notre système de reconnaissance des expressions faciales. Nous commençons d'abord par donner une brève description sur les bases de données ainsi que les critères d'évaluations utilisés dans les tests, puis nous présentons les résultats expérimentaux des trois approches proposées (approche basée descripteurs géométriques, approche basée descripteurs d'apparences, approche basée apprentissage profond). Nous dressons également un état comparatif qui permet d'analyser chaque approche proposée par rapport aux travaux existants dans la littérature. Nous terminerons ce chapitre par la présentation de l'application réalisée tout en donnant quelques captures d'écrans de chacune des trois approches proposées.

5.2 Base de données d'expressions faciales

L'un des aspects les plus importants pour évaluer les méthodes de reconnaissance d'expressions faciales est le choix d'une ou plusieurs bases de données. Afin d'obtenir l'évaluation la plus objective possible, les bases de données doivent être générales et indépendantes des hypothèses ou des restrictions liées à un domaine précis. Ainsi, les performances des méthodes peuvent être référencées et comparées plus facilement.

Pour tester la validité de notre système proposé, nous avons utilisé les quatre bases de données suivantes:

- **Cohn-Kanade (CK+) [LUC 10]:**

La base de données CK+ est destinée aux travaux d'analyse et de synthèse automatiques d'images, ainsi qu'aux études de perception comme la reconnaissance des expressions faciales. Elle contient 593 séquences vidéo de 123 sujets, âgés de 18 à 30 ans. Ces sujets sont constitués de 65% de femmes, 15% d'afro-américains et 3% d'asiatiques et de latinos. La base CK+ comporte deux types d'expressions, posées et spontanées. Les séquences d'images présentées commencent par l'expression neutre et se terminent par une dernière image présentant le pic de l'expression demandée. La taille de ces images était 640×490 pixels avec une précision de 8 bits pour les niveaux de gris. L'orientation de la caméra est frontale et les petits mouvements de la tête sont présents. La figure 5.1 (a) présente un exemple extrait de cette base simulant les sept émotions.

- **Japanese Female Facial Expression (JAFPE) [LYO 99]:**

La base JAFPE contient 213 images en niveau de gris. Ces images concernent dix femmes japonaises qui simulent sept expressions faciales (les six émotions universelles et

l'expression neutre). Les images sont présentées au format d'image .tiff et ont une résolution de 256×256 pixels. Toutes les images sont frontales. La figure 5.1 (b) montre un exemple d'images de la base de données JAFFE.

- **Bi-modal Face and Body Gesture (FABO) [FAB 05]:**

La base bimodale des visages et des gestes corporels (FABO) est destinée aux travaux d'analyse automatique du comportement émotionnel non verbal de l'humain. Elle a été créée à la Faculté de Technologie de l'information de l'Université de Technologie de Sydney (UTS) par Hatice Gunes et Massimo Piccardi en 2005. Cette base contient des vidéos reflétant les expressions du visage et du corps humain enregistrées simultanément par des caméras du visage et du corps.

La base FABO contient 23 sujets dont 10 d'Europe, 2 de Moyen-Orient, 3 d'Amérique latine, 7 d'Asie et 1 d'Australie. Elle se compose de 12 femmes et 11 hommes, âgés de 18 à 50 ans. La figure 5.1 (c) présente un échantillon de la base de données FABO.

- **Database for Facial Expressions (DaFEx) [DAF 05]:**

La base de données DaFEx comporte 1008 courtes vidéos contenant les sept expressions faciales (les six émotions universelles et l'expression neutre). Les expressions faciales ont été enregistrées par 8 acteurs italiens professionnels (4 hommes et 4 femmes). La figure 5.1 (d) présente un exemple extrait de la base de données DaFEx.

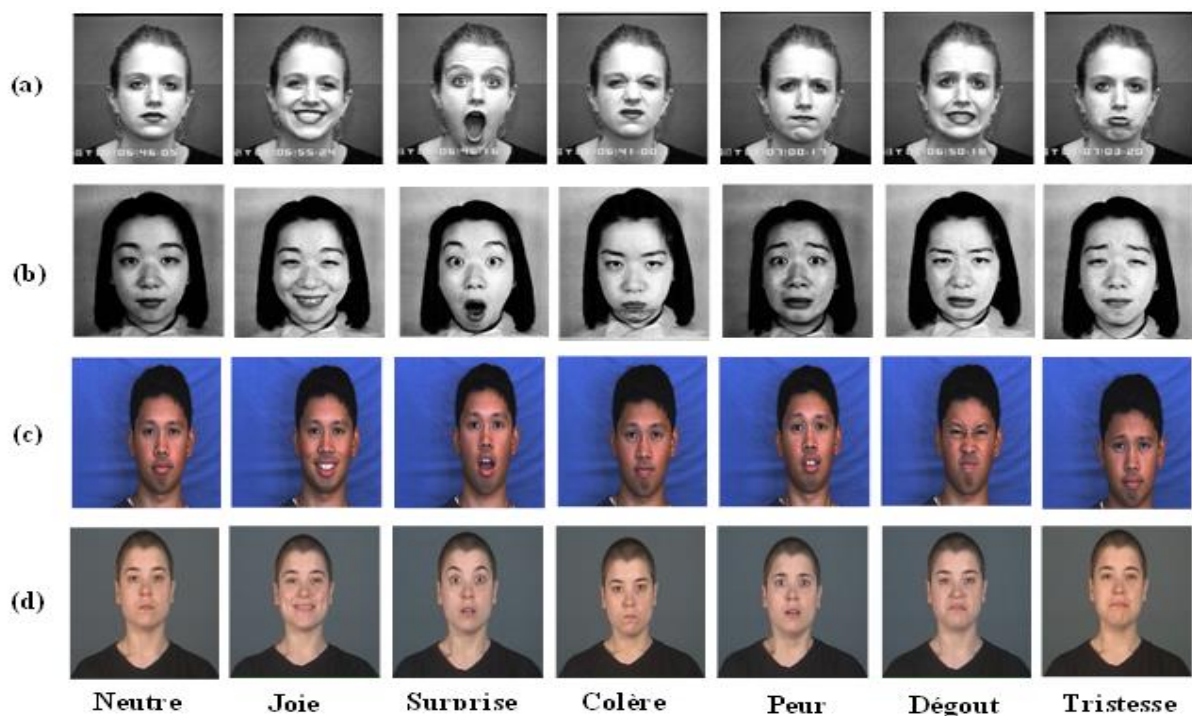


Figure 5.1 Base de données d'expressions faciales utilisée : (a) CK+, (b) JAFFE, (c) FABO, (d) DaFEx.

5.3 Critères d'évaluation

Pour évaluer l'efficacité de notre système proposé, nous avons utilisé les mesures de performance suivantes :

- Exactitude (ACC) : représente le rapport entre le nombre d'émotions correctes et le nombre total d'émotions.

$$Accuracy (ACC) = \frac{VP+VN}{VP+FP+VN+FN} \quad (5.1)$$

- Précision (PRE): représente la proportion entre les émotions correctement classées et l'ensemble de toutes les émotions classées.

$$Precision (PRE) = \frac{VP}{VP+FP} \quad (5.2)$$

- Sensibilité (Rappel) : représente le rapport entre les émotions correctement classées et l'ensemble de toutes les émotions classées attendues. Il indique la capacité du système à classer correctement toutes les émotions dans l'ensemble de données.

$$Sensitivity (SEN) = \frac{VP}{VP+FN} \quad (5.3)$$

- F1- mesure: représente la moyenne harmonique (pondéré) entre la précision et le rappel. Il est utilisé comme une mesure statistique pour évaluer les performances :

$$F1 - score (FS) = \frac{2*(Precision*Rappel)}{(Precision+Rappel)} \quad (5.4)$$

Où :

- VP (Vrai Positif) : indique le nombre d'émotions correctement classés,
- FP (Faux Positif) : représente le nombre d'émotion mal classés,
- VN (Vrai Négatif) : signifie le nombre d'émotions correctement classés qui ne sont pas classés comme étant des émotions analysées,
- FN (Faux Négatif) : représente le nombre d'émotions mal classés qui ne sont pas classés comme étant des émotions analysées.

5.4 Tests et résultats

Les tests et les résultats des trois approches proposées pour la reconnaissance des expressions faciales sont présentés dans les sections suivantes :

5.4.1 Tests et résultats de l'approche basée sur les caractéristiques géométriques pour la reconnaissance des expressions faciales

Dans cette expérimentation, nous avons choisi les deux bases de données FABO et DaFEx, puisqu'elles présentent les six émotions universelles et disposent d'un nombre important de poses acquises par quelques dizaines de personnes.

Pour tester l'effectivité de cette approche, le classificateur KNN est utilisé. Le choix de la valeur de K dépend de la nature, le type et la base de données utilisée. Cette valeur a un effet important sur la performance du classificateur KNN.

Dans notre expérimentation, après avoir effectué des tests avec différentes valeurs de K, nous avons observé que $K = 3$ est le meilleur choix pour notre travail dans les deux bases de données. La figure 5.2 montre l'impact des différentes valeurs de K sur le taux de reconnaissance pour les deux bases de données FABO et DaFEx respectivement.

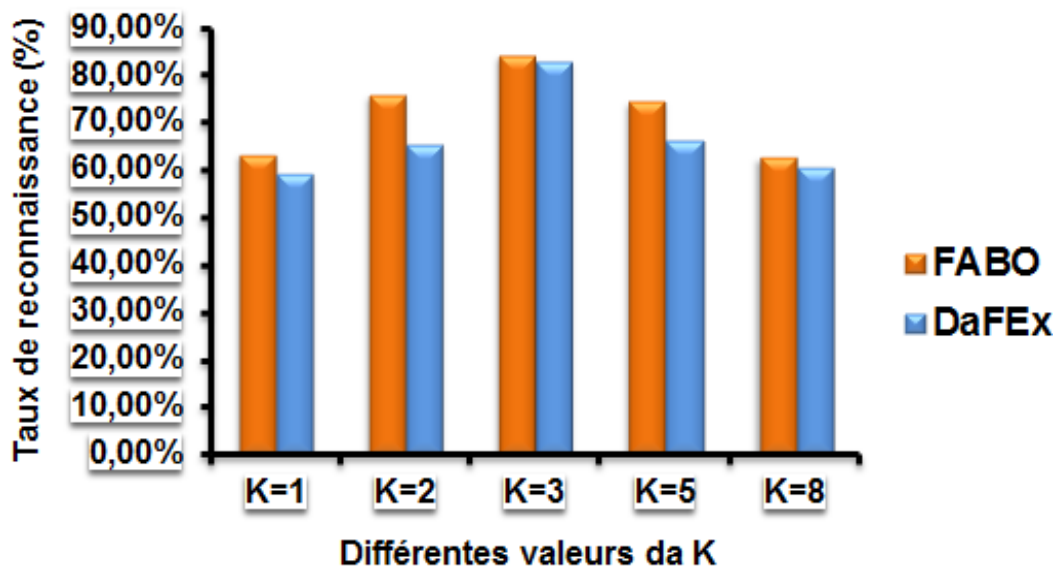


Figure 5.2 Impact des différentes valeurs de K sur la classification

A partir de la figure 5.2, nous pouvons constater que le classificateur KNN atteint des performances maximales lorsque la valeur de K est égale à 3. Après $K = 3$, le taux de reconnaissance diminue progressivement à 62,70 % et 60,84 % pour les deux bases de données FABO et DaFEx respectivement. .

Les résultats de test en utilisant le classificateur KNN avec $K = 3$ pour les deux bases de données FABO et DaFEx sont présenté par les deux matrices de confusion comme suit :

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	89,20%	5,80%	5%	0%	0%	0%
Dégout	4%	82,70%	1,30%	5%	6%	1%
Peur	3%	4,70%	78,30%	0%	6%	8%
Joie	0%	0%	9,10%	90,90%	0%	0%
Tristesse	1,80%	2,80%	0%	2,50%	84,50%	8%
Surprise	4%	6,00%	11,90%	0%	0%	78%

Tableau 5.1 Matrice de confusion pour 6-classes d'émotion de la base FABO (K=3)

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	87,30%	3,80%	6,70%	0%	2,50%	0%
Dégout	5,40%	81,60%	0,90%	2,10%	7%	3%
Peur	4%	3,10%	75,90%	0%	5,30%	11,70%
Joie	0%	0%	5,10%	91,90%	0%	3%
Tristesse	5,40%	4,60%	0%	3,80%	83,20%	3%
Surprise	2%	1%	22,30%	0%	0%	74,70%

Tableau 5.2 Matrice de confusion pour 6-classes d'émotion de la base DaFEx (K=3)

A partir des deux tableaux 5.1 et 5.2, il est possible d'observer que les meilleurs résultats de classification sont attribués aux émotions de *joie* et de *colère* avec respectivement 90,9 %, 89,2 % pour la base de données FABO et 91,9 %, 87,3 % pour la base de données DaFEx. Les émotions de *tristesse* et de *dégout* ont leurs taux de classification dans l'intervalle de 82 % à 84% pour la base FABO et de 81 % à 83 % pour la base DaFEx. Le mauvais résultat de classification dans les deux bases de données correspond à l'émotion de *surprise* qui est souvent confondue avec la peur. De plus, l'émotion de *peur* est assez souvent confondue avec la surprise.

Le tableau 5.3 montre les résultats obtenus pour le classificateur KNN en utilisant les mesures de performances :

Critères Evaluation BDD	Accuracy	Précision	Sensibilité	F1-Score
FABO	84,07%	84,18%	84,06%	84,12%
DaFEx	82,39%	82,69%	82,38%	82,54%

Tableau 5.3 Résultats obtenus pour le classificateur KNN (K = 3).

5.4.2 Tests et résultats de l'approche basée sur les caractéristiques d'apparences pour la reconnaissance des expressions faciales

Dans cette expérimentation, nous avons choisi les deux bases de données CK+ et JAFFE, car elles sont les plus utilisées et elles permettent de comparer plus facilement de nouvelles méthodes à celles de l'état de l'art.

Pour la base CK+, nous avons sélectionné 576 séquences d'images à partir de 96 sujets, avec 6 émotions de base par sujet. Pour obtenir une base de données avec 7 classes d'expressions, 192 images faciales neutres ont été ajoutées à la base de données contenant les 6 classes d'expressions. Pour la base JAFFE, nous avons choisi un ensemble de 210 séquences d'images à partir de 10 sujets (trois copies de chaque expression pour chaque individu). A partir de 210 images, 140 images sont utilisées pour l'apprentissage et les 70 images restantes sont utilisées pour le test.

Nous avons utilisé le classificateur statistique supervisé SVM pour faire la correspondance entre deux histogrammes (histogrammes des images d'apprentissage avec les histogrammes des images de test) et classer les vecteurs de caractéristiques dans l'une des six (sept) expressions faciales universelles.

Afin d'implémenter le classificateur SVM, nous avons opté pour la machine d'apprentissage SVM^{light} [JOA 99] et la fonction noyau RBF (Radial Basis functions).

Pour tester l'efficacité du nouveau descripteur 'LDDTCP' proposé, nous avons effectué les deux expérimentations suivantes :

5.4.2.1 Expérimentation 1 : Test et résultats en utilisant le visage entier

Les performances de la méthode proposée (LDDTCP) peuvent être affectées par le nombre de régions dans lesquelles les images sont partitionnées. Dans notre cas, nous avons divisé l'image de visage entier en régions (voir la figure 5.3) de taille 3×3 , 5×5 , 7×6 et 7×8 [CHE 22b].

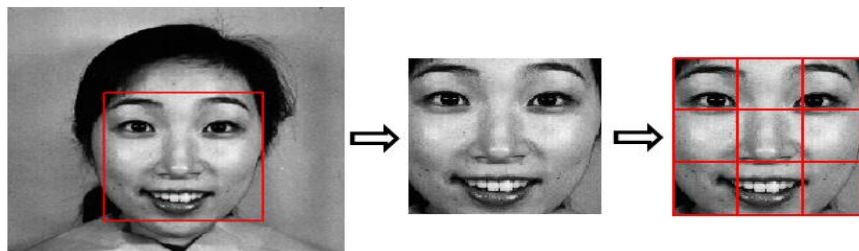


Figure 5.3 Division de l'image de visage entier en régions de taille 3x3

La figure 5.4 et 5.5 montre l'impact de la variation de la taille de la région sur le taux de reconnaissance du descripteur LDDTCP dans les deux bases de données CK+ et JAFFE pour 6 et 7 classes d'expressions respectivement [CHE 22b].

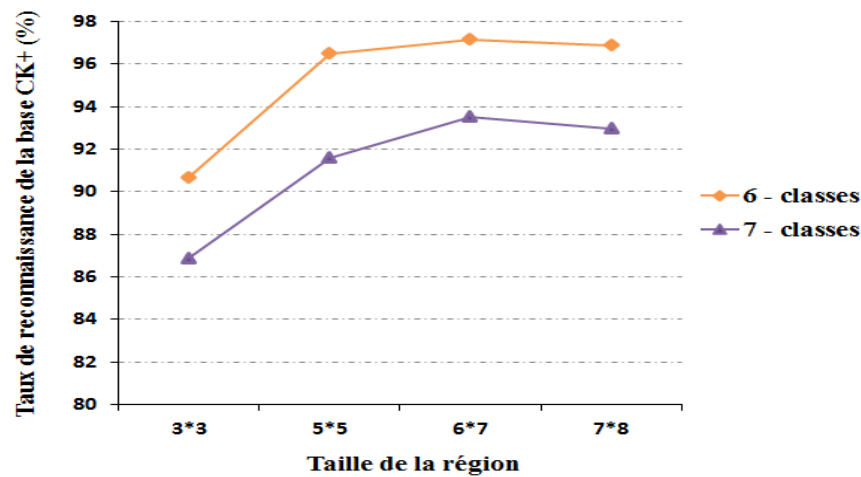


Figure 5.4 Taux de reconnaissance versus variation de la taille de région pour la base CK+

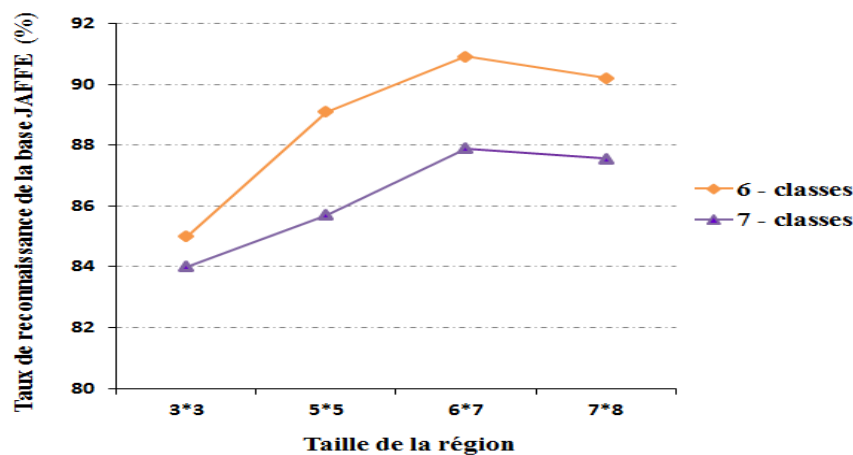


Figure 5.5 Taux de reconnaissance versus variation de la taille de région pour la base JAFFE

Il est à noter que le partitionnement d'une image dans un nombre variable de régions augmente le taux de classification. Par conséquent, la longueur du vecteur de caractéristiques et l'efficacité de calcul seront aussi augmentées.

Pour les deux bases de données CK+ et JAFFE, le taux de classification le plus élevé est obtenu pour les images partitionnées en 7×6 régions. D'un côté, pour les 6 classes d'expressions, le descripteur LDDTCP atteint un taux de reconnaissance optimal avec 97,18% pour la base CK+ et 90,92% pour la base JAFFE. D'un autre côté, pour les 7 classes d'expressions, le taux de reconnaissance est de 93,54% pour la base CK+ et de 87,89% pour la base JAFFE. L'augmentation de la taille de la région au-delà de 7×6 régions diminue progressivement le taux de reconnaissance pour les deux bases de données.

Les résultats de test du descripteur LDDTCP pour les deux bases de données CK+ et JAFFE en utilisant 6 classes d'expression et 7×6 régions sont présentés sous forme de matrices de confusion dans les tableaux 5.4 et 5.5 ci-dessus [CHE 22b].

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	98,4%	0%	0%	0,2%	0%	1,4%
Dégout	0%	97%	0%	1,9%	0%	1,1%
Peur	0%	1%	95,9%	0%	0%	3,1%
Joie	0,6%	0,7%	0%	98,7%	0%	0%
Tristesse	0%	2,2%	0%	0%	97,4%	0,4%
Surprise	0%	0%	3,9%	0,4%	0%	95,7%

Tableau 5.4 Matrice de confusion (%) de la base CK+ en utilisant 7×6 régions et 6 classes d'expression

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	92,5%	0%	0,5%	2,4%	0%	4,6%
Dégout	0,6%	90,6%	0%	2,7%	1,8%	4,3%
Peur	0%	3,8%	89,3%	0%	0%	6,9%
Joie	3,1%	3,8%	0%	92,9%	0%	0,2%
Tristesse	0%	8,9%	0%	0%	91,1%	0%
Surprise	0%	0%	7,8%	3,1%	0%	89,1%

Tableau 5.5 Matrice de confusion (%) de la base JAFFE en utilisant 7×6 régions et 6 classes d'expression

En général, nous observons que les émotions de *joie*, *colère*, *tristesse* et *dégoût* sont classées avec une grande précision pour les deux bases de données. Les taux de reconnaissance des émotions *peur* et *surprise* sont inférieurs à la moyenne, car de nombreuses émotions de peur sont confondues avec des émotions de surprise et vice versa.

Les tableaux 5.6 et 5.7 présentent les matrices de confusion du descripteur LDDTCP pour les bases de données CK+ et JAFFE en utilisant 7 classes d'émotion et 7×6 régions.

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise	Neutre
Colère	97,1%	0%	0%	0%	0%	0%	2,9%
Dégout	0%	95%	2,6%	0,3%	0%	2,1%	0%
Peur	0%	1,8%	96,2%	0%	0%	2%	0%
Joie	2,1%	2%	0%	90%	0%	5,9%	0%
Tristesse	1%	0%	0%	0%	91,2%	0%	7,8%
Surprise	0%	0,8%	1%	2,2%	0%	96%	0%
Neutre	7,5%	0%	0%	0%	3,2%	0%	89,3%

Tableau 5.6 Matrice de confusion (%) de la base CK+ en utilisant 7×6 régions et 7 classes d'expression

Vrai Label \ Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise	Neutre
Colère	91,9%	0%	0,2%	2,8%	0%	0%	5,1%
Dégout	0%	87,9%	5,6%	0%	0%	6,5%	0%
Peur	0%	5,4%	91,3%	0%	0%	3,3%	0%
Joie	3,2%	2,8%	0%	85,2%	0%	8,8%	0%
Tristesse	0%	2,1%	0%	1,9%	86,4%	0%	9,6%
Surprise	0,2%	1,8%	0%	9,1%	0%	88,9%	0%
Neutre	8,4%	0,5%	1%	0%	6,5%	0%	83,6%

Tableau 5.7 Matrice de confusion (%) de la base JAFFE utilisant 7×6 régions et 7 classes d'expression

D'après les tableaux 5.6 et 5.7, nous remarquons que les émotions de *colère* et de *peur* sont relativement bien détectées avec un taux de reconnaissance de 97,1%, 96,2% pour la base CK+ et de 91,9%, 91,3% respectivement pour la base de données JAFFE, ce qui indique que les réactions faciales lorsqu'une personne est en colère ou en peur sont plus uniformes que les autres émotions. Les émotions de *surprise* et de *dégoût* sont assez bien reconnues avec 96 %, 95 % pour la base CK+ et 88, 9 %, 87, 9 % pour la base JAFFE. Les émotions de *tristesse* et de *joie* ont leurs taux de classification dans l'intervalle de 90 % à 91 % pour la base de données CK+ et de 85 % à 86 % pour la base de données JAFFE. Le mauvais résultat de classification dans les deux bases de données correspond à l'émotion neutre qui est souvent confondue avec les émotions de colère et de tristesse et vice versa.

5.4.2.2 Expérimentation 2 : Test et résultats en utilisant les composantes faciales

La technique de division des images de visage dans différentes régions est également utilisée lorsque nous travaillons avec des composantes faciales. Dans cette expérimentation, nous avons divisé chacune des composantes faciales en blocs de taille : 1×2 , 2×3 , 3×4 et 4×5 blocs. La figure 5.6 montre un exemple de partitionnement des composantes faciales en blocs de taille 3×4 [CHE 22b].

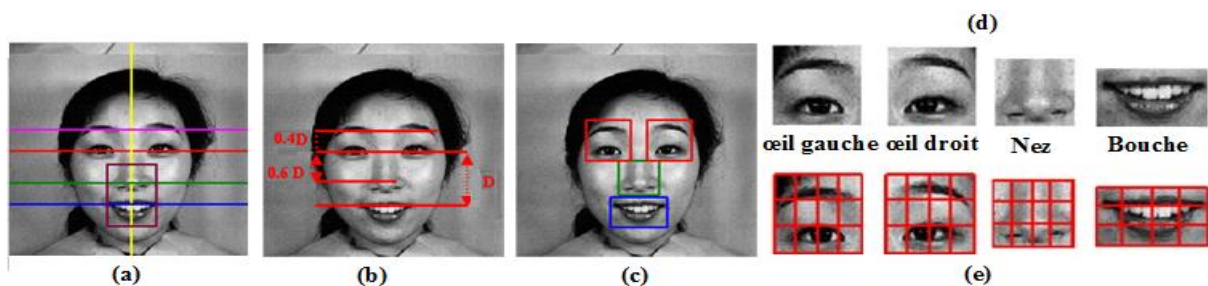


Figure 5.6 Procédure d'extraction des composantes faciales : (a) Détection des différents axes de visage, (b) Application du modèle géométrique, (c) Localisation des composantes faciales, (d) Extraction des composantes faciales, (e) Segmentation de chacune des composantes faciales en blocs de taille 3×4 [CHE 22b].

Le tableau 5.8 résume le taux de reconnaissance du descripteur LDDTCP pour les deux bases de données CK+ et JAFFE en utilisant quatre composantes faciales (œil gauche, œil droit, nez, bouche) et 6 - 7 classes d'expressions respectivement.

Descripteur LDDTCP		Taux de classification en utilisant différents nombres de régions pour chaque composante faciale			
		1×2	2×3	3×4	4×5
6 classe	CK+	91,2%	92,4%	97,93%	96,8%
	JAFFE	86,5%	90,7%	92,22%	90,9%
7 classe	CK+	88,5%	93,3%	95,32%	94,7%
	JAFFE	86,6%	87,1%	90,81%	89,5%

Tableau 5.8 Taux de reconnaissance (%) pour les bases de données CK+ et JAFFE en utilisant quatre composantes faciales et 6 - 7 classes d'expressions [CHE 22b].

A partir du tableau 5.8, nous pouvons remarquer que l'augmentation de la taille des blocs pour chaque composante faciale entraîne une augmentation dans le taux de reconnaissance pour les deux bases de données (CK+ et JAFFE). Nous pouvons aussi constater que le descripteur LDDTCP atteint des performances optimales lorsque les composantes faciales sont partitionnées en blocs de taille 3×4 . Pour les 6 classes d'expression, LDDTCP atteint un taux de reconnaissance élevé avec 97,93 % pour la base CK+ et 92,22 % pour la base JAFFE. En revanche, pour les 7 classes d'expressions, le taux de reconnaissance est de 95,32% pour la base CK+ et de 90,81% pour la base JAFFE.

Il est à noter que l'augmentation de la taille des blocs au-delà de 3×4 produit des informations redondantes et inutiles qui dégradent progressivement les performances.

Les tableaux 5.9 et 5.10 présentent les matrices de confusion du descripteur LDDTCP pour les bases de données CK+ et JAFFE en utilisant quatre composantes faciales avec 3×4 blocs et 6 classes d'expressions :

Vrai Label \ Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	97,8%	0%	0%	0,6%	0%	1,6%
Dégout	1%	98,2%	0%	0%	0,80%	0%
Peur	0%	1%	96,9%	0%	0%	2,1%
Joie	0,8%	0,2%	0%	98,8%	0%	0,2%
Tristesse	0,5%	0%	0%	0%	99,5%	0%
Surprise	0%	0%	0,9%	2,7%	0%	96,4%

Tableau 5.9 Matrice de confusion (%) de la base CK+ en utilisant quatre composantes faciales avec 3×4 blocs et 6 classes d'expressions

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	92,6%	0%	0%	2,6%	0%	4,8%
Dégout	3,7%	92,9%	0%	0%	3,4%	0%
Peur	0%	4,3%	90,9%	0%	0%	4,8%
Joie	0,9%	3,5%	0%	93,1%	0%	2,5%
Tristesse	1,9%	4,6%	0%	0%	93,5%	0%
Surprise	1,8%	1,3%	0,9%	5,7%	0%	90,3%

Tableau 5.10 Matrice de confusion (%) de la base JAFFE en utilisant quatre composantes faciales avec 3 x 4 blocs et 6 classes d'expressions

D'après les tableaux 5.9 et 5.10, nous pouvons constater que les taux de reconnaissance de la tristesse, de joie, de dégoût et de la colère sont supérieurs à la moyenne pour les deux bases de données. Le mauvais taux de classification correspond, encore une fois, aux émotions de peur et de surprise avec 96,9 % et 96,4 % respectivement pour la base CK+ et 90,9 % et 90,3 % respectivement pour la base JAFFE. Les tableaux 5.11 et 5.12 illustrent les matrices de confusion du descripteur LDDTCP pour les bases de données CK+ et JAFFE en utilisant quatre composantes faciales avec 3 x 4 blocs et 7 classes d'expressions :

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise	Neutre
Colère	98,9%	0,2%	0%	0%	0%	0,4%	0,5%
Dégout	0%	95,8%	2,6%	0,2%	0,2%	1,2%	0%
Peur	0%	0,8%	98,1%	0%	0%	0,9%	0,2%
Joie	0,8%	0,9%	0%	93,3%	0%	4,7%	0,3%
Tristesse	0,5%	0%	0%	0%	93,5%	0%	6%
Surprise	0,1%	1,6%	0%	1,8%	0,1%	96,4%	0%
Neutre	3,4%	0%	1,6%	0%	3,8%	0%	91,2%

Tableau 5.11 Matrice de confusion (%) de la base CK+ en utilisant quatre composantes faciales avec 3 x 4 blocs et 7 classes d'expressions

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise	Neutre
Colère	94,2%	0%	0,2%	1,9%	0%	0%	3,7%
Dégout	0%	90,8%	4,6%	0%	0,4%	4,2%	0%
Peur	0%	3,1%	93,9%	0%	0%	2,5%	0,5%
Joie	2,4%	2,6%	0%	88,7%	0%	5,3%	1%
Tristesse	0%	0%	0%	0%	89,8%	0%	10,2%
Surprise	0,8%	0,9%	0%	6,7%	0,3%	91,3%	0%
Neutre	4,3%	0,1%	1,7%	0%	7%	0%	86,9%

Tableau 5.12 Matrice de confusion (%) de la base JAFFE en utilisant quatre composantes faciales avec 3 x 4 blocs et 7 classes d'expressions

À partir des tableaux 5.11 et 5.12, nous pouvons observer que les meilleurs résultats de reconnaissance sont attribués aux émotions de colère et de peur avec respectivement 98,9 % et 98,1 % pour la base CK+ et 94,2 % et 93,9 % pour la base JAFFE. Les émotions de surprise, dégoût, joie et tristesse ont leurs taux de reconnaissance dans l'intervalle de 93 % à 96 % pour la base CK+ et de 88 % à 91 % pour la base JAFFE. Les taux de reconnaissance des émotions de joie et de tristesse sont inférieurs à la moyenne. Le mauvais résultat de classification dans les deux bases de données correspond à l'émotion neutre qui est confondue avec les émotions de colère, de peur et de tristesse et vice versa [CHE 22b].

5.4.2.3 Comparaison des résultats du descripteur LDDTCP avec les autres travaux de l'état de l'art

Pour mieux démontrer les avantages de la méthode proposée, nous avons comparé le descripteur LDDTCP avec d'autres descripteurs d'apparences utilisés pour la reconnaissance des expressions faciales tels que : LBP [SHA 09], LTP [TAN 10] , LDP [JAB 10a], Compound LBP (CLBP) [AHM 11], Gradient Directional Pattern (GDP) [AHM 12], Gradient Local Ternary Pattern (GLTP) [AHM 13], Directional Ternary Pattern (DTP) [AHM 18] et compressed DTP(cDTP) [AHM18].

Les tableaux 5.13 à 5.16 présentent une comparaison des taux de classification entre le descripteur LDDTCP et les autres méthodes de l'état de l'art pour les deux bases de données CK+ et JAFFE [CHE 22b].

Opérateurs	Taux de reconnaissance des différents nombre de régions		
	3x3	5x5	6x7
LBP	79,3%	89,7%	90,1%
LTP	87,3%	92,3%	93,6%
LDP	80,2%	91,9%	93,7%
CLBP	88,2%	94,4%	94,2%
GDP	86,8%	95,4%	95,9%
GLTP	90,5%	96,4%	97,2%
DTP	94,5%	97,1%	97,5%
cDTP	94,2%	96,8%	97,3%
Méthode proposée LDDTCP	90,7%	96,5%	97,18%

Tableau 5.13 Taux de classification des différents descripteurs de texture en utilisant la base CK+ et 6 classes d'expressions

Operateurs	Taux de reconnaissance des différents nombre de régions		
	3x3	5x5	6x7
LBP	73,8%	80,9%	83,3%
LTP	81,3%	88,5%	88,9%
LDP	75,7%	86,3%	88,4%
CLBP	82,1%	90,4%	89,2%
GDP	81,9%	90,4%	91,6%
GLTP	84%	90,6%	91,7%
DTP	90,3%	93,9%	95,8%
cDTP	89,9%	93,5%	95,2%
Méthode proposée LDDTCP	86,9%	91,6%	93,54%

Tableau 5.14 Taux de classification des différents descripteurs de texture en utilisant la base CK+ et 7 classes d'expressions

Operateurs	Taux de reconnaissance des différents nombre de régions		
	3x3	5x5	6x7
LBP	84,1%	87,6%	90,5%
LTP	84,3%	87,9%	90,9%
LDP	83,2%	88,9%	90,7%
DTP	87,5%	90,1%	92,5%
cDTP	87,1%	89,8%	91,2%
Méthode proposée LDDTCP	84,9%	89,1%	90,92%

Tableau 5.15 Taux de classification des différents descripteurs de texture en utilisant la base JAFFE et 6 classes d'expressions

Operateurs	Taux de reconnaissance des différents nombre de régions		
	3x3	5x5	6x7
LBP	81,5%	82,3%	85,3%
LTP	84,6%	85%	85,7%
LDP	83,3%	85,3%	85,9%
DTP	85,3%	86,9%	88,7%
cDTP	85,9%	86,5%	88%
Méthode proposée LDDTCP	84,8%	85,7%	87,89%

Tableau 5.16 Taux de classification des différents descripteurs de texture en utilisant la base JAFFE et 7 classes d'expressions

A partir des tableaux ci-dessus, nous pouvons constater que le descripteur LDDTCP proposé a atteint des performances supérieures en termes de taux de reconnaissance par rapport à

certaines descripteurs d'apparences tels que LBP, LTP, LDP, CLBP, GDP et GLTP pour les deux bases de données CK+ et JAFFE en utilisant 6 et 7 classes d'expressions. Cependant, le taux de classification du descripteur LDDTCP est inférieur aux descripteurs DTP et cDTP.

Nous pouvons aussi remarquer que le descripteur LDDTCP atteint des performances optimales lorsqu'on utilise les composantes faciales, comparativement aux taux de reconnaissance obtenus lorsqu'on utilise le visage entier pour les deux bases de données, car l'utilisation de composantes faciales conserve plus d'informations spatiales que l'utilisation du visage entier [CHE 22b].

5.4.3 Tests et résultats de l'approche basée sur l'apprentissage profond pour la reconnaissance des expressions faciales

Dans cette expérimentation, nous avons utilisé les deux bases de données CK+ et JAFFE. Pour la base CK+, nous avons sélectionné 720 séquences d'images comportant les six expressions de base. Parmi les 720 images choisies, 504 images sont utilisées pour l'apprentissage et les 216 images restantes sont utilisées pour le test. Pour la base JAFFE, nous avons choisi un ensemble de 180 séquences d'images à partir de 10 sujets (trois copies de chaque expression pour chaque individu). Parmi les 180 images, 130 images sont utilisées pour l'apprentissage et les 50 images restantes sont utilisées pour le test.

L'objectif principal de cette expérimentation est de concevoir un système de reconnaissance des expressions faciales qui combine les résultats des quatre classificateurs (CNN_{LBP} , CNN_{LDP} , CNN_{LGC} et CNN_{GDP}) pour atteindre des performances supérieures en termes de taux de reconnaissance par rapport aux autres systèmes de la littérature. Dans la section suivante, nous décrirons les résultats d'apprentissage, de test et de fusionnement :

5.4.3.1 Résultats d'apprentissage

Les quatre modèles CNN (CNN_{LBP} , CNN_{LGC} , CNN_{LDP} et CNN_{GDP}) sont entraînés à l'aide de l'optimiseur descente de gradient stochastique (SGD) avec une valeur de momentum de 0,9 pour les poids et les biais. L'optimiseur SGD est utilisé pour ajuster les pondérations et minimiser l'erreur quadratique moyenne entre la sortie de la couche Softmax (couche de perte) et la sortie désirée (ensemble des classes). Le taux d'apprentissage est fixé à 0,001. La norme L2 est appliquée pour régulariser les poids de la couche entièrement connectée avec une valeur de 0,005. Le taux de drop-out de chaque couche entièrement connectée est fixé à 0,5 [CHE 22a].

Après l'entraînement des quatre modèles CNN, différentes performances sont obtenues. La figure 5.7 présente les variations du taux de reconnaissance en fonction du nombre

d'époques durant la phase d'apprentissage des quatre modèles CNN pour la base de données CK+ [CHE 22a].

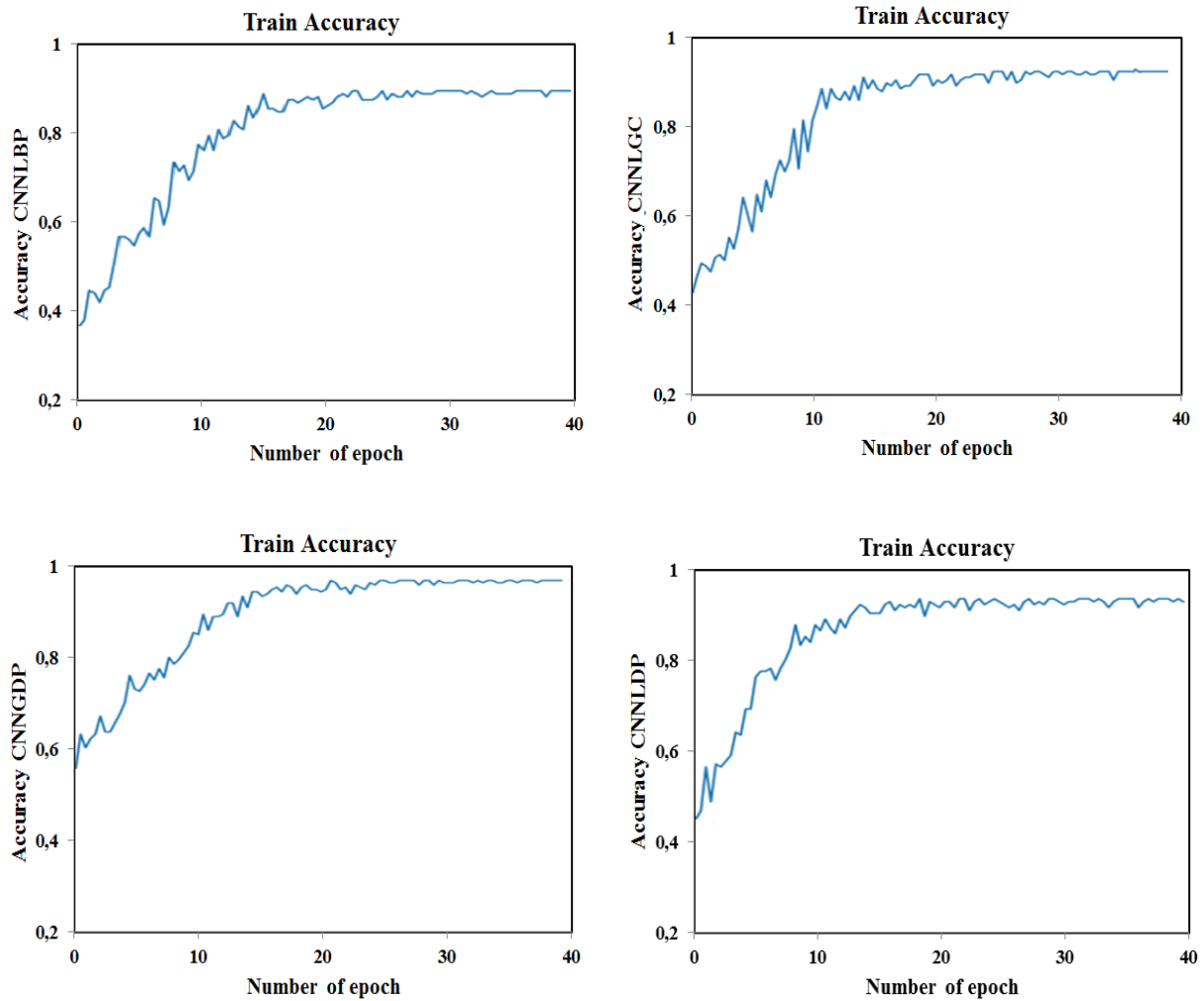


Figure 5.7 Variation du nombre d'époque versus le taux de reconnaissance des quatre modèles CNN pour la base CK+ [CHE 22a].

D'après la figure 5.7, nous pouvons remarquer que le taux de reconnaissance durant la phase d'apprentissage augmente progressivement dans les premières époques et se stabilise autour de l'époque 40 pour les quatre modèles CNN. Le meilleur taux de reconnaissance est attribué au CNN_{LDP} et CNN_{GDP} avec 96,5 % et 98,2 % respectivement. Nous pouvons également remarquer que le taux de reconnaissance des quatre modèles CNN en utilisant la base JAFFE se stabilise autour de l'époque 40 et il est inférieur de la base CK+.

Pour calculer le taux d'erreur des quatre modèles CNN, une fonction de perte doit être utilisée. La valeur de perte indique dans quelle mesure chaque modèle fonctionne après chaque itération d'optimisation. La figure 5.8 montre les variations du taux d'erreur en fonction du nombre d'époque durant la phase d'apprentissage des quatre modèles CNN en utilisant la base de données CK+ [CHE 22a].

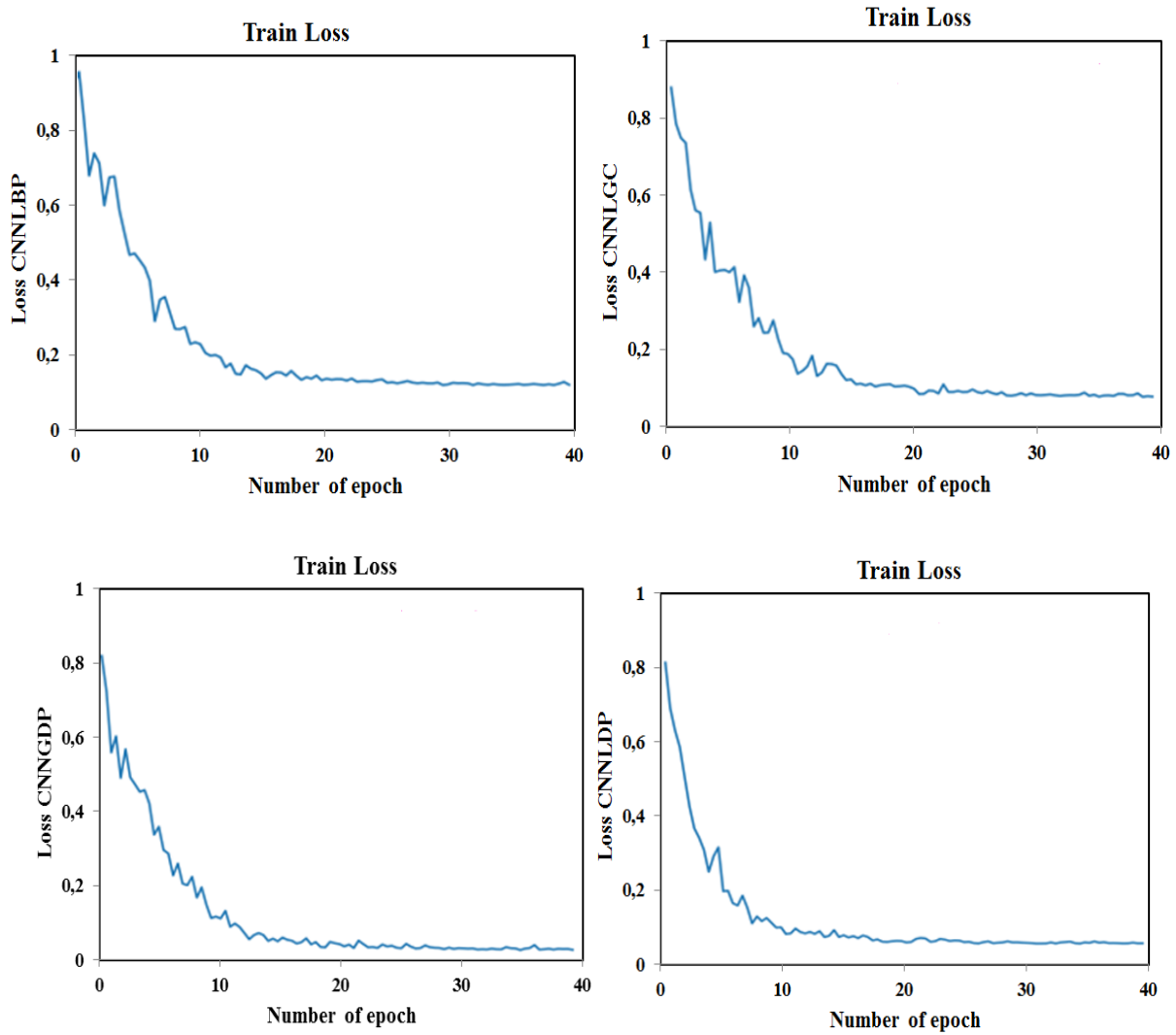


Figure 5.8 Variation du nombre d'époque versus le taux d'erreur des quatre modèles CNN pour la base CK+ [CHE 22a].

Nous observons que le taux d'erreur durant la phase d'apprentissage diminue progressivement dans les premières époques et se stabilise autour de l'époque 40 pour les quatre modèles CNN, ce qui indique que chaque modèle améliore progressivement ses performances jusqu'à l'époque 40 [CHE 22a].

5.4.3.2 Résultats de test

Les résultats du test obtenus en appliquant les quatre modèles CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et CNN_{LDP}) générés pendant la phase d'apprentissage pour les deux bases de données CK+ et JAFFE sont présentés comme suit :

A. Résultats de test en utilisant la base CK+

Les matrices de confusion des quatre modèles CNN pour la base CK+ utilisant 6 classes d'expressions sont présentées dans les tableaux 5.17 – 5.20 ci-dessous.

VL PL	Co	De	Pe	Jo	Tr	Su
Co	90,9	3,3	2,6	0	3,2	0
De	4,7	86	2,9	0	6,4	0
Pe	3,8	2,8	90,3	0	0	3,1
Jo	1,1	0	1,5	93,1	0	4,3
Tr	0	4,2	2,9	0	90	2,9
Su	1,9	0	2,4	3,1	0,3	92,3

Tableau 5.17 Résultats de test pour CNN_{LBP}

VL PL	Co	De	Pe	Jo	Tr	Su
Co	92,1	1,7	3,4	0	2,8	0
De	3,7	87	3,4	0	5,9	0
Pe	2,9	1,4	92,5	0	0	3,2
Jo	0,9	0	0,7	95,6	0	2,8
Tr	0	4,4	1,2	0,0	92	2,4
Su	0,5	0	0,8	3,2	0,8	94,7

Tableau 5.18 Résultats de test pour CNN_{LGC}

VL PL	Co	De	Pe	Jo	Tr	Su
Co	96,6	0,6	0,7	0,5	1,6	0,0
De	2	93,4	1,2	0,0	3,4	0,0
Pe	0,7	2,7	96	0	0	0,6
Jo	0	0	0	98	0	2
Tr	0	1,9	0	0	96,3	1,8
Su	1	0	1,3	1,8	0	95,9

Tableau 5.19 Résultats de test pour CNN_{GDP}

VL PL	Co	De	Pe	Jo	Tr	Su
Co	97,4	0,9	0,3	0	1,4	0
De	1,7	95	0,5	0	2,8	0
Pe	0,8	2,3	96,3	0	0	0,6
Jo	0	0	0,8	98	0	1,2
Tr	0	2,6	0	0	96,5	0,9
Su	0,9	0	0,7	0,9	0	97,5

Tableau 5.20 Résultats de test pour CNN_{LDP}

Remarque: Co: Colère, De: Dégout, Pe: Peur, Jo: Joie, Tr: Tristesse, Su: Surprise, VL: Vrai Label et PL: Predict Label

D'après les tableaux 5.17 à 5.20, nous pouvons observer que les émotions de joie et de surprise sont relativement bien détectées pour les quatre modèles CNN, ce qui indique que les réactions faciales lorsqu'une personne est heureuse et surprise sont plus uniformes que les autres émotions. Les émotions de la colère, la peur et la tristesse sont assez bien reconnues et leurs taux de classification se situent entre 90 % et 98 % pour les quatre modèles CNN. Le taux de reconnaissance du dégoût est inférieur à la moyenne, car de nombreuses émotions de dégoût sont confondues avec les émotions de colère, de peur et de tristesse et vice versa [CHE 22a].

B. Résultats de test en utilisant la base JAFFE

Les tableaux 5.21 à 5.24 présentent les matrices de confusion des quatre modèles CNN en utilisant la base de données JAFFE et 6 classes d'expressions.

VL PL	Co	De	Pe	Jo	Tr	Su
Co	89,9	3,5	2,4	0,1	4,1	0
De	6,6	79	7,8	0	5,3	1,3
Pe	4,9	5,8	83,8	0	5,5	0
Jo	2,3	0	3,8	88,8	0	5,1
Tr	1,7	5,9	3,8	0	85,1	3,5
Su	3,1	0	3,8	5,1	0	88

Tableau 5.21 Résultats de test pour CNN_{LBP}

VL PL	Co	De	Pe	Jo	Tr	Su
Co	91	3,3	1,9	0,4	3,4	0
De	6	81,4	6,9	0	5,7	0
Pe	3,7	4,1	86,2	0	4,9	1,1
Jo	1,9	0,0	2,4	90,1	0	5,6
Tr	0,8	4,6	3,7	0	87	3,9
Su	2,7	0	3,6	4,4	0	89,3

Tableau 5.22 Résultats de test pour CNN_{LGC}

VL PL	Co	De	Pe	Jo	Tr	Su
Co	94,7	2,3	0,8	0,1	2,1	0
De	3,5	87	5,2	0	4,3	0
Pe	2,2	3,8	90	0	3,6	0,4
Jo	0,9	0	1,2	93,4	0	4,5
Tr	0,5	3,1	2,4	0	91,7	2,3
Su	2	0	2,3	3,6	0	92,1

Tableau 5.23 Résultats de test pour CNN_{GDP}

VL PL	Co	De	Pe	Jo	Tr	Su
Co	96	1,5	0,9	0,1	1,5	0
De	2,1	89,3	4,7	0	3,9	0
Pe	1,9	4	91,9	0	2,2	0
Jo	0,4	0	1,4	95	0	3,2
Tr	0,1	2,7	2	0	93,1	2,1
Su	1,1	0	1,2	3,4	0	94,3

Tableau 5.24 Résultats de test pour CNN_{LDP}

À partir des tableaux 5.21 à 5.24, nous pouvons observer que les quatre modèles CNN sont capables de classer les émotions de colère et de joie avec une grande précision par rapport aux autres émotions. Les taux de reconnaissance de la peur, de la tristesse et de la surprise sont supérieurs à la moyenne pour les quatre modèles. Le mauvais résultat de classement correspond, une autre fois, à l'émotion de dégoût qui est souvent confondue avec les émotions de colère, de peur et de tristesse et inversement [CHE 22a].

Le tableau 5.25 résume les résultats obtenus des quatre modèles CNN pour les deux bases de données CK+ et JAFFE en utilisant les critères d'évaluation:

Critères Evaluation BDD / Modèle		Accuracy	Précision	Sensibilité	F1-Score
CK+	CNN _{LBP}	90,43%	90,49%	90,43%	90,44%
	CNN _{LGC}	92,31%	92,34%	92,32%	92,31%
	CNN _{GDP}	96,03%	96,04%	96,03%	96,05%
	CNN _{LDP}	96,78%	96,79%	96,78%	96,80%
JAFFE	CNN _{LBP}	85,76%	85,95%	85,77%	85,80%
	CNN _{LGC}	87,50%	87,62%	87,50%	87,51%
	CNN _{GDP}	91,48%	91,52%	91,49%	91,48%
	CNN _{LDP}	93,26%	93,28%	93,27%	93,27%

Tableau 5.25 Résultats de test des quatre modèles CNN pour les deux bases de données CK+ et JAFFE

5.4.3.3 Résultats de fusionnement en parallèle (agrégation par somme / produit)

Les résultats de fusionnement en parallèle en utilisant les méthodes d'agrégation de somme et de produit pour les deux bases de données CK+ et JAFFE sont présentés comme suit:

A. Résultats de fusionnement en utilisant la base CK+

Les tableaux 5.26 et 5.27 présentent les résultats de fusionnement en utilisant les méthodes d'agrégation de somme et de produit pour la base de données CK+ avec 6 classes d'expressions.

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	97,9%	0,8%	0,4%	0%	0,9%	0%
Dégout	2,8%	94%	0,3%	0%	2,9%	0%
Peur	1,3%	0,5%	97%	0%	0%	1,2%
Joie	0,5%	0%	0%	99,1%	0%	0,4%
Tristesse	0%	1,6%	0,5%	0%	97,1%	0,8%
Surprise	0%	0%	0,8%	0,9%	0%	98,3%

Tableau 5.26 Résultats de fusionnement en utilisant la règle somme et la base CK+

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	98,1%	1,2%	0%	0%	0,7%	0%
Dégout	2,8%	95,3%	0%	0%	1,9%	0%
Peur	1,1%	0%	97,9%	0%	0%	1%
Joie	0%	0%	0%	99,5%	0%	0,5%
Tristesse	0%	0,8%	0%	0%	97,8%	1,4%
Surprise	0%	0%	0,6%	0,4%	0%	99%

Tableau 5.27 Résultats de fusionnement en utilisant la règle produit et la base CK+

B. Résultats de fusionnement en utilisant la base JAFFE

Les matrices de confusion des résultats de fusionnement en utilisant les méthodes d'agrégation de somme et de produit pour la base de données JAFFE avec 6 classes d'expressions sont présentées dans les tableaux de 5.28 à 5.29.

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	96,9%	1,5%	0%	0%	1,6%	0%
Dégout	3,2%	88,1%	4,8%	0%	3,9%	0%
Peur	1,7%	2,6%	92%	0%	3,7%	0%
Joie	0%	0%	0,6%	95,8%	0%	3,6%
Tristesse	0%	2,9%	2,4%	0%	93,2%	1,5%
Surprise	1,3%	0%	1,6%	2,1%	0%	95%

Tableau 5.28 Résultats de fusionnement en utilisant la règle somme et la base JAFFE

Vrai Label Predict Label	Colère	Dégout	Peur	Joie	Tristesse	Surprise
Colère	97,2%	1,3%	0%	0%	1,5%	0%
Dégout	2,9%	89,5%	4,1%	0%	3,5%	0%
Peur	1,5%	1,9%	93,6%	0%	3%	0%
Joie	0%	0%	0,5%	96%	0%	3,5%
Tristesse	0%	2,5%	2,3%	0%	94%	1,2%
Surprise	1,1%	0%	1,3%	1,2%	0%	96,4%

Tableau 5.29 Résultats de fusionnement en utilisant la règle produit et la base JAFFE

Nous observons que les taux de classification dans la base de données JAFFE sont inférieurs à ceux de la base de données CK+. Nous pouvons également noter que la méthode d'agrégation en utilisant la règle de produit donne une meilleure précision que la méthode d'agrégation de somme pour les deux bases de données.

Le tableau 5.30 illustre les résultats de fusionnement obtenus en utilisant les méthodes d'agrégation de somme et de produit pour les deux bases de données CK+ et JAFFE.

Critères Evaluation BDD / Modèle		Accuracy	Précision	Sensibilité	F1-Score
CK+	Règle somme	97,23%	97,24%	97,23%	97,23%
	Règle produit	97,93%	97,95%	97,93%	97,92%
JAFFE	Règle somme	93,50%	93,52%	93,50%	93,49%
	Règle produit	94,45%	94,48%	94,45%	94,44%

Tableau 5.30 Résultats de fusionnement en utilisant la règle somme et produit pour les deux bases de données CK+ et JAFFE

5.4.3.4 Discussion

Les performances du système proposé peuvent être influencées en ajustant deux paramètres qui sont le nombre d'époques et la fonction d'activation utilisée. Les Figure 5.9 et 5.10 montrent l'impact de la variation du nombre d'époques sur le taux de reconnaissance des quatre modèles CNN (CNN_{LBP} , CNN_{LGC} , CNN_{LDP} et CNN_{GDP}) en se basant sur la fonction d'activation utilisée (ReLU / Identity) pour les bases de données CK+ et JAFFE respectivement [CHE 22a].

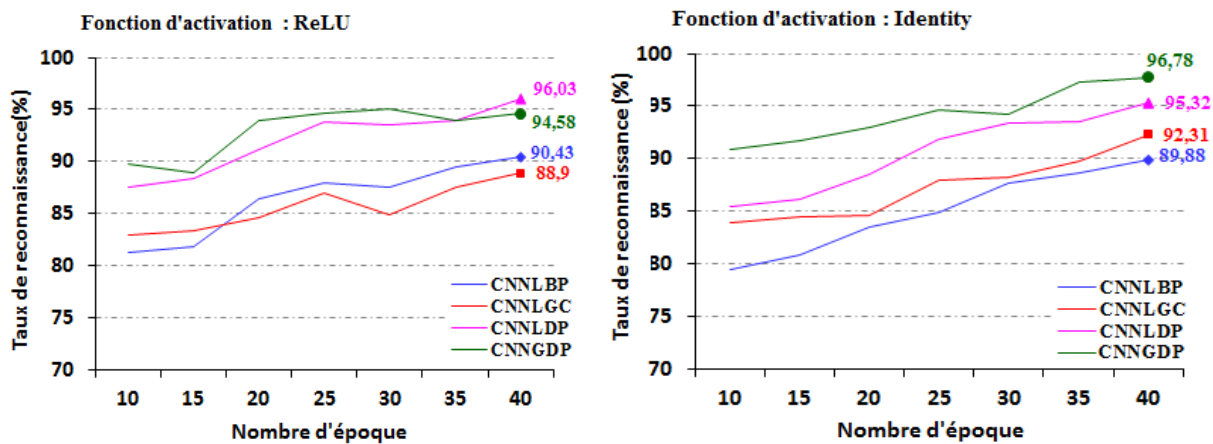


Figure 5.9 Variation du nombre d'époque versus le taux de reconnaissance pour la base CK+ en utilisant les fonctions d'activation ReLU et Identity

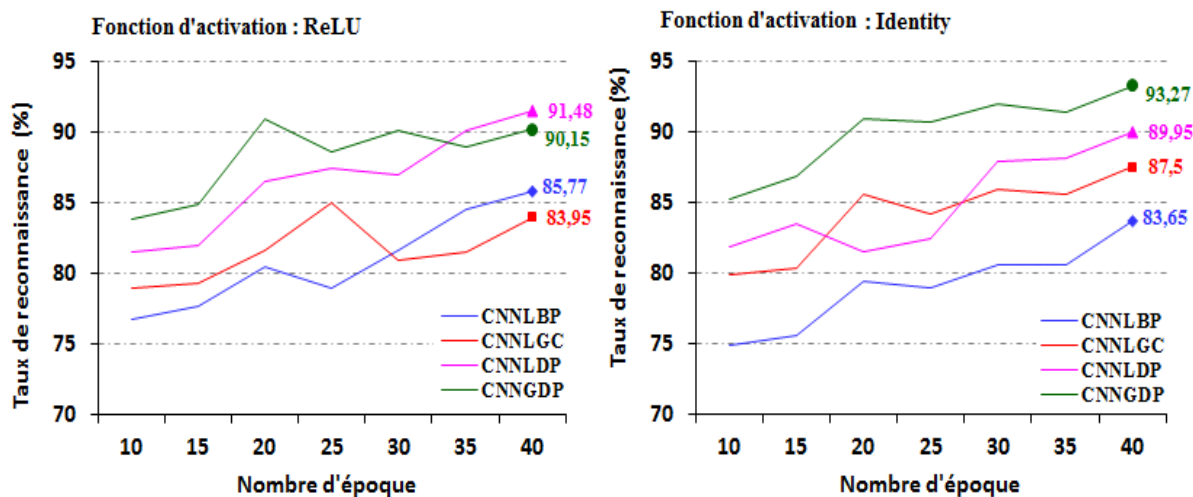


Figure 5.10 Variation du nombre d'époque versus le taux de reconnaissance pour la base JAFFE en utilisant les fonctions d'activation ReLU et Identity

A partir des figures 5.9 et 5.10, nous pouvons remarquer que l'augmentation du nombre d'époques entraîne une augmentation dans le taux de reconnaissance des quatre modèles CNN pour les deux bases de données (CK + et JAFFE) en utilisant les deux fonctions d'activation (ReLU et Identity). Nous pouvons aussi constater que les quatre classificateurs

CNN atteignent des performances optimales lorsque le nombre d'époques est égal à 40. L'augmentation du nombre d'époques au-delà de 40 produit un sur-ajustement des données qui dégradent progressivement les performances.

Il est à noter que les modèles CNN_{LBP} et CNN_{LDP} ont obtenu un taux de reconnaissance élevé en utilisant la fonction ReLU, tandis que les modèles CNN_{LGC} et CNN_{GDP} ont obtenu un taux de reconnaissance élevé en utilisant la fonction Identity pour les deux bases de données [CHE 22a].

5.4.3.5 Comparaison des résultats de l'approche de fusionnement avec les autres travaux de l'état de l'art

Le tableau 5.31 présente une étude comparative sur les performances de l'approche proposée et d'autres travaux de l'état de l'art pour les deux bases de données CK + et JAFFE.

Références	Architecture	Taux de reconnaissance	
		CK+	JAFFE
[YAN 17]	Weighted Mixture DNN (6 Class)	97,02%	92,21%
[LOP 17]	CNN + Data augmentation (6 Class)	96,76%	88,73%
[MOH 17]	CNN + Action Units (7 Class)	97,01%	/
[ALS 17]	Hybrid CNN + SIFT (7 Class)	99,1%	/
[CAI 18]	CNN + SBP (7 Class)	96,87%	95,24%
[FEN 18]	Spatial and temporal CNN + LBP Top (7 Class)	93,7%	/
[JAI 19]	DNN + Residual Block (7 Class)	93,24%	95,23%
[YAN 19]	CNN + LBP (7 Class)	95%	/
[RAV 20]	CNN + LBP (7 Class)	97,32%	77,27%
Méthode proposée	Fusion multimodal des CNN (CNN+ LBP,LGC,LDP,GDP) (6 Class)	Règle Produit	
		97,93%	94,45%

Tableau 5.31 Etude comparative entre l'approche proposée et les autres systèmes de reconnaissance des expressions faciales.

D'après le tableau 5.30, nous pouvons remarquer que l'approche proposée atteint des performances supérieures en termes de taux de reconnaissance par rapport à certains système de reconnaissance des expressions faciales pour les deux bases de données CK + et JAFFE.

5.5 Implémentation du système

Après avoir donné les différents résultats des trois approches utilisés pour modéliser notre système de reconnaissance des expressions faciales, on parlera dans cette partie de l'application réalisée tout en donnant quelques captures d'écrans de chacune des trois approches proposées.

5.5.1. L'application réalisée

Dans cette application, on suppose que les apprenants sont placés face à une webcam et connectés au système d'enseignement à distance (Système Tutoriel Intelligent) pour apprendre un cours donné ou résoudre un exercice. Une fois l'apprenant est connecté, le système de reconnaissance des expressions faciales commence à analyser les émotions ressenties par l'apprenant. A partir de l'observation et de l'identification de l'état émotionnel de l'apprenant, le tuteur affiche un message (par exemple un message d'encouragement en cas tristesse, un message de félicitation en cas de joie, etc) pour régulariser la qualité de l'apprentissage, puis les apprenants mettent à jour leur état émotionnel [CHE 22a].

5.5.1.1. Fenêtre Accueil

Cette fenêtre permet une première interaction de l'apprenant avec le système tutoriel intelligent. Elle contient cinq boutons : **Lire Cours**, **Résoudre Exercices**, **Système d'Analyse des Expressions Faciales**, **A propos** et **Quitter**. Ces boutons sont destinés à l'utilisateur du système.



Figure 5.11 Fenêtre accueil [CHE 22a].

5.5.1.2. Fenêtre de l'approche basée sur les caractéristiques géométriques pour la reconnaissance des expressions faciales

Cette fenêtre contient l'ensemble des traitements : Extraction des points caractéristiques facials, calcul du vecteur de caractéristiques et classification des expressions faciales par le classificateur KNN.

Calcul KPPV

Calcul Distance
Decision
Fermer

Oeil G : Dist1 16 Distance Oeil / Bouche D: Dist11 75

Distance Oeil / Sourcil G: Dist2 18 Dist Sourcil G Centre/Oeil G Centre: Dist12 23

Largeur Bouche: Dist3 31 Dist Sourcil D Centre/Oeil D Centre: Dist13 23

Longueur Ouverture Bouche: Dist4 33 Dist Oeil G Centre / Nez Centre: Dist14 44

Distance Oeil / Bouche G: Dist5 75 Dist Oeil D Centre / Nez Centre: Dist15 44

Largeur Ouverture Yeux G: Dist6 23 Dist Bouche Centre / Nez Centre: Dist16 36

Distance Sourcil D/G Centre: Dist7 49 Dist Oeil D coin D / Oeil G coin D: Dist17 25

Distance Oeil / Sourcil D: Dist8 18 Dist Oeil G Centre / Bouche Centre: Dist18 76

Largeur Ouverture Yeux D: Dist9 23 Dist Oeil D Centre / Bouche Centre: Dist19 77

Longueur Ouverture Yeux D: Dist10 13

Distance	NumClasse	NomClasse	NomImage	Image	LongOuvertY
22,538854598999	7	Happiness	H53.jpg	(GRAPHIC)	
19,353006362915	7	Happiness	H52.jpg	(GRAPHIC)	
19,2705211639404	7	Happiness	H51.jpg	(GRAPHIC)	
33,3357238769531	7	Happiness	H50.jpg	(GRAPHIC)	
19,6299705505371	7	Happiness	H49.jpg	(GRAPHIC)	
22,9701976776123	7	Happiness	H48.jpg	(GRAPHIC)	
13,0755577087402	7	Happiness	H47.jpg	(GRAPHIC)	
12,6915550231934	7	Happiness	H46.jpg	(GRAPHIC)	
14,8219957351685	7	Happiness	H45.jpg	(GRAPHIC)	

MIN OF Distance	NomClasse
19,5532760620117	Anger
8,81985282897949	Disgust
19,4281806945801	Fear
11,6654529571533	Happiness
10,8701581954956	Neutre
22,8707313537598	Sadness
18,4514675140381	Surprise

Enregistrer Fermer

Figure 5.12 Fenêtre approche basée descripteurs géométriques [CHE 22a].

5.5.1.3. Fenêtre de l'approche basée sur les caractéristiques d'apparences (descripteur LDDTCP) pour la reconnaissance des expressions faciales

Cette fenêtre permet d'afficher les différentes opérations effectuées pour le calcul du nouveau descripteur LDDTCP ($P_{LDDTCP-1}$).

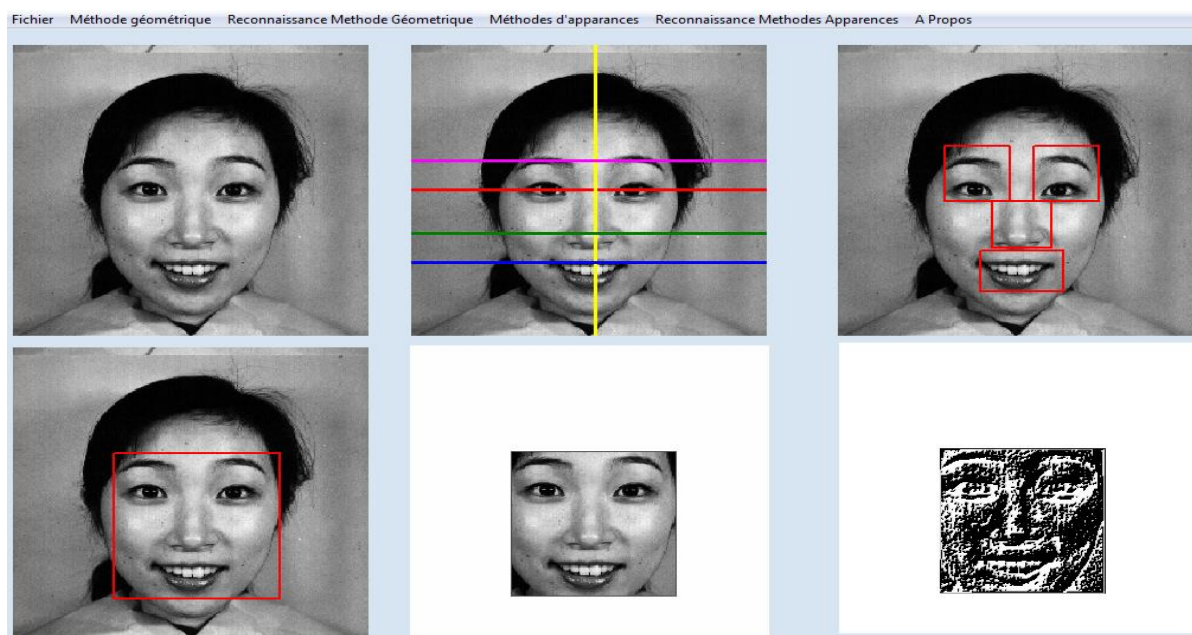


Figure 5.13 Fenêtre d'un exemple de calcul du descripteur $P_{LDDTCP-1}$ [CHE 22a].

5.5.1.4 Fenêtre de l'approche basée sur l'apprentissage profond pour la reconnaissance des expressions faciales

Cette fenêtre affiche l'architecture proposée pour les quatre modèles CNN (CNN_{LBP} , CNN_{LGC} , CNN_{LDP} , CNN_{GDP}).

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 132, 132, 16)	3904
batch_normalization (Batch Normalization)	(None, 132, 132, 16)	64
max_pooling2d (MaxPooling2D)	(None, 66, 66, 16)	0
conv2d_1 (Conv2D)	(None, 66, 66, 32)	25120
batch_normalization_1 (Batch Normalization)	(None, 66, 66, 32)	128
max_pooling2d_1 (MaxPooling2D)	(None, 33, 33, 32)	0
conv2d_2 (Conv2D)	(None, 33, 33, 64)	51264
batch_normalization_2 (Batch Normalization)	(None, 33, 33, 64)	256
max_pooling2d_2 (MaxPooling2D)	(None, 16, 16, 64)	0
conv2d_3 (Conv2D)	(None, 16, 16, 128)	204928
batch_normalization_3 (Batch Normalization)	(None, 16, 16, 128)	512
max_pooling2d_3 (MaxPooling2D)	(None, 7, 7, 128)	0
flatten (Flatten)	(None, 6272)	0
dense (Dense)	(None, 3200)	20073600
dropout (Dropout)	(None, 3200)	0
dense_1 (Dense)	(None, 1600)	5121600
dropout_1 (Dropout)	(None, 1600)	0
dense_2 (Dense)	(None, 6)	9606

Figure 5.14 Fenêtre de l'architecture des quatre modèles CNN [CHE 22a].

5.6 Conclusion

Dans ce chapitre, nous avons présenté les tests et les résultats expérimentaux des trois approches proposées, tout en menant une comparaison avec des travaux récents. Les résultats obtenus montrent que les approches proposées se classent mieux que des travaux existants. On peut aussi faire les remarques suivantes :

- L'approche basée sur l'apprentissage profond fournit des meilleurs résultats en termes de taux de reconnaissance (avec 97,93%) par rapport à l'approche basée sur les caractéristiques géométriques (avec 84,07%) et l'approche basée sur les caractéristiques d'apparences (LDDTCP) (avec 97,18%). Ceci est dû d'une part à l'utilisation des modèles CNN qui sont capable d'extraire les caractéristiques et de classer automatiquement les expressions faciales au même temps et d'autre part à la combinaison des quatre classificateurs CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} , CNN_{LDP}), ce qui permet de tirer parti des avantages et de la complémentarité de chacun d'eux pour augmenter la précision et la robustesse des systèmes de reconnaissance des expressions faciales.
- Les taux de classification dans la base de données JAFFE sont inférieurs à ceux de la base de données CK+, car dans la base de données JAFFE, certaines expressions sont

étiquetées de manière incorrecte ou exprimées de manière inexacte par les sujets [JAB 10a].

- L'utilisation de l'émotion neutre pour l'ensemble de données à 7 classes entraîne une diminution dans le taux de reconnaissance par rapport à l'ensemble de données à 6 classes, car l'émotion neutre est souvent confondue avec d'autres émotions comme la tristesse, la peur et la colère.

Conclusion générale et Perspectives

Un des plus importants objectifs dans les systèmes d'interactions homme-machine et particulièrement dans les STI est de développer la capacité d'identifier l'état émotionnel de l'apprenant. Cette capacité est un potentiel qui permet aux STI d'interagir naturellement avec l'apprenant et de communiquer avec lui en tenant compte de ses besoins. Elle permet aussi au tuteur de prévoir quand et comment il faut intervenir pour aider l'apprenant à accomplir sa tâche d'apprentissage dans des meilleures conditions.

Afin de pouvoir détecter l'état émotionnel de l'apprenant, les chercheurs ont recours à une panoplie de moyens de détection allant des simples questionnaires ou des échelles de mesure, à l'utilisation des mesures physiologiques (la fréquence cardiaque, le volume de la pression sanguine, la fréquence de la respiration, etc.), en passant par la reconnaissance des émotions en utilisant des mesures comportementales (les expressions faciales, la tonalité de la voix, la gestuelle ou la posture corporelle, etc.).

Parmi les différentes méthodes permettant de détecter les changements de l'état émotionnel de l'apprenant, les expressions faciales jouent un rôle fondamental dans la communication non verbale et constituent le moyen le plus couramment utilisés par les humains pour transmettre les états émotionnels internes.

L'analyse automatique des expressions faciales est un réel défi qui a suscité un intérêt croissant au cours des dernières décennies. Quoiqu'il en soit, l'être humain est particulièrement capable de détecter et interpréter, aisément, les expressions faciales qu'exprime autrui sur son visage. Cependant, la tâche de détection et d'extraction des différents composants qui caractérisent l'expression faciale est une activité très difficile à réaliser par les machines.

Le système de reconnaissance automatique des expressions faciale (SREF) consiste à classer des déformations faciales en classes abstraites, basées sur l'information visuelle. Il s'effectue généralement selon les trois étapes suivantes : La première étape consiste à détecter la présence d'un visage et d'en localiser sa position. Une fois détecté, l'étape suivante concerne l'extraction des caractéristiques du visage. A cet effet, différentes caractéristiques géométriques et/ou d'apparence peuvent être utilisées. L'étape finale prend comme entrée le vecteur de caractéristiques extrait précédemment pour effectuer la tâche de classification en utilisant une technique d'apprentissage.

Notre travail visait à concevoir un nouveau système pour détecter l'état émotionnel de l'apprenant lors d'une séance d'apprentissage en ligne à travers ces expressions faciales. Trois

différentes contributions ont été alors proposées pour la mise en place du système de reconnaissance des expressions faciales.

- La première contribution code les déformations faciales en se basant sur le calcul des distances entre des points caractéristiques. Ces points sont extraits à partir des régions d'intérêts (ROI) du visage tels que les yeux, le nez et la bouche.
- La deuxième contribution extrait les caractéristiques d'apparences du visage en se basant sur un nouveau descripteur appelé 'Local Directional Double Ternary Coding Patterns (LDDTCP)'. Ce descripteur encode la texture locale d'une image afin de trouver les informations les plus discriminantes qui décrivent les expressions faciales.
- La troisième contribution combine quatre modèles CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} et CNN_{LDP}) basée sur le calcul de quatre descripteurs de texture LBP, LGC, GDP et LDP pour extraire les caractéristiques faciales.

Une étude comparative entre les trois approches proposées a montré que la méthode basée sur l'apprentissage profond fournit de meilleurs taux de reconnaissance avec (97,93%) par rapport à l'approche basée sur les caractéristiques géométriques (avec 84,07%) et l'approche basée sur les caractéristiques d'apparences (LDDTCP) (avec 97,18%). Ceci est dû d'une part à l'utilisation des modèles CNN et d'autre part, à la combinaison des quatre classificateurs CNN (CNN_{LBP} , CNN_{LGC} , CNN_{GDP} , CNN_{LDP}).

Des études comparatives ont été aussi menées entre les trois approches proposées et les travaux existants dans la littérature. Les résultats expérimentaux ont montré la supériorité de l'approche basée sur le nouveau descripteur LDDTCP et l'approche basé sur le fusionnement des quatre modèles CNN par rapport aux méthodes existantes en termes de taux de reconnaissance, Néanmoins, les résultats restent insuffisantes pour l'approche basée sur les descripteurs géométriques.

Perspectives

La mise en place de notre système nous a ouvert un certain nombre de perspectives pour des travaux futurs. En voici quelques exemples :

➤ Amélioration des performances du système de détection de l'état émotionnel de l'apprenant

L'identification de l'état émotionnel de l'apprenant pendant son interaction avec un système d'apprentissage en ligne à travers des mesures comportementales comme les expressions faciales atteint des taux de reconnaissance élevés. Il serait intéressant de reconnaître les émotions des apprenants à travers des mesures physiologiques tel que la

fréquence cardiaque, le volume de la pression sanguine, la fréquence de la respiration, etc. En effet, contrairement aux indices comportementaux qui évaluent essentiellement des caractéristiques externes ou observables de l'apprenant, les indices physiologiques permettent de mesurer des réactions qui reflètent la dynamique interne du corps humain. Ces réactions sont plus spontanées, difficilement contrôlables, et indépendantes de l'apparence externe de l'apprenant.

Il serait également intéressant de combiner l'une des techniques des mesures comportementales avec l'une des techniques des mesures physiologiques ce qui permet d'intégrer les avantages de chacun d'eux afin d'améliorer la qualité du système de reconnaissance.

➤ **Amélioration des performances de l'approche géométrique pour la reconnaissance des expressions faciales**

Dans le cadre de cette thèse, nous avons supposé une vue frontale du visage. Les points faciaux qui sont détectés à partir d'une vue frontale décrivent la forme des caractéristiques (bouche, sourcil et yeux) et du contour du visage. La méthode de détection des points faciaux de Harris permet d'atteindre des taux de reconnaissance considérables. Cependant, dans le cas réel, cette méthode présente des limites. En effet, les sujets ne sont pas toujours face à la caméra et certaines expressions sont parfois accompagnées d'un mouvement de tête comme le rire ou la peur, ce qui perturbe le suivi des points faciaux et déforme complètement leurs dispositions. Afin d'améliorer la précision de la localisation des points faciaux, nous proposons d'utiliser la méthode de Shi-Tomasi [SHI 94] qui permet d'avoir des points sillons qui assurent la stabilité de l'étape de suivi basée sur le flux optique. Chaque muscle facial est représenté par une distance entre deux points. La distinction entre les différentes expressions faciales est basée sur la variation de ces distances par rapport à l'état neutre.

➤ **Amélioration des performances de l'approche basée sur l'apprentissage profond pour la reconnaissance des expressions faciales**

Dans notre application, nous avons utilisé les réseaux de neurones convolutionnels (CNN) dans l'approche basée sur l'apprentissage profond pour tester et évaluer le système de reconnaissance des expressions faciales. Nous envisageons d'utiliser d'autres méthodes de l'apprentissage profond comme les réseaux de neurones récurrents (RNN) ou les modèles génératifs profonds tels que les machines de Boltzmann (BM), les machines de Boltzmann Restreintes (RBM), les réseaux de croyance profonds (DBN), les réseaux antagonistes génératifs (GAN), les auto-encodeurs (AE), etc, afin d'améliorer les performances du système de reconnaissance.

➤ **Vers une reconnaissance des expressions faciales dans des conditions réelles**

Il serait intéressant de tester les approches proposées dans des conditions d'éclairage dégradées, dans un contexte de vidéo surveillance à faible résolution et sur des bases d'images à grande échelle impliquant des millions d'identités, afin de consolider les résultats obtenus.

Références bibliographiques

- [ABD 07] F. Abdat, C. Maaoui, and A. Pruski, “Suivi du gradient pour la localisation des caractéristiques faciales dans des images statiques”, in *Colloque GRETSI*, Troyes, France, 2007.
- [ABD 11] F. Abdat, C. Maaoui, and A. Pruski “Human-computer interaction using emotion recognition from facial expression”, In *Fifth UKSim European Symposium on Computer Modeling and Simulation (EMS)*, Madrid, Spain, pp. 196–201, 2011.
- [AHM 11] F. Ahmed E. Hossain, A.S.M. Hossain Bari, and ASM Shihavuddin, “Compound Local Binary Pattern (CLBP) for Robust Facial Expression Recognition”, in *12th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, Budapest, Hungary, pp. 391-395, 2011.
- [AHM 12] F. Ahmed, “Gradient directional pattern: a robust feature descriptor for facial expression recognition”, *Electronics letters*, 48(19), 2012.
- [AHM 13] F. Ahmed and E. Hossain, “Automated facial expression recognition using gradient-based ternary texture patterns”, *Chinese Journal of Engineering*, pp. 1–8, 2013.
- [AHM 18] F. Ahmed, and H. Kabir, “Facial expression recognition under difficult conditions: A comprehensive study on edge directional texture patterns”, *Int. J. Appl. Math. Comput. Sci*, 28(2), pp. 399–409, 2018.
- [ALS 17] M. Al-Shabi, W.P. Cheah and T. Connie, “Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator”, *Computer Vision and Pattern Recognition*, 2017, DOI: 10.1007/978-3-319-69456-6_12.
- [AND 83] J. R. Anderson, “A spreading activation theory of memory”, *Journal of verbal-learning and verbal behavior*, 22(3), pp. 261-295, 1983.
- [APD 17] D. Apdilah, M.Y. Simargolang, and R. Rahim, “A Study of Frei-Chen Approach for Edge Detection”, *IJSRSET*, 3(1), pp. 59-62, 2017.
- [ARN 11] L. Arnold, S. Rebecchi, S. Chevallier and H. Paugam-Moisy, “An introduction to deep learning”, in *Advances in Computational Intelligence and Machine Learning, ESANN'2011*, 2011.
- [ARR 07] I. Arroyo, K. Ferguson, J. Johns, T. Dragon, H. Meheranian, D. Fisher, A. Barto, S. Mahadevan, B.P. Woolf, “Repairing Disengagement With Non-Invasive Interventions”, *International Conference on Artificial Intelligence in Education*, Los Angeles, California, USA, IOS Press, pp. 195-202, 2007
- [ARR 09] I. Arroyo, B. P. Woolf, J. M. Royer and M. Tai, “Affective Gendered Learning Companions”, *International Conference on Artificial Intelligence in Education*, Brighton, England, IOS Press, 2009.
- [BAK 04] R. S. Baker, A. T. Corbett, K. R. Koedinger and A. Z. Wagner, “Off-task behavior in the cognitive tutor classroom: when students ‘game the system’ ”, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*, Vienna, Austria, pp. 383-390, 2004.
- [BAK 06] R. S. J. d. Baker, A. T. Corbett, K. R. Koedinger and I. Roll, “Generalizing detection of gaming the system across a tutoring curriculum”, *8th international conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, pp. 402-411, 2006.
- [BAL 11] P. Baldi, “Autoencoders, Unsupervised Learning, and Deep Architectures”, in *Unsupervised and Transfer Learning - Workshop held at ICML*, Bellevue, 2011.
- [BAS 08] S. Bashyal and Ganesh K. Venayagamoorthy, “Recognition of facial expressions using Gabor wavelets and learning vector quantization”, *Engineering Applications of Artificial Intelligence*, 21(7), pp.1056–1064, 2008.

- [BEL 12] K. Belghith, R. Nkambou, F. Kabanza and L. Hartman, “An intelligent simulator for telerobotics training”, *IEEE Transactions on Learning Technologies*, 5(1), pp. 11-19, 2012.
- [BEL 97] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, “Eigenfaces vs. fisherfaces : Recognition using class specific linear projection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), pp. 711–720, 1997.
- [BEN 07] C.E. Bencheriet, A. Boualleg and H. Tebbikh, “Segmentation de la Couleur de Peau par Seuillage selon Différents Espaces de Couleur”, in *3èmes Journées Internationales sur d’Informatique Graphique : JIG’2007*, INRIA Sophia-Antipolis, France, pp. 207-211, 2007.
- [BEN 14] K.S. Benli and M.T. Eskil, “Extraction and Selection of Muscle Based Features for Facial Expression Recognition”, *22nd International Conference on Pattern Recognition*, Stockholm, Sweden, pp. 1651–1656, 2014.
- [BEN 94] Y. Bengio, P. Simard and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult”, *IEEE Transactions on Neural Networks*, 5 (2), pp.157–166, 1994.
- [BLO 75] B. Bloom, “Taxonomie des objectifs pédagogiques”, *Domaine cognitif, Presses de l’Université du Québec*, 1975.
- [BLO 84] B. S. Bloom, “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring”, *Educational Researcher*, 13, pp. 3-16, 1984.
- [BOS 03] T. L. Bosch, “Emotions, speech and the ASR framework”, *Speech Communication*, 40(1), pp. 213-225, 2003, DOI: 10.1016/s0167-6393(02)00083-3.
- [BOT 02] A. Botino, “Real time head and facial features tracking from uncalibrated monocular views”, *5th Proceeding Asian Conference on Computer Vision ACCV*, Australia, pp. 23–25, 2002.
- [BRO 78] J. S. Brown and R. R. Burton, “Diagnostic models for procedural bugs in basic mathematical skills”, *Cognitive Science*, 2(2), pp. 155-191, 1978.
- [BUR 88] H. L. Burns and C. G. Capps, “Foundations of ITS: an Introduction”, In P. M. C. a. R. J. J. (Ed.), *Foundations of ITS*, Hillsdale, pp. 1-20, 1988.
- [BUS 04] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information”, In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI ’04)*, New York, NY, USA, pp. 205–211, 2004.
- [BUY 16] P. Buyssens et A. Elmoataz, “Réseaux de neurones convolutionnels multi-échelle pour la classification cellulaire”, *Reconnaissance des formes et intelligence artificielle*, Clermont-Ferrand, France, 2016.
- [CAI 18] J. Cai, Q. Chang., X. Lun-Tang, C. Xue and C. Wei, “Facial expression recognition method based on SparseBatch normalization CNN”, *37th Chinese control conference (CCC)*, Wuhan, China, pp. 9608-9613, 2018, DOI: 10.23919/ChiCC.2018.8483567.
- [CAI 21] J.Cai, Z. Meng, A.S. Khan, J. O’Reilly, Z. Li, S. Han and Y. Tong, “Identity free facial expression recognition using conditional generative adversarial network”, *Computer Vision and Pattern Recognition*, *arXiv:1903.08051*, 2021, DOI: :10.1109/ICIP42928.2021.9506593.
- [CAR 79] J. G. Carbonell, “Subjective understanding, computer models of belief system”, *PhD thesis*, Yale University New Haven, CT, USA, 1979.
- [CHA 04] S. Chaffar and C. Frasson, “Architecture pour des systèmes tuteurs émotionnellement intelligents”, *Technologies de l’Information et de la Connaissance dans l’Enseignement Supérieur et l’Industrie*, Compiègne, France. pp.223-228, 2004.
- [CHA 06] Y. Chang, C. Hu, R. Feris and M. Turk, “Manifold based analysis of facial expression”, *Image and Vision Computing*, 24(6), pp. 605–614, 2006.

- [CHA 09] M. Chaouachi, A. Heraz, I. Jraidi and C. Frasson, "Influence of Dominant Electrical Brainwaves on Learning Performance", *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Vancouver, Canada, pp. 2448- 2454, 2009.
- [CHA 10] S. Chaffar and C. Frasson, "Using Emotional Coping Strategies in Intelligent Tutoring Systems", *International Conference on Intelligent Tutoring Systems (ITS)*, 6095, pp. 285-287, 2010.
- [CHA 90] T.W. Chan and B. A. Baskin, "Learning Companion Systems". In C. Frasson & G. Gauthier (Eds.), *Intelligent Tutoring Systems: At the Crossroads of Artificial Intelligence and Education*, Norwood, New Jersey, pp. 6-33, 1990.
- [CHE 08] Y. Chen, Z. Fu, Y. Han, and R. Wang, "Independent component analysis of Gabor features for texture classification", *Optical Engineering*, 47(12), 2008.
- [CHE 19a] O. Chebah and M.T. Laskri, "Facial expression recognition for e-learning systems", In *4th International Conference on Networking and Advanced Systems (ICNAS)*, Annaba, Algeria, 2019
- [CHE 19b] O. Chebah and M.T. Laskri, "Facial expression recognition system", In *2nd Conference on Informatics and Applied Mathematics (IAM)*, Guelma, Algeria, 2019.
- [CHE 22a] O. Chebah and M.T. Laskri, "Facial expression recognition Facial expression recognition using convolution neural network fusion and texture descriptors representation", *International Journal of Computational Intelligence and Applications*, 21(1), 2250002 , 2022, DOI :10.1142/S146902682250002X.
- [CHE 22b] O. Chebah and M.T. Laskri, "Local directional double ternary coding pattern for facial expression recognition", *International Journal of Computational Vision and Robotics*, 2022.
- [CHI 99] C.J Lee and S.D. Wang, "Fingerprint feature extraction using Gabor filters", *Electronics Letters*, 35(4), pp. 288–290, 1999.
- [CHO 14] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1724–1734, 2014.
- [CLA 83] W.J. Clancey, "GUIDON", *Journal of Computer-Based Instruction*, 10(1), pp. 8-14, 1983.
- [CLA 92] W. J. Clancey, "Madel construction operators", *Artificial Intelligence*, 53(1), pp. 1-115, 1992.
- [CLO 00] G. Clore and K. Gasper "Feeling is believing: Some affective influences on belief", In N. H. Frijda, A. S. Manstead & S. Bem (Eds.), *Emotions and beliefs: how feelings influence thoughts*, Cambridge: Cambridge University Press, pp. 10-44, 2000.
- [COH 03] I. Cohen, N. Sebe, L. Chen, A. Garg et T.S. Huang, "Facial expression recognition from video sequences : temporal and static modelling", *Computer Vision and Image Understanding : Special issue on face recognition*, 91(1), pp. 160–187, 2003.
- [CON 04] C. Conati and H. Maclaren, "Evaluating A Probabilistic Model of Student Affect", *7th International Conference on Intelligent Tutoring Systems*, Maceio, Brazil, 3220, pp. 55-66 2004.
- [COO 01] T.F. Cootes, G.J. Edwards et C.J. Taylor, "Active appearance models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), pp. 681–685, 2001.
- [COO 92] T. Cootes et C. Taylor, "Active Shape Models Smart Snakes", In *British Machine Vision Conference*, Leeds, pp. 267–275, 1992.

- [COP 16] M. Copeland, “What’s the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?”, *NVIDIA*, 2016, <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>, 2016.
- [DAF 05] DaFEx: a database of kinetic facial expression, <https://i3.fbk.eu/resources/dafex-database-kinetic-facial-expressions>, 2005.
- [DAH 20] K. Dahmane, “Analyse d’images par méthode de Deep Learning appliquée au contexte routier en conditions météorologiques dégradées », Thèse de doctorat, Vision par ordinateur et reconnaissance de formes, Université Clermont Auvergne, France, 2020.
- [DAM 94] A. Damasio, “L’erreur de Descartes, ” *La raison des émotions*, Odile Jacob, 1994.
- [DAM 95] A. R. Damasio, “L’erreur de Descartes : La raison des émotions”, *Paris: Odile Jacob*, 1995.
- [DAR 72] C. Darwin, “The expression of the emotions in man and animals”, *NewYork Philosophical Library*, New York, 1872.
- [DAU 93] J. G. Daugman, “High confidence visual recognition of persons by a test of statistical independence”, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(11), pp. 1148-1161, 1993.
- [DAV 04] F. Davoine, B. Abboud et V.M. Dang, “Analyse de visages et d’expressions faciales par modèle actif d’apparence”, *Traitement du signal*, 21(3), pp.179-193, 2004.
- [DEE 20] “Deep Learning ou Apprentissage Profond : qu’est-ce que c’est ?”, *Blog datascientest*, <https://datascientest.com/deep-learning-definition>, 2020.
- [DEN 05] H.B. Deng, L.W. Jin, L.X. Zhen, J.C. Huang, “A new facial expression recognition method based on local gabor filter bank and PCA plus LDA”, *International Journal of Information Technology*, 11(11), pp. 86–96, 2005.
- [DEN 14] L. Deng, “A tutorial survey of architectures, algorithms, and applications for deep learning”, *In: APSIPA Transactions on Signal and Information Processing*, 3, 2014.
- [DEV 02] A. de Vicente and H. Pain, “Informing the Detection of the Students' Motivational State: An Empirical Study”, *6th International Conference on Intelligent Tutoring Systems*, 2363, pp. 933-943, 2002.
- [DME 05] S. K. D’Mello, S. D. Craig, B. Gholson, S. Franklin, R. W. Picard et A. C. Graesser, “Integrating Affect Sensors in an Intelligent Tutoring System”, *In Affective Interactions: The Computer in the Affective Loop, Proceedings of the 10th international conference on Intelligent User Interfaces (IUI)*, New York, pp. 7-13, 2005.
- [DME 06] S. K. D’Mello, S. D. Craig, J. Sullins, and A. C. Graesser, “Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor’s Mixed-Initiative Dialogue”, *International Journal of Artificial Intelligence in Education (IJAIED)*, 16(1), pp. 3-28, 2006.
- [DME 07] S. D’Mello, R. W. Picard and A. Graesser, “Toward an Affect-Sensitive AutoTutor”, *IEEE Intelligent Systems*, 22(4), pp. 53-61, 2007 DOI: 10.1109/MIS.2007.79
- [DUB 02] S. Dubuisson, F. Davoine et M. Masson, “A solution for facial expression representation and recognition”. *Signal Processing and Image Communication*, 17(9), pp. 657–673, 2002.
- [DUC 11] J.C. Duchi, E. Hazan et Y. Singer, “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization”, *Journal of Machine Learning Research*, 12, pp. 2121–2159, 2011.
- [DUT 96] D. Dutton, G. Conroy, “A review of machine learning”, *The Knowledge Engineering Review*, 12 (4): 341–367, 1996.
- [DUV 11] C. Duval, “Neuropsychologie des représentations mentales de soi et des autres : Effets de l’âge et exploration dans la démence sémantique”, *Thèse de doctorat*, Université de Caen, Caen, France, 2011.

- [EDD 96] S. R. Eddy, “Hidden markov models”, *Current opinion in structural biology*, 6(3), pp. 361–365, 1996.
- [EKM 93] P. Ekman and R. J. Davidson, “Voluntary smiling changes regional brain activity”, *Psychological Science*, 4(5), pp. 342–345, 1993.
- [EKM 03] P. Ekman et W. V. Friesen, “Unmasking the Face: a guide to recognizing emotions from facial clues”, *livre*, 2003.
- [EKM 69] P. Ekman, E.R. Sorenson and W.V. Friesen, “Pan-Cultural Elements in Facial Displays of Emotion”, *Science*, 164(3875), pp. 86–88, 1969.
- [EKM 78] P. Ekman and W. V. Friesen, “Facial action coding system”, *Consulting Psychologist Press*, 18(11), pp. 881–905, 1978.
- [EKM 82] P. Ekman, “Emotion in the Human Face, 2nd edition,” *New York: Cambridge University Press*, 1982.
- [ELA 16] S. Elaiwat, M. Bennamoun, F. Boussaid, “A spatio-temporal RBM-based model for facial expression recognition”, *Pattern Recognition*, 49: 152–161, 2016, DOI: 10.1016/j.patcog.2015.07.006.
- [ELF 18] C. Elfiorenza, S.Sarkar, S. Chattopadhyay, D. Subadra, “ Face Recognition and Emotion Detection Using Restricted Boltzmann”, *International Journal of Scientific & Engineering Research*, 9(10): 740-744, 2018.
- [ELK 17] S. El Kaddouhi, A. Saaidi, and M. Abarkan, “Eye detection based on the Viola-Jones method and corners points”, *Multimedia Tools and Application*, 2017, DOI 10.1007/s11042-017-4415-5.
- [ELK 18] I. El khadiri, A. Chahia, Y. El merabeta, Y. Ruichekb and R. Touahnia, “Local directional ternary pattern: A New texture descriptor for texture classification”, *Computer Vision and Image Understanding*, 2018, DOI: 10.1016/j.cviu.2018.01.004.
- [ELL 99] C. Elliott, J. Rickel, and J. Lester, “Lifelike pedagogical agents and affective computing: an exploratory synthesis”, *Artificial intelligence today*, 1600, pp. 195-212, 1999.
- [ESS 97] I. Essa et A. Pentland, “ Coding, Analysis, Interpretation, and Recognition of Facial Expressions”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), pp. 757–763, 1997.
- [FAB 05] FABO: A bimodal face end body gesture database, <http://mmv.eecs.qmul.ac.uk/fabo/>, 2005.
- [FAI 02] J. Faivre, C.Frasson, R. Nkambou, “Gestion Emotionnelle dans les Systèmes Tuteurs Intelligents,” *Colloque des Technologies de l’Information et de la Communication dans l’Enseignement d’ingénieurs et dans l’industrie (TICE)*, Villeurbanne, France. pp.73-80, 2002.
- [FEN 18] D. Feng, and F. Ren, “Dynamic Facial Expression Recognition based on Two-Stream-CNN with LBP-Top”, *5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, Nanjing, China, pp. 355-359, 2018.
- [FER 19] M. Ferecatu, “Intelligence Artificielle pour des données multimédia”, *Conservatoire National des Arts et Métiers, Laboratoire CEDRIC, Equipe Vertigo*, 2019.
- [FIS 36] R.A. Fisher, “The use of multiple measurements in taxonomic problems”, *Annals of eugenics*, 7(2), pp.179–188, 1936.
- [FOU 19] D. Fourure, “Réseaux de neurones convolutifs pour la segmentation sémantique et l’apprentissage d’invariants de couleur,” *Thèse de doctorat*, Vision par ordinateur et reconnaissance de formes, Université de Lyon, France, 2017.
- [FRI 86] N. Frijda, “*The Emotions*,” Cambridge University Press, 1986.
- [GAR 83] H. GARTNER, “Frames of Mind: The Theory of Multiple Intelligences”, *Basic Books*, New York, 1983.

- [GAS 06] F. Gasparini and R. Schettini, “Skin segmentation using multiple thresholding in Internet Imaging”, in *Proceedings of SPIE*, San Jose, California, USA, pp. 1–8, 2006.
- [GEN 21] “Générative Model ou Modèle Génératif : Tout comprendre », Blog datascientest, <https://datascientest.com/generative-model-ou-modele-generatif>, 2021.
- [GER 00] F. Gers, J. Schmidhuber and F.A. Cummins, “Learning to Forget: Continual Prediction with LSTM”, *Neural Computation*, 12 (10), pp. 2451–2471, 2000, DOI: 10.1162/089976600300015015.
- [GIG 14] P. Giguère, “GAN : Generative Adversarial Networks”, GLO- 4030/7030 Apprentissage par réseau de neurones profonds, Université de Laval, 2014.
- [GIL 09] S. Gil, “Comment étudier les émotions en laboratoire”, *Revue électronique de psychologie sociale*, 4, pp. 15-24, 2009.
- [GOL 95] D. GOLEMAN, “Emotional Intelligence”, *Bantam Books*, New York, 1995.
- [GOO 14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative adversarial nets”, *Advances in Neural Information Processing Systems 27: NIPS*, 2014.
- [GOO 16] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning”, *MIT Press*, 2016.
- [GHE 17] M. Gheisari, G. Wang, and M. Z. A. Bhuiyan, “A survey on deep learning in big data”, In *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 2, pp. 173–180, 2017.
- [GRA 03] A. C. Graesser and B. A. Olde, “How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down”, *Journal of Educational Psychology*, 95(3), pp. 524-536, 2003.
- [GRA 04] M. Gratch and S. Marsella, “A Domain-independent framework for modeling emotions”, *Journal of Cognitive Systems Research*, 5(4), pp. 269-306, 2004.
- [GRA 08] A. Graves, J. Schmidhuber, C. Mayer, M. Wimmer and B. Radig “Facial Expression Recognition with Recurrent Neural Networks”, *Proceedings of the International Workshop on Cognition for Technical Systems*, Santorini, Greece, pp. 1–6, 2008.
- [GRE 88] P. Greenspan, “Emotions & reasons: an inquiry into emotional justification,” *Routledge*, 1988.
- [GRU 93] T. R. Gruber, “A translation approach to portable ontologies”, *Knowledge Acquisition*, 5(2), pp. 199-220, 1993.
- [GUN 10] H. Gunes, and M. Pantic, “Automatic, dimensional and continuous emotion recognition”, *International Journal of Synthetic Emotions (IJSE)*, 1(1), pp. 68-99, 2010.
- [GUO 19] J.M Guo, P.C. Huang and L.Y. Chang, “Hybrid Facial Expression Recognition System Based on Recurrent Neural Network”, *16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Taipei, Taiwan, 2019.
- [HAI 21] P. Hairy, “Les réseaux de neurones récurrents pour les séries temporelles”, *Metal Blog : Le blog des experts de la métallurgie*, <https://metablog.ctif.com/2021/09/06/les-reseaux-de-neurones-recurrents-pour-les-series-temporelles/>, 2021.
- [HAM 05] Z. Hammal, N. Eveno, A. Caplier, and P. Coulon, “Parametric models for facial features segmentation”, *Signal Processing*, 86(2), pp. 399–413, 2005.
- [HAM 06] Z. Hammal, “Segmentation des Traits du Visage, Analyse et Reconnaissance des Expressions Faciales par les Modèles de Croyance Transférable”, *Traitement du signal et de l'image*, Université Joseph-Fourier - Grenoble I, 2006.
- [HAM 12] H. Hamdi, “Plate-forme multimodale pour la reconnaissance d’émotions via l’analyse de signaux physiologiques : Application à la simulation d’entretiens d’embauche”, *Thèse de doctorat, Modélisation et simulation*, Université d’Angers, France, 2012.

- [HAO 17] X.L. Hao and M. Tian, “Deep Belief Network Based on Double Weber Local Descriptor in Micro-expression Recognition”, *Advanced Multimedia and Ubiquitous Engineering*, 419-425, 2017, DOI: 10.1007/978-981-10-5041-1_68.
- [Harris 88] C. Harris and M. Stephens, “A combined corner and edge detector”, *In Proceeding of Fourth Alvey Vision Conference*, pp. 147-151, 1988.
- [HER 08] N. Hernandez, “La perception des visages et des expressions émotionnelles dans la pathologie autistique: Approche comportementale et fonctionnelle”, *Thèse de doctorat*, Université François-Rabelais, Tours, France, 2008.
- [HER 09] A. Heraz, I. Jraidi, M. Chaouachi and C. Frasson, “Predicting Stress Level Variation from Learner Characteristics and Brainwaves”, *14th International Conference on Artificial Intelligence in Education (AIED)*, Brighton, UK, pp. 722-724, 2009.
- [HIN 02] G.E. Hinton, “Training products of experts by minimizing contrastive divergence” *Neural computation*, 14(8):1771–1800, 2002.
- [HIN 06] G.E. Hinton, S. Osindero and Y.W. The, “A fast learning algorithm for deep belief nets”, *neural computation*, 18(7):1527–1554, 2006.
- [HIN 12] G. E. Hinton, “A Practical Guide to Training Restricted Boltzmann Machines”, *in Neural Networks: Tricks of the Trade: Second Edition*, Toronto, Springer Berlin Heidelberg, pp. 599-619, 2012.
- [HOC 97] S. Hochreiter, J. Schmidhuber J, “Long Short-Term Memory”, *Neural Computation*, 9 (8), pp.1735–1780, 1997.
- [HUA 15] B. Huang, Z. Ying, “Sparse Autoencoder for Facial Expression Recognition”, *UIC-ATC-ScalCom-CBDCom-IoP*, Beijing, China, 2015, DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.274
- [HUA 97] C.L. Huang et Y.M. Huang, “Facial Expression Recognition Using Model-Based Feature Extraction and Action Parameters Classification”, *Journal of Visual Communication and Image Representation*, 8 (3), pp. 278-290, 1997.
- [HUA 98] T. S. Huang, L. S. Chen, H. Tao, T. Miyasato et R. Nakatsu, “Bimodal emotion recognition by man and machine”, *In ATR Workshop on Virtual Communication Environments*, Kyoto, Japan, 1998.
- [HUB 62] D.H. Hubel et T.N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”, *The Journal of physiology*, 160(1), pp. 106-154, 1962, DOI: 10.1113/jphysiol.1962.sp006837.
- [ISE 00] A. M. Isen, “Positive Affect and Decision Making”, *Handbook of emotions (M. Lewis & J. Haviland-Jones edition)*, New York: Guilford, 2000.
- [IZA 71] C.E. Izard, “*The face of emotion*,” New York: Appleton-Century-Crofts, 1971.
- [IZA 93] C. E. Izard, “The Differential Emotions Scale: DES IV-A;[a Method of Measuring the Meaning of Subjective Experience of Discrete Emotions]: University of Delaware, 1993.
- [JAB 10a] T. Jabid, M. H. Kabir, and O. Chae, “Robust facial expression recognition based on local directional pattern”, *ETRI Journal*, 32(5), pp. 784–794, 2010a.
- [JAB 10b] T. Jabid, H. Kabir, and O. Chaei, “Local Directional Pattern (LDP) for face recognition”, *in Proceedings of the International Conference on Consumer Electronics (ICCE '10)*, Las Vegas, Nev, USA, pp. 329–330, 2010b.
- [JAI 18] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, M. Zareapoor, “Hybrid Deep Neural Networks for Face Emotion Recognition”, *Pattern Recognition Letters*, 115, pp. 101-106, 2018, DOI: 10.1016/j.patrec.2018.04.010
- [JAI 19] D. K. Jain, P. Shamsolmoali, and P. Sehdev, “Extended deep neural network for facial emotion recognition”, *Pattern Recognition Letters*, 120, pp. 69 -74, 2019, DOI: 10.1016/j.patrec.2019.01.008.
- [JAM 84] W. James, “What is an emotion?,” 1884.

- [JOA 99] T. Joachims, "Making large-Scale SVM Learning Practical, *Advances in Kernel Methods - Support Vector Learning*, B. Schalkopf and C. Burges and A. Smola (ed.)", MIT-Press, 1999.
- [JRA 09] I. Jraidi, A. Heraz, M. Chaouachi and C. Frasson, "New Architecture of a Multi Agent System which Measures the Learner Brainwaves to Predict his Stress Level Variation", *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Vancouver, Canada, pp. 2726-2733, 2009.
- [JRI 13] I. Jraidi, "Modélisation des émotions de l'apprenant et interventions implicites pour les Systèmes Tutoriels Intelligents", *Thèse de doctorat*, Université de Montréal, Canada, 2013.
- [KAI 06a] R. Kaiser, "Prototypical development of an affective component for an e-learning system", *Master Thesis*, Department of Computer Science, University of Rostock, Germany, 2006.
- [KAI 06b] R. Kaiser and K. Oertel, "Emotions in HCI: an affective e-learning system", *HCSNet workshop on Use of vision in Human-Computer Interaction*, Canberra, Australia, pp. 105-106, 2006.
- [KAN 19] K. Kang and X. Ma, "Convolutional Gate Recurrent Unit for Video Facial Expression Recognition in theWild", *Proceedings of the 38th Chinese Control Conference*, Guangzhou, China, pp. 7623- 7628, 2019.
- [KAP 19] A. Kaplan, M. Haenlein, "Siri, Siri, in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence", *Business Horizons*, 62(1): 15 -25, 2019.
- [KAP 03] A. Kapoor, Y. Qi, and R.W. Picard, "Fully automatic upper facial action recognition", In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, Nice, France, pp. 195–202, 2003.
- [KAP 07] A. Kapoor, W. Burleson and R. W. Picard, "Automatic prediction of frustration", *International Journal of Human-Computer Studies*, 65(8), pp. 724-736, 2007, DOI: 10.1016/j.ijhcs.2007.02.003.
- [KIM 19] D. H. Kim, W. J. Baddar, J. Jang, et Y. M. Ro, "Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition", *IEEE Transaction on Affective Computing.*, 10(2), pp. 223-236, 2019, DOI: 10.1109/TAFFC.2017.2695999.
- [KIN 15] D.P. Kingma et J. Ba, "Adam: A Method for Stochastic Optimization", *International Conference on Learning Representations (ICLR)*, San Diego, 2015.
- [KIN 14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", In *ICLR*, 2014.
- [KON 19] F. Kong, "Facial expression recognition method based on deep convolutional neural network combined with improved LBP features", *Personal and Ubiquitous Computing*, 23, pp. 531 -539, 2019.
- [KOR 01a] B. Kort, R. Reilly and R. Picard, "An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion", *IEEE International Conference on Advanced Learning Technology*, Madison, Wisconsin, pp. 43-48, 2001a.
- [KOR 01b] B. Kort, R. Reilly and R. W. Picard, "External representation of learning process and domain knowledge: affective state as a determinate of its structure and function", *Proceedings of Artificial Intelligence in Education Workshops (AI-ED)*, San Antonio, Texas, pp. 64–69, 2001b.
- [KOS 07] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines", *IEEE Transactions on Image Processing*, 16(1), pp.172–187, 2007.

- [KOT 97] C. Kotropoulos et I. Pitas, “Rule-based Face Detection in Frontal View”, *Proceeding of International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, pp. 2537-2540, 1997.
- [KRI 12] A. Krizhevsky, I. Sutskever, and G.E Hinton, “Imagenet classification with deep convolutional neural networks”, *In Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [KUA 15] L. Kuang, L. Yang, and Y. Liao, “An integration framework on cloud for cyber physical social systems big data”, *IEEE Transactions on Cloud Computing*, 2015.
- [KUM 14] V. Kumar, and A. S. A. Basha, “Facial expression recognition using wavelet and k-nearest neighbor”, *In IEEE 2nd International Conference on Current Trends in Engineering and Technology (ICCTET)*, Coimbatore, India ,pp. 48-52, 2014.
- [KUR 19] A.R. Kurupa, M. Ajitha, M. M. Ramon, “Semi-supervised facial expression recognition using reduced spatial features and Deep Belief Networks”, *Neurocomputing*, 367: 188-197, 2019
- [LAI 18] Y.H. Lai and S.H. Lai, “Emotion-Preserving Representation Learning via Generative Adversarial Network for Multi-view Facial Expression Recognition”, *13th IEEE International Conference on Automatic Face & Gesture Recognition*, 263- 270, 2018, DOI: 10.1109/FG.2018.00046.
- [LAK 21] D. Lakshmi and R. Ponnusamy, “Facial emotion recognition using modified HOG and LBP features with deep stacked autoencoders”, *Microprocessors and Microsystems*, 82, 103834, 2021, DOI:10.1016/j.micpro.2021.103834.
- [LAL 05] F. Laliberté, “Introduction aux expressions faciales prélude au projet MONNET”, *Présentation*, CRIM, 2005.
- [LAN 93] P. J. Lang, M. K. Greenwald, M. M. Bradley and A. O. Hamm, “Looking at pictures: Affective, facial, visceral, and behavioral reactions”, *Psychophysiology*, 30(3), pp. 261-273, 1993, DOI: 10.1111/j.1469-8986.1993.tb03352.x
- [LAN 95a] A. Lanitis, C.J. Taylor et T.F. Cootes, “An Automatic Face Identification System Using Flexible Appearance Models”, *Image and Vision Computing*, 13(5), pp. 393–401, 1995.
- [LAN 95b] P. J. Lang, “The emotion probe: Studies of motivation and attention”, *American Psychologist*, 50(5), pp. 372-385, 1995, DOI: 10.1037/0003-066x.50.5.372
- [LAS 02] M.T. Laski and D. Chefrour, “Who_Is :système d’identification des visages humains”, *ARIMA*, 1, pp. 39-61, 2001.
- [LEC 98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, 86(11), pp. 2278–2324, 1998.
- [LEE 19] M. K. Lee, D. Y. Choi and B. C.Song, “Facial Expression Recognition via Relation-based Conditional Generative Adversarial Network”, *International Conference on Multimodal Interaction (ICMI '19)*, Suzhou, China, 35–39, 2019, DOI: 10.1145/3340555.3353753.
- [LES 01] J. C. Lester, S. G. Towns, C. B. Callaway, J. L. Voerman and P. J. FitzGerald, “Deictic and emotive communication in animated pedagogical agents”, *Embodied conversational agent: MIT Press*, pp. 123-154, 2001.
- [LES 99] J.C. Lester, S.G. Towns and S.G.Fitzgerald, “Achieving affective impact:Visual emotive communication in lifelike pedagogical agents,” *International Journal of Artificial Intelligence in Education (IJAIED)*, 10(3), pp. 278-291,1999.
- [LEU 95] T. K. Leung, M.C Burl, et P. Perona, “Finding Faces in Cluttered Scenes Using Random Labeled Graph Matching”, *Proceeding of Fifth IEEE International Conference on Computer Vision*, Los Alamitos, CA, pp. 637-644, 1995.

- [LI 20] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey”, *IEEE Transactions on Affective Computing*, pp. 1949-3045, 2020, DOI: 10.1109/TAFFC.2020.2981446.
- [LIA 01] L. Liang, H. Ai et G. Xu, “Face Detection based on Template Matching and Neural Network Verification”, *Acta Electronica Sinica*, 2001
- [LIA 06] C.S. Chung, D.Y. Yeung, S. Liao, W. Fan, “Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features”, In *IEEE International Conference on Image Processing (ICIP)*, Atlanta, GA, USA, pp. 665- 668, 2006.
- [LIA 20] D. Liang, H. Liang, Z. Yu et Y. Zhang, “Deep convolutional BiLSTM fusion network for facial expression recognition”, *Visual. Computer*, 36(3), pp. 499-508, 2020, DOI: 10.1007/s00371-019-01636-3.
- [LIS 05] J. Liscombe, J. Hirschberg and J. J. Venditti, “Detecting Certainness in Spoken Tutorial Dialogues”, *Speech Com. Tech*, pp. 1837-1840, 2005.
- [LIU 12] L. Liu, L. Zhao, Y. Long, G. Kuang, P. Fieguth, “Extended local binary patterns for texture classification”, *Image and Vision Computing*, 30(2), pp. 86 - 99, 2012.
- [LIU 14] P. Liu, S. Han, Z. Meng and Y. Tong, “Facial Expression Recognition via a Boosted Deep Belief Network”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1805-1812, 2014.
- [LIU 14] Y. Liu, X. Hou, J. Chen, C. Yang, G. Su and W. Dou, “Facial Expression Recognition and Generation using Sparse Autoencoder”, *International Conference on Smart Computing*, 125-130, 2014.
- [LOP 17] A.T Lopes, E. De Aguiar, A.F. De Souza, and T. Oliveira-Santos, “Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order”, *Pattern Recognition*, 61, pp. 610–628, 2017.
- [LU 21] L. Lu, “Multi-angle face expression recognition based on generative adversarial network”, *Computational Intelligence Wiley*, Special issue, 1–18, 2021, DOI: 10.1111/coin.12437.
- [LUC 10] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression”, in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* , San Francisco, USA, pp. 94–101, 2010.
- [LUC 81] B. D. Lucas et T. Kanade, “An iterative image registration technique with an application to stereo vision”, *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, Vancouver, British, Columbia, pp. 674– 680, 1981.
- [LYO 98] M.J. Lyons et S. Akamatsu, “Coding facial expressions with gabor wavelets”, *Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, pp. 200-205, 1998.
- [LYO 99] M.J. Lyons, J. Budynek, and S. Akamatsu, “Automatic classification of single facial images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12), pp.1357–1362, 1999.
- [MAA 12] H.Maamatou, “Reconnaissance et mappage des expressions faciales vers un modèle virtuel”, Mémoire de Mastère de recherche, Systèmes Avancés en Génie Electrique, Université de Sousse, 2012.
- [MAN 92] G. Mandler, “Cognition and emotion: Extensions and clinical applications,” *Cognitive science and clinical disorders*, Academic Press, San Diego, USA, 1992.
- [MAY 95] J. D. Mayer, J. Allen and K. Beauregard, “Mood inductions for four specific moods: A procedure employing guided imagery vignettes with music”, *Journal of Mental Imagery*, 19, pp. 133-150, 1995.

- [MAY 97] J. D. Mayer and P. Salovey, “What is emotional intelligence ? *Emotional Development and Emotional Intelligence: Implications for Educators*”, New York: Basic books, pp.3-31, 1997.
- [MEH 68] A. Mehrabian, “Communication without words”, *Psychology Today*, 2 (4), pp.53-56, 1968.
- [MEP 98] MPEG Video and SNHC, “Text of ISO/IEC FDIS 14 496-3: Audio,” in *Atlantic City MPEG Mtg.*, Oct. 1998, Doc. ISO/MPEG N2503.
- [MER 20a] G. Merzougui, “Apprentissage profond”, *Support du cours*, Université Mustapha Benboulaïd de Batna, Spécialité Statistiques et Data Science (SDS), 2020.
- [MER 20b] G. Merzougui, “Les réseaux de neurones convolutionnels CNN”, *Support du cours*, Université Mustapha Benboulaïd de Batna, Spécialité Statistiques et Data Science (SDS), 2020.
- [MIN 86] M. Minsky, “The Society of Mind”, New York: Simon and Schuster, 1986.
- [MOH 17] M. Mohammadpour, H. Khaliliardali, S.M. RHashemi, and M. MalyanNezhadi, “Facial Emotion Recognition using Deep Convolutional Networks”, in *IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, Tehran, Iran, pp. 17-21, 2017.
- [MOK 18] M.Z. Mokri, “Classification des images avec les réseaux de neurones convolutionnels”, *these de doctorat*, 2018.
- [MOS 18] A. Mostafa, M.I. Khalil et H. Abbas, “ Emotion Recognition by Facial Features using Recurrent Neural Networks, *13th International Conference on Computer Engineering and Systems (ICCES)*, Cairo Egypt, pp. 417-422, 2018.
- [MOU 17] D.Y. Moualek, “Deep Learning pour la classification des images”, *Mémoire de master 2*, Université Abou Bakr Belkaid, Tlemcen, Algérie, 2017.
- [NAG 19] S. Nagpal, M. Singh, M. Vatsa, R. Singh and A. Noore, “Expression Classification in Children Using Mean Supervised Deep Boltzmann Machine”, *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, USA, 236-245, 2019, DOI: 10.1109/CVPRW.2019.00033.
- [NES 83] Y. Nesterov, “A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$ ”, In : *Doklady AN USSR*, 269, pp.543–547, 1983.
- [NG 02] A.Y. Ng and M.I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes”, *Advances in neural information processing systems*, 841-848, 2002.
- [NIE 15] S. Nie, Z. Wang and Q. Ji, “A generative restricted Boltzmann machine based method for high-dimensional motion data modeling”, *Computer Vision and Image Understanding*, 136: 14–22, 2015, DOI: 10.1016/j.cviu.2014.12.005
- [NUG 09] A. Nugier, “Histoire et grands courants de recherche sur les émotions”, *Revue électronique de Psychologie Sociale*, (4), pp.5–7, 2009.
- [NKA 04] R. Nkambou, and V. Héritier, “Reconnaissance émotionnelle par l’analyse des expressions faciales dans un tuteur intelligent affectif,” *Colloque des Technologies de l’Information et de la Connaissance dans l’Enseignement Supérieur et l’industrie (TICE)*, Compiègne, France. pp.149-155, 2004.
- [NKA 10] R. Nkambou, R. Mizoguchi et J. Bourdeau, “Advances in intelligent tutoring systems”, *Springer Science & Business Media*, 308-505, 2010.
- [NKA 97] R. Nkambou, G. Gauthier and C. Frasson, “Un modèle de représentation des connaissances relatives au contenu dans un système tutoriel intelligent”, *Sciences et Techniques Éducatives*, 4(3), 299-330, 1997.
- [OCH 04a] M. Ochs, “Systèmes Tutoriels Émotionnellement Intelligents”, *Master Thesis*, Université de Montréal, Montréal, QC, 2004.

- [OCH 04b] M. Ochs and C. Frasson, “Emotionally Intelligent Tutoring Systems (EITS)”, *The Florida Artificial Intelligence Research Society (FLAIRS) Conference*, Miami, Floride, Etats-Unis, pp. 251-255, 2004.
- [OER 07] K. Oertel, R. Kaiser, J. R. Voskamp and B. Urban, “AffectIX - An Affective Component as Part of an E-Learning-System”, *Proceedings of the 3rd international conference on Foundations of Augmented Cognition (FAC’07)*, Beijing, China, pp. 385-393, 2007.
- [OJA 02] T. Ojala, M. Pietikainen and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), pp. 971-987, 2002.
- [OJA 96] T. Ojala, M. Pietikäinen, D. Harwood, “A comparative study of texture measures with classification based on featured distributions”, *Pattern Recognition*, 29 (1), pp. 51-59, 1996.
- [ORE 03] K. O’Regan, “Emotion and E-Learning”, *Journal of Asynchronous Learning Networks (JALN)*, 7(3), pp. 78-92, 2003.
- [PAD 96] C. Padgett, G. Cottrell and R. Adolphs, “Categorical Perception in Facial Emotion Classification”, *In Proceedings of the 18th Annual Conference of the Cognitive Science Society*, University of California, San Diego, pp. 249-253, 1996.
- [PAN 04] M. Pantic and L.J.M. Rothkrantz, “Facial action recognition for facial expression analysis from static face images”, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(3), pp.1449–1461, 2004.
- [PAQ 11] G. Paquette, “Les environnements d’apprentissage intelligents: notes de cours et illustrations”, *INF5100, Télé-université au Québec, Département informatique*, 2011.
- [PEA 01] K. Pearson, “On lines and planes of closest fit to points in space”, *Philosophical Magazine*, 2(11), pp. 559-572, 1901.
- [PEK 11] R. Pekrun, T. Goetz, A. C. Frenzel, P. Barchfeld and R. P. Perry, “Measuring emotions in students’ learning and performance: The Achievement Emotions Questionnaire (AEQ)”, *Contemporary Educational Psychology*, 36(1), pp. 36-48, 2011.
- [PIC 01] R. Picard, E. Vyzas and J. Healey, “Toward Machine Emotional Intelligence: Analysis of Affective Physiological State”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), pp. 1175-1191, 2001.
- [PIC 97] R. Picard, “Affective Computing”, *MIT Press*, 1997.
- [PLU 80] R. Plutchik, “A general psycho-evolutionary theory of emotion”, *Theories of emotion*, Academic press, pp. 3-33, 1980, DOI: 10.1016/B978-0-12-558701-3.50007-7.
- [POR 20] S. Porcu, A. Floris and L. Atzori, “Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems”, *Electronics*, 9(11), 1892, 2020, DOI: :10.3390/electronics9111892
- [PRE 03] H. Prendinger, S. Mayer, J. Mori and M. Ishizuka, “Persona effect revisited. Using bio-signals to measure and reflect the impact of character-based interfaces”, *4th International Working Conference on Intelligent Virtual Agents (IVA-03)*, Kloster Irsee, Germany, pp. 283-291, 2003.
- [PRE 05] H. Prendinger and M. Ishizuka, “The empathic companion: A character-based interface that addresses users’ affective states”, *Applied Artificial Intelligence*, 19(3), pp. 267-285, 2005.
- [PSO 88] J. Psotka, L. D. Massey and S. A. Mutter, “Intelligent tutoring systems: lessons learned”, *L. Erlbaum Associates*, 1988.
- [QUI 68] M. R. Quillian, “Semantic Memory”, In M. Minsky (Ed.), *Semantic Information Processing MITPress*, Cambridge, pp. 216-270, 1968.
- [RAP 12] V. Rapp, T. Senechal, L. Prevost, K. Bailly, H. Salam and R. Segulier, “Combinaison de descripteurs hétérogènes pour la reconnaissance de micro-mouvements

- faciaux”, In *Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, Lyon, France. pp. 978-2-9539515-2-3, 2012.
- [RAV 20] R. Ravi, and al., “A Face Expression Recognition Using CNN & LBP”, in *Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp. 684 – 689, 2020., DOI: 10.1109/ICCMC48092.2020.ICCMC-000127
- [ROB 09] J. L. Robison, S. W. Mcquiggan and J. C. Lester, “Modeling Task-Based vs. Affect-based Feedback Behavior in Pedagogical Agents: An Inductive Approach”, *International Conference on Artificial Intelligence in Education*, Brighton, England, IOS Press, pp. 25-32, 2009.
- [ROW 98] H.A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1), pp. 23-38, 1998.
- [ROZ 03] P. Rozin, and A. B. Cohen, “Confusion infusions, suggestives, correctives, and other medicines”, *Emotion*, 3(1), pp. 92–96, 2003.
- [RUI 17] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, V. Palade, “Stacked Deep Convolutional Auto-Encoders for Emotion Recognition from Facial Expressions”, *International Joint Conference on Neural Networks (IJCNN)*, Anchorage, USA, 1586-1593, 2017, DOI: 10.1109/IJCNN.2017.7966040.
- [RUM 86] D.E. Rumelhart, G. E. Hinton et R. J. Williams, “Learning internal representations by error propagation”, *Parallel distributed processing :explorations in the microstructure of cognition*, 1, pp.318–362, 1986.
- [RUS 05] C. A. Russell, G. F. Wilson, M. M. Rizki, T. S. Webb and S. C. Gustafson, “Comparing classifiers for real time estimation of cognitive workload”, *Foundations of Augmented Cognition*. Las Vegas, NV, 2005.
- [RUS 80] J. A. Russell, “A circumplex model of affect”, *Journal of Personality and Social Psychology - PSP*, 39(6), pp. 1161-1178, 1980.
- [RUS 89] J. A. Russell, A. Weiss and G. A. Mendelsohn, “Affect grid: a single-item scale of pleasure and arousal”, *Journal of personality and social psychology*, 57(3), pp. 493-502, 1989.
- [SAL 20] W.M. Salih Abedi, A. T. Sadiq and I. Nadher, “Modified CNN-LSTM for Pain Facial Expressions Recognition”, *International Journal of Advanced Science and Technology*, 29 (3), pp. 304-312, 2020.
- [SAL 90] P. Salovey and J. D. Mayer, “Emotional intelligence”, *Imagination, Cognition and Personality*, 9(3), 185-211, 1990.
- [SAL 09] R. Salakhutdinov and G. Hinton, “Deep Boltzmann machines,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Clearwater, 2009.
- [SAN 02] K. Sandeep et A.N. Rajagopalan, “Human Face Detection in Cluttered Color Images Using Skin Color and Edge Information”, *Proceedings of the Third Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*, Ahmadabad, India, 2002.
- [SAX 96] D. Saxe and R. Foulds, “Toward Robust Skin Identification in Video Images”, *Proceeding of the Second International Conference on Automatic Face and Gesture Recognition*, Killington, VT, USA, pp. 379-384, 1996.
- [SCH 00] K.R. Scherer, “Emotion”, In M. Hewstone & W. Stroebe (Eds.), *Introduction to Social Psychology : A European perspective*, Oxford Blackwell Publishers, pp. 151–191, 2000.
- [SCH 05] K. R. Scherer, “What are emotions? and how can they be measured ?,” *Social Science Information*, 44(4), pp. 695–729, 2005.

- [SCH 98] H. Schneiderman and T. Kanade, “Probabilistic modeling of local appearance and spatial relationships for object recognition”, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, USA, pp. 45, 1998.
- [SEB 05] N. Sebe, I. Cohen, and T.S. Huang, “Multimodal Emotion Recognition”, In *Handbook of Pattern Recognition and Computer Vision*, pp. 387-409 (2005), DOI: 10.1142/9789812775320_0021.
- [SHA 09] C. Shan, S. Gong, and P. W. Mcowan, “Facial expression recognition based on Local Binary Patterns: A comprehensive study”, *Image and Vision Computing*, 27(6), pp. 803–816, 2009.
- [SHA 20] J. H. Shah, M. Sharif, M. Yasmin and S. L. Fernandes, “Facial expressions classification and false label reduction using LDA and threefold SVM”, *Pattern Recognition Letters*, 139, pp. 166-173, 2020.
- [SHB 15] R. Shbib and S. Zhou, “Facial expression analysis using active shape model”, *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(1), pp.9–22, 2015.
- [SHI 94] J. Shi, C. Tomasi, “Good features to track”, *Proceedings CVPR '94*, 1994 IEEE Computer Society Conference on, pp. 593- 600, 1994.
- [SHI 04] F. Shih and C. Chuang, “Automatic extraction of head and face boundaries and facial features”, *Information Sciences*, 158, pp.117–130, 2004.
- [SIM 91] P.K. Simpson, “Fuzzy min-max neural networks”, *IEEE International Joint Conference on Neural Networks*, Singapore, 1658–1669, 1991, DOI: 10.1109/IJCNN.1991.170647.
- [SLE 82] D. Sleeman, and J. S. Brown, “Intelligent Tutoring Systems”, *New York: Academic Press*, 1982.
- [SLO 87] A. Sloman and M. Croucher “Why Robots Will Have Emotions”, *Proceedings of IJCAI-87*, Vancouver, Canada, pp. 197-202, 1987.
- [SOB 96] K. Sobottka, et I. Pitas, “Face Localization and Feature Extraction Based on Shape and Color Information”, *Proceeding of 3rd IEEE International Conference on Image Processing*, Lausanne, Switzerland, pp. 483-486, 1996.
- [SOH 07] A. S. M Sohail, and P. Bhattacharya, “Classification of facial expressions using k-nearest neighbor classifier”, In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pp. 555–566, 2007.
- [SON 10] M. Song, D. Tao, Z. Liu, X. Li, and M. Zhou, “Image ratio features for facial expression Recognition Application”, *IEEE Transactions on systems, man, and cybernetics Part B : Cybernetics*, 40(3), pp. 779–788, 2010.
- [SRI 14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout A Simple Way to Prevent Neural Networks from Overfitting”, *Journal of Machine Learning Research*, 15 (56), pp.1929–1958, 2014.
- [STA 01] Standard mpeg-4 : ISO/IEC 14496-2, Information Technology – Coding of Audio-visual Objects, 2001.
- [STE 91] N. L. Stein and L. J. Levine, “Making sense out of emotion: The representation and use of goal-structured knowledge”, In W. Kessen, A. Ortony, & F. I. M. Craik (Eds.), *Memories, thoughts, and emotions: Essays in honor of George Mandler*, pp. 295-322, 1991.
- [STU 92] D. T. Stuss, C. A. Gow, and C. Hetherington, “ ‘no longer gage’ : Frontal lobe dysfunction and emotional changes,” *Journal of Consulting and Clinical Psychology*, 60(3), pp. 349 -359, 1992.
- [SUM 15] S. A. M. Al-Sumaidae, S. S. Dlay, W. L. Woo and J. A. Chambers, “Facial Expression Recognition using Local Gabor Gradient Code-Horizontal Diagonal

- Descriptor”, *2nd IET International Conference on Intelligent Signal Processing (ISP)*, London, UK, 2015, DOI: 10.1049/cp.2015.1766
- [SUN 17] W. Sun, H. Zhao, Z. Jin, “An efficient unconstrained facial expression recognition algorithm based on Stack Binarized Auto-encoders and Binarized Neural Networks”, *Neurocomputing*, 267: 385–395, 2017, DOI: 10.1016/j.neucom.2017.06.050.
- [TAN 10] X. Tan and B. Triggs, “Enhanced local texture feature sets for face recognition under difficult lighting conditions”, *IEEE transactions on image processing*, 19(6), pp. 1635–1650, 2010.
- [TIA 01] Y. Tian, T. Kanade et J.F. Cohn, “Recognizing action units for facial expression analysis”, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(2), pp. 97–115, 2001.
- [TOM 84] S. S. Tomkins, “Affect theory”, *Approaches to emotion*, Hillsdale, NJ: Erlbaum, 163, pp. 31-65, 1984.
- [TON 14] Y. Tong, R. Chen, and Y. Cheng, “Facial expression recognition algorithm using-LGC based on horizontal and diagonal prior principle”, *Optik Journal*, 125(16), pp. 4186–4189, 2014.
- [TUR 91] M.A. Turk, A.P. Pentland, “Face recognition using eigenfaces”, *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, 1991.
- [USM 17] M. Usman, S. Latif and J. Qadir, “Using Deep Autoencoders for Facial Expression Recognition”, *13th International Conference on Emerging Technologies (ICET)*, 1-6, 2017.
- [VAL 07] M.F. Valstar, H. Gunes et M. Pantic, “How to distinguish posed from spontaneous smiles using geometric features”, *In Proceedings of the 9th International Conference on Multimodal Interfaces*, pp. 38-45, 2007.
- [VAN 05] K. Vanlehn, C. Lynch, K. Schulze, J. A. Shapiro and R. Shelby, L. Taylor, D. Treacy, A. Weinstein and M. Wintersgill, “The Andes physics tutoring system: Lessons learned”, *International Journal of Artificial Intelligence in Education*, 15(3), pp. 147–204, 2005.
- [VAP 99] V.N. Vapnik, “An overview of statistical learning theory”, *IEEE Transaction on Neural Networks*, 10(5), pp. 988–999, 1999.
- [VAS 18] Vasudha and D. Kakkar, “Facial Expression Recognition with LDPP & LTP using Deep Belief Network », *5th International Conference on Signal Processing and Integrated Networks (SPIN)*, 503-508, 2018.
- [VIE 19] V. Vielzeuf, “Apprentissage Neuronal Profond pour l’Analyse de Contenus Multimodaux et Temporels”, *thèse de doctorat*, Université de Caen Normandie, 2019.
- [VIN 10] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion”. *JMLR*, 2010.
- [VIO 01] P. Viola and M. Jones, “Robust real-time object detection”, *In Second inter-national workshop on statistical and computational theories of vision – modeling, learning, computing, and sampling*, Vancouver, Canada, 2001.
- [WAN 14] X. Wang, V. Ly, R. Guo and C. Kambhamettu, “2D-3D Face Recognition via Restricted Boltzmann Machines”, *Proceedings of the 9th International Conference on Computer Vision Theory and Applications (VISAPP)*, 574-580, 2014, DOI: 10.5220/0004736505740580.
- [WAN 18a] X. Wang, W. Li, G. Mu, D. Huang and Y. Wang, “Facial Expression Synthesis by U-Net Conditional Generative Adversarial Networks”, *International Conference on Multimedia Retrieval*, Yokohama, Japan, 283-290, 2018, DOI: 10.1145/3206025.3206068

- [WAN 18b] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, “Deep learning for smart manufacturing :Methods and applications”. *Journal of Manufacturing Systems*, 48 , pp.144–156, 2018.
- [WAN 19] Y. Wang, Y. Li, Y. Song and X. Rong, “The Application of a Hybrid Transfer Algorithm Based on a Convolutional Neural Network Model and an Improved Convolution Restricted Boltzmann Machine Model in Facial Expression Recognition”, *IEEE Access*, 7 : 184599- 184610, 2019, DOI: 10.1109/ACCESS.2019.2961161.
- [WAN 21] J. Wang, “Improved Facial Expression Recognition Method Based on GAN”, *Hindawi*, Special issue, 2021, DOI: doi.org/10.1155/2021/2689029.
- [WEI 00] R. Weiss, and R. Palumbo, “Emotion and Learning”, *Training & Development*, 54(11), pp. 44-48, 2000.
- [WEN 87] E. Wenger, “Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge”. Los Altos, CA,1987.
- [WU 13] Y. Wu, Z. Wang and Q. Ji, “Facial Feature Tracking under Varying Facial Expressions and Face Poses based on Restricted Boltzmann Machines”, *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, 3450-3459, 2013, DOI: 10.1109/CVPR.2013.443.
- [WU 17] Y. Wu and W. Qiu, “Facial expression recognition based on improved deep belief networks”, *AIP Conference Proceedings*, 1864: 020130-1–020130-4, 2017, DOI: 10.1063/1.4992947
- [YAN 02] M.H. Yang, D. J. Kriegman, and N. Ahuja. “Detecting faces in images: A survey”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1), pp 34–58, 2002.
- [YAN 17] B. Yang, J. Cao, R. Ni and Y. Zhang, “Facial Expression Recognition using Weighted Mixture Deep Neural Network Based on Double-channel Facial Images”, *IEEE Access*, 6 (2017), pp. 4630-4640, 2017, DOI:10.1109/ACCESS.2017.2784096.
- [YAN 19] Y. Yan, C. Li, Y. Lu, F. Zhou, Y. Fan and M. Liu “Design and Experiment of Facial Expression Recognition Method based on LBP and CNN”, *14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Xi'an, China, pp. 602-607, 2019.
- [YAN 94] G. Yang et T. S. Huang, “Human Face Detection in Complex Background”, *Pattern Recognition*, 27(1), pp. 53–63, 1994.
- [YOL 19] G. Yolcu, I. Oztel, S. Kazan, C. Oz, K. Palaniappan, T.E. Lever and F. Bunyak, “Facial expression recognition for monitoring neurological disorders based on convolutional neural network”, *Multimedia Tools and Applications*, 78(22), pp. 31581-31603, 2019, DOI: 10.1007/s11042-019-07959-6.
- [YU 18] Z. Yu, G. Liu, Q. Liu, et J. Deng, “Spatio-temporal convolutional features with nested LSTM for facial expression recognition”, *Neurocomputing*, 317, pp. 50-57, 2018, DOI: 10.1016/j.neucom.2018.07.028.
- [ZEN 07] Z. Zeng, J.Tu, M. Lu, T. S. Huang, B. Pianfetti, D. Roth et S. Levinson, “Audio visual affect recognition”, *IEEE Transactions on Multimedia*, 9(2), pp. 424–428, 2007.
- [ZEN 18] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li and A. M. Dobaie, “Facial Expression Recognition via Learning Deep Sparse Autoencoders”, *Neurocomputing*, 273 (C) :643–649 2018, DOI: 10.1016/j.neucom.2017.08.043.
- [ZHA 07] G. Zhao, and 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), pp. 915–928, 2007, DOI: 10.1109/TPAMI.2007.1110.
- [ZHA 98] Z. Zhang, M. Lyons, M. Schuster et S. Akamatsu, “Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron”, *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 454–459, 1998.

- [ZHE 98] Z. Zhengyou, M. Lyons, M. Schuster, and S. Akamatsu, “Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron”, In *3rd IEEE International Conference on Automatic Face and Gesture Recognition (FG 98)*, Nara, Japan, pp. 454–459, 1998, DOI: 10.1109/AFGR.1998.670990.
- [ZHU 22] X. Zhu, Z. He, L. Zhao, Z. Dai and Q. Yang, “A Cascade Attention Based Facial Expression Recognition Network by Fusing Multi-Scale Spatio-Temporal Features”, *Sensors*, 22, pp. 1350, 2022, DOI: 10.3390/s22041350.

Liste des publications

- O. Chebah and M.T. Laskri, “Facial expression recognition for e-learning systems”, In *4th International Conference on Networking and Advanced Systems (ICNAS)*, Annaba, Algeria, 2019.
- O. Chebah and M.T. Laskri, “Facial expression recognition system”, In *2nd Conference on Informatics and Applied Mathematics (IAM)*, Guelma, Algeria, 2019.
- O. Chebah and M.T. Laskri, “Facial expression recognition Facial expression recognition using convolution neural network fusion and texture descriptors representation”, *International Journal of Computational Intelligence and Applications*, 21(1), 2250002 , 2022, DOI :10.1142/S146902682250002X.
- O. Chebah and M.T. Laskri, “Local directional double ternary coding pattern for facial expression recognition”, *International Journal of Computational Vision and Robotics*, 2022.