UQAC

UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

THÈSE DE DOCTORAT

Reconnaissance des expressions faciales *pour* l'assistance ambiante

Auteur :
Yacine YADDADEN

Superviseurs :
Abdenour BOUZOUANE
Mehdi ADDA
Sébastien GABOURY

Cette thèse est présentée comme exigence partielle du Doctorat

en sciences et technologies de l'information

Laboratoire d'intelligence ambiante pour la reconnaissance d'activités
Département d'informatique et de mathématiques

20 juin 2019

Moi, Yacine YADDADEN, déclare que cette thèse intitulée, « Reconnaissance des expressions faciales *pour* l'assistance ambiante » et le travail qui y est présenté sont les miens. Je confirme que :

— Ce travail a été fait entièrement ou principalement lors de la candidature pour un diplôme de recherche à cette université.

— Si une partie de cette thèse a déjà été soumise pour un diplôme ou toute autre qualification à cette université ou dans une autre institution, cela a été clairement indiqué.

— Lorsque j'ai consulté le travail publié par d'autres, cela est toujours clairement attribué.

— Là où j'ai cité le travail d'autrui, la source est toujours donnée. À l'exception de ces citations, cette thèse est entièrement mon propre travail.

— J'ai reconnu toutes les principales sources d'aide.

— Lorsque la thèse est basée sur un travail effectué par moi-même conjointement avec d'autres, j'ai clairement précisé ce qui a été fait par les autres et ce que j'ai moi-même contribué.

Signé :
____

Date :
____

« Le succès n'est pas final, l'échec n'est pas fatal : c'est le courage de continuer qui compte. »

Winston Churchill

UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

Résumé

Département d'informatique et de mathématiques

Doctorat

Reconnaissance des expressions faciales *pour* l'assistance ambiante

par Yacine YADDADEN

Au cours de ces dernières décennies, le monde a connu d'importants changements démographiques et notamment au niveau de la population âgée qui a fortement augmenté. La prise d'âge a comme conséquence directe non seulement une perte progressive des facultés cognitives, mais aussi un risque plus élevé d'être atteint de maladies *neurodégénératives* telles qu'Alzheimer et Parkinson. La perte des facultés cognitives cause une diminution de l'autonomie et par conséquent, une assistance quotidienne doit être fournie à ces individus afin d'assurer leur bien-être. Les établissements ainsi que le personnel spécialisé censés les prendre en charge représentent un lourd fardeau pour l'économie. Pour cette raison, d'autres solutions moins coûteuses et plus optimisées doivent être proposées.

Avec l'avènement des nouvelles technologies de l'information et de la communication, il est devenu de plus en plus aisé de développer des solutions permettant de fournir une assistance adéquate aux personnes souffrant de déficiences cognitives. Les maisons intelligentes représentent l'une des solutions les plus répandues. Elles exploitent différents types de *capteurs* pour la collecte de données, des algorithmes et méthodes d'apprentissage automatique pour l'extraction/traitement de l'information et des *actionneurs* pour le déclenchement d'une réponse fournissant une assistance adéquate. Parmi les différentes sources de données qui sont exploitées, les images/vidéos restent les plus riches en termes de quantité. Les données récoltées permettent non seulement la reconnaissance d'activités, mais aussi la détection d'erreur durant l'exécution de tâches/activités de la vie quotidienne.

La reconnaissance automatique des émotions trouve de nombreuses applications dans notre vie quotidienne telles que l'interaction homme-machine, l'éducation, la sécurité, le divertissement, la vision robotique et l'*assistance ambiante*. Cependant, les émotions restent un sujet assez complexe à cerner et de nombreuses études en psychologie et sciences cognitives continuent d'être effectuées. Les résultats obtenus servent de base afin de développer des approches plus efficaces. Les émotions humaines peuvent être perçues à travers différentes *modalités* telle que la voix, la posture, la gestuelle et les *expressions faciales*. En se basant sur les travaux de Mehrabian, les expressions faciales représentent la modalité la plus pertinente pour la reconnaissance automatique des émotions. Ainsi, l'un des objectifs de ce travail de recherche consistera à proposer des méthodes permettant l'identification des six émotions de base à savoir : la joie, la peur, la colère, la surprise, le dégoût et la tristesse. Les méthodes proposées exploitent des données d'entrée *statiques* et *dynamiques*, elles se basent aussi sur différents types de descripteurs/représentations (*géométrique*, *apparence* et *hybride*).

Après avoir évalué les performances des méthodes proposées avec des bases de données *benchmark* à savoir : **JAFFE**, **KDEF**, **RaFD**, **CK+**, **MMI** et **MUG**. L'objectif principal de ce travail de recherche réside dans l'utilisation des expressions faciales afin d'améliorer les performances des systèmes d'assistance existants. Ainsi, des expérimentations ont été conduites au sein de l'environnement intelligent **LIARA** afin de collecter des données de validation, et ce, en suivant un protocole d'expérimentation spécifique. Lors de l'exécution d'une tâche de la vie quotidienne (*préparation du café*), deux types de données ont été récoltés. Les données **RFID** ont permis de valider la méthode de reconnaissance automatique des actions utilisateurs ainsi que la détection automatique d'erreurs. Quant aux données faciales, elles ont permis d'évaluer la contribution des expressions faciales afin d'améliorer les performances du système d'assistance en termes de détection d'erreurs. Avec une réduction du taux de fausses détections dépassant les 20%, l'objectif fixé a été atteint avec succès.

Remerciements

Le travail présenté dans cette thèse s'inscrit dans le cadre des activités de recherche du Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (**LIARA**). J'aimerais remercier toutes les personnes qui ont contribué, de près ou de loin, à l'achèvement de ce projet et qui m'ont aidé tout au long. J'aimerais en premier lieu exprimer toute ma gratitude à mes deux directeurs de thèse Abdenour Bouzouane et Mehdi Adda qui m'ont fait confiance en me confiant ce projet. Ils ont su me guider et m'orienter durant ce travail de recherche, ils m'ont toujours soutenu et se sont montrés plus que patients avec moi. J'ai appris énormément à leurs côtés, car ils m'ont toujours permis d'être autonome et m'ont encouragé à prendre des initiatives.

Je tiens aussi à remercier mon codirecteur Sébastien Gaboury ainsi que Bruno Bouchard qui ont contribué au bon déroulement de ce projet. Ils ont tout fait pour fournir un environnement de travail propice à la recherche. Pour cela, je les remercie du fond du cœur.

Enfin, j'aimerais remercier mes parents pour leur soutien à toute épreuve. En particulier ma mère qui m'a toujours encouragé à aller le plus loin possible dans tout ce que j'entreprends. Mes sincères remerciements aussi à mes frères et ma sœur qui ont toujours trouvé les mots pour me redonner le sourire et me remonter le moral dans les moments difficiles.

Table des matières

iii

Résumé **viii**

Remerciements **ix**

1 Introduction Générale **1**

1.1 Contexte et motivation 1

1.2 Émotions et expressions faciales 3

1.2.1 Définition des émotions 4

1.2.2 Théories sur les émotions 4

Théorie de l'*universalité* 5

Théorie *physiologique* 6

1.2.3 Représentation des émotions 6

Approche *dimensionnelle* 6

Approche *catégorielle* 8

1.2.4 Reconnaissance automatique des émotions et applications . . . 8

1.2.5 Les expressions faciales 11

1.3 Problématiques et objectifs 12

1.3.1 Reconnaissance automatique des expressions faciales 13

Entrée 14

Pré-traitements 15

Génération des caractéristiques 15

Sélection des attributs 17

Modèle de Classification 18

Sortie 19

1.3.2 Environnement intelligent et assistance ambiante 19

1.4 Contributions de la thèse 22

1.4.1 Reconnaissance automatique d'expressions faciales 22

1.4.2 Contribution à l'assistance ambiante 25

1.4.3 Conférences Internationales 26

1.4.4 Journaux Internationaux 27

1.5 Méthodologie de recherche 28

1.5.1 Bases de données publiques 29

Base de données *statique* 30

Base de données *dynamique* 31

1.5.2 Environnement d'expérimentation 32

1.5.3 Collecte de données 33

1.5.4 Protocole d'évaluation 34

1.6 Organisation de la thèse 35

2 One-Class and Bi-Class SVM Classifier Comparison for Automatic Facial Expression Recognition 39
Abstract . . . 40
2.1 Introduction . . . 40
2.2 Background . . . 41
2.2.1 Automatic Facial Expression Recognition . . . 41
2.2.2 Multi-class Support Vector Machine . . . 41
2.2.3 Related Works . . . 41
2.3 Proposed Approach . . . 42
2.3.1 Geometric-based Features . . . 42
2.3.2 Feature Ranking & Reduction . . . 42
2.3.3 Value Normalization . . . 43
2.3.4 Multi-class Support Vector Machine . . . 43
2.4 Experimentation & Evaluation . . . 43
2.5 Results & Discussion . . . 43
2.5.1 Comparison Between Multi-class SVM Architectures . . . 43
2.5.2 Evaluation of the Static AFER Approach . . . 44
2.6 Concluding Remarks . . . 44
Acknowledgment . . . 44
References . . . 45

3 Facial Sub-regions for Automatic Emotion Recognition using Local Binary Patterns 47
Abstract . . . 48
3.1 Introduction . . . 48
3.2 Background . . . 49
3.2.1 Fundamentals . . . 49
3.2.2 Related Works . . . 49
3.3 Proposed Approach . . . 50
3.3.1 Pre-processing . . . 50
3.3.2 extended Local Binary Patterns . . . 50
3.3.3 Principal & Independent Component Analysis . . . 50
3.3.4 Multi-Class Support Vector Machine . . . 51
3.4 Experimentation & Evaluation . . . 51
3.5 Results & Discussion . . . 51
3.6 Comparison between **PCA** and **ICA** . . . 51
3.7 Evaluation of the static **AFER** approach . . . 52
3.8 Concluding Remarks . . . 52
Acknowledgment . . . 52
References . . . 52

4 Hybrid-based Facial Expression Recognition Approach for Human-Computer Interaction 55
Abstract . . . 56
4.1 Introduction . . . 56
4.2 Background . . . 56
4.2.1 Fundamentals . . . 57
4.2.2 Related Works . . . 57
4.3 Proposed Approach . . . 57
4.3.1 Pre-processing . . . 57
4.3.2 Geometric-based Features . . . 57

4.3.3 Appearance-based Features 58
4.3.4 Feature Selection 59
4.3.5 Classification 59
4.4 Experimentation & Evaluation 60
4.5 Results & Discussion 60
4.6 Concluding Remarks 60
Acknowledgment 61
References 61

5 An Automatic Facial Expression Recognition Approach using an Efficient Spatio-Temporal Representation 63
Abstract 64
5.1 Introduction 65
5.2 Background 67
5.3 Related Works 68
5.3.1 Static Systems 68
5.3.2 Dynamic Systems 69
Sequence-based 69
Frame-based 70
5.3.3 Limitations & Challenges 71
5.4 Proposed Approach 72
5.4.1 Feature Extraction 72
5.4.2 Feature Selection 74
5.4.3 Classification 75
SVM classifier 76
***k*-NN** classifier 76
Other classifiers 77
5.5 *Frame-based* Version 77
5.6 Experimentation & Evaluation 79
5.6.1 Benchmark Datasets 79
5.6.2 Validation Strategy 80
5.7 Obtained Results 80
5.7.1 Machine Learning Techniques 80
5.7.2 **CK+** dataset (six classes) 81
5.7.3 **MMI** dataset 81
5.7.4 **MUG** dataset 82
5.7.5 **CK+** dataset (seven classes) 83
5.7.6 Sequence-based vs Frame-based 84
5.7.7 Cross-dataset Evaluation 85
5.8 Discussion 85
5.9 Conclusion 87
Acknowledgment 87
References 88

6 User Action and Facial Expression Recognition for Error Detection System in an Ambient Assisted Environment 91
Abstract 92
6.1 Introduction 92
6.2 Background 93
6.2.1 Automatic facial expressions recognition 93
6.2.2 Deep learning 94

6.2.3 Smart environment & ambient assistance . . . 96
6.3 Automatic facial expression recognition . . . 97
6.4 Proposed *ambient assisting system* . . . 98
6.4.1 Action recognition module . . . 98
6.4.2 Error detection module . . . 99
6.5 Experimentation & validation . . . 100
6.5.1 Validation of the static AFER approach . . . 100
6.5.2 Validation of the ARM . . . 100
6.5.3 Experimental protocol . . . 101
6.6 Results and analysis . . . 102
6.6.1 Performance of the static AFER approach . . . 102
JAFFE dataset . . . 102
RaFD dataset . . . 102
KDEF dataset . . . 103
MMI dataset . . . 103
CK+ dataset . . . 103
Other evaluations . . . 103
6.6.2 ARM performance . . . 104
6.6.3 EDM performance . . . 105
6.7 Discussion . . . 105
6.7.1 Static AFER approach . . . 105
6.7.2 ARM . . . 106
6.7.3 EDM . . . 106
6.8 Conclusion and further works . . . 106
Acknowledgment . . . 106
References . . . 107

7 Conclusion Générale **109**
7.1 Objectifs accomplis . . . 111
7.2 Comparaison des approches développées . . . 114
7.3 Limitations connues . . . 116
7.4 Perspectives et travaux futurs . . . 117

Bibliographie **121**

Table des figures

1.1 Théories *physiologiques* : James-Lange [29] et Cannon-Bard [6]. 5
1.2 Le modèle de Plutchik [53]. 7
1.3 Les muscles faciaux et les nerfs de contrôle [32]. 11
1.4 Système de reconnaissance automatique. 14
1.5 Aperçu de l'environnement intelligent **LIARA**. 20
1.6 Échantillon d'images de base de données **JAFFE**. 30
1.7 Échantillon de séquence (*joie*) de base de données **CK+**. 31
1.8 Description de la *protocole d'évaluation* adopté. 34
2.1 (a) Original image (b) Image with facial fiducial points. 42
2.2 Overview of the proposed static AFER approach. 42
2.3 Accuracy of three different multi-class SVM architectures (JAFFE). . . 43
2.4 Accuracy per used multi-class SVM architecture. 44
2.5 % of attributes per used multi-class SVM architecture. 44
3.1 Overview of the proposed static AFER approach. 49
3.2 Overview of the basic and extended LBP extraction process. 50
3.3 Convergence of the accuracy when using **PCA** (*six emotions*). 51
3.4 Convergence of the accuracy when using **ICA** (*six emotions*). 52
3.5 Evaluating the used *dimension reduction* techniques (*six emotions*). . . . 52
4.1 Overview of the proposed *hybrid-based* AFER approach. 58
4.2 (a) Original image (b) Image with facial fiducial points. 58
4.3 (a) 2-D DWT decomposition (b) 2-D DWT regions. 59
4.4 The proposed classification architecture using multi-class SVM. 59
4.5 Affects of the three types of features on the accuracy (*six emotions*). . . 60
4.6 Comparison between feature selection versions (*six emotions*). 60
5.1 Basic building blocks of a common AFER system. 67
5.2 Overview of our dynamic AFER approach (*sequence-based* version). . . 72
5.3 Face detection & shape extraction. 73
5.4 *Spatio-temporal* representation of the *input*. 73
5.5 Overview of our *dynamic* AFER approach (*frame-based* version). 78
5.6 Accuracy of the proposed *sequence-based* approach. 80
5.7 Comparison between *sequence-based* and *frame-based* approach. 84
6.1 Overview of a common **AFER** system. 94
6.2 Components of a basic ambient assisting system. 96
6.3 Preprocessing workflow on an image from **JAFFE** dataset. 97
6.4 The proposed CNN architecture *optimized* for **AFER**. 98
6.5 Overview of the proposed *ambient assisting* system. 99
6.6 Description of the proposed user **ARM**. 99
6.7 Description of the proposed **EDM**. 100
6.8 Validation process of the proposed *static* **AFER** approach. 101
6.9 The LIARA lab & used objects with RFID tags. 102
6.10 Comparison between the proposed CNN and **LeNet-5**. 103
6.11 Affect of various image enhancement techniques. 104

6.12 Accuracy using the five datasets with various image enhancement techniques. 104
6.13 Basic action recognition accuracy per subject. 104
6.14 Basic action recognition accuracy during *coffee making*. 105
6.15 **EDM** evaluation using **ARM** information. 105
6.16 EDM evaluation using AFER information. 105

7.1 Comparaison avec base de données *statique*. 115
7.2 Comparaison avec base de données *dynamique*. 116

Liste des tableaux

1.1 Les bases de données de type *statique*. 30
1.2 Les bases de données de type *dynamique*. 32
2.1 Image facial expression datasets. 43
2.2 Computing time evaluation in training and predicting (JAFFE). 44
2.3 Accuracy for each emotion on the three benchmark datasets using three different multi-class SVM classifiers. 45
2.4 Comparison with *state-of-the-art* methods in terms of accuracy. 45
3.1 Image facial expression datasets. 51
3.2 Accuracy for each emotion on the three benchmark datasets (*six* & *seven* classes). 53
3.3 Comparison with *state-of-the-art* methods in terms of accuracy. 53
4.1 Image facial expression datasets. 60
4.2 Accuracy for each emotion on the three benchmark datasets (*six* & *seven* classes). 61
4.3 Comparison with *state-of-the-art* methods in terms of accuracy. 61
5.1 Benchmark facial expression datasets. 79
5.2 Confusion matrix using **CK+** (*six classes*) dataset. 81
5.3 Comparison with *state-of-the-art* methods using **CK+** (*six classes*). 81
5.4 Confusion matrix using **MMI** dataset. 82
5.5 Comparison with *state-of-the-art* methods using **MMI**. 82
5.6 Confusion matrix using **MUG** dataset. 82
5.7 Comparison with *state-of-the-art* methods using **MUG**. 83
5.8 Confusion matrix using **CK+** (*seven classes*) dataset. 83
5.9 Comparison with *state-of-the-art* methods using **CK+** (*seven classes*). . . 83
5.10 Comparison between *sequence-based* and *frame-based* using various metrics. 84
5.11 Computational time for the *sequence-based* and *frame-based* versions. . . 85
5.12 Cross-dataset evaluation for both *sequence-based* and *frame-based* versions. 85
6.1 Benchmark facial expression datasets. 101
6.2 AFER approach accuracy comparatively to state-of-the-art methods (**JAFFE**). 102
6.3 Confusion matrix using the **JAFFE** dataset. 102
6.4 AFER approach accuracy comparatively to state-of-the-art methods (**RaFD**). 102
6.5 Confusion matrix using **RaFD** dataset. 103
6.6 AFER approach accuracy comparatively to state-of-the-art methods (**KDEF**). 103
6.7 Confusion matrix using **KDEF** dataset. 103
6.8 AFER approach accuracy comparatively to state-of-the-art methods (**MMI**). 103
6.9 Confusion matrix using **MMI** dataset. 103

6.10 AFER approach accuracy comparatively to state-of-the-art methods (**CK+**). . . . 104
6.11 Confusion matrix using **CK+** dataset. . . . 104
6.12 Confusion matrix for action recognition in a subject independent way. 105

Liste des abréviations

LIARA	**L**aboratoire d'**I**ntelligence **A**mbiante pour la **R**econnaissance d'**A**ctivités
FACS	**F**acial **A**ction **C**oding **S**ystem
AU	**A**ction **U**nit
ASM	**A**ctive **S**hape **M**odel
AAM	**A**ctive **A**ppearance **M**odel
FE	**FE**ar (*peur*)
SU	**SU**rprise (*surprise*)
HA	**HA**ppiness (*joie*)
DI	**DI**sgust (*dégoût*)
AN	**AN**ger (*colère*)
SA	**SA**dness (*tristesse*)
NE	**NE**utral (*état neutre*)
CO	**CO**ntempt (*mépris*)
RFID	**R**adio **F**requency **ID**entification
CASAS	**C**enter for **A**dvanced **S**tudies in **A**daptive **S**ystems
DOMUS	laboratoire de **DOM**otique et informatique mobile l'**U**niversité de **S**herbrooke
COACH	**C**ognitive **O**rthosis for **A**ssisting a**C**tivities in the **H**ome

Je dédie cette thèse à mes très chers parents . . .

Chapitre 1

Introduction Générale

1.1 Contexte et motivation

Avec l'amélioration des conditions de vie, ainsi que le développement de la médecine au cours de ces dernières décennies, la population âgée a connu une forte augmentation. Ainsi, la population canadienne de plus de 85 ans est passée de 169 637 en **1988** à 821 490 individus en **2018** [1]. L'une des conséquences directes du vieillissement est la diminution progressive des facultés cognitives et l'augmentation des risques de souffrir de maladies *neurodégénératives* telles qu'Alzheimer et Parkinson. La population âgée touchée éprouve des difficultés et devient incapable d'exécuter les tâches les plus basiques de la vie quotidienne. Les établissements ainsi que le personnel spécialisé mis en place par le gouvernement pour la prise en charge de cette tranche d'âge sont un fardeau pour l'économie. De plus, certains individus désirent rester chez eux et sont réticents à l'idée d'être placés dans des résidences spécialisées.

Afin d'assurer le bien-être des personnes âgées tout en leur permettant de rester chez elles, de nouveaux domaines de recherche ont vu le jour tel que l'intelligence ambiante. L'objectif principal est de proposer des solutions alternatives moins onéreuses et plus efficaces. En effet, avec l'avènement du numérique et le développement des nouvelles technologies, le monde se dirige de plus en plus vers une utilisation massive des technologies de l'information et de la communication telles que les ordinateurs, téléphones intelligents, tablettes numériques, etc. Ces derniers permettent chaque jour de collecter d'énormes quantités de données relatives à l'utilisateur, son environnement et son quotidien. Plus récemment encore, l'apparition de

[1] `https://www.statcan.gc.ca/`

nouvelles techniques d'apprentissage automatique a provoqué un important bouleversement [36], notamment dans la vision artificielle [37]. En effet, ces techniques ont permis de mettre en avant le potentiel latent de ces dispositifs par l'analyse approfondie et automatique des données collectées quotidiennement [3, 22]. L'un des enjeux majeurs de ces dernières années consiste dans développement de nouvelles méthodes pour l'analyse et l'extraction automatique d'informations utiles et pertinentes à partir de ces données brutes.

Parmi les solutions technologiques qui ont été mises en place, les *maisons intelligentes* sont les plus répandues. Ce sont des environnements qui ont été équipés de différents dispositifs permettant de collecter des données, de les analyser avant de fournir une assistance adéquate. Parmi ces environnements, nous pouvons citer le Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (**LIARA**)[2] et le laboratoire de DOmotique et en informatique Mobile à l'Université de Sherbrooke (**DOMUS**)[3]. Le rôle principal des méthodes d'intelligence ambiante réside dans l'analyse et le traitement des données de bas et haut niveau collectées au sein de l'environnement intelligent.

Le type de données de haut niveau le plus commun et qui fournit d'importantes quantités d'information reste l'audiovisuel. En effet, les récents progrès en traitement de signaux ainsi que l'apparition des nouvelles méthodes d'apprentissage automatique ont favorisé le développement des solutions d'*assistance ambiante*. Elles exploitent la vision par ordinateur [40] afin de comprendre et d'interpréter de façon automatique ce qui se passe à partir d'une scène [26]. L'information ainsi extraite permet de détecter les situations nécessitant une assistance adéquate. D'autres chercheurs ont préféré cibler des activités spécifiques telles que le *lavage des mains* [47] ou bien le *brossage des dents* [52]. Ce genre de systèmes intelligents est destiné principalement aux personnes âgées atteintes de déficience cognitive.

Les émotions représentent une des composantes principales d'une interaction. En effet, elles jouent un rôle primordial dans notre vie de tous les jours, notamment dans la communication non verbale. Se basant sur les travaux de Mehrabian [46], les

[2] `http://www.uqac.ca/liara/index.htm`

[3] `https://www.usherbrooke.ca/domus`

expressions faciales contribuent le mieux à la transmission du message comparativement au timbre de voix et aux traces verbales. Pour cette raison, différents travaux ont été menés afin de développer des méthodes de reconnaissance des émotions à partir des expressions faciales. Cependant, le développement de méthodes efficaces reste un challenge et cela est dû à la complexité des expressions faciales.

Dans le cadre de ce projet de recherche, nous nous sommes intéressés à la contribution des expressions faciales afin d'augmenter l'efficacité des systèmes d'assistance ambiante. En premier lieu, nous avons travaillé sur la reconnaissance automatique des émotions à travers les expressions faciales. Chacune des méthodes proposées a été évaluée à l'aide de base de données *benchmark*. Ensuite, nous avons conduit des expérimentations au sein du **LIARA** afin de collecter des données lors d'une tâche de la vie quotidienne. Les données ont servi pour la validation des deux modules composant le système d'assistance proposé à savoir : la reconnaissance d'activité et la détection d'erreurs.

1.2 Émotions et expressions faciales

Les émotions font partie intégrante de notre vie. En effet, elles se manifestent dans notre quotidien et contribuent à l'amélioration des interactions entre les individus. De plus, elles suscitent l'attention des chercheurs dans les domaines de la psychologie, car elles permettent une meilleure compréhension de l'être humain. Ainsi, différentes recherches et expérimentations ont été menées. Avec l'avènement des nouvelles technologies de l'information et de la communication, d'autres domaines d'application sont apparus pour exploiter la *reconnaissance automatique des émotions*. Dans ce qui suit, nous nous intéresserons aux bases et fondamentaux des émotions et plus précisément aux expressions faciales, car c'est la *modalité* qui a été adoptée dans le cadre de ce travail de recherche.

1.2.1 Définition des émotions

Initialement, les émotions ont été étudiées du point de vue psychologique. Ainsi, elles sont définies comme des sentiments qui entraînent des changements *physiques* et *psychologiques* qui influencent la pensée et le comportement. Elles sont aussi définies par d'autres chercheurs comme des états psychologiques complexes qui impliquent quatre composantes principales [24, 49], à savoir :

Composante physiologique : qui est représentée par les changements physiques (*internes*) tels que la fréquence cardiaque, le flux sanguin et la température corporelle.

Composante subjective : qui définit la façon dont l'émotion est ressentie par l'individu. Elle est propre à chaque individu.

Composante cognitive : qui est relative à la façon dont l'individu interprète et comprend la scène ainsi que la manière dont il évalue le stimulus extérieur.

Composante expressive/comportementale : qui est directement liée aux changements physiques (*externes*) tels que les *expressions faciales*, la voix et la gestuelle. Elles peuvent aussi être perçues comme des *modalités* qui servent à l'identification des émotions.

Les émotions peuvent aussi être perçues comme étant des états motivationnels [8]. Elles sont constituées d'impulsion, de désir ou d'aversion et elles impliquent des changements de motivation. Elles poussent, également, l'individu à modifier sa relation avec un objet, un état du monde, un état de soi ou bien à maintenir une relation existante malgré des obstacles [7].

1.2.2 Théories sur les émotions

Plusieurs études et travaux ont été réalisés sur le domaine des émotions. C'est un domaine qui a toujours attiré l'intérêt des chercheurs en psychologie. Par conséquent, dans ce qui suit nous aborderons les deux théories les plus connues dans le domaine des émotions.

Théorie de l'*universalité*

Elle est aussi associée à la *théorie néo-darwinienne* car Charles Darwin (1872) fut l'un des premiers à s'intéresser aux phénomènes émotionnels et à leur expression chez l'homme et l'animal [14]. Ces travaux s'inscrivent dans la continuité de sa théorie de l'évolution. Ainsi, il a mis en évidence l'importance et le rôle des émotions dans la survie des espèces. De plus, il a émis l'hypothèse que les émotions sont inscrites dans le système nerveux humain, ce qui les rend *universelles*.

Dans ses travaux, Darwin s'est intéressé aux expressions faciales comme *modalité*. Ainsi, il a classé les changements au niveau des expressions faciales en sept groupes distincts. En reprenant son travail, d'autres chercheurs contemporains ont essayé de confirmer par des expérimentations la théorie de l'*universalité* en définissant un certain nombre d'émotions dites de base. Paul Ekman [16] est une référence dans le domaine des émotions et des expressions faciales. Il a émis l'hypothèse selon laquelle les émotions de base ont des caractéristiques uniques. Elles sont primaires, fondamentales et sont indépendantes des cultures et des origines ethniques.

Les émotions sont déterminantes lors d'une interaction non-verbale. En effet, elles jouent un rôle communicatif permettant ainsi de réguler le comportement du récepteur ou de l'informer sur l'état émotionnel de l'individu.

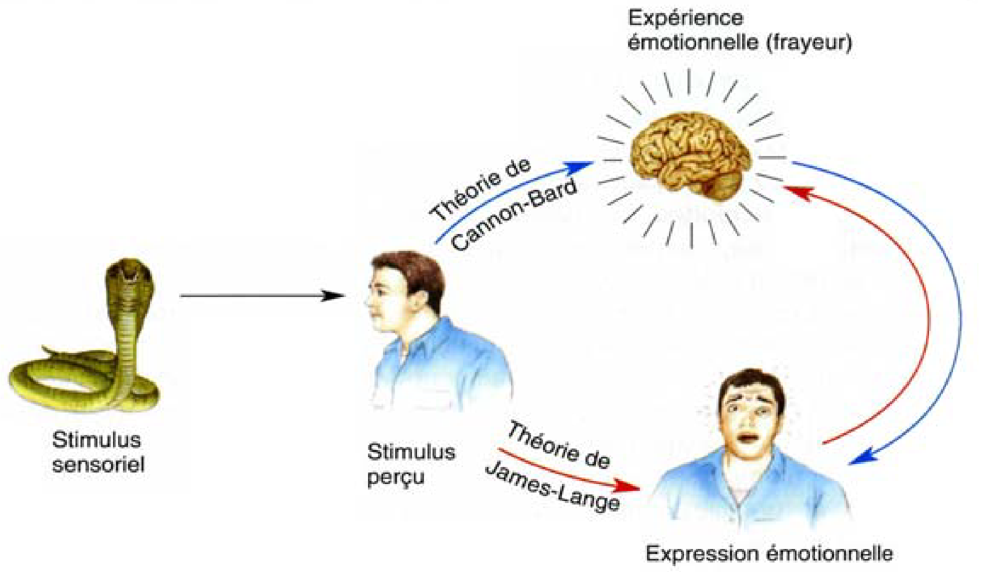


FIGURE 1.1 – Théories *physiologiques* : James-Lange [29] et Cannon-Bard [6].

Théorie *physiologique*

L'hypothèse défendue par le psychologue américain William James (1884) [29] et soutenue par le physiologiste danois Carl Lange (1885) met en évidence le rôle déterminant des réactions émotionnelles dans l'activation des expériences émotionnelles. Ainsi, ils défendent l'hypothèse selon laquelle les réponses du système nerveux périphérique (changement de température corporelle, larmes, augmentation de la fréquence cardiaque et la sueur) sont à l'origine de l'expérience émotionnelle.

Tout à fait à l'opposé de cette théorie, Walter Cannon (1927) [6] a avancé une autre théorie, qui a été soutenue par Phillip Bard (1934), selon laquelle l'expérience émotionnelle est à l'origine des changements corporels et physiologiques. Il a été prouvé à travers des expériences que l'état émotionnel persiste même après que les réponses du système nerveux périphérique se soient progressivement dissipées.

À partir de la Figure 1.1, nous pouvons voir une illustration des deux théories physiologiques décrites. Ainsi, nous remarquons qu'à partir d'un *stimulus* externe, qui consiste en la vue d'un serpent, il y a deux processus qui peuvent être déclenchés. Le premier (James-Lange) implique que les réactions corporelles et physiologiques vont être à l'origine de l'expérience émotionnelle. Inversement, le second (Cannon-Bard) implique que c'est l'expérience émotionnelle qui va déclencher les différentes réponses physiologiques.

1.2.3 Représentation des émotions

Dans la littérature, il y a généralement deux types de représentations des émotions qui sont couramment utilisées. La première est *dimensionnelle* ou *continue* alors que la seconde est *catégorielle* ou *discrète* . Ces représentations ont des caractéristiques spécifiques que nous détaillerons dans ce qui suit.

Approche *dimensionnelle*

C'est une représentation *continue* sur plusieurs axes ou dimensions. Parmi les plus connues, nous pouvons citer le modèle de Russel [54]. Il a proposé une représentation des émotions qui n'est pas catégorique, mais *bidimensionnelle* suivant deux axes distincts. Le premier axe est celui du *plaisir* (plaisir/peine) alors que le second est dédié à la quantification de la *force du ressenti* (positive/négative). L'intérêt de ce genre de représentation réside dans le fait qu'elle permet de situer l'ensemble des émotions *catégoriques* sur les deux axes. Se basant sur ce travail, Cowie et al. [12] ont proposé l'outil *Feeltrace* pour l'annotation des émotions à l'aide d'un modèle *bidimensionnelle* similaire.

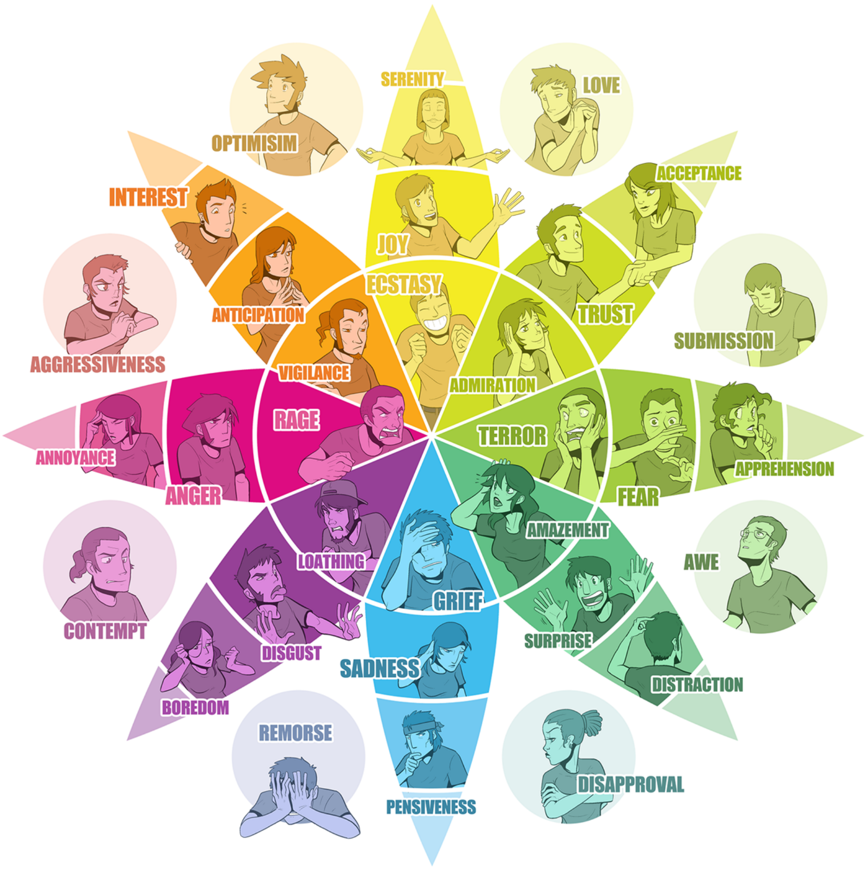


FIGURE 1.2 – Le modèle de Plutchik [53].

Dans la Figure 1.2 ci-dessus est illustré le modèle proposé par Plutchik [53]. C'est une représentation qui est souvent qualifiée d'*hybride* ou de *mixte*. En effet, son modèle inclut *huit* émotions de base (peur, surprise, joie, dégoût, colère, tristesse, acceptation et anticipation) qui sont représentées à l'aide d'un cercle de sorte à avoir quatre ensembles d'émotions opposées sans utiliser une approche *dimensionnelle*. De plus, la représentation proposée inclut aussi des émotions secondaires.

Approche *catégorielle*

D'autres théoriciens se sont penchés sur un autre type de représentation qui comporte une classification des émotions suivant un certain nombre de groupe. En effet, cette catégorisation est faite suivant le concept d'émotion de base. Ainsi, des recherches ont pu mettre en évidence la présence d'un certain nombre d'émotions primaires et fondamentales. Chacune d'elle possède des caractéristiques qui lui sont propres et qui permettent de la distinguer des autres émotions. Chacun des théoriciens qui soutiennent et utilisent cette représentation ont introduit un certain nombre d'*émotions de base* comme défini ci-dessous :

Ekman [17] : peur, surprise, joie, dégoût, colère et tristesse.

Tomkins [57] : peur, surprise, joie, dégoût, colère, détresse, mépris et honte.

Izar [27] : peur, surprise, joie, dégoût, colère, détresse, mépris, culpabilité, intérêt et honte.

Cette représentation est adoptée dans la plupart des travaux dédiés à la reconnaissance automatique des émotions. Ainsi, les bases de données que nous avons utilisées afin d'évaluer les méthodes proposées adoptent la représentation *catégorielle* proposée par Ekman [17].

1.2.4 Reconnaissance automatique des émotions et applications

Avec l'avènement des nouvelles technologies de l'information et de la communication, il a été possible de développer des méthodes et systèmes permettant de reconnaître de façon automatique l'état émotionnel d'un individu. Ces systèmes exploitent des *modalités* spécifiques comme source d'information pour l'identification de l'émotion exprimée. Généralement, il y a trois types de sources d'information qui sont utilisées [55, 28].

1. **Source audio :** le périphérique d'acquisition dans ce cas est le *microphone* permettant d'enregistrer le signal sonore. À partir duquel, il est possible d'extraire l'*expression verbale* et les *traces verbales* qui permettent d'identifier l'émotion ressentie par l'individu.

2. **Source visuelle :** dans ce cas, le périphérique d'acquisition consiste en la *caméra* qui enregistre soit de simples images ou bien des séquences (vidéos). Ces données ainsi acquises peuvent contribuer à la détection de mouvements de la tête, la reconnaissance des expressions faciales, gestuelles et la posture.
3. **Source audiovisuelle :** il est aussi possible d'exploiter les deux sources en même temps afin d'avoir une plus grande quantité d'information et ainsi améliorer les performances du système de reconnaissance automatique des émotions. Ce genre de système est dit *multimodal* puisqu'il exploite plus d'une source d'information par contraste à celui qui est *unimodal*.

La reconnaissance automatique des émotions trouve plusieurs applications dans différents domaines parmi lesquels :

Interaction homme-machine : les approches classiques d'interaction deviennent obsolètes et il y a une forte demande de nouvelles approches innovatrices et plus intuitives. Les smartphones, consoles de jeux vidéo et télévisions intelligentes incluent l'utilisation de capteurs audio (*microphones*) et visuel (*caméras*) afin d'améliorer l'interaction. Plus précisément, les signaux audiovisuels collectés sont analysés afin de reconnaître l'état émotionnel de l'utilisateur et d'adapter le contenu multimédia selon son humeur [28].

Éducation : avec l'avènement des technologies de l'information et de la communication et en particulier internet, l'enseignement a bousculé les barrières avec ce que nous appelons aujourd'hui l'e-learning qui permet de s'inscrire et de suivre des cours à distance. Afin d'adapter le contenu selon le niveau de l'utilisateur, il est nécessaire de détecter et de reconnaître de façon automatique ses réactions qui sont liées à son état émotionnel [59].

Sécurité : ce sont les applications liées à la vidéosurveillance, les données acquises par les caméras sont analysées afin de détecter et reconnaitre les actions des personnes ainsi que leurs expressions faciales. Cela permet dans la cadre de la sécurité de détecter les comportements suspects dans les lieux publics (aéroports et gares) [33, 21].

Divertissement : les films, séries télévisées et jeux vidéo font partie intégrante de notre vie quotidienne. Afin d'évaluer leur qualité, et la manière dont ils sont accueillis par le public, il est possible de se baser sur les différentes émotions qu'elles provoquent. En effet, si le film n'inspire que des émotions négatives (ennui, tristesse et colère), son évaluation sera impactée en conséquence [25].

Vision robotique : les robots humanoïdes sont de plus en plus répandus. Afin d'être plus proches de l'être humain, la capacité de reconnaissance automatique d'émotions est nécessaire. En effet, les robots ayant cette capacité sont capables de percevoir l'état émotionnel de la personne en face et de réagir en conséquence [19, 41].

Assistance ambiante : les personnes âgées qui sont en perte d'autonomies requièrent une assistance quasi permanente dans leur vie quotidienne. Pour des soucis économiques, la technologie a été mise à leur disposition afin de leur fournir l'assistance dont ils ont besoin tout en leur permettant de rester chez eux. C'est bien moins coûteux que de les placer dans des établissements spécialisés ou bien d'engager du personnel pour s'en occuper. Un des points clés de l'assistance, en particulier pour les personnes âgées qui ne sont pas à l'aise avec la technologie, est de trouver le meilleur moyen pour interagir avec elles tout en anticipant leurs besoins. La détection et la supervision de leur état émotionnel permettent de s'assurer de leur bien-être psychologique [43].

Dans le cadre de ce travail de recherche, nous nous concentrerons sur le développement de méthode *unimodales* exploitant les expressions faciales comme source d'information. Même si les différentes méthodes proposées dans cette thèse peuvent être exploitées dans différents domaines, nous nous focaliserons sur leur utilisation dans le cadre de l'*assistance ambiante* afin d'améliorer les performances des systèmes existants en termes d'assistance.

1.2.5 Les expressions faciales

Comme décrit dans les sections précédentes, il y a plusieurs *modalités* qui peuvent être exploitées pour la reconnaissance automatique des émotions. Cependant, toutes les sources d'information ne se valent pas en termes de pertinence. Selon Mehrabian [46], la majeure partie de l'information (55%) lors d'une interaction est transmise à travers les expressions faciales, alors que le reste est partagé entre la partie vocale (38%) et verbale (7%). Nous basant sur ce travail, nous avons choisi d'explorer l'utilisation les expressions faciales comme données d'entrée pour les différentes méthodes proposés.

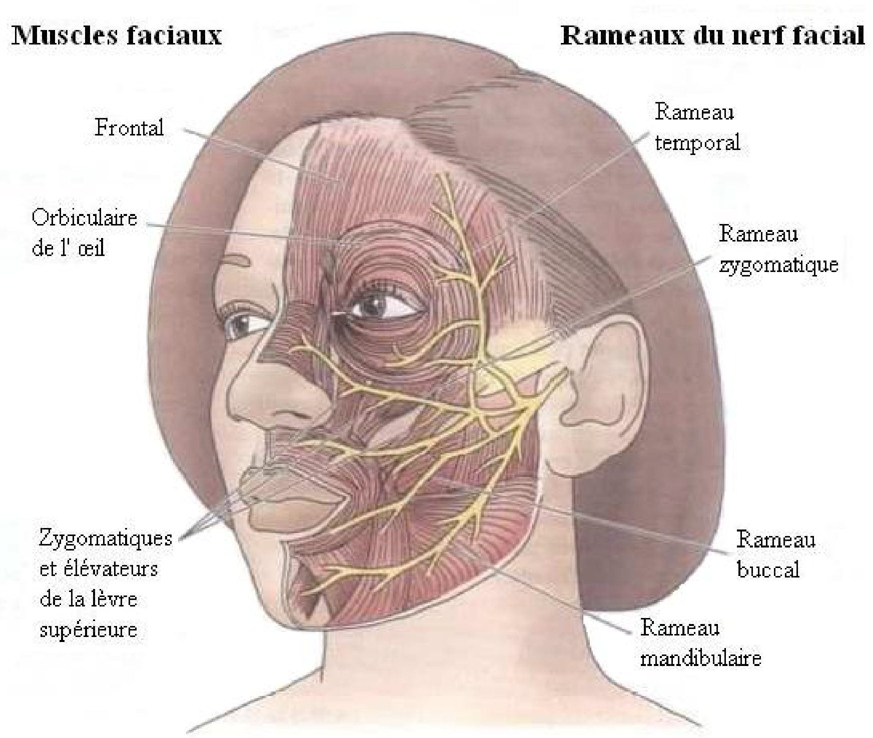


FIGURE 1.3 – Les muscles faciaux et les nerfs de contrôle [32].

Les expressions faciales sont considérées comme un élément clé de la communication non-verbale. Elles peuvent être définies comme les variations et changements dans le visage, perceptibles visuellement, et qui sont dû à l'activation (volontaire ou non) d'un ou plusieurs des *quarante-quatre* muscles composant le visage (250000 expressions possibles) [32]. Dans la Figure 1.3 sont représentés les différents types de muscles faciaux ainsi que les nerfs permettant leur contrôle.

La plupart des recherches qui ont été menées dans le domaine des émotions se sont basées sur les expressions faciales pour la représentation des émotions. Ainsi, Ekman a associé les émotions des bases aux expressions faciales. Avec l'aide de Friesen, il a aussi proposé un système appelé **FACS** (*Facial Action Coding System*) permettant le codage des expressions faciales en quantifiant le mouvement des muscles du visage [18]. Le système exploite des **AU** (*Action Unit*) pour la caractérisation des différentes expressions faciales. Chacune des *quarante-quatre actions unitaires* correspond au mouvement d'un muscle facial. Une expression faciale qui correspond à une émotion spécifique est codée par la combinaison d'un certain nombre d'*actions unitaires* en plus de l'intensité de contraction du muscle qui est codée quant à elle sur *cinq* niveaux. Il y a encore de nombreuses méthodes de reconnaissance automatique des expressions faciales qui utilisent ce système.

1.3 Problématiques et objectifs

Avec l'importante croissance de la population âgée et l'augmentation des cas de démence, l'un des enjeux majeurs de ces dernières décennies réside dans le développement de nouveaux systèmes permettant d'assurer le bien-être de cette partie de la population. Diverses initiatives ont été entreprises pour l'exploitation des technologies de l'information et de la communication afin de développer des systèmes d'assistance ambiante. Cependant, ils sont loin d'être efficaces comparés à une assistance humaine. De plus, la prise en charge de personnes souffrantes de déficience cognitive est une tâche délicate, voire critique. Pour cette raison, les deux modules ; *reconnaissance d'activité* et *détection d'erreurs* doivent être impérativement optimisés, car l'efficacité du système d'assistance en dépend. Divers prototypes ont été proposés [5, 20], notamment au sein du **LIARA**. Ils sont principalement basés sur l'utilisation de la technologie **RFID** (*Radio Frequency IDentification*) et malgré les résultats prometteurs obtenus, leurs performances restent à améliorer. D'autres chercheurs ont pris l'initiative d'utiliser la vision par ordinateur, en plus de capteurs de bas niveau afin d'augmenter l'efficacité de l'assistance fournie [47, 52].

L'objectif principal de ce travail de recherche consiste à développer un système d'assistance combinant les capteurs de bas niveau (technologie **RFID**) et haut niveau (vision par ordinateur). Pour cela, nous nous sommes d'abord intéressés aux expressions faciales pour les importantes quantités d'information qu'elles fournissent. Nous nous sommes ainsi focalisés dans un premier temps sur le développement de méthodes de reconnaissance automatique des émotions à partir des expressions faciales. Chacune des méthodes proposées a été évaluée à l'aide de bases de données *benchmark* afin de s'assurer de son efficacité. Nous avons ensuite mené des expérimentations au sein du **LIARA** afin de récolter les données nécessaires à la validation de notre système d'assistance (modules de reconnaissance d'activité et détection d'erreurs). Sa particularité principale réside dans l'étude de l'apport des expressions faciales pour l'amélioration de la détection d'erreurs. Dans ce qui suit, nous aborderons les différents éléments nécessaires à la compréhension des objectifs que nous nous sommes fixés dans le cadre de cette thèse.

1.3.1 Reconnaissance automatique des expressions faciales

Dans le cadre de notre travail, nous avons été amenés à exploiter les technologies de l'information et de la communication afin de concevoir des systèmes de reconnaissance automatique. Ce genre de système est composé des mêmes blocs qu'un système de reconnaissance de forme standard [34]. Comme illustré dans la Figure 1.4, le système intègre quatre blocs distincts, en plus de l'entrée et sortie, qui se chargent d'effectuer des tâches bien spécifiques.

La *reconnaissance automatique des expressions faciales* fait appel à un système ayant le même schéma et composition que celui présenté dans la Figure 1.4. Cependant, chacun des différents blocs a des caractéristiques qui lui sont propres et qui coïncident avec la tâche à effectuer, à savoir la reconnaissance automatique d'émotions à travers les expressions faciales. Dans ce qui suit, nous expliquerons brièvement chacun des différents blocs.

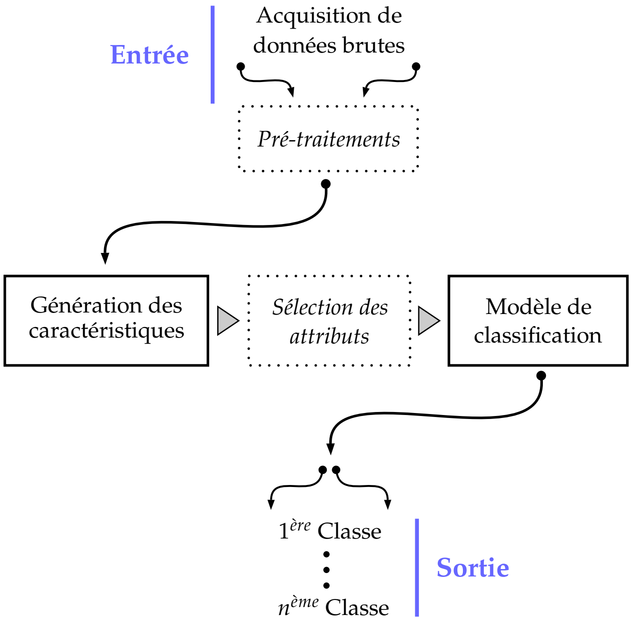


FIGURE 1.4 – Système de reconnaissance automatique.

Entrée

Dans le contexte de la *reconnaissance automatique des expressions faciales*, les données d'entrée peuvent être sous deux formes distinctes. La première est la plus simple et elle est dite *statique*. Elle exploite de simples images où sont représentés les visages de personnes exprimant une émotion précise [61]. L'autre type de données d'entrée est dit *dynamique* et consiste en l'utilisation de séquences d'images [63]. En tenant compte de l'aspect temporel, il permet de détailler les différentes phases de transitions propres aux expressions faciales (*début*, *sommet* et *fin*). Cette représentation est avantageuse en raison des informations et détails supplémentaires fournis. Néanmoins, il y a un inconvénient à cause de l'augmentation de la complexité de traitement.

Dans le cadre de ce travail, nous nous sommes principalement focalisés sur le développement de méthodes *statiques* [61, 64, 68, 66, 67, 65] car l'objectif final est l'implémentation dans des systèmes embarqués. En effet, afin d'optimiser leur fonctionnement dans des systèmes avec des ressources matérielles limitées, il est nécessaire de réduire la complexité de traitement. Néanmoins, nous avons aussi travaillé sur des méthodes *dynamiques* en proposant une représentation *spatio-temporelle* [63, 62]. L'objectif étant de trouver le bon équilibre entre précision et complexité.

Pré-traitements

Généralement les données d'entrée sont brutes et peuvent nécessiter des *pré-traitements* afin de faciliter l'exécution des étapes suivantes [42]. Néanmoins, cette étape n'est pas indispensable, car il est possible de traiter les données d'entrée telles quelles au détriment d'un bon taux de reconnaissance. Comme illustré dans la Figure 1.4, les données d'entrée alimentent un bloc de *pré-traitements* afin de les formater, rehausser leur qualité et les préparer pour les traitements suivants. Différentes opérations peuvent y être appliquées au niveau de la forme et de la texture.

Génération des caractéristiques

Une des étapes les plus critiques dans les systèmes de reconnaissance automatique reste l'*extraction de caractéristiques*. Cette opération consiste en la génération d'une *représentation pertinente* à l'aide de descripteurs spécifiques qui dépendent du type de données d'entrée. En tenant compte des algorithmes qui ont été proposés récemment, nous distinguons deux approches possibles pour la caractérisation des données d'entrée :

1. **Manuelle :** c'est au concepteur du système de reconnaissance de définir le type de descripteur à exploiter afin de générer une représentation adéquate. Différents types de caractéristiques peuvent être utilisés, parmi lesquels : les *motifs binaires locaux* [50] ou l'*histogramme de gradient orienté* [13]. L'inconvénient principal de ce genre de descripteurs réside dans le manque de constance, c'est-à-dire qu'un descripteur peut être efficace avec une certaine base de données, mais ne le sera pas forcément avec une autre.
2. **Automatique :** avec l'apparition des algorithmes d'apprentissage profond, de nouvelles approches d'extraction de caractéristiques inspirées de modèles biologiques ont été proposées [38]. Elles permettent de générer des représentations pertinentes de façon automatique et adaptative. Les résultats présentés dans les récents travaux attestent de leurs efficacités.

Dans le contexte de la *reconnaissance automatique des expressions faciales*, nous recensons trois types de descripteurs qui sont très répandus dans la littérature :

1. **Géométrique :** ce type de descripteurs se base sur l'utilisation de *points de caractéristiques du visage*. Il existe différents algorithmes qui peuvent être exploités afin de générer ce genre de représentations tels que l'**ASM** (*Active Shape Model*) [11] ou l'**AAM** (*Active Appearance Model*) [10]. Le principal avantage de ce type de descripteur réside dans son insensibilité aux effets de contraste et de luminosité. Dans le cadre de ce travail de recherche, nous ferons principalement appel à la technique proposée par Kazemi et Sullivan [31]. Nous avons utilisé ce genre de descripteur dans le cadre des travaux suivants : [61, 63, 67].
2. **Apparence :** ce genre de descripteurs exploite les caractéristiques de textures. De nombreuses techniques peuvent être exploitées afin d'extraire une représentation pertinente à partir de l'ensemble de pixels de l'image. Parmi ces différentes techniques, nous pouvons citer des coefficients issus de la *transformée en ondelettes* [56], les *motifs binaires locaux* [50], l'*histogramme de gradient orienté* [13] ou les filtres de Gabor [45]. Nous avons utilisé ce genre de descripteur dans le cadre des travaux suivants : [66, 65].
3. **Hybride :** ce dernier descripteur exploite la fusion des informations issues des deux représentations précédentes (*géométrique* et *apparence*). Le principal avantage de ce genre de descripteurs réside dans le fait de combiner deux représentations pertinentes afin d'améliorer le taux de reconnaissance. Il y a deux façons de fusionner les représentations et qui consistent en :

   (a) **Amont :** où la combinaison des descripteurs est réalisée par une simple *concaténation* des vecteurs de caractéristiques.

   (b) **Aval :** où chacune des deux représentations est traitée indépendamment et la fusion est réalisée au niveau du bloc de *classification*.

   Dans le cadre de cette thèse, nous avons proposé une méthode *hybride* où les deux représentations (*géométrique* et *apparence*) sont combinées en *aval* [66].

Sélection des attributs

Selon les données d'entrée et la technique utilisée pour la génération des caractéristiques, la taille de la représentation obtenue sous forme d'un vecteur de caractéristiques peut être volumineuse. La taille du descripteur affecte directement les performances du système de reconnaissance automatique en termes de précision et de rapidité. En effet, la représentation initiale peut contenir des attributs redondants et d'autres qui peuvent être perçus comme du bruit. Afin de remédier à cette contrainte, il est possible d'ajouter un bloc *optionnel* au système afin de se débarrasser de ces attributs. Nous distinguons deux types de techniques qui peuvent être utilisés afin de réduire la taille du vecteur de caractéristiques :

1. **Score :** où une technique est exploitée afin d'accorder à chaque attribut un certain score qui dépend de critères spécifiques. Ensuite, les attributs sont classés par ordre décroissant suivant le score attribué. Afin de sélectionner un certain nombre d'attributs, nous devons définir, généralement de façon *empirique*, une valeur de *seuillage*. Il y a de nombreuses méthodes qui peuvent être exploitées. Parmi lesquelles, nous pouvons citer celle proposée dans certains de nos travaux et qui se base sur l'utilisation de la *variance* comme critère [61, 63]. Il est aussi possible d'utiliser des techniques d'apprentissage *supervisé* tel quel les *arbres extrêmement aléatoires* [66, 67].

2. **Transformation :** où le vecteur de caractéristiques est complètement modifié et l'information est réarrangée. Le principe consiste en l'application d'une projection de la représentation initiale afin d'en générer une nouvelle où l'information est réarrangée afin de faciliter son utilisation. Dans le cadre de cette thèse, nous avons pu comparer les performances de deux techniques qui sont très utilisées à savoir l'*analyse en composantes principales* et *indépendantes* [65]. Il est aussi possible d'exploiter une des nouvelles techniques d'*apprentissage profond* à savoir les *auto-encodeurs* [22] qui permettent entre autres de réduire la dimension de la représentation initiale.

Modèle de Classification

Le dernier bloc d'un système de reconnaissance automatique est de loin le plus critique. C'est celui qui est chargé de la tâche de *reconnaissance*. Dans le contexte de ce travail de recherche, nous avons utilisé différentes techniques qui peuvent être classées en deux catégories distinctes selon l'apprentissage [60, 15]

1. **non-Supervisé :** c'est une technique d'intelligence artificielle et plus précisément un problème d'*apprentissage automatique*. Il permet de partitionner les échantillons dans un certain nombre de segments (ou *cluster*). Cette opération est effectuée sur des échantillons non étiquetés. La technique la plus commune et qui est utilisée dans le cadre de cette thèse reste le *k-moyennes* (voir Chapitre 5). L'apprentissage non-supervisé permet aussi l'estimation de densité de distribution ainsi que la réduction de densité en exploitant l'*analyse en composantes principales* (voir Chapitre 3).
2. **Supervisé :** par contraste aux techniques *non-supervisé*, les algorithmes *supervisés* ont comme objectif la *classification* des échantillons suivant un certain nombre de *classes* prédéfinies. Ces algorithmes opèrent sur des échantillons étiquetés sur lesquels est réalisée une *phase d'entrainement* afin de générer un *modèle*. En exploitant le modèle généré, il est possible de reconnaître et de classifier les échantillons non-étiquetés. Dans le cadre de ce travail de recherche, nous avons utilisé diverses techniques parmi lesquelles : *machine à vecteurs de support* [61, 63, 67, 66], *perceptron multi-couche* [61], *k plus proches voisins* [61, 63], *arbres de décision* [61] et *forêt d'arbres décisionnels* [68].

Nous avons également utilisé, dans le cadre de cette thèse, un autre type d'algorithme qui est d'actualité et qui consiste en l'*apprentissage profond*. Nous nous sommes focalisés sur une technique en particulier ; le *réseau de neurones à convolution* avec une architecture *optimisée* afin de développer une méthode de reconnaissance automatique des expressions faciales [68, 64]. Le principal avantage de ce genre de techniques réside dans le fait que les représentations sont générées de façon automatique et non prédéfinies par le concepteur.

Sortie

C'est le dernier bloc du système. Après avoir appliqué l'algorithme ou modèle d'apprentissage pour l'identification de l'émotion à partir de l'expression faciale, la sortie est composée d'un certain nombre de classes nominales. Se basant sur les travaux d'Ekman et Friesen [17], la plupart des méthodes existantes permettent de reconnaître de façon automatique les *six émotions de base* à savoir : *peur* (FE), *surprise* (SU), *joie* (HA), *dégoût* (DI), *colère* (AN), *tristesse* (SA) et l'*état neutre* (NE). Ainsi, dans le cas des systèmes *statiques*, nous dénombrons *sept* classes de sortie. Quant aux systèmes *dynamiques*, ils disposent de *six* classes de sortie. Le nombre de classes dépend aussi de la base de données d'évaluation utilisée (se référer aux Tables 1.1 et 1.2).

1.3.2 Environnement intelligent et assistance ambiante

L'objectif principal de ce travail de recherche réside dans la conception de système de reconnaissance automatique des émotions à travers les expressions faciales. Néanmoins, le champ d'application des approches proposées demeure l'*assistance ambiante*. Ainsi et afin de valider l'utilisation de ces approches, nous avons mis en place un protocole d'expérimentation afin de récolter des données à partir d'un environnement intelligent.

Dans le contexte de notre travail, les environnements intelligents peuvent être perçus comme des laboratoires d'expérimentation ayant les mêmes caractéristiques que les *maisons intelligentes*. En effet, ils disposent de trois éléments primordiaux à savoir :

1. **Capteurs :** ce sont les dispositifs d'*entrée* qui permettent de récolter les données d'*interaction* entre l'utilisateur et l'environnement intelligent. Il y a différents types de données qui peuvent être collectées [48], à savoir : *1) catégoriques* qui sont issues de capteurs binaires et **RFID**, *2) séries temporelles* issues de capteurs de température, luminosité, humidité et inertie, *3) signaux analogiques* qui peuvent provenir de capteurs de signe vitaux, *4) audio* qui

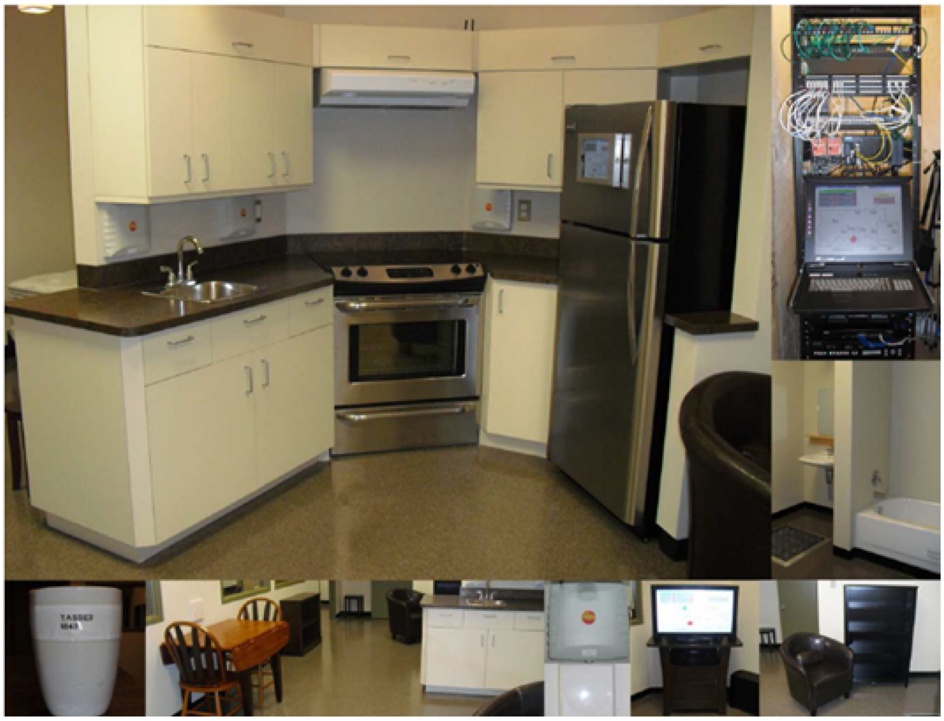


FIGURE 1.5 – Aperçu de l'environnement intelligent **LIARA**.

permettent la détection de la voix et autres sons dans l'environnement intelligent, *5) image et vidéo* proviennent de caméras numériques permettant la reconnaissance d'actions.

2. **Unité de traitement :** une fois que les données ont été collectées par les différents capteurs, elles sont envoyées à un autre dispositif afin de les traiter. Ce dernier est généralement un serveur qui permet d'extraire de l'information utile à partir des données brutes fournies par les différents capteurs. Pour cela, il exploite différents algorithmes d'apprentissage automatique. Il y a, généralement, deux types distincts de tâches qui sont exécutées à savoir : *1) reconnaissance d'activité* qui permet de reconnaître les tâches de la vie quotidienne qui sont exécutées par l'utilisateur et à l'aide desquelles il est possible d'apprendre les habitudes de l'utilisateur afin de l'assister plus efficacement, *2) détections d'erreurs* qui permet de détecter de façon automatique les anomalies durant l'exécution des tâches de la vie quotidienne.

3. **Actionneurs :** ce sont des dispositifs qui permettent de fournir l'assistance suivant ce qu'ils reçoivent de l'*unité de traitement*. De plus, il y a plusieurs

manières de fournir une assistance à l'utilisateur, ça peut se faire via des : *1) messages vocaux* qui sont émis en utilisant des haut-parleurs placés dans l'environnement intelligent à des endroits stratégiques, *2) séquences vidéo* qui permettent d'illustrer le message d'assistance à travers des vidéos/images sur un écran, *3) séquences audiovisuelle* qui incluent des vidéos/images et des messages vocaux afin d'augmenter la clarté du message d'assistance, *4) envoi de notifications* qui permet de contacter via un appel ou message texte le personnel adéquat afin de fournir une assistance, surtout en cas d'urgence.

Il existe de nombreux environnements intelligents où des expérimentations sont régulièrement conduites afin de collecter des données utiles. Parmi ces nombreux environnements, nous pouvons citer **CASAS** (*Center for Advanced Studies in Adaptive Systems*) [9] qui est destiné à fournir une assistance aux personnes souffrant de démence. Il est équipé de *soixante-dix* détecteurs de mouvement et des dispositifs permettant d'alerter les infirmiers en cas de problème. Ici au Québec, nous sommes en mesure de citer deux différents environnements intelligents à savoir le **DOMUS** (*laboratoire de DOMotique et informatique mobile l'Université de Sherbrooke*) [5] et le **LIARA** [20] qui disposent de pratiquement la même architecture et qui utilisent principalement la technologie **RFID**. En effet, le **DOMUS** exploite plus de *deux-cents* capteurs parmi lesquels *vingt* antennes **RFID**, différents capteurs infrarouges et détecteurs de mouvements. Quant au **LIARA**, il exploite aussi principalement la technologie **RFID** mais aussi des capteurs électromagnétiques, de force, ultrasoniques, etc. Ces capteurs sont répartis entre les différentes pièces de l'environnement intelligent à savoir : le séjour, la salle de bain, la cuisine et chambre à coucher.

D'autres recherches ont été menées pour la conception de systèmes d'assistances qui seront intégrés dans les environnements intelligents. Ainsi, Mihailidis et al. [47] ont proposé un système appelé **COACH** (*Cognitive Orthosis for Assisting aCtivities in the Home*) permettant d'assister les personnes souffrant de déficience cognitive lors de l'activité de *lavage des mains*. D'autres comme Peters et al. [52] ont proposé un système permettant de fournir une assistance lors de l'activité de *brossage des dents*.

1.4 Contributions de la thèse

Le présent travail de recherche a pour objectif principal de proposer un système d'assistance ambiante efficace. Il est destiné à être intégré dans un environnement intelligent afin de fournir une assistance adéquate aux personnes âgées, notamment celles qui souffrent de déficience cognitive. Suivant l'architecture des systèmes d'assistance, il est composé de modules de reconnaissance d'activité et de détection d'erreurs. Le système proposé est basé sur l'utilisation de la technologie **RFID**. Cependant, la spécificité de ce travail de recherche réside dans l'utilisation des expressions faciales afin d'améliorer les performances de la détection automatique d'erreurs. Pour cette raison, les contributions apportées peuvent être décrites suivant deux axes distincts, mais dépendants à savoir : la reconnaissance automatique d'expressions faciales et l'assistance ambiante.

1.4.1 Reconnaissance automatique d'expressions faciales

Depuis l'avènement des technologies de l'information et de la communication, les chercheurs se sont toujours intéressés à la reconnaissance automatique des émotions. Même si le domaine commence à dater, il reste néanmoins d'actualité surtout avec l'apparition des nouveaux algorithmes d'apprentissage profond [36, 3]. Dans ce contexte, nous avons proposé dans le cadre de cette thèse différentes méthodes ayant chacune des spécificités qui lui sont propres.

Nous nous sommes principalement concentrés sur le traitement de données d'entrée *statique* pour la facilité d'implémentation dans des systèmes embarqués avec ressources matérielles limitées. Ainsi, nous avons commencé par proposer diverses méthodes *statiques* utilisant différents types de descripteurs : *géométriques* [61], *apparence* [65] et *hybride* [66]. Toutes ces méthodes se basent sur l'utilisation de techniques d'apprentissage automatique classiques.

En premier lieu, nous nous sommes intéressés aux descripteurs *géométriques*. L'intérêt de ce type de descripteurs est qu'ils sont insensibles aux effets de contrastes et de luminosité. Pour cela, de nombreux chercheurs ont proposé des méthodes basées sur l'utilisation de l'**ASM** et de l'**AAM**. D'autres chercheurs ont utilisé des techniques pour l'extraction des points caractéristiques du visage. Ensuite, ils ont calculé des distances spécifiques en s'inspirant du **FACS** proposé par Ekman [18]. Dans notre cas, nous avons proposé d'utiliser la méthode qui se base sur la technique de Kazemi et Sullivan [31] pour l'extraction des *soixante-huit* points caractéristiques du visage. Ensuite, toutes les distances euclidiennes possibles sont calculées avant d'utiliser une méthode de sélection des attributs pour définir le vecteur de caractéristiques le plus optimisé. Deux méthodes de sélection des attributs ont été utilisées, l'une est basée sur la *variance* [61] et l'autre sur les *arbres extrêmement aléatoires* [67]. En ce qui concerne la partie reconnaissance, nous avons fait appel à deux différentes techniques d'apprentissage automatique : *k plus proches voisins* [61] et *machine à vecteurs de support* [67]. Le principal avantage des méthodes proposées réside dans le fait que les distances calculées ne sont pas choisies *manuellement* mais de façon automatique à l'aide de critères de sélection spécifiques.

Parmi les principales limites des caractéristiques *géométriques* figure l'étroite dépendance aux techniques d'extraction des points caractéristiques du visage. En effet, ces techniques offrent de bonnes performances avec des images frontales, mais restent très sensibles à l'échelle et aux changements d'angle. Les descripteurs de types *apparence* ne souffrent pas de ces limitations. Ils se basent sur les caractéristiques et opèrent sur l'ensemble des pixels de l'image. Dans le cadre de cette thèse, nous nous sommes intéressés à l'utilisation d'une version *étendue* des *motifs binaires locaux* [65]. C'est un descripteur proposé par Ojala et al. [50] et qui a été utilisé de manière efficace pour la reconnaissance faciale. La méthode proposée commence par extraire les *cinq* sous-régions spécifiques du visage (les yeux, le nez et la bouche). Ensuite, la représentation basée sur les *motifs binaires locaux* est générée à partir de chacune des sous-régions. Afin d'avoir un seul et unique vecteur de caractéristiques, les différentes représentations sont concaténées. L'une des contributions de cette méthode réside dans la comparaison entre deux techniques de sélection des attributs (par *transformation*) : l'*analyse en composantes principales* et *indépendantes*. En ce qui concerne la partie reconnaissance, elle est réalisée en utilisant un classifieur *multiclasse* basé sur les *machine à vecteurs de support*.

Après avoir évalué les deux méthodes avec les deux types des descripteurs, nous avons constaté que les informations fournies par chacun d'entre eux sont différentes, mais complémentaires. Pour cela, nous avons proposé une méthode *hybride* combinant les deux types de descripteurs précédents [66]. En ce qui concerne la représentation *géométrique*, nous avons repris le travail que nous avons fait précédemment [61, 67]. Par contre, pour le descripteur de types *apparence*, nous avons utilisé les coefficients générés par la *transformée en ondelettes*. Afin de réduire la taille des deux représentations résultantes, nous avons utilisé la technique des *arbres extrêmement aléatoires*. La fusion des deux représentations est réalisée en *aval*, en combinant les sorties de deux classifieurs *multiclasse* basé sur les *machine à vecteurs de support*. Comme prédit, les résultats obtenus en combinant les deux types de descripteurs sont meilleurs que ceux obtenus par chacun d'eux.

Avec l'apparition, relativement récente, des nouvelles techniques d'apprentissage automatique, il est possible de générer de façon automatique des représentations pertinentes à partir de données brutes. Contrairement aux approches classiques où il est nécessaire de définir au préalable le type de descripteur à utiliser, les méthodes basées sur les techniques d'*apprentissage profond* permettent d'extraire et de sélectionner les informations utiles de façon automatique. Dans le cadre de ce travail de recherche, nous avons proposée une nouvelle architecture de *réseau de neurones à convolution* [64] inspirée de **LeNet-5** proposée par LeCun et al. [38]. L'architecture proposée est non seulement *optimisée* pour la reconnaissance des expressions faciales, mais elle est *allégée* afin de faciliter son implémentation sur des systèmes embarqués limités en ressources matérielles. Afin d'assurer les bonnes performances de l'architecture, nous avons aussi inclue des opérations de pré-traitement. Les résultats obtenus sont bien meilleurs que ceux des méthodes précédentes confirmant ainsi l'efficacité des techniques d'*apprentissage profond*.

Toutes les méthodes proposées et décrites ci-dessus opèrent sur des données d'entrée *statique*. La raison est due au fait que ces méthodes sont destinées à être implémentées sur des systèmes embarqués avec des ressources matérielles limitées. Cependant, nous nous sommes quand même intéressés aux méthodes *dynamiques* traitant des vidéos et séquences d'images. Ainsi, nous avons proposé une méthode qui permet d'extraire une représentation *spatio-temporelle* efficace à base de descripteurs *géométriques*. L'un des principaux inconvénients des méthodes *dynamiques* réside dans la nécessité de *normalisation* des séquences afin qu'elles aient le même nombre d'images. La méthode que nous avons proposée permet de remédier à ce problème, en générant une représentation de même taille peu importe la séquence d'entrée. L'autre avantage apporté réside dans la réduction de la complexité, car non seulement le vecteur de caractéristique initial est réduit, mais nous avons aussi appliqué des méthodes de sélection des attributs afin de réduire encore sa taille. Dans le travail [63], nous avons utilisé une méthode basée sur la *variance* [61]. Alors que dans le travail [62], nous avons utilisé les *arbres extrêmement aléatoires*. La partie reconnaissance a été réalisée en utilisant un classifieur *multiclasse* basé sur les *machines à vecteurs de support*.

1.4.2 Contribution à l'assistance ambiante

Ce travail de recherche s'inscrit principalement dans la thématique de l'*assistance ambiante*. Ces dernières décennies, le monde a connu une augmentation significative de la population âgée et ainsi une augmentation majeure des cas de maladies dégénératives telle qu'Alzheimer. La principale conséquence de ces maladies est la *déficience cognitive* qui cause entre autres des pertes de mémoire. Ainsi, les sujets atteints sont dans l'incapacité d'exécuter les tâches les plus basiques de la vie quotidienne d'où la nécessité d'une constante assistance.

Au sein de notre laboratoire le **LIARA**, différents travaux ont été réalisés pour fournir des systèmes d'*reconnaissance* à intégrer dans un environnement intelligent. Ainsi, Fortin-Simard et al. [20] ont proposé un système pour la localisation des objets (ayant des tags **RFID**) dans l'environnement intelligent en se servant de la force du signal reçu par les antennes. Le système proposé permet la reconnaissance automatique des activités en cours de réalisation ainsi que la détection automatique des erreurs en se servant uniquement des données de localisation. Belley et al. [2] ont aussi proposé un autre système basé sur la technologie **RFID**, mais la particularité de ce système réside dans le fait qu'il se focalise sur une tâche particulière et qui est la préparation du petit déjeuner.

Dans le cadre de cette thèse, nous avons présenté différentes méthodes de *reconnaissance automatique des expressions faciales*. Un des challenges que nous nous sommes fixés réside dans le fait de trouver un moyen d'exploiter ces méthodes dans le cadre de l'*assistance ambiante*. Ainsi, nous avons proposé un système permettant de fournir une assistance pertinente lors de l'exécution d'une tâche spécifique de la vie quotidienne à savoir la *préparation du café* [68]. Nous avons utilisé la méthode proposée par Bilodeau et al. [4] pour la localisation des objets au sein de l'environnement intelligent et qui est basée sur la technologie **RFID**. À partir de la position des différents objets, notre système permet de reconnaître les différentes actions de bases durant l'activité. Les données d'entrée sont sous forme de séquences de coordonnées cartésiennes qui représentent la variation des positions des objets au cours du temps. Tout comme pour la reconnaissance automatique des expressions faciales à partir de données *dynamiques*, nous avons été confrontés au problème des longueurs différentes des séquences. Par conséquent, nous avons appliqué la même technique que précédemment [62] afin d'avoir des vecteurs de caractéristiques de même taille. À partir des descripteurs obtenus, nous avons été en mesure de reconnaître des actions utilisateurs durant l'activité de la vie quotidienne. En se basant sur l'analyse des actions reconnues, il est possible de détecter la présence d'erreurs commises. Après évaluation du système proposé à l'aide des données récoltées lors des expérimentations aux **LIARA**, les résultats obtenus confirment l'apport des expressions faciales à l'amélioration de la détection d'erreurs [68]. En effet, le taux de fausse détection s'est vu réduit de plus de **20**%, ce qui est considérable.

1.4.3 Conférences Internationales

Dans le cadre de cette thèse, nous avons été amenés à présenter notre avancement ainsi que nos méthodes proposées lors de différentes *conférences internationales* traitant de différents domaines tels que la reconnaissance de forme et les systèmes intelligents. Ci-dessous une liste de nos contributions que que nous avons soumises et présentées lors de manifestations de vulgarisation scientifique :

Y. Yaddaden, A. Bouzouane, M. Adda, S. Gaboury, B. Bouchard, "A new approach of facial expression recognition for ambient assisted living," in *Proceedings of the 9th ACM international conference on PErvasive Technologies Related to Assistive environments (PETRA'16)*, ACM, 2016, pp. 14 :1-14 :8. (Statut : Présenté)

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "Facial expression recognition from video using geometric features," in *Proceedings of the 8th International Conference on Pattern Recognition Systems*, IET, 2017, pp. 1-6. (Statut : Présenté)

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "Facial expressions based error detection for smart environment using deep learning," in *Ubiquitous Intelligence Computing (UIC)*, IEEE, 2017, pp. 1-7. (Statut : Présenté)

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "One-Class and Bi-Class SVM Classifier Comparison for Automatic Facial Expression Recognition," in *International Conference on Applied Smart Systems (ICASS'18)*, IEEE, 2018, pp. 1-6. (Statut : Présenté)

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "Facial Sub-regions for Automatic Emotion Recognition using Local Binary Patterns," in *International Conference on Signal, Image, Vision and their Applications (SIVA'18)*, IEEE, 2018, pp. 1-6. (Statut : Présenté)

1.4.4 Journaux Internationaux

Nous avons également étendu, affiné et approfondi nos méthodes en effectuant des expérimentations supplémentaires afin de soumettre nos travaux à des revues scientifiques internationales ayant une réputation éprouvée. Ci-dessous une liste de papiers publiés ou en cours de traitement :

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "User action and facial expression recognition for error detection system in an ambient assisted environment," in *Journal of Expert Systems with Applications*, Elsevier, vol. 112, 2018, pp. 173-189. (Statut : Publié)

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "An Automatic Facial Expression Recognition Approach using an Efficient Spatio-Temporal Representation," in *Journal of Ambient Intelligence and Humanized Computing*, Springer, 2018, pp. 1-26. (Statut : Soumis)

1.5 Méthodologie de recherche

Le contexte global de cette thèse est l'*assistance ambiante*. Par conséquent, l'objectif principal est le développement d'un système permettant de fournir une assistance efficace destinée aux personnes âgées et souffrant de déficience cognitive. Comme décrit précédemment, un système d'assistance est composé de deux principaux modules réalisant la *reconnaissance d'activité* et la *détection d'erreurs*. La contribution principale de ce travail réside dans l'apport des *expressions faciales* à l'amélioration de la détection d'erreurs. Ainsi, nous avons suivi une méthodologie spécifique dans la conduite de ce projet qui se focalise de façon générale sur l'*assistance ambiante* et de façon plus spécifique à la *reconnaissance automatique des expressions faciales*.

En premier lieu, nous nous sommes intéressés au domaine des émotions où les expressions faciales représentent la source d'information la plus pertinente [46]. Nous avons proposé différentes méthodes pour la reconnaissance automatique des émotions à travers les expressions faciales. La majorité sont *statiques* en raison de leur faible complexité et la facilité d'implémentation dans des systèmes embarqués avec des ressources matérielles limitées. Nous avons également travaillé sur le développement de méthodes *dynamique* avec une représentation *spatio-temporelle* efficace. Avec l'avènement des nouvelles techniques d'apprentissage profond, nous avons proposé une méthode basée sur l'utilisation des *réseaux de neurones à convolution*. L'ensemble des méthodes proposées a été évalué avec différentes bases de données *benchmark* et suivant un protocole d'évaluation spécifique. Dans ce qui suit seront détaillés les différentes bases de données utilisées ainsi que le protocole d'évaluation adopté.

Après avoir confirmé l'efficacité des méthodes proposées, nous nous sommes focalisés sur l'*assistance ambiante* en proposant un système permettant de fournir une assistance adéquate et mettant à contribution les expressions faciales. Le système développé inclut deux composants principaux réalisant la *reconnaissance d'activité* et la *détection d'erreurs*. Le premier module est basé sur l'utilisation de la technologie **RFID** et le deuxième se base sur l'analyse des actions identifiées. La contribution dans le cadre de ce système réside dans la contribution des expressions faciales pour améliorer les performances de la détection d'erreurs. Afin de valider le système proposé, nous avons mené des expérimentations au sein du **LIARA** afin de récolter les données nécessaires. Dans ce qui suit est détaillé le protocole d'expérimentation adopté.

1.5.1 Bases de données publiques

La reconnaissance automatique des expressions faciales est un domaine de recherche qui est d'actualité, surtout avec l'avènement des nouvelles technologies de l'information et de la communication. Néanmoins, c'est un domaine qui ne date pas d'hier. En effet, la communauté scientifique s'est toujours intéressée à tout ce qui touche de près ou de loin à l'être humain, et surtout, à ce qui permettrait de comprendre, voire prédire son comportement. Par conséquent, diverses bases de données ont été constituées par différents laboratoires, et la plupart sont accessibles au grand public pour l'évaluation de nouvelles méthodes.

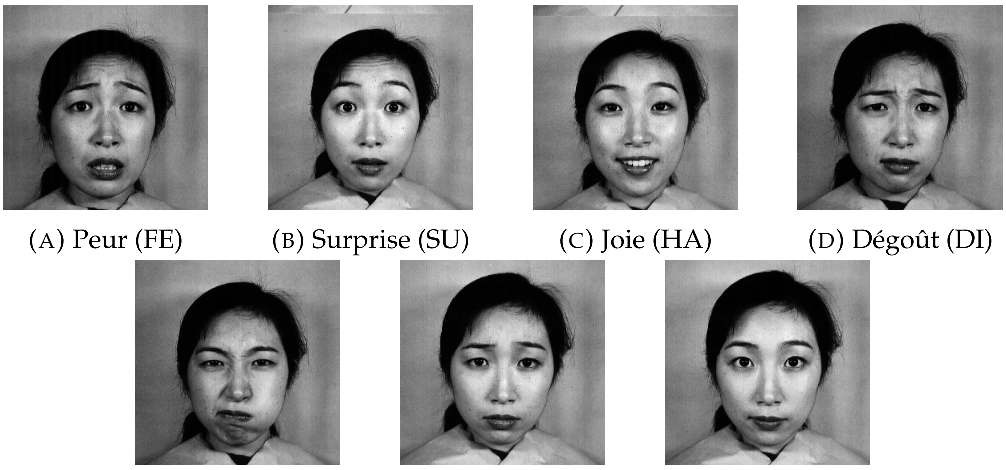
Dans le cadre de nos travaux de recherche, nous nous sommes basés sur diverses bases de données afin d'évaluer les différentes méthodes proposées. Les bases de données que nous avons a utilisé sont reparties suivant deux catégories distinctes dépendant du type de système (*statique* et *dynamique*).

Base de données *statique*

La majorité des méthodes proposées dans le cadre de cette thèse exploitent de simples images *statiques* comme données d'entrée. Ainsi, dans le cadre de ce travail de recherche, nous avons utilisé trois bases de données *statiques* becnhmark dont les caractéristiques sont détaillées dans la Table 1.1.

TABLE 1.1 – Les bases de données de type *statique*.

Base de Données		**JAFFE**	**KDEF**	**RaFD**
Émotions	FE	30	140	67
	SU	30	140	67
	HA	29	140	67
	DI	28	140	67
	AN	30	140	67
	SA	30	140	67
	NE	30	140	67
	Total	**207**	**980**	**469**
Résolution		256 × 256	562 × 762	681 × 1024



(A) Peur (FE) (B) Surprise (SU) (C) Joie (HA) (D) Dégoût (DI)

(E) Colère (AN) (F) Tristesse (SA) (G) Neutre (NE)

FIGURE 1.6 – Échantillon d'images de base de données **JAFFE**.

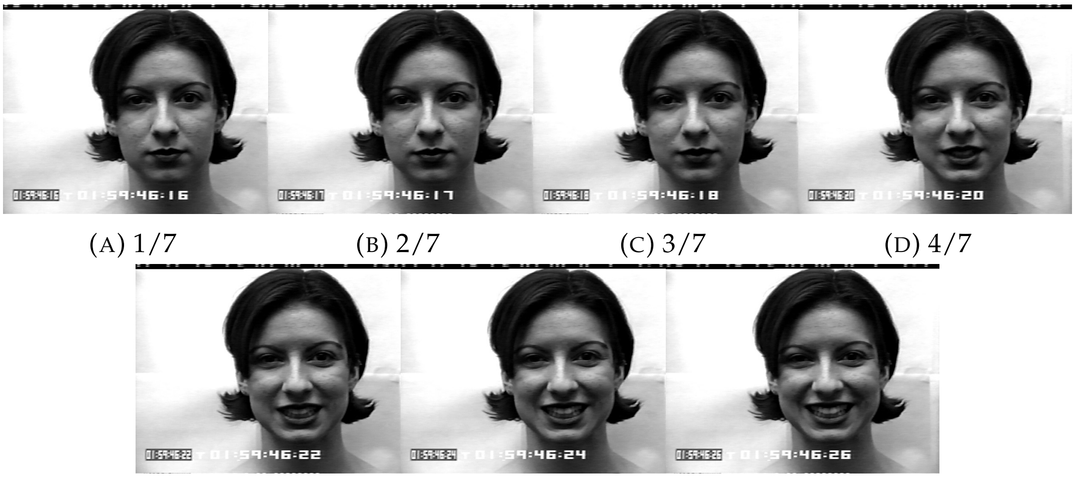
1. **JAFFE** ou **JA**panese **F**emale **F**acial **E**xpression [45], elle a été constituée par Michael Lyons, Miyuki Kamachi et Jiro Gyoba. Chacune des images qui composent la base de données représente une femme d'origine japonaise exprimant une des six émotions de base (ou l'état neutre). Au total, les auteurs ont fait appel à *dix* participantes.

2. **RaFD** ou **R**adboud **F**aces **D**atabase [35], elle a été constituée en faisant appel à *soixante-sept* participants de différentes origines (caucasien et Marocaine hollandaise), genre (homme et femme), tranche d'âge (adulte et enfant). Chacune des images de la base de données représente le visage d'un individu exprimant une des six émotions de base en plus de l'état neutre et du *mépris* (CO).

3. **KDEF** ou **K**arolinska **D**irected **E**motional **F**aces [44], elle a été constituée par Daniel Lundqvist, Anders Flykt et Arne Öhman. *Soixante-dix* individus ont participé aux expérimentations afin de collecter les images de la base de données. Chacune des images représente un sujet/individu exprimant une des six émotions de base en plus de l'état neutre.

Base de données *dynamique*

Après avoir proposé diverses méthodes *statiques* de reconnaissance automatique des émotions à travers les expressions faciales, nous nous sommes intéressés à l'utilisation de vidéos ou séquences d'image. Ainsi, nous avons travaillé sur des méthodes dites *dynamiques* [63]. Tout comme pour les méthodes *statiques*, nous avons utilisé trois bases de données *dynamiques*. Les caractéristiques de ces bases de données *benchmark* sont détaillées dans la Table 1.2.



(A) 1/7 (B) 2/7 (C) 3/7 (D) 4/7

(E) 5/7 (F) 6/7 (G) 7/7

FIGURE 1.7 – Échantillon de séquence (*joie*) de base de données **CK+**.

1. **CK+** ou **C**ohn-**K**anade *extended* [30], elle a été constituée à l'aide de *cent vingt-trois* participants avec différentes tranches d'âge (18 à 50 ans), genre (69% femmes et 31% hommes) et origine ethnique (81% euroaméricains, 13% afro-américains et 6% d'autres groupes éthiques). Dans chacune des séquences, le participant exprime une des six émotions de base en plus du *mépris* (CO).
2. **MMI** ou **M**an-**M**achine **I**nteraction [51], cette base de données a été constituée à l'aide de *soixante-quinze* participants. Dans chacune des séquences est représenté le visage du participant exprimant une des six émotions de base.
3. **MUG** ou **M**ultimedia **U**nderstanding **G**roup [1], elle a été constituée à l'aide de *quatre-vingt-six* participants de différents genres (35 femmes et 51 hommes) et tranches d'âge (20 à 35 ans). Dans chacune des séquences, le participant exprime une des six émotions de base.

TABLE 1.2 – Les bases de données de type *dynamique*.

Base de Données		**CK+**	**MMI**	**MUG**
Émotions	FE	25	28	127
	SU	83	38	173
	HA	69	42	175
	DI	59	31	153
	AN	45	28	167
	SA	28	32	136
	CO	18	–	–
	Total	**327**	**199**	**931**
Résolution		640 × 490	768 × 576	896 × 896
Transitions		*Début-Sommet*	*Début-Sommet-Fin*	*Début-Sommet-Fin*
Images/Séquence		6 *à* 71	30 *à* 243	11 *à* 179

1.5.2 Environnement d'expérimentation

L'un des objectifs de notre travail de recherche est de mettre en application et d'exploiter les différentes méthodes de reconnaissance automatique des expressions faciales dans le cadre de l'*assistance ambiante*. Par conséquent, il est nécessaire de conduire des expérimentations au sein d'un environnement intelligent. Ainsi, notre choix s'est naturellement porté sur le **LIARA**.

Comme décrit précédemment, le **LIARA** est un environnement intelligent qui reproduit les mêmes conditions qu'une maison intelligente moderne. Il est équipé de différents types de *capteurs*, *actionneurs* et une *unité de traitement* pour la prise de décision. Dans le cadre de nos expérimentations, nous nous sommes focalisés sur l'utilisation de la technologie **RFID** [68]. Les différents objets qui se trouvent dans la cuisine du **LIARA** sont équipés de tags **RFID** passifs et il y a *quatre* différentes antennes permettant de recevoir le signal à partir des tags. Afin de reconnaître à tout moment la localisation des différents objets, nous avons utilisé la méthode proposée par Bilodeau et al. [20] qui exploite la force du signal.

1.5.3 Collecte de données

Dans la littérature, il n'existe pas de bases de données combinant des données **RFID** permettant l'identification des actions utilisateur dans le cadre de la reconnaissance d'activité et celle permettant la reconnaissance des expressions faciales. Par conséquent, les expérimentations que nous avons conduites au sein du **LIARA** ont pour principal objectif de collecter la base de données permettant la validation du système proposé [68].

Nous avons défini un protocole d'expérimentation qui se déroule dans la cuisine du **LIARA** et la tâche de la vie quotidienne choisie est la *préparation du café*. L'ensemble des individus qui ont participé aux expérimentations sont des étudiants et membres du **LIARA**. En réalisant la tâche demandée (préparation du café), un dispositif avec une caméra a été utilisé afin de récolter les changements d'expression faciale tout au cours de l'activité. Les données concernant les changements des positions de différents objets sont téléchargées à partir du serveur du **LIARA** où la méthode de Bilodeau et al. [20] est appliquée.

En résumé, nous nous retrouvons avec une base de données subdivisée en deux parties distinctes. La première est dédiée à l'analyse des expressions faciales (vidéos ou séquences d'image) et la seconde est dédiée à la reconnaissance des actions de base de l'activité de la vie quotidienne (coordonnées cartésiennes). Cette base de données a été exploitée dans le cadre de la dernière partie de la thèse afin de définir la contribution des expressions faciales à l'*assistance ambiante* (voir Chapitre 6).

1.5.4 Protocole d'évaluation

Les différentes méthodes proposées et présentées dans le cadre de ce travail de recherche incluent l'utilisation d'algorithmes de *classification*. Comme définit plus haut, ce sont des techniques d'*apprentissage supervisées* qui nécessitent une phase d'entrainement avec des données connues afin de générer un *modèle*. Ce dernier permettra d'effectuer de la reconnaissance automatique.

Les différentes méthodes de reconnaissance proposées, que ça soit dans le cadre de la reconnaissance automatique des expressions faciales ou l'assistance ambiante, nécessitent une évaluation afin de vérifier leur performances. Dans la littérature, il existe différentes méthodes d'évaluation, mais dans le cadre de ce travail de recherche, nous nous focaliserons sur l'utilisation de *validation croisée*. Afin d'utiliser cette méthode, il est nécessaire de définir le paramètre k qui définit le nombre de *sous-échantillons* ou *sous-ensembles* (se référer à la Figure 1.8).

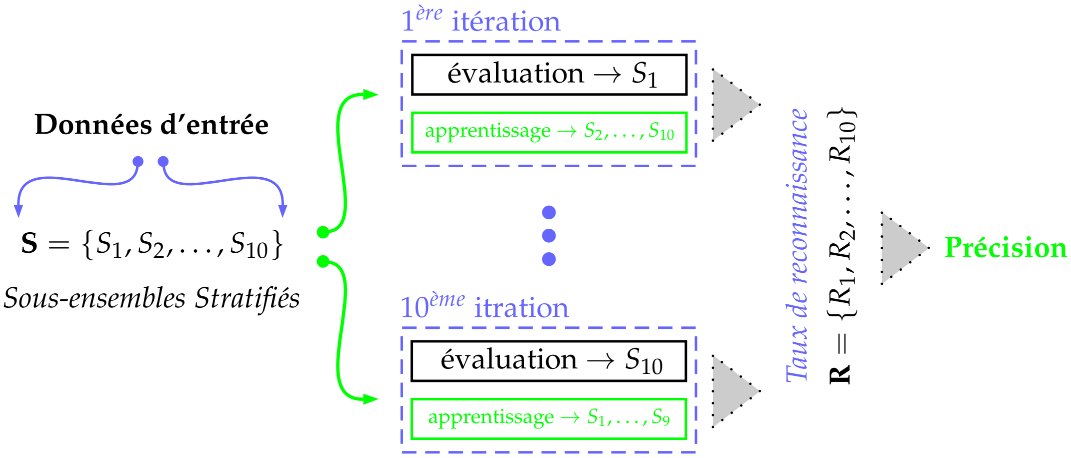

$1^{ère}$ itération
évaluation → S_1
apprentissage → $S_2, \ldots, S_{10}$
Données d'entrée
$\mathbf{S} = \{S_1, S_2, \ldots, S_{10}\}$
Sous-ensembles Stratifiés
$10^{ème}$ itration
évaluation → S_{10}
apprentissage → $S_1, \ldots, S_9$
Taux de reconnaissance
$\mathbf{R} = \{R_1, R_2, \ldots, R_{10}\}$
Précision

FIGURE 1.8 – Description de la *protocole d'évaluation* adopté.

Comme illustré dans la Figure 1.8, il y a plusieurs étapes à suivre pour la réalisation de la validation croisée. Ainsi, durant la première étape, la base de données d'entrée est subdivisée est $k = 10$ sous-ensembles *stratifiés* $\mathbf{S} = \{S_1, S_2, \ldots, S_{10}\}$. L'opération d'évaluation est réalisée en dix itérations distinctes. Chacune d'elle traite un sous-ensemble S_i avec $i = 1, 2, \ldots, k = 10$. Durant chaque itération, un sous-ensemble S_i est utilisé pour l'évaluation et le reste $S_{j \neq i}$ est dédié à l'apprentissage. À la fin de chaque itération, nous obtenons un taux de reconnaissance R_i. Afin de calculer le taux de reconnaissance finale, il faut appliquer la formule 1.1.

$$\mathbf{R} = \frac{\sum_{i=1}^{i=10} R_i}{k} \tag{1.1}$$

R représente la performance du modèle en termes de reconnaissance calculé à partir de la moyenne des différents taux de reconnaissance obtenus de chacune des différentes itérations. Nous avons utilisé cette méthode de validation dans l'ensemble des méthodes proposées dans le cadre de ce travail de recherche.

1.6 Organisation de la thèse

Durant la conduite de ce projet de recherche, nous avons eu l'occasion d'aborder différents aspects de la *reconnaissance automatique des expressions faciales* ainsi que de l'*assistance ambiante*. Le reste de la présente thèse s'organise en cinq parties distinctes. Chacune d'elle traite d'un élément clé et est présentée sous forme d'article de journal ou de conférence internationale. Dans ce qui suit, nous présenterons brièvement le contenu des différents chapitres.

Dans le Chapitre 2 nous abordons l'utilisation des caractéristiques *géométriques* associées à une technique d'apprentissage particulière qui consiste en la *classification mono-classe*. Ainsi, un système de reconnaissance automatique des expressions faciales *statiques* est proposé et validé avec trois bases de données *benchmark* (**JAFFE**, **RaFD** et **KDEF**). Une comparaison pertinente entre deux classfieurs (machine à vecteurs de support *mono-classe* et *bi-classes*) sera établie.

Le Chapitre 3 est dédié à l'exploitation de descripteurs d'*apparence*. À partir des données d'entrée sous forme d'image, des sous-régions spécifiques du visage sont extraites auxquelles est appliquée la technique *étendue* des *motifs binaires locaux*. Nous avons aussi comparé les performances de deux techniques de réduction de dimension à savoir l'*analyse en composantes principales* et *indépendantes*. La méthode proposée dans ce chapitre a été validée en utilisant trois bases de données *benchmark* (**JAFFE**, **RaFD** et **KDEF**).

Ayant travaillé avec deux types de représentations (*géométrique* et *apparence*) dans les deux chapitres précédents, nous nous sommes intéressés à leur combinaison. Ainsi, dans le Chapitre 4 est proposée une méthode *hybride* qui exploite les caractéristiques *géométriques* [61] combinées aux coefficients issus de la *transformée en ondelettes*. Après l'évaluation de la méthode en utilisant trois bases de données *benchmark* (**JAFFE**, **RaFD** et **KDEF**), il a été confirmé que la méthode *hybride* présente de meilleures performances.

Après avoir travaillé sur la conception de systèmes dit *statiques*, nous nous sommes intéressés à l'exploitation de vidéo ou séquence d'images comme données d'entrée. Ainsi, le Chapitre 5 est dédiée à la présentation d'une méthode *dynamique* qui exploite des caractéristiques *géométriques* afin de générer une représentation *spatio-temporelle*. L'évaluation de cette méthode s'est faite en exploitant des bases de données spécifiques contenant des séquences d'images (**CK+**, **MMI** et **MUG**).

Avec l'avènement des récentes techniques d'*apprentissage profond*, nous avons proposé dans Chapitre 6 une nouvelle architecture de *réseaux de neurones à convolution* optimisée pour la reconnaissance automatique des expressions faciales. Elle a été inspirée des travaux de LeCun et al. [38] avec l'architecture **LeNet-5**. Nous avons pu valider l'architecture proposée à l'aide de cinq bases de données *benchmark* (**JAFFE**, **RaFD**, **KDEF**, **CK+**, **MMI**). Dans le contexte de l'*assistance ambiante*, nous avons aussi proposé une méthode pour la reconnaissance automatique des actions de base d'une activité de la vie quotidienne en utilisation la technologie **RFID**. Nous avons aussi proposé une méthode pour la détection automatique des erreurs et anomalies en utilisant les expressions faciales. Afin de valider le système d'assistance proposé, nous avons conduit des expérimentations au sein même du **LIARA**.

Pour finir, dans la Conclusion Générale est présenté la synthèse des différentes contributions de ce travail de recherche. Nous exposerons également les différentes limites des méthodes et systèmes d'assistance proposés. Nous présenterons quelques perspectives et les travaux futurs que nous comptons entreprendre.

Chapitre 2

Titre :

Comparaison entre Classificateurs Machine à Vecteurs de Support Mono et Bi-classe pour la Reconnaissance Automatique des Expressions Faciales

Résumé - *Ce chapitre est consacré à la reconnaissance statique et automatique des expressions faciales en utilisant des caractéristiques géométriques. Ainsi, à partir de données statiques sous forme d'images, une méthode introduite par Kazemi et Sullivan [31] est appliquée afin d'extraire de façon automatique soixante-huit points caractéristiques du visage. Afin de générer une représentation pertinente, l'ensemble des distances euclidiennes possibles sont calculées à partir des différents points caractéristiques du visage. La taille de la représentation obtenue est réduite en utilisant une technique basée sur les arbres extrêmement aléatoires. Elle permet de sélectionner uniquement les attributs les plus pertinents. Pour la dernière étape de reconnaissance, nous avons évalué les performances de deux types de classifieurs à machine de vecteurs de support mono et bi-classe. Afin d'évaluer la méthode proposée, nous avons utilisé trois bases de données statiques :* ***JAFFE****,* ***KDEF*** *et* ***RaFD****. Les résultats obtenus attestent de l'efficacité des classifieurs à machine de vecteurs de support mono-classe, non seulement en termes de reconnaissance, mais aussi de rapidité.*

Mots clés :

Reconnaissance des Expressions Faciales, Machine à Vecteurs de Support Bi-classe, Machine à Vecteurs de Support Mono-classe, Caractéristiques Géométriques, Arbres Extrêmement Aléatoires

Contributions associées :

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "One-Class and Bi-Class SVM Classifier Comparison for Automatic Facial Expression Recognition," in *International Conference on Applied Smart Systems (ICASS'18)*, IEEE, 2018, pp. 1-6. (Statut : Présenté)

Y. Yaddaden, A. Bouzouane, M. Adda, S. Gaboury, B. Bouchard, "A new approach of facial expression recognition for ambient assisted living," in *Proceedings of the 9th ACM international conference on PErvasive Technologies Related to Assistive environments (PETRA'16)*, ACM, 2016, pp. 14 :1-14 :8. (Statut : Présenté)

2018 International Conference on Applied Smart Systems (ICASS'2018)
24-25 November 2018, Medea, ALGERIA

One-Class and Bi-Class SVM Classifier Comparison for Automatic Facial Expression Recognition

[1]Yacine Yaddaden, [1,2]Mehdi Adda, [1]Abdenour Bouzouane, [1]Sébastien Gaboury & [1]Bruno Bouchard

[1]Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA),
Département d'informatique et de mathématique,
Université du Québec à Chicoutimi (UQAC). Chicoutimi, Québec, Canada.
[2]Département de mathématiques, d'informatique et de génie,
Université du Québec à Rimouski (UQAR). Rimouski, Québec, Canada.

Email: {yacine.yaddaden1, abdenour_bouzouane, sebastien_gaboury}@uqac.ca, mehdi_adda@uqar.ca

Abstract—**Facial expressions might be seen as a relevant and useful source of information. Indeed, they allow understanding and even identifying people behavior based on the emotional changes. Therefore, automatic facial expression recognition has been widely solicited in the context of smart cities and homes. However, recognizing human emotion automatically through facial expressions remains challenging. Moreover, *multi-class* Support Vector Machine classifiers have been widely employed and in most cases, the proposed architectures are based on the use of *bi-class* classifiers. In this paper, we propose an approach that exploits selected *geometric-based* features using the Extremely Randomized Trees method while the recognition is handled by three distinct *multi-class* Support Vector Machine architectures namely *bi-class* (*One-against-One* and *One-against-All*) and *one-class* classifiers. We also investigate the performance of the three different architectures by performing a comparison in terms of accuracy and computation time. The carried experiment on three benchmark datasets attests to the efficiency of the *one-class* classifier since the proposed approach yields 92.68%, 85.83% and 93.33% with the JAFFE, RaFD and KDEF datasets, respectively.**

Index Terms—**Facial Expression Recognition, *Bi-Class* Support Vector Machine, *One-Class* Support Vector Machine, Geometric-based Features, Extremely Randomized Trees**

I. Introduction

The recent advances in information and communication technology have attracted the interests of researchers to design and deploy smart devices and solutions in the context of smart cities and homes. For the sake of providing an adequate assistance, the used smart devices exploit information retrieved directly from the user and its environment. In this context, facial expressions are widely employed. Indeed, they might be considered as relevant indicators of the human behavior and emotional state. Therefore, they have been exploited in different fields such as human-computer interaction [1], video surveillance, e-learning, ambient assistance [2] and entertainment. Several works have been conducted in the field of facial expressions. Thus, Mehrabian [3] highlighted the fact that during a common interaction, the biggest amount of information ($\approx 55\%$) is transferred through facial expressions while the rest is handled by the verbal ($\approx 7\%$) and vocal ($\approx 38\%$) parts. Moreover, Ekman & Friesen [4] have also contributed by stating about the presence of *six basic emotions* namely happiness, fear, anger, surprise, disgust and sadness.

Based on various psychological studies and advances in information and communication technology, several Automatic Facial Expression Recognition (AFER) methods have been proposed. They allow identifying the human emotional state based on information retrieved from facial expressions. One of the main components of such systems remains the *recognition* that exploits a *machine learning* technique. In this context, various techniques might be employed and one of the most used remains the Support Vector Machine (SVM) classifier. However, it still challenging to design an efficient and adequate *multi-class* classifier architecture. Usually, existing methods employ *bi-class* SVM combination to achieve a *multi-class* recognition. Even if these architectures yield relatively good performance in terms of accuracy, they remain greedy in terms of computation time. Therefore, another type of SVM classifier might be exploited namely *one-class* classifier that has been effectively used for handwritten signature verification [5].

In this paper, we propose an approach to identify the *six basic emotions* from frontal face images. Similarly to any pattern recognition system, we begin by extracting or generating a relevant *representation* that highlights the useful and discriminant information. The generated representation exploits *geometric-based* features that have proved their efficiency in previous works [6] [7]. In order to select the most discriminant attributes, we employ a feature *ranking* technique based on Extremely Randomized Trees (**ExtRa Trees**) method before applying a *threshold* defined *empirically*. The recognition is achieved using three different *multi-class* SVM architectures namely *bi-class* (*One-against-One* (*O-a-O*) and *One-against-All* (*O-a-A*)) and *one-class* classifiers. Thereby, we investigate the performance of each one and measure the accuracy and computation time when using three benchmark facial expression datasets namely **JAFFE** [8], **KDEF** [9] and **RaFD** [10].

The remainder of this paper is structured as follows. In Section II, we introduce fundamentals about common AFER systems and the SVM-based classifiers either *bi-class* and *one-class*. We also present some existing AFER methods that

978-1-5386-6866-5/18/$31.00 ©2018 IEEE

employ SVM classifiers. The proposed *static* AFER approach is detailed in Section III. The conducted experiment and evaluations are described in Section IV before discussing and interpreting the obtained results in Section V. Finally, concluding remarks are presented in Section VI.

II. Background

This section is dedicated to generalities about common AFER systems. Each one of the different components of such system is detailed. Moreover, we also present both SVM classifiers either *bi-class* and *one-class*. Finally, we present some related works and exiting AFER methods.

A. Automatic Facial Expression Recognition

The main purpose of an AFER system lies on identifying the human emotional state in an *automatic* way. This might be achieved using information and communication technology. An AFER system consists of the same components as a common *pattern recognition* system. Moreover, we distinguish two different AFER systems depending on the type of *input* that might be either *static* (single image) [6] and *dynamic* (image sequence) [7]. The first component of a basic AFER system consists in the *feature extraction* allowing to generate a relevant representation of the input by highlighting the useful information. In the field of AFER, three distinct types of descriptors might be employed: 1) *geometric-based* are descriptors computed based on previously extracted facial fiducial points, 2) *appearance-based* might be seen as textural information generated from the entire image, 3) *hybrid-based* merge the information provided by both previous feature types in order to increase the recognition rate. Usually, the size of the obtained representation might affect the AFER system by increasing both complexity and computing time. Therefore, a *feature selection* component is added in order to get rid of the redundant and noisy attributes. Finally, the *recognition* is achieved using a *supervised* machine learning technique.

B. Multi-class Support Vector Machine

One of the main objectives of this work remains the evaluation and comparison of different *multi-class* SVM classifier architectures in the context of AFER. Therefore, we introduce fundamentals for both *bi-class* and *one-class* SVM classifiers.

The currently used SVM classifier has been introduced by Cortes & Vapnik [11]. It consists in a *supervised* machine learning technique that performs a separation between two distinct classes after a *training* or *learning* phase. Indeed, the purpose of such a phase consists in finding the *maximum-margin hyper-plane* that ensures an optimal *separation* between two distinct classes. We consider a labeled dataset $\{(x_1, y_1), (x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$ represents a feature vector and $y_i \in \{\pm 1\}$ the corresponding label. Based on the defined dataset, a *model* is generated during the *learning* phase and might be defined by the equation 1 as a *decision function* $f_{bi}(x)$.

$$f_{bi}(x) = sgn(\sum_{i=1}^{i=n} \alpha_i y_i K(x, x_i) + b) \tag{1}$$

α_i represent the Lagrange multipliers, K the *kernel* function and $b \in \mathbb{R}$ the *bias*.

The *bi-class* SVM, as described above, aims to find the optimal separation between two distinct classes in a *supervised* way. Based on this, Schölkopf et al. [12] proposed a *one-class* SVM in order to separate a specific and single-class from any other possible classes. It employs an *unsupervised* learning algorithm to distinguish the defined class from the other *outliers*. Unlike the *bi-class* SVM, the *one-class* classifier aims to find a *hyper-sphere* that contains most training data with a minimum volume. Following the previous definition, the *one-class* SVM might be defined by the equation 2 as a *decision function* $f_{one}(x)$.

$$f_{one}(x) = sgn(\sum_{i=1}^{i=n} \alpha_i K(x, x_i) - \rho) \tag{2}$$

ρ represents the maximal distance between the origin and the defined *hyper-sphere*. The *one-class* SVM classifier is usually employed in the context of *novelty detection* and adapted when the available samples of the chosen class is relatively low. In the context of AFER, the objective lies on distinguishing between the *six basic emotions*. Therefore, the adequate SVM-based classifier is constructed by combining *bi-class* or *one-class* SVM classifiers.

C. Related Works

Over the years, various AFER methods have been proposed and among them, several ones employ an SVM-based classifier. Thus, Yaddaden et al. [6] and Samara et al. [1] proposed two distinct methods that employ the same type of *geometric-based* features. Indeed, in both methods were exploited extracted facial fiducial points (*sixty-eight* and *forty-nine*, respectively) to compute all the possible Euclidean distances. The recognition is achieved using two different classifiers namely *k*-Nearest Neighbors (*k*-NN) and *multi-class* SVM. Using another type of descriptors, Liew & Yairi [13] proposed to employ the Local Binary Patterns (LBP) along with a *multi-class* SVM classifier. Using the same type of classifiers, Ali et al. [14] proposed to exploit another *appearance-based* feature namely the Histogram of Oriented Gradient (HOG). Liu et al. [15] extended the previous methods by combining both descriptors before applying the Principal Component Analysis (PCA) for the dimension reduction and a *multi-class* SVM classifier for the recognition. Li et al. [16] used the *low-dimensional Gabor features* and four different statistical metrics computed from each *Gabor-Level Co-occurrence Matrix*. Then, the generated representation fed a *k*-NN classifier for the recognition. Shokrani et al. [17] employed the Pyramid HOG (PHOG) in order to generate a relavant representation before feeding a *multi-class* SVM classifier to achieve the recognition. Recently, a new trend of machine learning has emerged and allows reaching high performances in terms of accuracy. Thus, Yaddaden et al. [2], [18] proposed a Convolutional Neural Network (CNN) architecture. It belong to *Deep Learning* techniques and allows

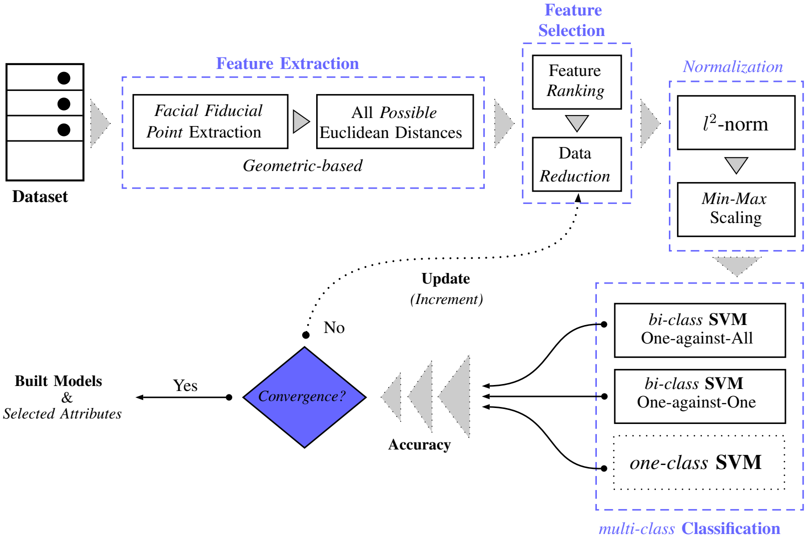


Fig. 1. Overview of the proposed *static* AFER approach.

to get rid of *hand-crafted* features by generating automatically a relevant representation of the input.

Although the numerous existing and regularly proposed AFER methods, designing such systems still challenging. Moreover, the most critical component in an AFER system remains the *recognition* and from the above brief review, *multi-class* SVM is widely used. In this paper, we propose to investigate the performance of different SVM-based classifiers namely *bi-class* (*O-a-O* and *O-a-A*) and *one-class* in the context of AFER.

III. Proposed Approach

In this section, we detail the different components of the proposed *static* AFER approach (see Fig. 1).

A. Geometric-based Features

In the context of this work and in order to implement the proposed *AFER* approach, we employ *geometric-based* features to generate a relevant representation. Thus, as achieved in our previous works [6] [7], we used an efficient method proposed by Kazemi & Sullivan [19] in order to extract *sixty-eight* facial fiducial points (see Fig. 2). This method might be defined by the equation 3:

$$\hat{S}^{t+1} = \hat{S}^{t} + r_t(I, \hat{S}^{t}) \tag{3}$$

Where $\hat{S}^t$ represents the current estimation of the face shape of the input image I. Moreover, the face shape is defined by $S = \{P_1, P_2, \ldots, P_N\} \in \mathbb{R}^{2N}$ where each P_i represents a facial fiducial point. In summary, we denote $|S| = N = 68$ facial fiducial points and each one is represented by *Cartesian co-ordinates* $P_i(x_i, y_i)$. In the equation 3, the *regression function* $r_t()$ aims to adjust, *iteratively* the face shape until convergence using a cascade of decision trees trained using the gradient boosting technique.

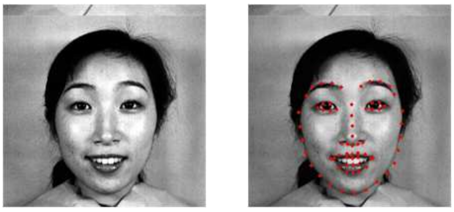


Fig. 2. (a) Original image (b) Image with facial fiducial points.

After successfully extracting the *sixty-eight* facial fiducial points, the next step consists in generating a relevant representation. Thus, $|V| = 2278$ possible *Euclidean distances* are computed and employed as feature vector.

B. Feature Ranking & Reduction

Due to the large size of the obtained feature vector that contains $|V| = 2278$ attributes, a *feature selection* stage is added in order to reduce its size. To our knowledge, we distinguish two different types of *feature selection* techniques: 1) *transformation-based* are the most widely used and one of the most common remains the PCA that allows to generate *principal components* to highlight the useful information, 2) *ranking-based* consist in evaluating the relevancy of each attribute before applying a specific *threshold*.

In the context of this work, we proposed to employ a *ranking-based* technique namely **ExtRa Trees** [20]. It might be seen as a *machine learning* algorithm that builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure. In our case, we employ the *gini index* as impurity measure. The constructed *model* in a *supervised* way generates a *feature importance* vector that contains different score values corresponding to each attribute. The different attributes are sorted in a *decreasing* way following their score values before applying a *threshold* to select a specific percentage of them.

C. Value Normalization

From our previous work [7], we found that the *data value range* might affect the *recognition* stage performance. Therefore, in the context of this work, we apply *value normalization* to the different descriptor vectors. Moreover, two distinct normalization techniques namely the *l^2-norm* and *min-max scaling* are applied. Both techniques are defined by the two equations: 4 and 5, respectively.

$$x' = \frac{x}{z} \;\; where \;\; z = \|x\|_2 = \sqrt{\sum_{i=1}^{i=n} x_i^2} \tag{4}$$

$$x' = \left(\frac{x - x_{min}}{x_{max} - x_{min}}\right) \times (R_{max} - R_{min}) + R_{min} \tag{5}$$

Where x' is the normalized form of the the input feature vector x. x_{min} and x_{max} represent the *minimum* and *maximum* values of x. R_{min} and R_{max} represent the *value range* and are set to $[R_{min} = 0, \; R_{max} = 255]$.

D. Multi-class Support Vector Machine

As explained previously in Section II, SVM classifiers are either *bi-class* or *one-class*. In the context of AFER, we have to distinguish between six different classes (namely the *six basic emotions*). Therefore, different solutions have been proposed in order to overcome this limitation by combining the SVM classifiers to construct a single one.

The most used SVM architectures in the AFER field remains the *bi-class* [1] [15] [17]. For $N > 2$ classes, Two different combination might exploited: 1) *One-against-All* consists of N distinct *bi-class* SVM classifiers and each one defines a separation between a specific class and the others, 2) *One-against-One* contains $^{N\times(N-1)}/_2$ distinct *bi-class* SVM classifiers and each one defines a separation between a specific class and another one. In both cases, the decision is generated by comparing the recognition probability of each individual *bi-class* SVM and the others. The highest value defines the predicted class. In this work, we employ a *linear kernel* since it has proved its efficiency in our previous work [7].

Even if the *one-class* SVM classifier has been widely employed in other fields such as handwritten signature verification [5], it has not been sufficiently employed in the context of AFER. Unlike the *bi-class* SVM, the learning phase is achieved in an *unsupervised* way meaning that there is no need of labeled dataset. The main objective lies on finding a *hyper-sphere* to enclose samples belonging to a specific and single class. In order to construct a *multi-class* SVM classifier, we need N *one-class* SVM classifiers and each one defines a specific class (one of the *six basic emotions*). In this case, the *decision function* is defined by the *k*-NN algorithm. In order to identify the class of an unlabeled sample, each one of the N *one-class* SVM-based *model* provides the distance between the sample and the *hyper-sphere*. With $k = 1$, the class corresponding to the smallest distance defines the class of the unlabeled sample to identify. For each *one-class* SVM classifier, we employ the *radial basis function kernel* since the conducted experiment attests of its efficiency.

IV. Experimentation & Evaluation

The main aim of this work consists in comparing the performance of different *multi-class* SVM classifiers (namely *bi-class* and *one-class*) in terms of accuracy and compution time. In the context of this evaluation, we exploit three different benchmark datasets. In Table I is detailed the different datasets namely JApanese Female Facial Expression (**JAFFE**) [8], Karolinska Directed Emotional Faces (**KDEF**) [9] and Radboud Faces Database (**RaFD**) [10]. From the three used datasets, we retain only the *six basic emotions* namely happiness (**HA**), fear (**FE**), anger (**AN**), surprise (**SU**), disgust (**DI**) and sadness (**SA**).

TABLE I
Image facial expression datasets.

Dataset		**JAFFE** [8]	**KDEF** [9]	**RaFD** [10]
Emotions	FE	30	140	67
	SU	30	140	67
	HA	29	140	67
	DI	28	140	67
	AN	30	140	67
	SA	30	140	67
	Total	**177**	**840**	**402**
Resolution		256 × 256	562 × 762	681 × 1024

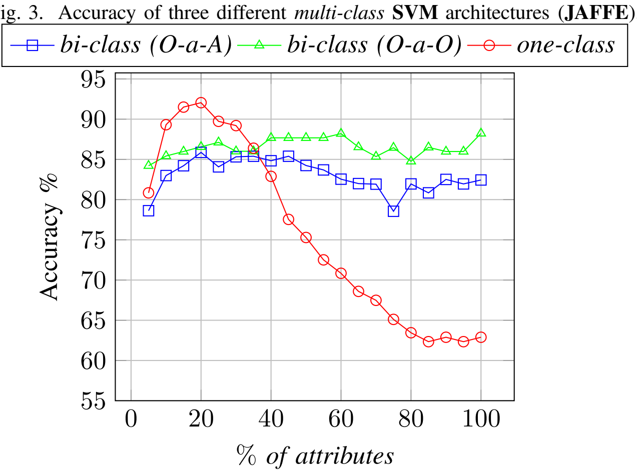
The performed evaluation is achieved according to the *ten-folds cross-validation* strategy. Thus, each dataset is divided into ten *stratified* groups. Then, ten sub-evaluations are performed and during each one, nine groups are employed during the *learning* phase while the last one is employed to evaluate the *model*. Finally, the recognition rate is computed by averaging the ten different accuracies.

V. Results & Discussion

The conducted experiment has generated several and different results. Thus, we organized them following the type of performed evaluation.

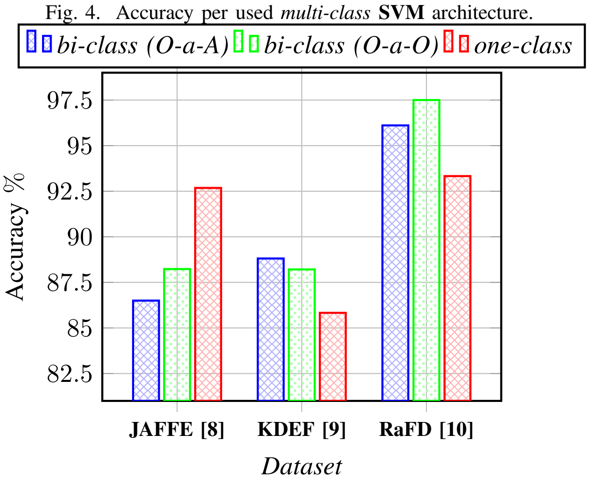
A. Comparison Between Multi-class SVM Architectures

Fig. 3. Accuracy of three different *multi-class* **SVM** architectures (**JAFFE**).



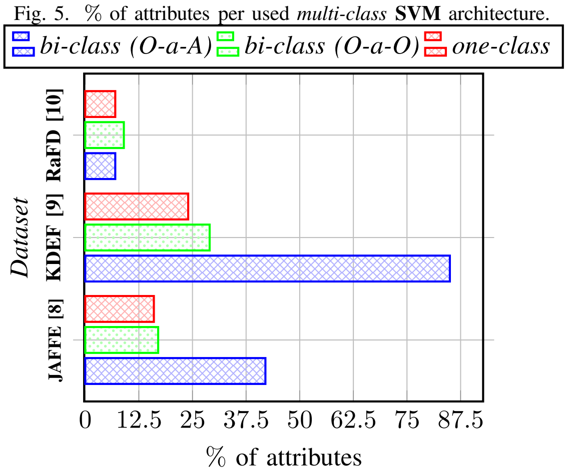
In Fig. 3 is illustrated the accuracy of the three different *multi-class* SVM-based architectures with different percentage of attributes. We notice that both *bi-class* architectures converge when increasing the number of attributes. The *one-class* architecture achieves the best performance when employing a relatively small number of attributes before decreasing. We might conclude that the *one-class* architecture is efficient with a small percentage of relevant attributes and adding noisy or redundant ones might affect negatively its performance.

Fig. 4. Accuracy per used *multi-class* **SVM** architecture.



In Fig. 4 is represented the *top-accuracy* for each one of the three *multi-class* SVM-based architectures when evaluated with the three facial expression benchmark datasets. We clearly notice from the graph that the *one-class* architecture reaches the best accuracy with the smallest dataset (**JAFFE**). In the case of the other datasets (namely **KDEF** and **RaFD**), both *bi-class* architectures achieve better performance. Fig. 5 represents the *top-accuracy* regarding to the employed number of attributes. In all cases, the *one-class* architecture uses the smallest number of attributes.

Fig. 5. % of attributes per used *multi-class* **SVM** architecture.



In Table II is illustrated the *computing time* during both phases (*training* and *predicting*) for the different *multi-class* SVM-based architectures. We clearly notice that the *one-class* architecture takes less time for building the *model* and the predicting time is estimated to $\approx 1\ ms$. We might conclude that the *one-class* architecture is more computationally efficient than both *bi-class* architectures.

TABLE II
COMPUTING TIME EVALUATION IN *training* AND *predicting* (**JAFFE**).

	bi-class (O-a-A)	*bi-class (O-a-O)*	*one-class*
Training	1630.66 *ms*	44.28 *ms*	**14.70** *ms*
Predicting	0.29 *ms*	5.50 *ms*	**1.01** *ms*

B. Evaluation of the Static AFER Approach

In Table III is presented the recognition rate of each one of the *six basic emotions* when employing the *bi-class* and *one-class* architectures. The evaluation is performed using three benchmark facial expression datasets (namely **JAFFE** [8], **KDEF** [9] and **RaFD** [10]). We notice that the *one-class* architecture achieves relatively good results, particularly with the **JAFFE** dataset. The highest recognition rate is attributed to both happiness (**HA**) and disgust (**DI**) (see **RaFD** dataset) while the lowest results are attributed to both surprise (**SU**) and sadness (**SA**) (see **KDEF** dataset). The *one-class* architecture is more efficient with the **JAFFE** dataset than the others.

The last evaluation consists in comparing the performance, in terms of accuracy, of the proposed *static* AFER approach with existing *state-of-the-art* methods (as illustrated in Table IV). We notice that the proposed *multi-class* SVM architecture based on *one-class* SVM classifiers performs better than the existing methods using *bi-class* SVM classifiers. Only the method proposed by Yaddaden et al. [2] outperforms the proposed *static* AFER approach. However, such high performance has a cost and consists in the computation time and the dependency on high-performance hardware configuration.

Even if the *one-class* architecture does not reach the same high accuracy as a *Deep Learning* architecture [2], [18], it remains an interesting choice in the context of AFER since it achieves relatively good performance while being computationally efficient.

VI. CONCLUDING REMARKS

In this paper, we introduced a *static* AFER approach that allows identifying the *six basic emotions*. However, the main contribution of this work remains the investigation of *one-class* SVM classifier performance and comparing it with the *bi-class* SVM (*O-a-O* and *O-a-A*). We notice that *one-class* SVM classifier achieves relatively good performance in terms of accuracy and it might be more computationally efficient than the other architectures.

ACKNOWLEDGMENT

This work is financially supported by the Canadian Foundation for Innovation (CFI) and the Université du Québec à Rimouski (UQAR). We also acknowledge every person and organization for providing the benchmark facial expression datasets namely **JAFFE** [8], **KDEF** [9] and **RaFD** [10].

TABLE III
ACCURACY FOR EACH EMOTION ON THE THREE BENCHMARK DATASETS USING THREE DIFFERENT *multi-class* **SVM** CLASSIFIERS.

Dataset	**Classifier**	**Architecture**	**FE**	**SU**	**HA**	**DI**	**AN**	**SA**	***Overall***
JAFFE [8]	*bi-class* SVM	*One-against-All*	76.67%	93.33%	96.55%	78.57%	90.00%	83.33%	86.41%
		One-against-One	90.00%	86.67%	96.55%	78.57%	86.67%	90.00%	88.08%
	one-class SVM	–	**93.33%**	**93.33%**	**96.55%**	**92.86%**	**83.33%**	**96.67%**	**92.68%**
KDEF [9]	*bi-class* SVM	*One-against-All*	92.86%	91.43%	99.29%	91.43%	88.57%	89.29%	88.81%
		One-against-One	75.00%	88.57%	96.43%	89.29%	90.71%	89.29%	88.21%
	one-class SVM	–	**83.86%**	**82.14%**	**90.00%**	**90.71%**	**87.14%**	**82.14%**	**85.83%**
RaFD [10]	*bi-class* SVM	*One-against-All*	98.51%	95.52%	100.00%	98.51%	98.51%	86.57%	96.27%
		One-against-One	100.00%	100.00%	100.00%	98.51%	97.01%	89.55%	97.51%
	one-class SVM	–	**88.06%**	**91.04%**	**98.51%**	**98.51%**	**92.54%**	**91.04%**	**93.28%**

TABLE IV
COMPARISON WITH *state-of-the-art* METHODS IN TERMS OF ACCURACY.

Dataset	***Method***	***Feature type***	***Classification***	***Accuracy***
JAFFE [8]	Yaddaden et al. [6]	*Geometric-based*	*k*-NN	92.29%
	Liu et al. [15]	*Appearance-based*	*bi-class* SVM	90.00%
	Yaddaden et al. [2]		CNN	95.30%
	Liew & Yairi [13]		*bi-class* SVM	55.70%
	Shokrani et al. [17]			82.60%
	Li et al. [16]		*k*-NN	91.13%
	Proposed	*Geometric-based*	*one-class* SVM	**92.68%**
KDEF [9]	Yaddaden et al. [2]	*Appearance-based*	CNN	90.62%
	Samara et al. [1]	*Geometric-based*	*bi-class* SVM	81.84%
	Yaddaden et al. [6]		*k*-NN	79.69%
	Liew & Yairi [13]	*Appearance-based*	*bi-class* SVM	74.70%
	Ali et al. [14]			70.50%
	Proposed	*Geometric-based*	*one-class* SVM	**85.83%**
RaFD [10]	Shokrani et al. [17]	*Appearance-based*	*bi-class* SVM	89.59%
	Yaddaden et al. [2]		CNN	97.57%
	Ali et al. [14]		*bi-class* SVM	79.50%
	Li et al. [16]		*k*-NN	88.41%
	Proposed	*Geometric-based*	*one-class* SVM	**93.33%**

REFERENCES

[1] A. Samara, L. Galway, R. Bond, and H. Wang, "Affective state detection via facial expression analysis within a human–computer interaction context," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, 2017.

[2] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "User action and facial expression recognition for error detection system in an ambient assisted environment," *Expert Systems with Applications*, vol. 112, pp. 173–189, 2018.

[3] A. Mehrabian, *Communication without words*, 2nd ed., 1968, pp. 51–52.

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[5] Y. Guerbai, Y. Chibani, and B. Hadjadji, "The effective use of the one-class svm classifier for handwritten signature verification based on writer-independent parameters," *Pattern Recognition*, vol. 48, no. 1, pp. 103–113, 2015.

[6] Y. Yaddaden, A. Bouzouane, M. Adda, and B. Bouchard, "A new approach of facial expression recognition for ambient assisted living," in *Proceedings of PETRA*. ACM, 2016, p. 14.

[7] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "Facial expression recognition from video using geometric features," in *Proceedings of the ICPRS*. IET, 2017, pp. 1–6.

[8] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *International Conference on AFGR*. IEEE, 1998, pp. 200–205.

[9] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces - kdef, cd rom from department of clinical neuroscience, psychology section, karolinska institutet," 1998.

[10] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[13] C. F. Liew and T. Yairi, "Facial expression recognition and analysis: a comparison study of feature descriptors," *IPSJ Transactions on Computer Vision and Applications*, vol. 7, pp. 104–120, 2015.

[14] M. A. Ali, Z. Hanqi, and K. I. Ali, "An approach for facial expression classification," *International Journal of Biometrics*, vol. 9, no. 2, pp. 96–112, 2017.

[15] Y. Liu, Y. Li, X. Ma, and R. Song, "Facial expression recognition with fusion features extracted from salient facial areas," *Sensors*, vol. 17, no. 4, p. 712, 2017.

[16] R. Li, P. Liu, K. Jia, and Q. Wu, "Facial expression recognition under partial occlusion based on gabor filter and gray-level cooccurrence matrix," in *International Conference on CICN*. IEEE, 2015, pp. 347–351.

[17] S. Shokrani, P. Moallem, and M. Habibi, "Facial emotion recognition method based on pyramid histogram of oriented gradient over three direction of head," in *International eConference on ICCKE*. IEEE, 2014, pp. 215–220.

[18] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "Facial expressions based error detection for smart environment using deep learning," in *International Conference on UIC*. IEEE, 2017, pp. 1–7.

[19] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the CVPR*, 2014, pp. 1867–1874.

[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, Apr 2006.

Chapitre 3

Titre :

Reconnaissance Automatique des Émotions en utilisant les Motifs Binaires Locaux à partir de sous-régions faciales

Résumé - *Ce chapitre est consacré à la reconnaissance statique et automatique des expressions faciales en utilisant des caractéristiques d'apparence. En effet, une représentation pertinente est générée à partir des données d'entrée en utilisant une variante du descripteur en motifs binaires locaux introduits par Ojala et al. [50] sur des sous-régions faciales spécifiques. Afin de réduire la taille de la représentation générée, deux différentes techniques de transformation de données sont exploitées à savoir l'analyse en composantes principales et indépendantes. De plus, les deux techniques sont évaluées en termes de taux de reconnaissance afin de définir la plus adéquate à utiliser. La partie reconnaissance est réalisée à l'aide d'un classifieur à machine de vecteurs de support multi-classes. Afin d'évaluer les performances de la méthode proposée, nous avons utilisé trois bases de données statiques :* ***JAFFE****,* ***KDEF*** *et* ***RaFD****. Les résultats obtenus confirment l'efficacité du descripteur en motifs binaires locaux qui ont été largement exploités pour la reconnaissance automatique faciale. Nous avons aussi comparé les performances de l'approche proposée avec des méthodes existantes.*

Mots clés :

Reconnaissance des Expressions Faciales, Motifs Binaires Locaux, Analyse en Composantes Principale, Analyse en Composantes Indépendante, Machine à Vecteurs de Support Multi-classes

Contributions associées :

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "Facial Sub-regions for Automatic Emotion Recognition using Local Binary Patterns," in *International Conference on Signal, Image, Vision and their Applications (SIVA'18)*, IEEE, 2018, pp. 1-6. (Statut : Présenté)

Facial Sub-regions for Automatic Emotion Recognition using Local Binary Patterns

[1]Yacine Yaddaden, [1,2]Mehdi Adda, [1]Abdenour Bouzouane, [1]Sébastien Gaboury & [1]Bruno Bouchard

[1]Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA),
Département d'informatique et de mathématique,
Université du Québec à Chicoutimi (UQAC). Chicoutimi, Québec, Canada.
[2]Département de mathématiques, d'informatique et de génie,
Université du Québec à Rimouski (UQAR). Rimouski, Québec, Canada.

Email: {yacine.yaddaden1, abdenour_bouzouane, sebastien_gaboury}@uqac.ca, mehdi_adda@uqar.ca

***Abstract*—Facial expressions are considered as a relevant source of information in order to automatically detect and recognize the human emotional state changes. Over the last decades, numerous methods have been proposed. One of the most popular and widely used descriptors remains the Local Binary Patterns. However, the performance of the existing methods in terms of accuracy varies and needs to be improved. In this paper, we introduce an approach that exploits specific facial sub-regions in order to generate a *spatial representation* using the Local Binary Patterns technique. Moreover, two different dimension reduction techniques (namely Principal Component Analysis and Independent Component Analysis) are used to improve the generated representation. The recognition of the *six basic emotions* is achieved using a *multi-class* Support Vector Machine classifier. The obtained results after validation with three benchmark datasets attests to the efficiency of the proposed approach since it yields 96.63%, 94.25% and 86.19% with the JAFFE, RaFD and KDEF datasets, respectively.**

***Index Terms*—Facial Expression Recognition, Local Binary Pattern, Principal Component Analysis, Independent Component Analysis, *Multi-class* Support Vector Machine**

I. Introduction

With the advance of Information and Communications Technology, *affective computing* raises a considerable interest among the research community. Indeed, being able to automatically detect and recognize the human emotional state changes represents an asset in several fields such as human-computer interaction, e-learning, healthcare and entertainment. Therefore, various new methods are regularly proposed in order to enhance the recognition performance. Automatic emotion recognition might be achieved through several types of *modality* or information source such as body gesture, speech and facial expressions. Due to the diversity of modalities, choosing the most suitable one is challenging. Fortunately, Mehrabian [1] has been able to quantify the contribution of three different modalities during a common interaction. Indeed, he estimated that facial expressions contribute by 55% while verbal and vocal part contribute by 7% and 38%, respectively. According to the amount of information they might provide, facial expressions represent a relevant choice. Moreover, Ekman & Friesen [2] made a breakthrough in the field by stating about the presence of *six basic emotions* namely happiness, fear, anger, surprise, disgust and sadness.

Based on the psychological studies, several Automatic Facial Expression Recognition (AFER) methods have been developed either *static* [3] of *dynamic* [4] that include temporal information. One of the most popular and widely used descriptors remains the Local Binary Patterns (LBP) introduced by Ojala et al. [5], [6]. However, the obtained performance remains relatively low even if LBP descriptors have a considerable potential for texture characterization. Therefore, designing efficient AFER approaches based on LBP descriptors represents a challenging task.

In this paper, we introduce a *static* AFER approach exploiting LBP descriptors. Indeed, from the input face image is extracted four specific facial sub-regions namely left-eye, right-eye, mouth and nose. Then, the LBP is applied to each extracted facial sub-region before generating corresponding histograms that emphasis LBP representation. Usually, in order to improve the obtained representation and reduce its size for computer efficiency, several *dimension reduction* techniques might be exploited. In the context of this work, we investigate how two distinct techniques, namely Principal Component Analysis (PCA) and Independent Component Analysis (ICA), might impact the AFER approach performance. Finally, the enhanced and dimensionally reduced representation feeds a *multi-class* Support Vector Machine (SVM) architecture for recognition of the *six basic emotions*. For the sake of verifying the efficiency of our approach, we used three benchmark facial expression datasets.

The rest of the paper is organized as follows. In Section II, we introduce fundamentals about AFER systems and present existing methods especially those based on LBP descriptors. An overview of the proposed AFER approach is presented in Section III and each component is detailed. In Section IV, we detail the experimental protocol used for the validation. The obtained results are analyzed and discussed in Section V. Finally, in Section VI are presented concluding remarks and potential improvements we are planning to achieve.

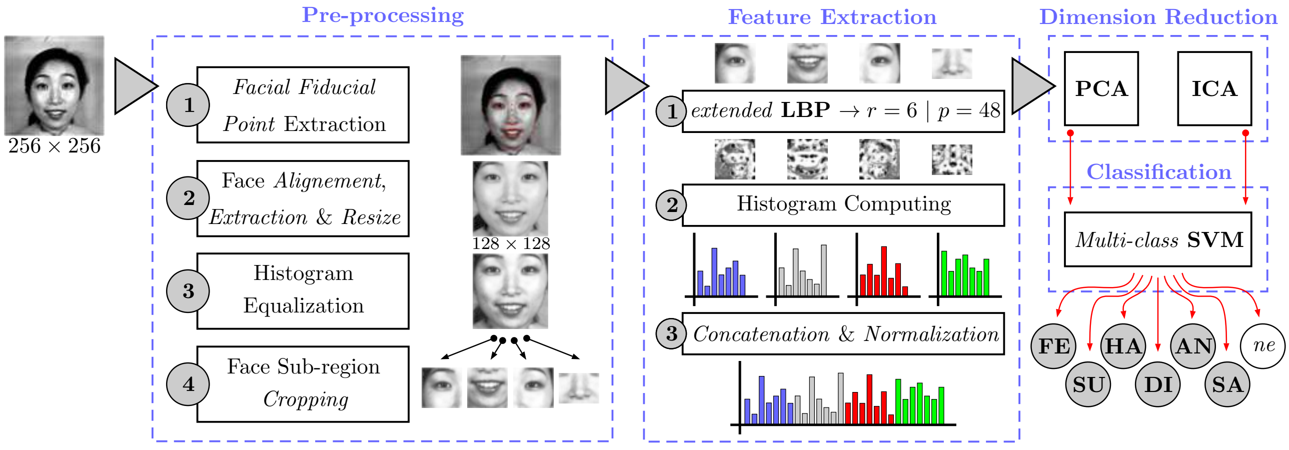


Fig. 1. Overview of the proposed *static* AFER approach.

II. Background

In this section, we detail the basic architecture of a common AFER system. Moreover, we present existing methods with a particular focus on the use of LBP descriptors.

A. Fundamentals

The human emotional state might be detected and recognized using an AFER system after analysis of facial images. Moreover, the basic architecture of an AFER system consists of the same processing stages as a common pattern recognition system. The category of the AFER system is defined by the type of used *input*. Indeed, it might be either *static* [3] using simple images or *dynamic* [4] that exploit image sequences that include temporal changes. The first stage perform *feature extraction* that basically consists in generating a relevant representation. In the context of AFER, we distinguish three different types of descriptors: 1) *geometric-based* that exploit Facial Fiducial Points (FFPs) to generate a representation, 2) *appearance-based* highlight textural information extracted from the entire image, 3) *hybrid-based* combine the information provided by both previous features. The *feature selection* or *dimension reduction* stage is not mandatory even if it allows to improve the computer efficiency by removing noisy and redundant attributes while keeping the most discriminant ones. The last stage namely *classification* employs a *supervised* machine learning technique in order to perform recognition after a *training phase*. Based on the psychological studies [1], the *output* consists of the *six basic emotions*.

B. Related Works

In order to improve the performance of exiting AFER methods, several ones are regularly proposed. Yaddaden et al. [3] proposed an AFER method that extract *sixty-eight* FFPs using the Kazemi & Sullivan [7] technique. Then, it computes all the possible Euclidean distances before applying several classification techniques where the best results were achieved using a k-Nearest Neighbors (k-NN) classifier. Ali et al. [8] exploited an *appearance-based* descriptor namely Histogram of Oriented Gradient (HOG). The recognition part is achieved using a Sparse Representation Classifier (SRC). Other authors have been interested in a new trend of machine learning namely *Deep Learning* that get rid of *hand-crafted* features. Thus, Yaddaden et al. [9] introduced a Convolutional Neural Network (CNN) architecture allowing to *autonomously* generate a relevant representation before classification.

In the context of this work, we focus on existing AFER which exploit LBP descriptors since our approach employs the same descriptor. In most cases, the studied methods include pre-processing steps such as face detection and extraction using Viola & Jones [10] technique. Vupputuri & Meher [11] proposed a method that applies a basic LBP [5] to each extracted face region before computing the corresponding histogram. For each facial expression class (emotion) is generated a distinct *template* by averaging histograms from the same class. The recognition is performed using a Kullback Leibler (KL) *divergence* distance classifier. Liew & Yairi [12] evaluated different descriptors and classification techniques in the context of AFER. They used LBP descriptors along with Linear Discriminant Analysis (LDA), *multi-class* SVM and AdaBoost classifiers. They have also optimized the classification by adding feature selection and *dimension reduction* techniques namely Random Feature Selection (RFS) and PCA. Cao et al. [13] proposed to divide the face image into a grid where each cell is 8×8 pixel size. Then, the LBP is applied to each one before generating the corresponding histogram as feature vectors. The concatenated descriptors feed a *multi-class* SVM classifier for recognition. Similarly, Zhao & Zhang [14] introduced a method that divide the face image into *forty-two* blocks on which is applied LBP and computed the corresponding histograms. Moreover, they applied different dimension reduction techniques such as PCA and Discriminant Kernel Locally Linear Embedding (DKLLE). The recognition is performed using a k-NN classifier. Other authors proposed the fusion of descriptors in order to enhance the recognition. Thus, Liu et al. [15] combined LBP and HOG representation before applying PCA for dimension reduction. The recognition is achieved using a *multi-class* SVM classifier.

III. Proposed Approach

In this section, an overview of the proposed *static* AFER approach is presented in addition to details relative to each one of its components.

A. Pre-processing

In order to ease the processing and optimize the recognition, several *pre-processing* steps are performed. As illustrated in Fig. 1, the first step consists in FFPs extraction using the Kazemi & Sullivan [7] technique. As shown in Equation 1, it exploits a *regression function* $r_t()$ in order to adjust, *iteratively*, a predefined face shape $\hat{S}^t$ using the input image I. The results consists in $|S| = N = 68$ distinct FFPs defined by $S = \{P_1, P_2, \ldots, P_N\} \in \mathbb{R}^{2N}$ where $P_i(x_i, y_i)$ represents the *Cartesian co-ordinates*.

$$\hat{S}^{t+1} = \hat{S}^t + r_t(I, \hat{S}^t) \tag{1}$$

The extracted FFPs are used to perform the next two pre-processing operations namely face *alignment* and *cropping*. The extracted face region is resized to 128 × 128 pixels. In order to get rid of illumination and contrast issues, we perform *histogram equalization*. Moreover, it is well-known that the biggest effect of an emotion is focused on specific face regions. Therefore, we extracted four different facial sub-regions (namely *left-eye*, *right-eye*, *mouth* and *nose*) by dividing the resultant face image (as shown in Fig. 1).

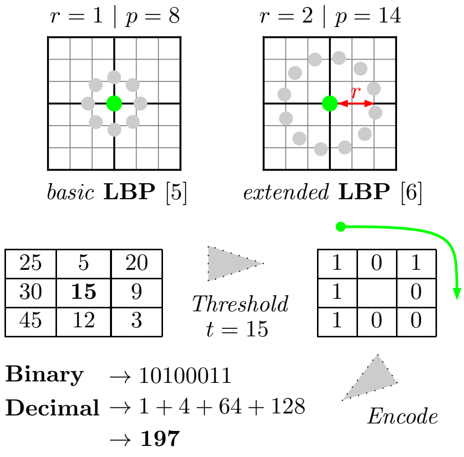


Fig. 2. Overview of the *basic* and *extended* **LBP** extraction process.

B. extended Local Binary Patterns

The next stage consists in *feature extraction* and allows generating relevant *representations* from the previously extracted facial sub-regions. In the context of this work, we employ an *appearance-based* feature that consists in LBP. The first version has been introduced by Ojala et al. [5] in 1996. As shown in Fig. 2, the LBP representation of an image is computed in a pixel-level and aims to highlight the texture information. Basically, the *basic* LBP version operates on a 3×3 pixels neighborhood where the values are in range of $[0, 255]$. As shown in Fig. 2, we distinguish two main steps. The first one applies a *threshold* to the neighbored pixels $x_{r,i}$ (r represents the radius and in the *basic* LBP, it is equal to 1) depending on the central pixel value $x_{0,0}$ as illustrated in the Equation 2.

$$f(x_{r,i}) = \begin{cases} 1, & \text{if } x_{r,i} \geqslant x_{0,0} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The next step *encodes* each *binary* pixel value depending on its position and following the Equation 3. p represents the number of points to take into account for the LBP computing and in the case of the *basic* version, it is equal to 8.

$$LBP_{p,r} = \sum_{i=0}^{p-1} f(x_{r,i}) \times 2^i \tag{3}$$

The *extended* version has also been proposed by Ojala et al. [6]. Moreover, the main difference with the *basic* version remains the extend of the r and p values to consider. In the context of this work, we computed the *extended* LBP version to each facial sub-region with $r = 6$ and $p = 48$ as parameters that we defined *empirically* (see Fig. 1). From the *extended* LBP representations of the different sub-regions are computed the corresponding *histograms*. Then, they are *concatenated* in order to form a single descriptor vector. Finally, a *min-max normalization* technique is performed to set the data value range between 0 and 1.

C. Principal & Independent Component Analysis

The size of each generated descriptor vectors from the previous stage is estimated to $n = 9036$. Due to the large size of the descriptor vectors, two distinct *unsupervised* dimension reduction techniques are performed.

The first one namely PCA [16] analyzes the N descriptor vectors or *observations* defined by $x = \{x_1, x_2, \ldots, x_n\}$ in order to extract the important and *relevant* information. It is represented as new *orthogonal* variables called *principal components*. m different components might be generated and each one is defined by $z = \{z_1, z_2, \ldots, z_m\}$. The PCA computing is detailed in the below Equation 4.

$$z_j = a_j^T x = \sum_{i=1}^{n} a_{i,j} \times x_i \tag{4}$$

Where $a = \{a_1, a_2, \ldots, a_n\}$ and for the first component z_1, the *variance* $var(z_1)$ is *maximized*. From the j^{th} component (with $j > 1$) several conditions have to be satisfied:

1) the $var(z_j)$ is *maximized*,
2) subject to the *co-variance* $cov(z_j, z_k) = 0$ $(j > k \geqslant 1)$,
3) $a_j^T a_j = 1$.

Similarly to the first technique, the ICA aims to find new *basis* (components) to represent the input data. Basically, it is employed to separate a multivariate signal into *additive sub-components* that are *maximally independent*. We assume that

$s = \{s_1, s_2, \ldots, s_m\}$ exist and they are independent signals (components). They might be retrieved using the Equation 5:

$$s_j = b_j^T x = \sum_{i=1}^{n} b_{i,j} \times x_i \quad (5)$$

Where $b = \{b_1, b_2, \ldots, b_n\}$ is called the *mixing matrix*. Moreover, both b and s are *unknown* and are meant to be defined. In the context of this work, we exploited the **FastICA** implementation [17].

Even if both techniques allow dimension reduction, they remain quite different. Indeed, the PCA removes only correlation while the ICA also get rid of high order dependence. Moreover, the extracted components using the PCA are *orthogonal* and not *equally important* unlike ones extracted using the ICA. The PCA has already proven its efficiency for face recognition when used along with LBP features [18]. In this work, we employ both techniques in the context of AFER and we investigate the performance and impact of each one.

D. Multi-Class Support Vector Machine

The last stage of the proposed *static* AFER approach aims to to perform recognition. Indeed, we employ a popular *supervised* machine learning technique namely SVM. Initially, it has been introduced by Cortes & Vapnik [19] as a *binary* classification technique. Basically, it allows finding a *hyperplane* that ensure an optimal *separation* between two distinct classes. It might be defined by $y_i = sign(\langle \mathbf{w}, x_i \rangle + b)$ where the *maximum-margin hyperplane* is represented by $(\mathbf{w}, b)$, the feature vectors by $x_i \in \mathbb{R}^d$ and labels by $y_i \in \{\pm 1\}$. In the context of this work, we aim to distinguish between the *six basic emotions*. Therefore, we exploit a *One-Against-All* architecture that trains six different *linear binary*-SVM classifiers, each one corresponds to a specific emotion namely fear (**FE**), surprise (**SU**), happiness (**HA**), disgust (**DI**), anger (**AN**), sadness (**SA**) and neutral state (*ne*).

IV. Experimentation & Evaluation

In order to evaluate the performance of the proposed *static* approach, we exploited three benchmark facial expression datasets namely JApanese Female Facial Expression (**JAFFE**) [20], Radboud Faces Database (**RaFD**) [20] and Karolinska Directed Emotional Faces (**KDEF**) [21]. As shown in Table I, each dataset contains images that represent participants expressing the *six basic emotions* in addition to the *neutral state*.

TABLE I
Image facial expression datasets.

Dataset		**JAFFE** [20]	**KDEF** [21]	**RaFD** [22]
Emotions	FE	30	140	67
	SU	30	140	67
	HA	29	140	67
	DI	28	140	67
	AN	30	140	67
	SA	30	140	67
	NE	30	140	67
	Total	**207**	**980**	**469**
Resolution		256 × 256	562 × 762	681 × 1024

Moreover, the evaluation and experimentation are performed following the *ten-folds cross-validation* strategy. Indeed, the dataset is divided into ten *stratified* groups. During each one of the ten iteration, nine groups are used for training and the last one for evaluation. The final accuracy is obtained by averaging all the obtained accuracies.

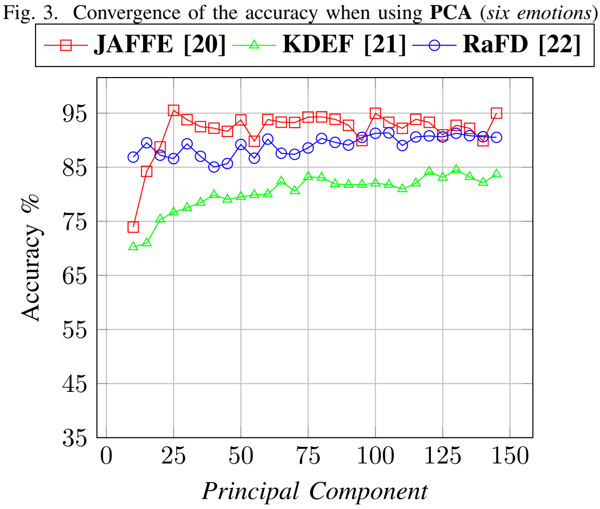
V. Results & Discussion

In this section, we present and discuss the obtained results after evaluation. Moreover, the validation process is divided into two distinct parts. In the first one, we investigate the performance and impact of both *dimension reduction* techniques namely PCA and ICA. In the last part, we present the performance of the proposed *static* AFER approach in terms of accuracy.

A. Comparison between *PCA* and *ICA*

In the Fig. 3 and 4 is presented the obtained accuracy when evaluating with the three benchmark facial expression datasets namely **JAFFE** [20], **KDEF** [21] and **RaFD** [22]. The value of the accuracy varies depending on the number of used *principal* and *independent* components in the case of PCA and ICA, respectively.

Fig. 3. Convergence of the accuracy when using **PCA** (*six emotions*).



We notice from Fig. 3 that the PCA converge faster than the ICA in Fig. 4. Indeed, with only 25 *principal* components, the PCA allows achieving high accuracy and reach a certain convergence while the ICA needs 50 *independent* components. We also notice from our experimentation that the PCA needs less processing time than the ICA which makes it more computer efficient.

In Fig. 5 is represented the highest accuracy achieved by the proposed *static* AFER approach when using the two *dimension reduction* techniques and involving the three benchmark facial expression datasets namely **JAFFE** [20], **KDEF** [21] and **RaFD** [22] with *six basic emotions*. We clearly notice that the PCA allows reaching the highest accuracy with the **JAFFE** and **KDEF** datasets with 96.63% and 86.19%. The

ICA achieves the best performance with **RaFD** dataset by reaching 94.25% accuracy. In summary, the PCA technique performs better since the average value of the recognition rates $Avg(\mathbf{PCA})$ is higher (with 92.29%) than $Avg(\mathbf{ICA})$ (with 90.97%).

Fig. 4. Convergence of the accuracy when using **ICA** (*six emotions*).

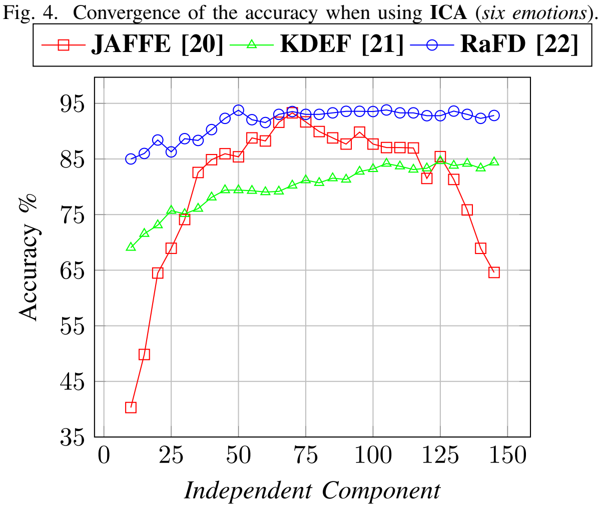
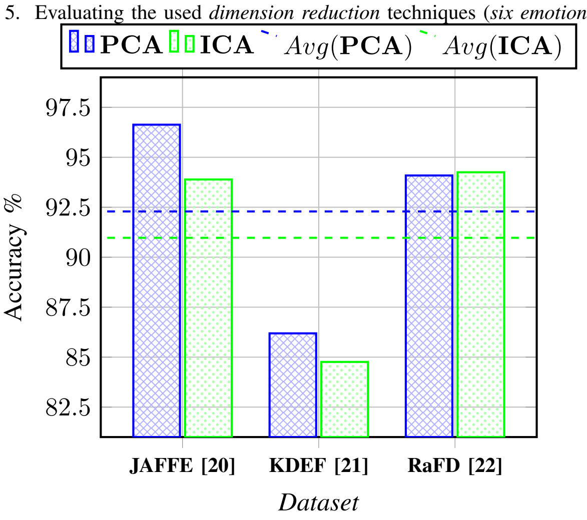


Fig. 5. Evaluating the used *dimension reduction* techniques (*six emotions*).



B. Evaluation of the static **AFER** approach

The other part of the evaluation is dedicated to measuring the performance of the proposed *static* AFER approach. In Table II is presented the different recognition rates for each emotion when evaluating with the three benchmark facial expression datasets (namely **JAFFE** [20], **KDEF** [21] and **RaFD** [22]) and using the PCA and ICA *dimension reduction* techniques. Thus, the two first rows with **JAFFE** and **KDEF** datasets correspond to the PCA while the last on with the **RaFD** dataset corresponds to using the ICA.

We notice that in almost all cases, the recognition rate is higher than 80.00% which is a relatively good result. However, the proposed *static* AFER approach has some difficulties to recognize the fear (**FE**) and sadness (**SA**) emotions with an estimated recognition rate (see **KDEF** dataset) of 71.43% and 74.29%, respectively. In summary, the **JAFFE** dataset yields the best performance with an *overall* value of 96.63% followed by **RaFD** and **KDEF** datasets with 94.28% and 86.19%, respectively.

The last evaluation presents a comparison between the proposed *static* AFER approach and existing *state-of-the-art* methods in terms of accuracy. Indeed, in Table III is presented the obtained recognition rates for the three benchmark facial expression datasets (namely **JAFFE** [20], **KDEF** [21] and **RaFD** [22]). For each one, we have two accuracy values for *six* and *seven* different emotions. In each case, the chosen *dimension reduction* techniques corresponds to the one which achieves the highest accuracy.

We notice that our approach yields the best performance in almost all the cases when evaluating with the **JAFFE** dataset. Indeed, only the method proposed by Cao et al. [13] outperforms ours with a small difference. It might be explained by the fact that the authors applied the LBP to more face regions than our approach. It makes their method a bit more complex than ours, which focuses only in specific facial sub-regions. When evaluating with the **KDEF** dataset, we also notice that our approach achieves relatively good performance. The methods proposed by Ali et al. [8] and Yaddaden et al. [9], [23] yield higher accuracy but they are more complex and require a specific hardware setup especially for the latter one. Similarly, our approach provides good results with the **RaFD** dataset and only the method proposed by Yaddaden et al. [9], [23] (based on a CNN) outperforms ours in terms of accuracy.

In summary, the proposed *static* AFER approach using the *extended* LBP descriptors performs relatively well when compared with existing and *state-of-the-art* methods.

VI. Concluding Remarks

In this paper, we proposed a *static* AFER approach based on the *extended* LBP. It exploits on four specific facial sub-regions. We have also investigated the effect of the PCA and ICA as *dimension reduction* techniques. The obtained results attest of the efficiency of our approach comparing to *state-of-the-art* methods. We are working on improvements in order to combine with other descriptors and process *image sequences*.

Acknowledgment

This work is financially supported by the Canadian Foundation for Innovation (CFI) and the Université du Québec à Rimouski (UQAR). We also acknowledge the providers of the used benchmark facial expression datasets namely **JAFFE** [20], **KDEF** [21] and **RaFD** [22].

References

[1] A. Mehrabian, *Communication without words*, 2nd ed., 1968, pp. 51–52.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[3] Y. Yaddaden, A. Bouzouane, M. Adda, and B. Bouchard, "A new approach of facial expression recognition for ambient assisted living," in *Proceedings of the PETRA*. ACM, 2016, p. 14.

TABLE II
ACCURACY FOR EACH EMOTION ON THE THREE BENCHMARK DATASETS (*six* & *seven* CLASSES).

Dataset	**FE**	**SU**	**HA**	**DI**	**AN**	**SA**	*ne*	***Overall***
JAFFE [20]	96.67%	90.00%	100.00%	92.86%	100.00%	93.33%	100.00%	96.12%
	96.67%	**100.00%**	**100.00%**	**96.43%**	**93.33%**	**93.33%**	–	**96.63%**
KDEF [21]	71.43%	84.29%	96.43%	88.57%	87.86%	74.29%	86.43%	84.19%
	77.86%	**87.14%**	**95.00%**	**85.71%**	**89.29%**	**82.14%**	–	**86.19%**
RaFD [22]	92.54%	95.52%	95.52%	98.51%	82.09%	88.06%	100.00%	93.18%
	97.01%	**95.52%**	**98.51%**	**98.51%**	**95.52%**	**80.60%**	–	**94.28%**

TABLE III
COMPARISON WITH *state-of-the-art* METHODS IN TERMS OF ACCURACY.

Dataset	***Method***	***Feature type***	***Classification***	***∑ Emotions***	***Accuracy***
JAFFE [20]	Yaddaden et al. [3]	*Geometric*	*k*-NN	Seven	92.29%
	Liu et al. [15]	*Appearance* (LBP + HOG)	PCA + SVM	Six	90.00%
	Yaddaden et al. [9]	*Appearance*	CNN	Seven	95.30%
	Zhao & Zhang [14]	*Appearance* (LBP)	PCA + *k*-NN		92.43%
			DKLLE + *k*-NN		95.85%
	Vupputuri & Meher [11]		KL *divergence*		95.24%
	Cao et al. [13]		SVM		96.28%
	Liew & Yairi [12]		PCA + LDA		63.30%
			SVM		55.70%
	Proposed		PCA + SVM		**96.19%**
				Six	**96.63%**
KDEF [21]	Yaddaden et al. [3]	*Geometric*	*k*-NN	Seven	79.69%
	Ali et al. [8]	*Appearance* (HOG)	*SRC*	Six	90.67%
	Liew & Yairi [12]	*Appearance* (LBP)	PCA + LDA	Seven	72.00%
			SVM		74.70%
	Yaddaden et al. [9]	*Appearance*	CNN		90.62%
	Proposed	*Appearance* (LBP)	PCA + SVM		**84.18%**
				Six	**86.19%**
RaFD [22]	Yaddaden et al. [9]	*Appearance*	CNN	Seven	97.57%
	Ali et al. [8]	*Appearance* (HOG)	*SRC*	Six	82.17%
	Proposed	*Appearance* (LBP)	ICA + SVM	Seven	**93.16%**
				Six	**94.25%**

[4] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "Facial expression recognition from video using geometric features," in *Proceedings of the ICPRS*. IET, 2017, pp. 1–6.

[5] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[6] 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*, vol. 24, no. 7, pp. 971–987, 2002.

[7] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the CVPR*, 2014, pp. 1867–1874.

[8] M. A. Ali, Z. Hanqi, and K. I. Ali, "An approach for facial expression classification," *International Journal of Biometrics*, vol. 9, no. 2, pp. 96–112, 2017.

[9] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "User action and facial expression recognition for error detection system in an ambient assisted environment," *Expert Systems with Applications*, vol. 112, pp. 173–189, 2018.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the CVPR*, vol. 1. IEEE, 2001, pp. 511–518.

[11] A. Vupputuri and S. Meher, "Facial expression recognition using local binary patterns and kullback leibler divergence," in *Proceedings of the ICCSP*. IEEE, 2015, pp. 0349–0353.

[12] C. F. Liew and T. Yairi, "Facial expression recognition and analysis: a comparison study of feature descriptors," *IPSJ Transactions on Computer Vision and Applications*, vol. 7, pp. 104–120, 2015.

[13] N. T. Cao, A. H. Ton-That, and H. I. Choi, "Facial expression recognition based on local binary pattern features and support vector machine," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 06, p. 1456012, 2014.

[14] X. Zhao and S. Zhang, "Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding," *EURASIP journal on Advances in signal processing*, vol. 2012, no. 1, p. 20, 2012.

[15] Y. Liu, Y. Li, X. Ma, and R. Song, "Facial expression recognition with fusion features extracted from salient facial areas," *Sensors*, vol. 17, no. 4, p. 712, 2017.

[16] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[17] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[18] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[20] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *International Conference on AFGR*. IEEE, 1998, pp. 200–205.

[21] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces - kdef, cd rom from department of clinical neuroscience, psychology section, karolinska institutet," 1998.

[22] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[23] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "Facial expressions based error detection for smart environment using deep learning," in *International Conference on UIC*. IEEE, 2017, pp. 1–7.

Chapitre 4

Titre :

Approche Hybride de Reconnaissance Automatique des Expressions Faciales pour l'interaction Homme-Machine

Résumé - *Ce chapitre est consacré à la reconnaissance statique et automatique des expressions faciales en utilisant des caractéristiques hybrides. À partir de données d'entrée statiques, une représentation pertinente est générée en combinant deux types de descripteurs géométriques et d'apparences. Dans un premier lieu, les caractéristiques géométriques sont extraites de la même façon que sur notre précédent travail [61]. En utilisant la méthode introduite par Kazemi et Sullivan [31], soixante-huit points caractéristiques du visage sont extraits de façon automatique avant de générer l'ensemble des distances euclidiennes possibles. Quant aux caractéristiques de type apparence, nous avons utilisé la transformée en ondelettes discrètes introduite initialement par Alfred Haar [23]. À partir de chacun des deux types de descripteurs constituant la représentation, sont sélectionnés les attributs les plus pertinents en utilisant une technique basée sur les arbres extrêmement aléatoires. La fusion des deux descripteurs se fait en aval. En effet, deux classifieurs à machine de vecteurs de support multi-classes sont proposés et combinés avec une règle spécifique. L'évaluation de la méthode proposée est réalisée en utilisant trois bases de données statiques :* ***JAFFE****,* ***KDEF*** *et* ***RaFD****. Les résultats obtenus sont prometteurs et mettent en avant l'efficacité du descripteur hybride en comparaison avec les deux autres.*

Mots clés :

Reconnaissance des Expressions Faciales, Transformée en Ondelettes Discrète, Caractéristiques Hybrides, Arbres Extrêmement Aléatoires, Machine à Vecteurs de Support Multi-classe

Contributions associées :

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "Hybrid-based Facial Expression Recognition Approach for Human-Computer Interaction," in *International Workshop on Multimedia Signal Processing (MMSP'18)*, IEEE, 2018, pp. 1-6. (Statut : Présenté)

Hybrid-based Facial Expression Recognition Approach for Human-Computer Interaction

[1]Yacine Yaddaden, [1,2]Mehdi Adda, [1]Abdenour Bouzouane, [1]Sébastien Gaboury & [1]Bruno Bouchard

[1]Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA),
Département d'informatique et de mathématique,
Université du Québec à Chicoutimi (UQAC). Chicoutimi, Québec, Canada.
[2]Département de mathématiques, d'informatique et de génie,
Université du Québec à Rimouski (UQAR). Rimouski, Québec, Canada.

Email: {yacine.yaddaden1, abdenour_bouzouane, sebastien_gaboury}@uqac.ca, mehdi_adda@uqar.ca

***Abstract*—Human-Computer Interaction represents an important component in each device designed to be used by humans. Moreover, improving interaction leads to a better user experience and effectiveness of the designed device. One of the most intuitive ways of interaction remains emotions since they allow to understand and even predict the human behavior and react to it. Nevertheless, emotion recognition still challenging since emotions might be complex and subtle. In this paper, we introduce a new *hybrid-based* approach to identify emotions through facial expressions. We combine two different feature types that are *geometric-based* (from facial fiducial points) and *appearance-based* (from Discrete Wavelet Transform coefficients). Each one provides specific information about the *six basic* emotions to identify. Furthermore, we propose to use a *multi-class* Support Vector Machine architecture for classification and Extremely Randomized Trees as feature selection technique. Carried experimentation attests to the effectiveness of our approach since it yields 96.11%, 91.79% and 99.05% with three benchmark facial expression datasets namely JAFFE, KDEF and RaFD.**

I. Introduction

During the last decades, several technological advances have been witnessed in the field of Information & Communication Technologies. Indeed, various devices have been developed for diverse applications, especially for human assistance. One of the most critical component when designing such devices lies in the way they interact with the user. Moreover, improving the Human-Computer Interaction (HCI) component leads to a better user experience and a more effective device. The most spread HCI interfaces of interaction are the keyboard and mice. Although their effectiveness, they lack intuitiveness and are far from natural human interaction techniques. Emotions are intuitive and allow understanding and even predicting the human behavior in order to react to it.

Emotions might be seen as a universal way of communication. Therefore, they raised interest and motivation for their use in interaction. Moreover, several *modalities* might be exploited such as body gesture, speech and facial expressions. However, Mehrabian [1] estimated that during a conversation, facial expressions contribute by 55% to the message effect while the verbal and vocal parts contribute by 7% and 38%, respectively. Thus, this study attests to the relevancy of using facial expressions for human interactions. Ekman [2] conducted several studies in emotion recognition through facial expressions. One of his biggest breakthrough in this field remains his statement about the presence of *six basic* emotions namely happiness, fear, anger, surprise, disgust and sadness. Moreover, his psychological works and studies represent the basic of almost all existing Automatic Facial Expressions Recognition (AFER) methods.

In this paper, we introduce a new *hybrid-based* AFER approach to recognize the six basic emotions. Thus, we propose to combine two distinct feature types and each one contributes with specific information to the recognition. The generated representations or feature vectors consist of all possible Euclidean distances computed using previously extracted facial fiducial points and Discrete Wavelet Transform (DWT) coefficients. Moreover, we exploit a feature selection technique based on Extremely Randomized Trees (ExtRa-Trees) in order to select the most relevant attributes. Since the proposed *hybrid-based* AFER approach relies on two distinct representations, a specific classification architecture based on a *multi-class* Support Vector Machine (SVM) is employed. In addition, three different pre-processing steps are applied on the input image in order to enhance the performance in terms of accuracy. The main contribution of this work lies on the *combination* of two distinct feature types that increase considerably the recognition rate.

The rest of the paper is organized as follows. In Section II is presented fundamentals and existing AFER methods. A detailed description of the proposed *hybrid-based* AFER approach is introduced in Section III. In Section IV, the experimental protocol is detailed before presenting the obtained results and interpretations in Section V. Finally, concluding remarks and future works are presented in Section VI.

II. Background

In this section, we introduce generalities about AFER systems and their components. Moreover, we present various existing AFER methods and discuss their limitations and challenges to overcome.

A. Fundamentals

An AFER system might be defined as the operation that identify emotion through facial expressions. Moreover, it consists of the same building blocks as a common *pattern recognition* system. Depending on the type of input (*static* [3] or *dynamic* [4]), each component is designed to perform a specific task. Thus, *feature extraction* aims to generate a relevant representation of the input through descriptor vectors. In AFER field, three different feature types might be considered: 1) *geometric-based* exploit facial fiducial points to generate specific distances, 2) *appearance-based* consider textural information from the whole image using specific image transformation techniques such as DWT, 3) *hybrid-based* combine the information provided by both of the two other features types. *Feature selection* is not mandatory even if it might be advantageous since it reduces the feature vector sizes by removing irrelevant and redundant attributes. The last component namely *classification* employs a *supervised* machine learning technique to perform recognition.

B. Related Works

Several works have been made in the field of AFER. Among the existing methods, Uçar et al. [5] proposed the use of *Discrete Curvelet Transform* coefficients. After applying the transform, the proposed method computes statistical measures such as the entropy, the mean and the standard deviation from the obtained coefficients. Moreover, the classification is achieved using a combination of Spherical Clustering (SC) and Online Sequential Extreme Learning Machine (OSELM). Liu et al. [6] proposed a method that combines two representations generated by Histogram of Oriented Gradient (HOG) and Local Binary Patterns, respectively. Then, both representations are concatenated to form a single feature vector on which is applied the Principal Component Analysis (PCA). The resultant descriptor vector feeds a multi-class SVM classifier. Samara et al. [7] proposed a method that extracts facial fiducial points (*forty-nine* in total) and then computes all the possible *Euclidean distances*. The resultant feature vector feeds a multi-class SVM classifier. Similarly, Yaddaden et al. [3] introduced a method that extracts *sixty-eight* facial fiducial points and computes all the possible *Euclidean distances*. The authors also proposed a *variance*-based feature selection technique to reduce the feature vector size. Finally, they employ a *k*-Nearest Neighbors (*k*-NN) classifier for recognition. In [8], Ali et al. introduced a method that exploits HOG descriptors combined with a Sparse Representation Classifier (SRC). Li et al. [9] proposed a method that employs *low-dimensional Gabor features* and four different statistical metrics computed from each *Gabor-Level Co-occurrence Matrix*. The recognition is achieved using a *k*-NN classifier. Jiang et Jia [10] proposed an AFER method based on the use of Two-Dimensional Local Discriminative Component Analysis (2-D LDCA) algorithm for feature extraction and classification. Other works focus on the use of a new trend of machine learning techniques namely *Deep Learning*. Yaddaden et al. [11] introduced a Convolutional Neural Network (CNN) architecture and Lopes et al. [12] proposed a combination of *seven* binary-CNN. Moreover, both methods include pre-processing stages and data augmentation to handle *overfitting*. The main advantage of such methods remains the fact that they get rid of *hand crafted* features.

All methods described above exploit a single feature type either *geometric-based* or *appearance-based*. Combining both feature types might be challenging since it leads to an increase of both computational load and processing time.

III. Proposed Approach

In this section, we introduce and describe the proposed *hybrid-based* AFER approach that performs recognition from images. Moreover, we detail each one of its different components.

A. Pre-processing

The input image might not be in perfect condition. Indeed, various factors might affect the images such as the acquisition system (camera), the environmental conditions (brightness, contrast and illumination) or the user behavior (orientation). Therefore, we propose to apply three different *pre-processing* techniques and each one provides a specific enhancement.

Spatial normalization aims to correct the face alignment using the eye's pupils positions. This operation requires the detection of the face region using the Viola et Jones technique [13]. Then, it extracts the facial fiducial points using a relatively recent technique proposed by Kazemi et Sullivan [14]. Using the facial fiducial points around each eye, we compute eye's pupils positions and draw a line that connect them. This line is compared to a perfect horizontal line to calculate the correction angle. According to the obtained value, the image is rotated to provide a perfect aligned image. On the resulting face image is performed *face region extraction* before resizing it to 128 × 128 pixels. In order to overcome the limitation due to the environmental conditions (brightness, contrast and illumination), we apply an image enhancement technique namely *histogram equalization*. Basically, it aims to increase the global contrast using the image histogram by readjusting it in order to provide a uniform distribution of the pixel intensity values.

B. Geometric-based Features

One of the main contributions introduced by our approach relies on a new combination of two feature types namely *geometric-based* and *appearance-based*. Various existing techniques might be employed to extract *geometric-based* descriptors such as Active Shape Model (ASM) and Active Appearance Model (AAM). However, we have been motivated by the use of a more recent technique proposed by Kazemi et Sullivan [14] for its computer efficiency. Moreover, the following Equation 1 defines how it performs facial fiducial points extraction.

$$\hat{S}^{t+1} = \hat{S}^{t} + r_t(I, \hat{S}^{t}) \tag{1}$$

Equation 1 considers two different parameters: the current estimation of the face shape $\hat{S}^t$ and the input image I. The

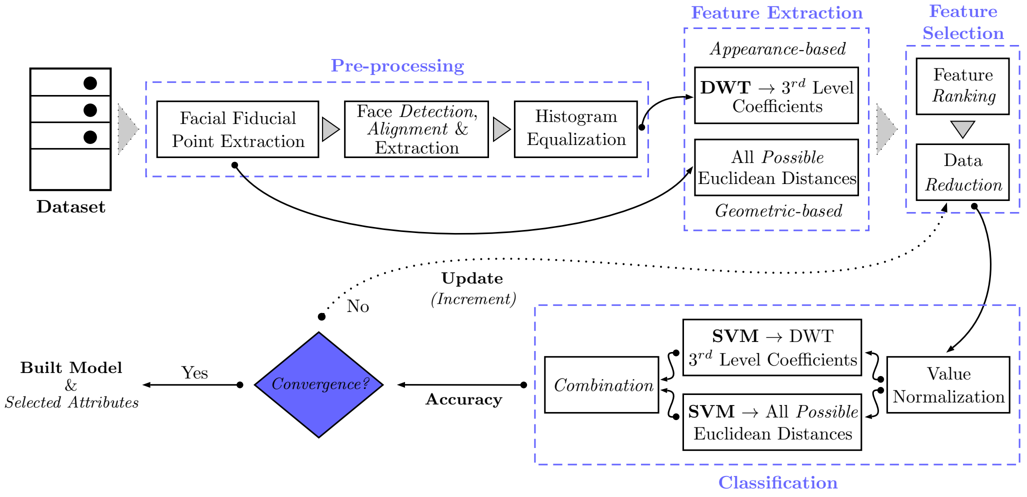

Dataset
Pre-processing
Facial Fiducial Point Extraction
Face Detection, Alignment & Extraction
Histogram Equalization
Feature Extraction
Appearance-based
DWT → 3rd Level Coefficients
All Possible Euclidean Distances
Geometric-based
Feature Selection
Feature Ranking
Data Reduction
Update (Increment)
No
Built Model & Selected Attributes
Yes
Convergence?
Accuracy
Combination
SVM → DWT 3rd Level Coefficients
SVM → All Possible Euclidean Distances
Value Normalization
Classification

Fig. 1. Overview of the proposed *hybrid-based* AFER approach.

shape face is defined by $S = \{P_1, P_2, ..., P_N\} \in \mathbb{R}^{2N}$ and each P_i represents a facial fiducial point. The face shape S contains $|S| = N = 68$ facial fiducial points and each one is represented by *Cartesian co-ordinates* $P_i(x_i, y_i)$ (see Figure 2). The *regression function* $r_t()$ adjusts, *iteratively*, the face shape until convergence using a cascade of decision trees trained using the gradient boosting technique.

The *geometric-based* feature vector is constructed by computing all the possible *Euclidean distances* using the extracted facial fiducial points. Since $|S| = N = 68$, the size of the resulting descriptor vector is estimated to $|V_{geo}| = 2278$.

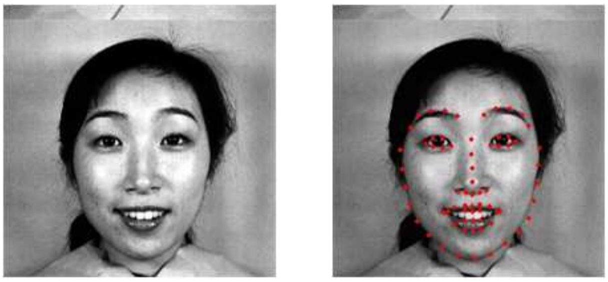


Fig. 2. (a) Original image (b) Image with facial fiducial points.

C. Appearance-based Features

Several techniques might be employed to provide an *appearance-based* representation of the input image. In the context of our approach, we have been motivated by the use of a common *multi-resolution* transform named DWT and introduced by Alfréd Haar. Moreover, performing a DWT decomposition on a two-dimensional (2-D) image requires 2-D scaling function $\varphi(x, y)$ and three 2-D wavelets $\psi^H(x, y)$, $\psi^V(x, y)$ and $\psi^D(x, y)$ [15]. Each one represents the product of two one-dimensional functions, but we only keep the four which produce a separable scaling and directionally sensitive wavelet functions (see Equation 2).

$$\begin{cases} \varphi(x,y) = \varphi(x)\varphi(y) \\ \psi^H(x,y) = \psi(x)\varphi(y) \\ \psi^V(x,y) = \varphi(x)\psi(y) \\ \psi^D(x,y) = \psi(x)\psi(y) \end{cases} \tag{2}$$

The main purpose of these wavelets consists in measuring the intensity variations of the input image along different directions. ψ^H measures variations along columns, ψ^V corresponds to variations along rows and ψ^D responds to variations along diagonals. Before defining the straightforward 2-D DWT, we introduce the scaled and translated basis functions (see Equation 3) :

$$\begin{cases} \varphi_{j,m,n}(x,y) = 2^{j/2}\varphi(2^j x - m, 2^j y - n) \\ \psi^i_{j,m,n} = 2^{j/2}\psi^i(2^j x - m, 2^j y - n), \quad i = H, V, D \end{cases} \tag{3}$$

In Equation 3, i refers to the directional wavelets. The 2-D DWT of an image $I(x, y)$ with size $M \times N$ might be defined by:

$$\begin{cases} W_\varphi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y)\varphi_{j_0,m,n}(x,y) \\ W^i_\psi(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y)\psi^i_{j,m,n}(x,y) \end{cases} \tag{4}$$

In Equation 4, j_0 represents an arbitrary starting scale and the $W_\varphi(j_0, m, n)$ coefficients define an approximation of $I(x, y)$ at scale j_0. The $W^i_\psi(j, m, n)$ coefficients provide details in horizontal, vertical and diagonal direction depending on $i \in \{H, V, D\}$ for scales $j \geq j_0$. Usually, $j_0 = 0$ and $N = M = 2^j$ so, $j = 0, 1, 2, \ldots, J - 1$ and $m = n = 0, 1, 2, \ldots, 2^j - 1$.

The 2-D DWT requires the use of specific wavelets and for our approach, we chose the *Haar Wavelet* since it yields the best performance. In Figure 3 is shown the application of a *three-level* 2-D DWT decomposition. In our case, we exploit only the coefficients from the last level provided by the directional details $W_{\psi}^{H_3}(j,m,n)$, $W_{\psi}^{V_3}(j,m,n)$ and $W_{\psi}^{D_3}(j,m,n)$. Thus, the resulting feature vector contains $|V_{app}| = 768$ attributes.

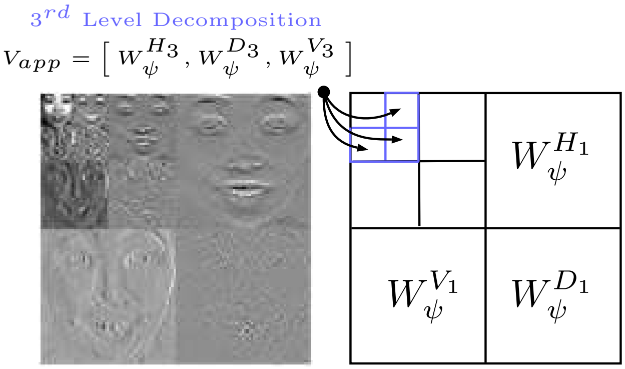

3^{rd} Level Decomposition
$V_{app} = [\, W_{\psi}^{H3}, W_{\psi}^{D3}, W_{\psi}^{V3} \,]$
$W_{\psi}^{H_1}$
$W_{\psi}^{V_1}$
$W_{\psi}^{D_1}$

Fig. 3. (a) 2-D DWT decomposition (b) 2-D DWT regions.

D. Feature Selection

The resultant descriptor vectors might be quite large since they contain a total of $V = V_{geo} + V_{app} = 2278 + 768 = 3046$ attributes. Moreover, the feature vectors might include noisy and redundant attributes that leads to *overfitting*, training time increase and accuracy reduction. Therefore, *feature selection* is required to select the most relevant attributes and reduce the size of the descriptor vectors.

In the context of our approach, we propose a feature selection technique that performs attribute *ranking* and vectors size *reduction* (see Figure 1). We employed an **ExtRa-Trees** based technique that is a variant of *Random Forests*. As defined in [16], **ExtRa-Trees** consist in a *machine learning* algorithm that builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure. Unlike other tree-based techniques, it splits nodes by choosing cut-points fully at random and it uses the whole learning sample to grow the trees. Moreover, we used the **ExtRa-Trees** classifier along with the *gini index* as impurity measure. The classifier is trained in a *supervised* way for each feature type with the corresponding labels to generate a *model*. In order to perform feature selection, we employ the generated *feature importance* vector that contains different *score* values corresponding to each attribute. The different attributes are sorted in a *decreasing* way following their score values before applying a *threshold* to select a specific percentage of attributes. Moreover, its value is adjusted (incremented) until convergence in terms of accuracy as shown in Figure 1.

We proposed two different versions of the feature selection technique. According to the way the relevant attributes are selected, we distinguish the *same-version* where the threshold value is the same for both feature vectors (*geometric-based* and *appearance-based*). Its main advantage remains its speed since it exploits a smaller range of values. The other one consists in *diff-version* where the optimal threshold value is computed for each feature vector. It allows to achieve better performance in terms of accuracy but remains slower than the first version.

E. Classification

The core component of the proposed *hybrid-based* AFER approach consists in *classification*. It allows recognizing *unlabeled* facial expression images after performing *learning* on a training set. Thus, we employ a *supervised* machine learning technique namely an SVM. Moreover, we have been motivated by using such classification technique for its robustness and generalization capability. Basically, the algorithm constructs a *hyperplane* that provides an optimal *separation* between two distinct classes. Thus, SVM algorithm belongs to the category of *binary* classifiers. A basic *linear* Binary-SVM (Bi-SVM) might be defined by $y_i = sign(\langle \mathbf{w}, x_i \rangle + b)$ where the *maximum-margin hyperplane* is represented by $(\mathbf{w}, b)$, the feature vectors by $x_i \in \mathbb{R}^d$ and labels by $y_i \in \{\pm 1\}$.

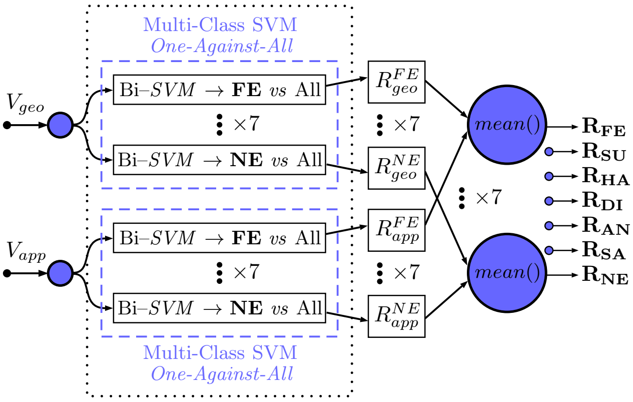


Fig. 4. The proposed classification architecture using *multi-class* SVM.

One of the most critical issues we might face remains the fact that we have to perform recognition over six different classes and the SVM is a *binary* classifier. In order to overcome this limitation, we employ a *multi-class* SVM that consists in combining Bi-SVM. Moreover, there are two different architectures: 1) *One-Against-One* that consists in performing training on each possible pair of classes, 2) *One-Against-All* is more recent and involves the construction of an SVM classifier for each class. In our case, we employ the latter one since it is faster an less complex. As shown in Figure 4, we construct two *multi-class* SVM classifiers and each one corresponds to a specific feature type V_{geo} and V_{app}. Each *multi-class* SVM classifier generates *seven* responses (%) namely R_{geo}^{i} and R_{app}^{i} where $i \in \{FE, SU, HA, DI, AN, SA, NE\}$. The combination is done following this formula $R_i = {(R_{geo}^{i} + R_{app}^{i})}/{2}$ and the decision is generated by taking the highest probability $arg_{max}(R_i)$.

In order to improve our approach performance, we propose another processing before classification. As shown in Figure 1, the *value normalization* employs the *min-max normalization* technique and aims to apply transformation changing the range of data to a pre-defined one. Indeed, machine learning techniques are somehow *sensitive* to the data value range of the feature vectors.

IV. Experimentation & Evaluation

In the context our experimentation, we focus on the use of three different benchmark facial expression datasets. JApanese Female Facial Expression (**JAFFE**) dataset [17] contains only Japanese female subjects. Radboud Faces Database (**RaFD**) dataset [18] contains 67 models. The last dataset consists in the Karolinska Directed Emotional Faces (**KDEF**) [19]. Each dataset includes the *six basic* emotions in addition to the *neutral* state (see Table I).

TABLE I
Image facial expression datasets.

Dataset		**JAFFE** [17]	**KDEF** [19]	**RaFD** [18]
Emotions	FE	30	140	67
	SU	30	140	67
	HA	29	140	67
	DI	28	140	67
	AN	30	140	67
	SA	30	140	67
	NE	30	140	67
	Total	**207**	**980**	**469**
Resolution		256 × 256	562 × 762	681 × 1024

For the validation of the proposed *hybrid-based* AFER approach, we carried the experimentation following the *10-folds cross-validation* strategy. Thus, the input dataset is divided into ten *stratified* subsets and the evaluation is achieved through ten iterations. During each one, nine subsets are exploited to *train* the classifier and the last one is used for evaluating the generated model in terms of accuracy. The final accuracy is computed by averaging the different values from each iteration.

V. Results & Discussion

Figure 5 represents a comparison between the different used feature types (namely *geometric-based*, *appearance-based* and *hybrid-based*) in terms of accuracy when evaluated with three distinct benchmark datasets (*six emotions*). In all cases, the highest accuracy is reached using the *hybrid-based* features. Indeed, it allows to achieve 96.11%, 91.79% and 99.05% accuracy with the **JAFFE**, **KDEF** and **RaFD** datasets, respectively.

In Figure 6 is illustrated a comparison between the two versions of the employed feature selection technique. The *diff-version* performs better since it allows to reach higher accuracy while using less attributes. Indeed, it allows to construct feature vectors with 934, 2617 and 426 attributes while the *same-version* provides vectors with 1614, 2162 and 761 attributes (with the **JAFFE**, **KDEF** and **RaFD**, respectively). However, the *diff-version* is more greedy in terms of processing time.

One of the important criteria when designing an AFER approach remains the processing speed. The classification stage might appear complex since it consists of several binary SVM classifiers, it is actually quite fast. Indeed, the training of *seven* samples takes ≈ 4.64 *ms* while predicting a single sample takes ≈ 0.07 *ms* (HP ProBook i7 CPU & 8Go RAM).

Fig. 5. Affects of the three types of features on the accuracy (*six emotions*).

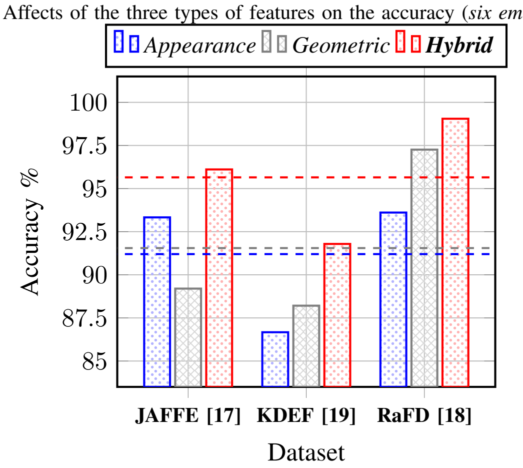


Table II represents the obtained accuracy for each emotion when evaluated with the different benchmark datasets. The second row (in bold) illustrates the evaluation when recognizing the *six basic* emotions while the first one includes the *neutral* state. In almost all the cases, the accuracy is higher than 90.00% and it even reaches 100.00%.

Fig. 6. Comparison between *feature selection* versions (*six emotions*).

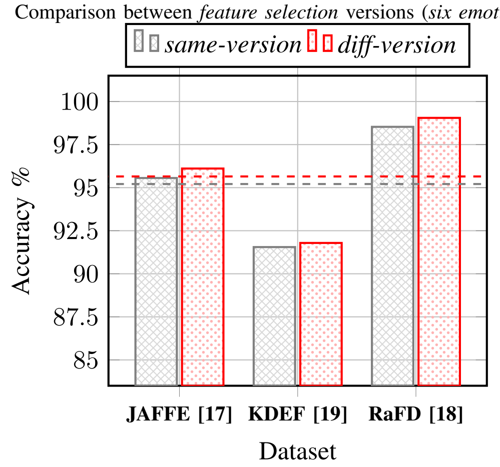


Table III illustrates a comparison with *state-of-the-art* AFER methods. We notice that our approach yields the highest accuracy: 96.19%, 91.79% and 99.05% (with the **JAFFE**, **KDEF** and **RaFD**, respectively). Even if *Deep Learning* techniques [11], [12] allow achieving high performance with large datasets. However, they require high performance hardware, are greedy in terms of computational load and processing time. The proposed approach is more computer efficient.

VI. Concluding Remarks

In this paper, we proposed a new *hybrid-based* AFER approach that combines two distinct feature types. We introduced a *multi-class* SVM based classification architecture to handle the feature vectors. The obtained results attest to the effectiveness of the proposed approach since it yields better performance in terms of accuracy. Moreover, we are still working on improving it and make it more efficient.

TABLE II
ACCURACY FOR EACH EMOTION ON THE THREE BENCHMARK DATASETS (*six* & *seven* EMOTIONS).

Dataset	**FE**	**SU**	**HA**	**DI**	**AN**	**SA**	**NE**	***Overall***
JAFFE [17]	100.00% **100.00%**	93.33% **93.33%**	96.55% **96.55%**	96.43% **96.43%**	93.33% **96.67%**	96.67% **93.33%**	96.67% **–**	96.14% **96.06%**
KDEF [19]	77.86% **80.00%**	91.43% **91.43%**	97.86% **98.57%**	94.29% **93.57%**	92.14% **93.57%**	87.14% **93.57%**	97.14% **–**	91.12% **91.79%**
RaFD [18]	97.01% **100.00%**	100.00% **100.00%**	100.00% **100.00%**	100.00% **98.51%**	95.52% **97.01%**	88.06% **98.51%**	92.54% **–**	98.51% **99.00%**

TABLE III
COMPARISON WITH *state-of-the-art* METHODS IN TERMS OF ACCURACY.

Dataset	***Method***	***Feature Type***	***Classifier***	$\sum$ ***Emotions***	***Accuracy***
JAFFE [17]	Liu et al. [6]	*Appearance*	SVM	Six	90.00%
	Yaddaden et al. [3]	*Geometric*	*k*-NN	Seven	92.29%
	Jiang et Jia [10]	*Appearance*	2-D LDCA		84.94%
	Uçar et al. [5]		OSELM-SC		94.65%
	Lopes et al. [12]		*seven* Binary-CNN		86.74%
	Proposed	*Hybrid*	SVM	Six	**96.11%**
				Seven	**96.19%**
KDEF [19]	Samara et al. [7]	*Geometric*	SVM	Seven	81.84%
	Yaddaden et al. [11]	*Appearance*	CNN		90.62%
	Ali et al. [8]		SRC	Six	82.17%
	Yaddaden et al. [3]	*Geometric*	*k*-NN	Seven	79.69%
	Proposed	*Hybrid*	SVM	Six	**91.79%**
				Seven	**91.12%**
RaFD [18]	Ali et al. [8]	*Appearance*	SRC	Six	90.67%
	Jiang et Jia [10]		2-D LDCA	seven	94.52%
	Li et al. [9]		*k*-NN		88.41%
	Proposed	*Hybrid*	SVM	Six	**99.05%**
				Seven	**96.16%**

ACKNOWLEDGMENT

Thi work is financially supported by Canadian Foundation for Innovation (CFI) and Université du Québec à Rimouski (UQAR). We also acknowledge every person and organization for providing the used benchmark facial expression datasets namely **JAFFE** [17], **KDEF** [19] and **RaFD** [18].

REFERENCES

[1] A. Mehrabian, *Communication without words*, 2nd ed., 1968, pp. 51–52.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[3] Y. Yaddaden, A. Bouzouane, M. Adda, and B. Bouchard, "A new approach of facial expression recognition for ambient assisted living," in *Proceedings of the PETRA*. ACM, 2016, p. 14.

[4] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "Facial expression recognition from video using geometric features," in *Proceedings of the ICPRS*. IET, 2017, pp. 1–6.

[5] A. Uçar, Y. Demir, and C. Güzeliş, "A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering," *Neural Computing and Applications*, vol. 27, no. 1, pp. 131–142, 2016.

[6] Y. Liu, Y. Li, X. Ma, and R. Song, "Facial expression recognition with fusion features extracted from salient facial areas," *Sensors*, vol. 17, no. 4, p. 712, 2017.

[7] A. Samara, L. Galway, R. Bond, and H. Wang, "Affective state detection via facial expression analysis within a human–computer interaction context," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, Dec 2017.

[8] M. A. Ali, Z. Hanqi, and K. I. Ali, "An approach for facial expression classification," *International Journal of Biometrics*, vol. 9, no. 2, pp. 96–112, 2017.

[9] R. Li, P. Liu, K. Jia, and Q. Wu, "Facial expression recognition under partial occlusion based on gabor filter and gray-level cooccurrence matrix," in *International Conference on CICN*. IEEE, 2015, pp. 347–351.

[10] B. Jiang and K. Jia, "Robust facial expression recognition algorithm based on local metric learning," *Journal of Electronic Imaging*, vol. 25, no. 1, pp. 1–8, 2016.

[11] Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, and B. Bouchard, "Facial expressions based error detection for smart environment using deep learning," in *International Conference on UIC*. IEEE, 2017, p. 7.

[12] 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*, vol. 61, pp. 610–628, 2017.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the CVPR*, vol. 1. IEEE, 2001, pp. 511–518.

[14] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the CVPR*, 2014, pp. 1867–1874.

[15] R. C. Gonzalez and R. E. Woods, "Wavelets and multiresolution processing," in *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006, pp. 462–519.

[16] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, Apr 2006.

[17] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *International Conference on AFGR*. IEEE, 1998, pp. 200–205.

[18] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[19] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces - kdef, cd rom from department of clinical neuroscience, psychology section, karolinska institutet," 1998.

Chapitre 5

Titre :

Reconnaissance Automatique des Expressions Faciales en utilisant une Représentation Spatio-Temporelle Efficace

Résumé - *Ce chapitre est consacré à la reconnaissance dynamique et automatique des expressions faciales en utilisant des caractéristiques géométriques. À la différence des chapitres précédents, il sera question de vidéos et séquences d'image qui seront utilisées comme données d'entrée. En effet, l'aspect temporel des séquences sera mis à contribution afin d'améliorer les performances. Ainsi, une représentation spatio-temporelle efficace est proposée. Comme pour les travaux précédents, la méthode proposée par Kazemi et Sullivan [31] est appliquée afin d'extraire de façon automatique soixante-huit points caractéristiques du visage. Ensuite, l'ensemble des distances euclidiennes sont générées pour former la représentation spatiale. La partie temporelle est générée à l'aide de trois différentes mesures statistiques. La technique basée sur les arbres extrêmement aléatoires est appliquée afin de réduire la taille de la représentation. Pour la partie reconnaissance, quatre différents classifieurs ont été utilisées et leurs performances ont été comparées. L'évaluation a été faite en utilisant trois bases de données dynamiques :* ***CK+***, ***MMI*** *et* ***MUG***. *Les résultats obtenus attestent de l'efficacité de la représentation spatio-temporelle proposée en termes de reconnaissance et de rapidité.*

Mots clés :

Reconnaissance d'émotions, Expressions Faciales, Séquence d'images, Caractéristiques Géométriques, Représentation Spatio-Temporelle

Contributions associées :

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "An Automatic Facial Expression Recognition Approach using an Efficient Spatio-Temporal Representation," in *Journal of Ambient Intelligence and Humanized Computing*, Springer, 2018, pp. 1-26. (Statut : Soumis)

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "Facial expression recognition from video using geometric features," in *Proceedings of the 8th International Conference on Pattern Recognition Systems*, IET, 2017, pp. 1-6. (Statut : Présenté)

Noname manuscript No.
(will be inserted by the editor)

An Automatic Facial Expression Recognition Approach using an Efficient Spatio-Temporal Representation

Yacine Yaddaden · Mehdi Adda · Abdenour Bouzouane · Sébastien Gaboury · Bruno Bouchard

Received: date / Accepted: date

Abstract Automatic facial expression recognition represents an active field of research since it allows to recognize and understand the human emotional state. However, it remains a challenging task, especially when processing image sequences. Indeed, even if they provide a large amount of information by

Y. Yaddaden
Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA),
Département d'informatique et de mathématique,
Université du Québec à Chicoutimi (UQAC),
555 Boulevard de l'Université, Chicoutimi, Québec G7H 2B1, Canada.
E-mail: yacine.yaddaden1@uqac.ca

M. Adda
Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA),
Département de mathématiques, d'informatique et de génie,
Université du Québec à Rimouski (UQAR),
300 Allée des Ursulines, Rimouski. Québec G5L 3A1, Canada.
E-mail: mehdi_adda@uqar.ca

A. Bouzouane
Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA),
Département d'informatique et de mathématique,
Université du Québec à Chicoutimi (UQAC),
555 Boulevard de l'Université, Chicoutimi, Québec G7H 2B1, Canada.
E-mail: abdenour_bouzouane@uqac.ca

S. Gaboury
Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA),
Département d'informatique et de mathématique,
Université du Québec à Chicoutimi (UQAC),
555 Boulevard de l'Université, Chicoutimi, Québec G7H 2B1, Canada.
E-mail: sebastien_gaboury@uqac.ca

B. Bouchard
Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA),
Département d'informatique et de mathématique,
Université du Québec à Chicoutimi (UQAC),
555 Boulevard de l'Université, Chicoutimi, Québec G7H 2B1, Canada.
E-mail: bruno_bouchard@uqac.ca

considering the temporal aspect, they remain greedy in terms of computational time and hardware resource consumption. In this paper, we introduce an approach that exploits statistical metrics to describe temporal variations of *geometric-based* features. The classification is achieved through a *linear multi-class support vector machine* that recognizes the six basic emotions from the *spatio-temporal* representation. We evaluated our approach with three benchmark datasets. The obtained results are promising since it yields **94.65**%, **93.98**% and **75.59**% recognition rates with **CK+**, **MUG** and **MMI**, respectively. The accuracy is not the only improvement, our approach is computationally efficient since it reduces the computational time in both training and predicting phases.

Keywords Emotion Recognition · Facial Expressions · Image Sequences · Geometric-based Features · Sequence-based · Spatio-temporal Representation · Frame-based

1 Introduction

It is well-known that human beings are, by nature, social and eager for interactions and communication. Furthermore, emotions are considered as *non-verbal* communication, and people rely on them in their daily life. Moreover, the technological advances and spread of Information and Communication Technologies (ICT) made the automatic emotion recognition possible. Thus, it has found applications in several fields such as Human-Computer Interaction (HCI), healthcare, ambient assistance (Yaddaden et al. 2018), entertainment and social robots. Emotions might be expressed and perceived through various *modalities* such as body gestures, facial expressions and speech (Castellano et al. 2008). Besides, automatic emotion recognition systems might either be *unimodal* or *multimodal*. While unimodal systems rely on a single source of information, the multimodal systems combine several modalities in order to enhance the recognition rate. Due to the diversity of the modalities, choosing the appropriate one appears tough since each one provides a different type of information. Moreover, Mehrabian (1968) stated that during an interaction, facial expressions provide 55% of the information while the verbal and vocal part contribute with 7% and 38%, respectively. Based on this, facial expressions seem to be the appropriate choice as a modality.

Several works and studies have been conducted in the field of facial expressions. The one achieved by Ekman and Friesen (1971) represents one of the most important. Indeed, they have stated, for the first time, about the presence of *six basic emotions* namely fear, surprise, happiness, disgust, anger and sadness. Each emotion might be described by a specific facial expression. The same authors have also introduced a new system to identify the six basic emotions using the movement of facial muscles and called Facial Action Coding System (FACS) (Ekman and Rosenberg 2005). These works represent the basis of Automatic Facial Expression Recognition (AFER). Depending on the *input*, we might distinguish two different types of AFER systems. The *static* ones perform the recognition using simple images while *dynamic* systems exploit image sequences. By considering the temporal aspect, *dynamic* systems allows to extract and process larger amounts of information. Even if they might improve the accuracy, they remain greedy in terms of processing time and computational load. Thus, improving the computationally efficiency of such systems represents a real challenge and if achieved, it might be a considerable asset for real-time applications such as video surveillance.

In this paper, we introduce a *dynamic* AFER approach to recognize autonomously emotions through facial expressions represented by image sequences. The proposed approach consists of three main stages. During the *feature extraction* stage, face detection and shape extraction are performed. Then, the facial fiducial points from the face shape are used to compute all possible *Euclidean distances* as a *spatial representation.* In order to describe the temporal variation of each distance, three different *statistical metrics* are employed. The extracted feature vector from each image sequence might have a large size. Therefore, a *feature selection* stage is required in order to reduce the feature vectors size by keeping only the most relevant *attributes.* Among the existing methods, we employ the one based on the Extremely Randomized Trees (ExtRa-Trees) classifier. The last stage namely *classification* allows to recognize the expressed emotion among the six basic ones. We employ and compare two different *supervised* machine learning techniques namely *linear multi-class* Support Vector Machine (SVM) for its robustness and generalization capability and k-Nearest Neighbors (k-NN) for its simplicity and easiness to implement. In order to confirm the efficiency of our proposed *dynamic* AFER approach, we evaluated it with three distinct benchmark datasets. The proposed *spatio-temporal representation* that corresponds to the *sequence-based* version of our approach is compared with the *frame-based* one. We evaluated them in terms of accuracy and time computing efficiency. In summary, the contributions introduced in this paper might be defined as follows:

- An efficient *dynamic* AFER approach to recognize emotions through facial expressions in image sequences,
- A new relevant *spatio-temporal* representation to describe the facial deformations,
- Two distinct versions of our approach are proposed namely *sequence-based* and *frame-based.*

The rest of the paper is organized as follows. In Sections 2 and 3, we present AFER fundamentals and describe briefly the main existing methods. Sections 4 and 5 detail the *sequence-based* and the *frame-based* versions of the proposed approach. Section 6 describes the experimental protocol and the benchmark datasets that we exploited for the evaluation. In Sections 7 and 8 are presented the obtained results and a discussion, respectively. Concluding remarks, perspectives and future work are presented in Section 9.

2 Fundamentals

A basic AFER system consists of the same building blocks as a common pattern recognition system (Konar et al. 2015). As illustrated in Fig. 1, the first block represents the *input* and depending on its type, we distinguish two categories of AFER systems (Fasel and Luettin 2003; Zhao et al. 2003). *Static* systems allow recognizing the facial expression from a single image. In most cases, this type of systems is less greedy in terms of computational time and resource consumption. However, it is not representative of the reality since a facial expression includes a temporal aspect and might be described by three distinct *transitional phases*: 1) *onset* is the first transitional state and highlights the beginning, 2) *apex* represents the reach of the highest intensity, 3) *offset* might also be seen as a transitional state before retrieving the *neutral* and initial state. Moreover, only image sequences might emphasis these three phases since they describe the facial deformations. *Dynamic* systems have been introduced in order to handle such *data representation* (Cohen et al. 2003). Besides, two sub-categories might be defined depending on the way of processing the input (see Fig. 1). *Sequence-based* systems exploit the whole image sequence to generate a relevant representation while *frame-based* ones process each image of the sequence independently (Cohen et al. 2003).

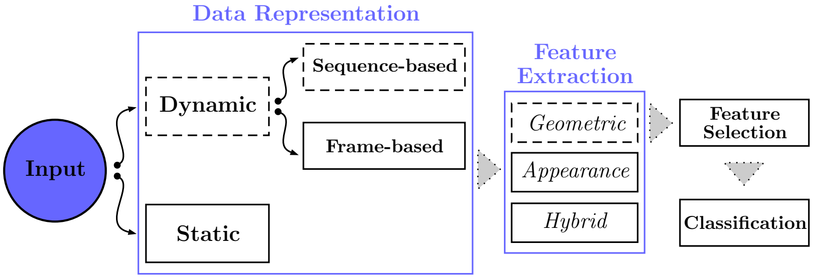


Fig. 1 Basic building blocks of a common AFER system.

Depending on the adopted data representation, the *feature extraction* block is adapted to process such input. As shown in Fig. 1, we distinguish three different types of descriptors in the context of AFER (Konar et al. 2015; Fasel and Luettin 2003). *Geometric-based* features rely on facial fiducial points to generate high-level representations such as Active Appearance Model (AAM) (Cootes et al. 2001) and Active Shape Model (ASM) (Cootes et al. 1995). Other methods exploit *appearance-based* descriptors that operate on the entire image to generate a texture representation after applying different techniques such as the two-dimensional Discrete Wavelet Transform (DWT) or the Local Binary Pattern (LBP) (Wang and He 1990). Obviously, the last feature type combines the two previous ones and named *hybrid-based* descriptors. Even if it might improves the system's accuracy, it also increases the complexity as well as the processing time and computational load. The *feature selection* block is not mandatory, but it allows to reduce the size of the feature vectors by keeping only the most relevant attributes while getting rid of noisy and redundant ones. The last block namely *classification* might be defined as a *supervised* machine learning technique. It needs to perform *learning* on *labeled* samples to construct a *model* required to recognize the class of *unlabeled* samples.

3 Related Works

As explained previously, we distinguish two distinct types of AFER systems depending on the input. Therefore, our study of existing methods is divided into two distinct parts corresponding to *static* and *dynamic* AFER systems, respectively.

3.1 Static Systems

Based on the previous description (see Section 2) of a common AFER system architecture, Maximiano da Silva and Pedrini (2016, 2015) proposed two different *static* methods. They introduced a method that employs facial fiducial points as geometric-based features combined with an SVM classifier. Moreover, they extended their method with coefficients generated by Gabor filters. Their second method includes pre-processing steps namely face region extraction and image enhancement with *histogram equalization*. Then, several appearance-based features are extracted using LBP, Histogram of Oriented Gradients (HOG) and Gabor filters. In order to reduce the size of the resulting descriptor vector, the authors applied Principal Component Analysis (PCA). The classification is achieved using an Artificial Neural Network (ANN) classifier. Using the same type of features, Guo et al. (2017) proposed a method that divides the input image into a grid of blocks after detecting and extracting the face region. In each block is applied the *extended* LBP, but the main contribution remains the dimension reduction by applying covariance matrix transform to the Karhunen-Loeve transform of the feature vector. The resulting descriptor vector feeds an SVM classifier. Ghimire et al. (2017) introduced a method that exploit hybrid-based features. It defines a set of twenty-nine local facial regions using the extracted facial fiducial points. The most relevant ones are selected by applying an exhaustive search technique. From each selected regions is computed the LBP as appearance-based features. The geometric-based ones are extracted using the normalized central moments to capture shape information. Both feature vectors are concatenated before feeding an SVM classifier.

3.2 Dynamic Systems

Unlike *static* AFER systems, *dynamic* ones process image sequences instead of single images. As explained in Section 2, this category of systems might be divided into two sub-categories namely *sequence-based* and *frame-based*.

3.2.1 Sequence-based

Sequence-based systems process the whole image sequence in order to generate a relevant *spatio-temporal* representation. Long and Bartlett (2016) presented an AFER method that employs a spatio-temporal base learning using a sparse coding algorithm. The learned bases are exploited to compute high-dimensional features before performing spatio-temporal pyramid max pooling based on spatial pyramid matching. It aims to provide a more compact and discriminatory representation that feeds a *linear* SVM classifier. Lei et al. (2017) introduced a method that exploits the constrained local model to track and align the facial fiducial points. Then, it extracts two different descriptors: 1) Spatial-Temporal Motion LBP that generates LBP histogram through three orthogonal planes, 2) Gabor Multi-orientation Fusion Histogram. The two feature vectors are scaled to $[\,-1,\ +1\,]$ before being concatenated using a specific equation and feeding an SVM classifier. Wei et al. (2015) proposed an AFER method using two distinct feature types. The first one corresponds to three extracted image patches (namely left-eye, right-eye and mouth). The other one consists of the movements of the fiducial points between the current frame and the previous one. Both feed an AutoEncoder (AE) network with the structured regularization (SR) for the feature learning. Basically, it allows transforming the input into outputs with the least possible amount of distortion before performing recognition using a softmax layer. Gupta et al. (2018) introduced a method exploiting *deep learning* techniques (LeCun et al. 2015). It starts by detecting and extracting the face region from every frame using the Viola & Jones technique Viola and Jones (2001). The recognition is performed using a Deep Convolutional Neural Network (DCNN). Moreover, their architecture includes a pre-trained AE combined with a predictor that relies on a semi-supervised learning. Due to the difference in terms of image sequences length, the authors had to perform *normalization* on each sequence. Xijian and Tardi (2017) introduced an AFER method that exploits two distinct types of spatio-temporal features. The first one consists in the Motion History Image (MHI) combined with the Optical Flow (OF) algorithm. In order to highlight the information contained in each sub-region of the face image, the authors employed the entropy to generate adequate weights. The other feature type consists in Quantised Local Zernike Moment (QLZM) combined with Motion Change Frequency (MCF). Due to the high dimensionality of the extracted descriptors, the authors used a two-dimensional PCA to generate a more compact representation. The resultant representation feeds an SVM classifier for the recognition. Liu et al. (2016) proposed to exploit three different descriptors generated using HOG, Scale-Invariant Feature Transform

(SIFT) and DCNN. Then, three different image set models are exploited: linear subspace, co-variance matrix and Gaussian distribution for their capability of capturing data variation from image sequences. Riemannian kernels are used to enable the classifier to operate in an extrinsic feature space without computing the coordinates of data in the original space. Their method achieves recognition using a fusion schema of the three decisions provided by an SVM, Logistic Regression and Partial Least Squares classifiers. Niki and Anastasios (2014) introduced a method based on the use of facial fiducial points obtained using AAM. For each facial expression is defined a manifold that is composed of several subsets. Then, the authors apply the *clustering* algorithm k-means to generate a group of subsets per manifold. The mean expression vector is computed using PCA before feeding an SVM classifier for the recognition.

3.2.2 Frame-based

The second sub-category of *dynamic* AFER systems processes *independently* the different frames of an image sequence before recognizing the facial expressions. Xijian and Tardi (2015) proposed a method that exploits two different descriptors. The first one consists in the Pyramid HOG of Three Orthogonal Planes (PHOG-TOP) applied to four facial sub-regions (namely eyebrows, forehead, nose and mouth). The other descriptor corresponds to the dense optical flow that emphasis the temporal information in the image sequence. Both descriptors are concatenated with specific weights. The recognition is achieved using a One-Against-One multi-class SVM classifier. Lee et al. (2016) introduced a method that begins by extracting the facial fiducial points in order to align and crop the face region. Then, peak face expressions are selected from the image sequences to generate the Intra Class Variation (ICV). The feature vector is obtained by computing the difference between the peak face expression and ICV faces after applying Gabor filtering. The recognition part of their method is performed using a Sparse Representation Classifier (SRC). Agarwal et al. (2016) proposed a method that performs AFER in video stream. They used the Viola and Jones (2001) technique to detect, extract and resize face region. Then, on each image sequence is performed salient region selection using information theoretic correlation analysis. From the obtained salient regions are selected the salient planes from which is computed LBP. The recognition is achieved using an SVM classifier. Kamarol et al. Kamarol et al. (2017) introduced a framework that performs intensity recognition in addition to AFER. From each frame is extracted the AAM and then applied k-NN technique to generate feature representation. A set of weighted votes are defined and a specific weight is attributed to each frame depending on its expression class. The AFER is achieved using Hidden Markov Model (HMM) and intensity estimation using change-point detection. Yaddaden et al. (2017) proposed a method that begins by *normalizing* the image sequence length with a specific technique. Then, from each frame is extracted sixty-eight facial fiducial points used for computing all possible *Euclidean distances*. The authors also proposed a new *variance*-based feature selection technique. The resultant features vectors are concatenated before feeding a *linear* SVM classifier. Peng and Yin (2018) introduced a framework that extracts the facial fiducial points using the AAM before defining local patches. Then, feature extraction is performed by applying SIFT technique on each local patch before concatenating them to form a single descriptor vector. The recognition is achieved using a Kernel Discriminant Analysis (KDA) classifier aiming to find a nonlinear projection directions that separate the different facial expression classes.

3.3 Limitations & Challenges

The study of the existing methods highlights several issues and limitations related to *dynamic* AFER systems. One of the most recurrent one is the necessity of *sequence normalization*. Indeed, the number of frames per image sequence might vary from a sample to another and most classification techniques impose to use feature vectors of the same size. Therefore, several existing methods define a specific number of frames and according to its value, they perform *sequence normalization* (Gupta et al. 2018; Yaddaden et al. 2017). Basically, it consists in duplicating frames for shorter image sequences and removing frames from longer ones. Even if this operation overcomes the main issue, it affects the initial input data by adding or removing information. Other authors preferred considering only the initial frame (*neutral* state) and the highest intensity frame (*apex*) (Lee et al. 2016). It overcomes the classification limitation to the detriment of the image sequence integrity.

Unlike *static* systems, *dynamic* ones are supposed to process image sequences instead of single images. In order to obtain a *spatio-temporal* representation, several methods extract feature vectors from each frame of an image sequence before concatenating them to form a single descriptor vector (Peng and Yin 2018; Lei et al. 2017; Yaddaden et al. 2017). Even if these methods consider all the frames of an image sequence, a large feature vector size might be seen as an issue. In most cases, an AFER system is required to be computationally efficient and fast for real-time applications. Thus, concatenating feature vectors might introduce some drawbacks especially if the number of frames per image sequence is high.

In summary, the main requirements when it comes to design a *dynamic* AFER system consists in:

- Respect of the *sequence integrity* by not adding or removing information,
- A *spatio-temporal* representation including *spatial* and *temporal* features,
- Fast *processing* and *recognition* for real-time applications,
- A high *accuracy* or *recognition rate* compared to *state-of-the-art* methods,
- It has to be *robust* when evaluated with different benchmark datasets.

The above list of criteria is not exhaustive. However, we used it as basis for the design of our *dynamic* AFER approach.

4 Proposed Approach

As shown in Fig. 2, the proposed *dynamic* AFER approach is composed of the same building blocks as a common pattern recognition system (see Section 2). In the context of our approach, the considered *input* is a facial expression dataset with a variable number of frames per image sequence.

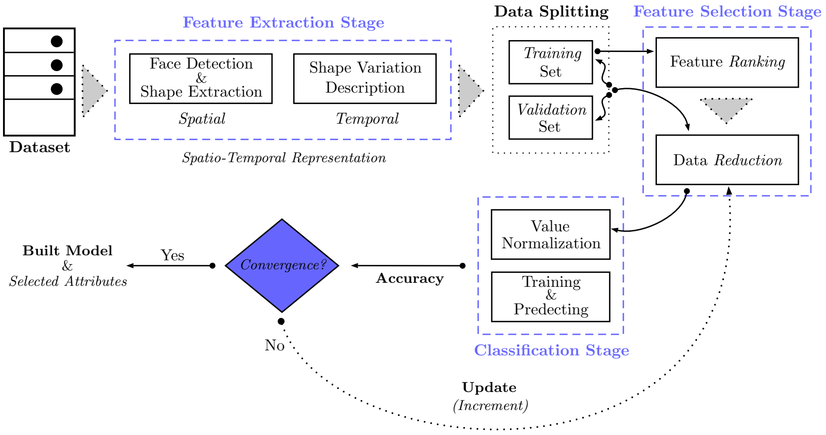


Fig. 2 Overview of our *dynamic* AFER approach (*sequence-based* version).

In the following, we detail each component of the proposed *dynamic* AFER approach namely *feature extraction*, *feature selection* and *classification.*

4.1 Feature Extraction

As discussed in Section 2, we distinguish three types of features in the AFER field. In our case, we have been motivated by the use of a specific one. Indeed, *geometric-based* features have the advantage of being invariant to face translation, rotation and illumination. Before computing the feature vectors, we exploit the Viola and Jones (2001) technique for face region detection (see Fig. 3). As illustrated in Fig. 2, we need two different representations namely *spatial* and *temporal.* The first one requires to extract the face shape. Different methods might be used such as AAM and ASM but in our case, we chose the use of a more recent technique introduced by Kazemi and Sullivan (2014). It allows to extract sixty-eight facial fiducial points as shown in Fig. 3. Basically, their technique relies on the following Equation 1.

$$\hat{S}^{t+1} = \hat{S}^t + r_t(I, \hat{S}^t) \tag{1}$$

It takes as inputs the current estimation of the face shape $\hat{S}^t$ and the input face image I. The shape face is defined by $S = \{P_1, P_2, \ldots, P_N\} \in \mathbb{R}^{2N}$ and each element P_i represents a specific facial fiducial point. We denote $|S| = N =$

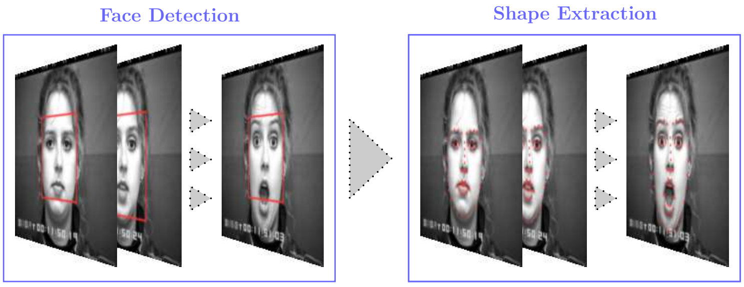


Fig. 3 Face detection & shape extraction.

68 facial fiducial points and each one is represented by *Cartesian co-ordinates* $P_i(x_i, y_i)$. The main goal of Equation 1 is to, *iteratively*, adjust the face shape until reaching convergence. This is achieved using the *regression function* $r_t()$ that takes as argument $\hat{S}^t$ and I. Basically, it consists in a cascade of decision trees trained using the gradient boosting approach.

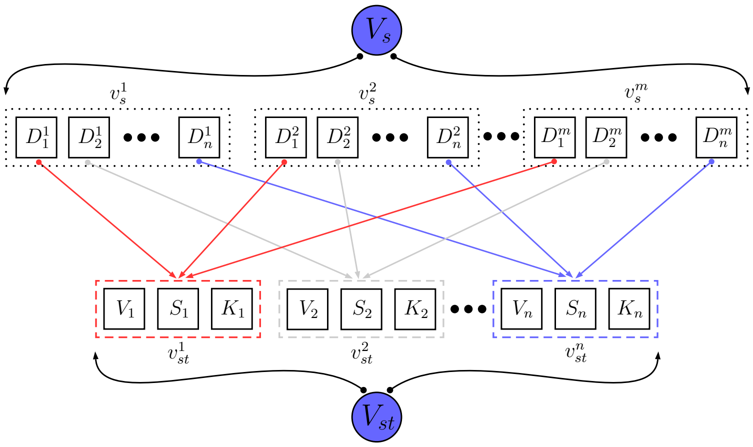


Fig. 4 *Spatio-temporal* representation of the *input*.

Inspired from previous works in (Yaddaden et al. 2016, 2017), the *spatial* representation of the input is achieved by computing all the possible *Euclidean distances* (see Equation 2). Thus, for each frame of an image sequence, we have $|S| = N = 68$ facial fiducial points and it implies $n = 2278$ possible distances (see Equation 3).

$$D_{Euclidean}(P_a, P_b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \tag{2}$$

$$n = \frac{N \times (N-1)}{2} \tag{3}$$

As said before, an image sequence consists of a variable number frames m. Therefore, we have to extract $|V_s| = m$ feature vectors and each one corresponds to a specific frame of the image sequence (see Fig. 4). Then, the *spatial* representation might be defined by $V_s = \{v_s^1, v_s^2, \ldots, v_s^m\}$ where $v_s^i = \{D_1^i, D_2^i, \ldots, D_n^i\}$ is the descriptor vector of the i^{th} frame and might takes $i = 1, 2, \ldots, m$ as values. Moreover, each vector v_s^i contains all the possible *Euclidean distances* represented by D_j^i and corresponds to the j^{th} distance and might takes the following values $j = 1, 2, \ldots, n = 2278$. This representation is inspired from the FACS (Ekman and Rosenberg 2005) but instead of using specific distances corresponding to AUs, it exploits all the possible distances.

Using only *spatial* representation might be enough to perform recognition as proposed in (Yaddaden et al. 2017). However, one has to deal with several issues and the most critical remains the variable number of frames per image sequence. Therefore, *sequence normalization* is required in order to be able to exploit a *supervised* machine learning technique. Even by doing this, the approach will not be cost-effective in terms of computational time and resource consumption. Due to this issue, we propose a new representation to describe temporal variations of each *Euclidean distance* D_j^i using *statistical metrics*. As shown in Fig. 4, the *spatio-temporal* representation is defined by $V_{st} = \{v_{st}^1, v_{st}^2, \ldots, v_{st}^n\}$. Moreover, each vector v_{st}^j contains a *statistical metric* triplet that represents the temporal variation of the j^{th} *Euclidean distance*. Nevertheless, the *statistical metrics* might be summarized as follows: *variance* (see Equation 4), *skewness* (see Equation 5) and *kurtosis* (see Equation 6).

$$V_j = \sigma^2(D_j) = \frac{\sum\limits_{i=1}^{m}(D_j^i - \overline{D}_j)^2}{m} \tag{4}$$

$$S_j = \gamma_1(D_j) = \frac{\sum\limits_{i=1}^{m}\frac{(D_j^i - \overline{D}_j)^3}{m}}{\sigma^3(D_j)} \tag{5}$$

$$K_j = \gamma_2(D_j) = \frac{\sum\limits_{i=1}^{m}\frac{(D_j^i - \overline{D}_j)^4}{m}}{\sigma^4(D_j)} \tag{6}$$

Finally, the resultant *spatio-temporal* representation of the *input* is defined by the feature vector V_{st}. Moreover, the descriptor vector has $|V_{st}| = 3 \times n = 6834$ different attributes.

4.2 Feature Selection

As defined previously, the size of the obtained *spatio-temporal* representation is estimated to $|V_{st}| = 6834$. Moreover, it is well-known that a high number of *attributes* introduces several issues such as: *overfitting* caused by the presence of redundant and noisy attributes, increase *training time*, *accuracy* reduction since some attributes might be misleading when generating the model. Therefore, we have been motivated by adding a *feature selection* stage (see Fig. 2) in order to reduce the *spatio-temporal* representation size.

Based on existing works, different feature selection techniques might be exploited. To our knowledge, the most recurrent one remains the PCA (Niki and Anastasios 2014; Maximiano da Silva and Pedrini 2015) and might be seen as a statistical procedure that uses an orthogonal transformation to convert a set of possibly correlated attributes into a set of linearly uncorrelated variables called principal components with improved discrimination propriety. Other techniques consist in *ranking* attributes by importance. Basically, they generate a specific *score* value for each attribute and the higher it is, the most important is the attribute. Based on the obtained scores, specific attributes are selected using a *threshold* value that defines the percentage of attributes to keep. In the context of our approach, we exploited **ExtRa-Trees** technique that is a variant of *Random Forests* (Louppe et al. 2013). As defined in (Geurts et al. 2006), **ExtRa-Trees** algorithm builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure. Its two main differences with other tree-based ensemble methods are that it splits nodes by choosing cut-points fully at random and that it uses the whole learning sample (rather than a bootstrap replica) to grow the trees.

In our case, we propose to use the **ExtRa-Trees** technique along with *gini index* as impurity measure. In a *supervised* way, it exploits the feature vectors with their corresponding labels to build an **ExtRa-Trees** model. Then, the *feature importance* is generated with *score* values corresponding to each attributes. According to the feature importance vector, the different attributes are sorted decreasingly. As shown in Fig. 2, a *threshold* is set to select a specific percentage. Its value is updated (incremented) as long as the accuracy does not reach convergence.

4.3 Classification

Classification represents the last component of a *dynamic* AFER approach and it allows the identification of the different facial expressions. It relies on a *supervised* machine learning technique. Thus, it requires a *training* or *learning* phase with labeled samples. Moreover, machine learning techniques are somehow *sensitive* to the data value range of the feature vectors. Therefore, we propose to apply a value normalization technique to improve the performance in terms of accuracy. We employ the *min-max normalization* that basically provides a transformation changing the range of data to a predefined one (see Equation 7).

$$v_{norm} = \left(\frac{v - V_{min}}{V_{max} - V_{min}} \right) \times (R_{max} - R_{min}) + R_{min} \tag{7}$$

Each element of the original feature vector is represented by v and its actual range of values is defined by $[V_{min}, V_{max}]$. The $[R_{min}, R_{max}]$ represents the new range of values for the normalized feature vector. In our case, we set $R_{min} = 0$ and $R_{max} = 1$. To our knowledge, several machine learning techniques might be exploited, but we focus on two specific ones. These techniques consist in ***k*-NN** and **SVM** even if we also used other techniques in order to obtain an objective comparison in terms of recognition rates.

4.3.1 ***SVM*** *classifier*

An SVM classifier might be defined as a *supervised* machine learning technique introduced by Cortes and Vapnik (1995). Basically, the algorithm constructs a *hyperplane* whose purpose is to provide an optimal *separation* between two different classes. Thus, SVM belongs to the category of *binary* classifier since it allows distinguishing between two different classes. For *linearly* separable datasets, many hyperplanes might be exploited to separates the two classes of data. However, the best choice remains the one whose the distance from it to the nearest data point is maximized and it is called *maximum-margin hyperplane*. As defined in (Shalev-Shwartz and Ben-David 2014), a basic *linear* SVM classifier might be defined by $y_i = sign(\langle \mathbf{w}, x_i \rangle + b)$ where the maximum-margin hyperplane is represented by $(\mathbf{w}, b)$, the feature vectors by $x_i \in \mathbb{R}^D$ and labels by $y_i \in \{\pm 1\}$. Moreover, SVM algorithm might also classify *nonlinearly* separable datasets using the *kernel* trick and allowing to fit the maximum-margin hyperplane in a transformed feature space.

In our case, we employed a *linear* SVM classifier for its robustness and generalization capabilities. Nevertheless, we face a *multi-class* classification problem as we need to distinguish between M facial expressions. Fortunately, this limitation might be overcome by combining SVM classifiers. Moreover, two distinct strategies are commonly used: 1) *One-Against-One* consists in training an SVM classifier for each possible pair of classes (see Equation 3), 2) *One-Against-All* is the most recent and involves the construction of an SVM classifier for each class. For the proposed *dynamic* AFER approach, we chose the latter one as it implies the training of fewer SVM classifiers. Moreover, the best performance in terms of accuracy was achieved when setting the *cost* variable to $C = 1$.

4.3.2 ***k-NN*** *classifier*

The k-NN is considered as one of the simplest machine learning techniques. Moreover, it belongs to the category of *instance-based* techniques (Aha et al. 1991). Basically, its *learning phase* consists in storing the training samples, the *prediction* of unlabeled instance is achieved by searching for the closest neighbors in the training set. Thus, its main advantage over other methods lies in its simplicity and fully understandability since it does not generate a *black-box* and complex model. In order to build a k-NN classifier, two parameters have to be defined and consist in: 1) k the number neighbors to consider for classification, 2) ρ that represents the *distance metric* used to compute the distance between training instances and the one to classify. We might describe the k-NN classification by the *majority* label among $\{y_{\pi_i(x)} : i \leq k\}$ where y are the labels, $\pi_i(x)$ the reordering of training set instances following the distance $\rho(x, x_i)$ and k the number of neighbor instances to consider.

In our case, we employed the k-NN classifier as presented in (Yaddaden et al. 2016, 2017). Indeed, the number of neighbors to consider has been set to $k = 1$ and the chosen distance metric ρ consists in *Cosine* distance (see Equation 8). In (Yaddaden et al. 2016), several other distances such as *Euclidean* and *Manhattan* have been compared but the best performance was achieved using *Cosine* distance.

$$D_{Cosine}(V_a, V_b) = \frac{\sum_{i=1}^{n} V_a[\,i\,] \times V_b[\,i\,]}{\sqrt{\sum_{i=1}^{n} V_a[\,i\,]^2} \times \sqrt{\sum_{i=1}^{n} V_b[\,i\,]^2}} \tag{8}$$

D_{Cosine} is computed between two feature vectors $|V_a| = n$ and $|V_b| = n$.

4.3.3 Other classifiers

In the purpose of presenting an objective comparison in terms of accuracy, we selected two common and widely used classification techniques. The first one consists in the Decision Tree (DT) classifier using the common algorithm **C4.5** with *gini index* as *impurity measure.* The other technique consists in a Multi-Layer Perceptron (MLP) that employs the well-known *backpropagation* algorithm during the learning phase.

5 *Frame-based* Version

In the previous Section 4, we described the *sequence-based* version of our approach. As we distinguish two types of *dynamic* AFER systems (see Section 2), we also propose the *frame-based* version. The main objective is to investigate and compare between each version's performance and define the most suitable one to use. Yaddaden et al. (2017) presented a *frame-based* approach that extracts from each frame of an image sequence the spatial representation and then concatenate them to obtain a global one. However, the image sequences have to be *normalized* in order to have the same number of frames per image sequence. In the context of this work, the proposed approach has to deal with the variation of image sequence lengths.

Fig. 5 describes the *frame-based* version. Similarly to the *sequence-based* version, it exploits the same type of *input.* During the *feature extraction*, the *frame-based* version generates a relevant representation of the input using only *spatial* information. The obtained representation is divided into *training* and *validation* sets. Then, we have to select the *highest intensity* frames from each image sequence in training set using the Algorithm 1. It takes the *spatial* representation s_{in} of the image sequence and its corresponding label l_{in} as arguments. It begins by applying a *clustering* technique that is the k-means

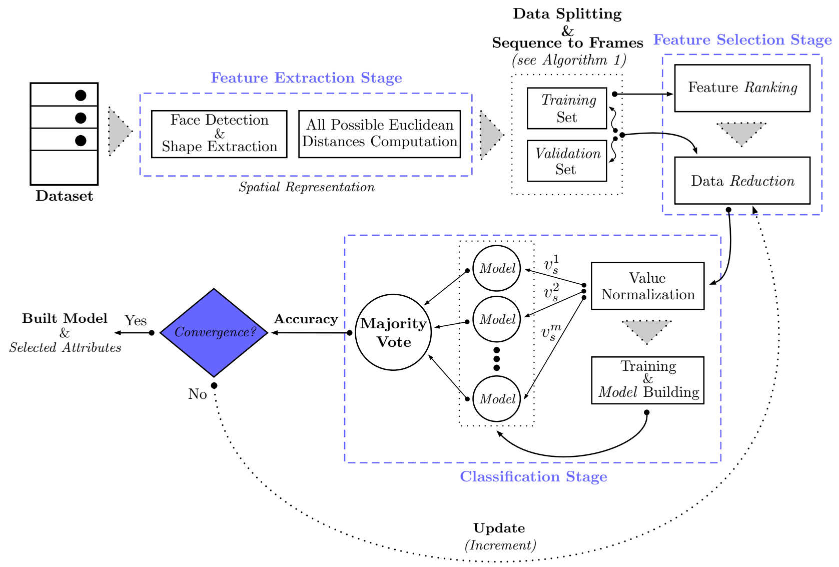


Fig. 5 Overview of our *dynamic* AFER approach (*frame-based* version).

Algorithm 1: *Selection* of frames from an image sequence.

```
Data: Image sequence descriptor $s_{in}$ and label $l_{in}$
Result: Selected Frame descriptors $s_{out}$ and labels $s_{out}$
begin
    $l_{cluster} \leftarrow k\text{-}means(s_{in},\ |clusters| = 2)$;
    create an empty array $s_{out}$;
    create an empty array $l_{out}$;
    for $i = 1$ to $len(l_{cluster})$ do
        if $l_{cluster}[\,i\,] \neq l_{cluster}[\,0\,]$ then
            append $s_{in}[\,i\,]$ to $s_{out}$;
            append $l_{in}$ to $l_{out}$;
        end
    end
end
```

algorithm. It allows assigning a label to each frame depending on its intensity that is stored in $l_{cluster}$ array. Moreover, the only parameter to define consists in the number of clusters that is in our case $|clusters| = 2$. Given the fact that in training set the first frame in the image sequence represents the *neutral* state, we store all the frames that do not belong to the same cluster as the first frame $l_{cluster}[\,0\,]$ in the s_{out} array. We do the same for the corresponding labels and store them in the l_{out} array. In summary, s_{out} contains the feature vectors of an image sequence that are the most representative of a facial expression.

Similarly to the *sequence-based* version described previously in Section 4, the size of the descriptor vectors is reduced using **Extra-Trees** technique. In the *classification stage*, we perform the training using only images (selected frames) after applying the *min-max normalization* technique. As shown in Fig.

5, to identify the facial expression from an image sequence, we classify each frame *independently* before using a *majority vote* function. According to the obtained accuracy, the *threshold* value is adjusted (incremented) until reaching a certain convergence.

6 Experimentation & Evaluation

In this section, we present the validation process for both versions of the proposed *dynamic* AFER approach. We describe the used benchmark datasets and we detail the *validation strategy* and the different performed evaluations.

6.1 Benchmark Datasets

In the context of this work, we employed three distinct benchmark facial expression datasets which are publicly available. Each one consists of image sequences representing a facial expression. As shown in Table 1, we selected the Cohn-Kanade Extended (**CK+**), Multimedia Understanding Group (**MUG**) and Man-Machine Interaction (**MMI**) datasets. The **CK+** (Kanade et al. 2000) dataset is the most popular, it contains 327 labeled image sequences of 123 different subjects expressing seven distinct emotions: happiness (HA), anger (AN), disgust (DI), sadness (SA), fear (FE), surprise (SU) and contempt (CO). The **MMI** (Pantic et al. 2005) dataset contains 199 labeled image sequences of 31 different subjects expressing the six basic emotions. The last dataset consists in **MUG** (Aifanti et al. 2010) dataset and contains the highest number of image sequences, a total of 931. Moreover, 86 subjects were involved to construct the datasets (51 are men and 35 women).

Table 1 Benchmark facial expression datasets.

Dataset		**CK+**	**MMI**	**MUG**
Emotions	HA	69	42	175
	AN	45	28	167
	DI	59	31	153
	SA	28	32	136
	FE	25	28	127
	SU	83	38	173
	CO	18	–	–
	Total	**327**	**199**	**931**
Resolution		640 × 490	768 × 576	896 × 896
States		*Onset-Apex*	*Onset-Apex-Offset*	*Onset-Apex-Offset*
$\sum$ Frames		6 *to* 71	30 *to* 243	11 *to* 179

6.2 Validation Strategy

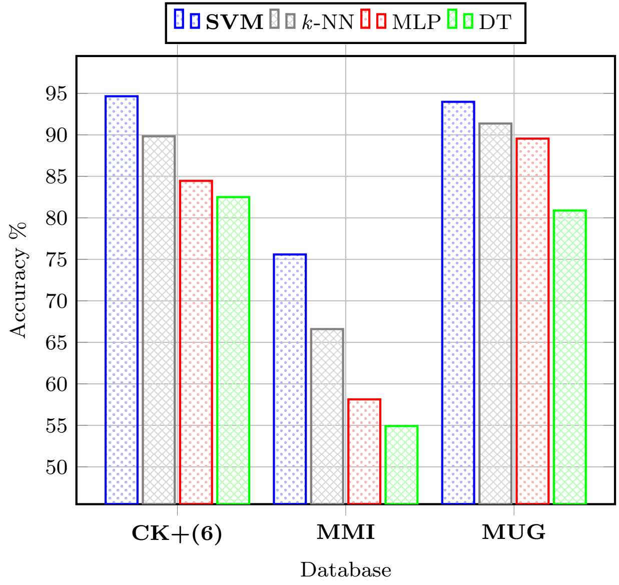
The evaluation of both versions of the proposed *dynamic* AFER approach is performed following the *ten-folds cross-validation* strategy. Thus, the input dataset is divided into ten *stratified* subsets and the evaluation is performed through ten iterations. During each one, nine subsets are used to *train* the classifier and the last one is exploited to evaluate the generated *model* in terms of accuracy. Finally, the resultant accuracy is computed by *averaging* the different values from each iteration. Furthermore, we achieved the evaluation through different steps. We begin by comparing the use of different machine learning techniques and their effects on accuracy. Then, we compare both versions of the proposed *dynamic* AFER approach with the existing methods in terms of accuracy. Finally, we perform a more thorough comparison between the *sequence-based* and *frame-based* versions using several evaluation metrics.

7 Obtained Results

In this section, we describe all the obtained results. Moreover, each one of the different evaluation aspects is detailed independently.

7.1 Machine Learning Techniques

Fig. 6 Accuracy of the proposed *sequence-based* approach.



The Fig. 6 represents the performance of the *sequence-based* version using four different classifiers (namely **SVM**, ***k*-NN**, MLP and DT) and evaluated using three different benchmark facial expression datasets (namely **CK+**, **MMI** and **MUG**).

7.2 **CK+** dataset (six classes)

Table 2 Confusion matrix using **CK+** *(six classes)* dataset.

	FE	**SU**	**HA**	**DI**	**AN**	**SA**
FE	**84.00**	0.00	8.00	0.00	0.00	8.00
SU	0.00	**98.80**	0.00	0.00	0.00	1.20
HA	0.00	0.00	**100.00**	0.00	0.00	0.00
DI	0.00	0.00	0.00	**98.31**	0.00	1.69
AN	0.00	4.44	0.00	0.00	**86.67**	8.89
SA	3.57	3.57	3.57	3.57	3.57	**82.14**
Overall = **91.65**%						

Table 2 represents the obtained *confusion matrix* when evaluating the *sequence-based* version using the **CK+** dataset (*six basic emotions*).

Table 3 Comparison with *state-of-the-art* methods using **CK+** *(six classes)*.

Type	**Approach**	**Features**	**Classifier**	**Accuracy**
Static	Guo et al. (2017)	*Appearance*	SVM	92.30%
Dynamic	Long and Bartlett (2016)			81.40%
	Wei et al. (2015)	*Hybrid*	AE + SR	91.90%
	Kamarol et al. (2017)	*Geometric*	HMM	82.40%
	Yaddaden et al. (2017)		SVM	92.54%
	Proposed (*Sequence*)			**94.64**%
			k-NN	**89.43**%
	Proposed (*Frame*)		SVM	**82.68**%

In Table 3 is compared the proposed *dynamic* AFER approach to other *state-of-the-art* methods in terms of accuracy using the SVM and *k*-NN with the **CK+** dataset (*six basic emotions*).

7.3 **MMI** dataset

Table 4 represents the obtained *confusion matrix* when evaluating the *sequence-based* version using the **MMI** dataset.

Table 4 Confusion matrix using **MMI** dataset.

	FE	**SU**	**HA**	**DI**	**AN**	**SA**
FE	**42.86**	21.43	7.14	7.14	0.00	21.43
SU	7.89	**78.95**	0.00	5.26	0.00	7.89
HA	4.76	0.00	**95.24**	0.00	0.00	0.00
DI	3.23	3.23	3.23	**77.42**	9.68	3.23
AN	0.00	7.14	0.00	7.14	**75.00**	10.71
SA	12.50	6.25	6.25	3.12	0.00	**71.88**
Overall = **73.56**%						

In Table 5 is compared the proposed *dynamic* AFER approach to other *state-of-the-art* methods in terms of accuracy using the SVM and k-NN with the **MMI** dataset.

Table 5 Comparison with *state-of-the-art* methods using **MMI**.

Type	**Approach**	**Features**	**Classifier**	**Accuracy**
Dynamic	Gupta et al. (2018)	*Appearance*	CNN + AE	65.57%
	Lei et al. (2017)		SVM	71.92%
	Liu et al. (2016)			74.63%
	Xijian and Tardi (2015)			74.30%
	Lee et al. (2016)		SRC	70.12%
	Proposed (*Sequence*)	*Geometric*	k-NN	**66.60**%
			SVM	**75.59**%
	Proposed (*Frame*)			**70.11**%

7.4 **MUG** dataset

Table 4 represents the obtained *confusion matrix* when evaluating the *sequence-based* version using the **MUG** dataset.

Table 6 Confusion matrix using **MUG** dataset.

	FE	**SU**	**HA**	**DI**	**AN**	**SA**
FE	**86.61**	8.66	0.00	0.00	1.57	3.15
SU	4.62	**93.64**	0.58	0.58	0.58	0.00
HA	1.14	0.57	**98.29**	0.00	0.00	0.00
DI	0.00	0.00	0.00	**99.35**	0.65	0.00
AN	2.40	0.60	0.00	1.20	**92.81**	2.99
SA	1.47	1.47	0.00	1.47	4.41	**91.18**
Overall = **93.65**%						

In Table 7 is compared the proposed *dynamic* AFER approach to other *state-of-the-art* methods in terms of accuracy using the SVM and k-NN with the **MUG** dataset.

Table 7 Comparison with *state-of-the-art* methods using **MUG**.

Type	**Approach**	**Features**	**Classifier**	**Accuracy**
Static	Maximiano da Silva and Pedrini (2016)	*Geometric*	SVM	90.54%
	Maximiano da Silva and Pedrini (2015)	*Apperance*	ANN	89.23%
Dynamic	Agarwal et al. (2016)		SVM	78.57%
	Peng and Yin (2018)	*Hybrid*	KDA	93.32%
	Niki and Anastasios (2014)	*Geometric*	SVM	92.52%
	Proposed (*Sequence*)			**94.19**%
			k-NN	**91.37**%
	Proposed (*Frame*)		SVM	**79.20**%

7.5 CK+ dataset (seven classes)

In Table 8 is compared the proposed *dynamic* AFER approach to other *state-of-the-art* methods in terms of accuracy using the SVM and *k*-NN with the **CK+** dataset (*six basic emotions* and *contempt*).

Table 8 Confusion matrix using **CK+** *(seven classes)* dataset.

	FE	**SU**	**HA**	**DI**	**AN**	**SA**	**CO**
FE	**84.00**	0.00	12.00	0.00	0.00	4.00	0.00
SU	0.00	**98.80**	0.00	0.00	0.00	0.00	1.20
HA	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00
DI	0.00	0.00	0.00	**96.61**	1.69	1.69	0.00
AN	0.00	2.22	0.00	2.22	**86.67**	8.89	0.00
SA	0.00	3.57	3.57	3.57	3.57	**75.00**	10.71
CO	5.56	0.00	0.00	0.00	0.00	16.67	**77.78**
Overall = **88.41**%							

In Table 9 is compared the proposed *dynamic* AFER approach to other *state-of-the-art* methods in terms of accuracy using the SVM and *k*-NN with the **CK+** dataset (*six basic emotions* and *contempt*).

Table 9 Comparison with *state-of-the-art* methods using **CK+** *(seven classes)*.

Type	**Approach**	**Features**	**Classifier**	**Accuracy**
Static	Ghimire et al. (2017)	*Hybrid*	SVM	91.95%
Dynamic	Xijian and Tardi (2015)	*Appearance*		83.70%
	Xijian and Tardi (2017)			88.30%
	Gupta et al. (2018)		CNN + AE	90.52%
	Lee et al. (2016)		SRC	92.34%
	Proposed (*Sequence*)	*Geometric*	*k*-NN	**87.59**%
			SVM	**92.68**%
	Proposed (*Frame*)			**79.00**%

7.6 Sequence-based vs Frame-based

Fig. 7 illustrates the comparison between both *sequence-based* and *frame-based* versions of the proposed *dynamic* AFER approach in terms of accuracy. It is achieved using the SVM classifier with the three benchmark datasets.

Fig. 7 Comparison between *sequence-based* and *frame-based* approach.

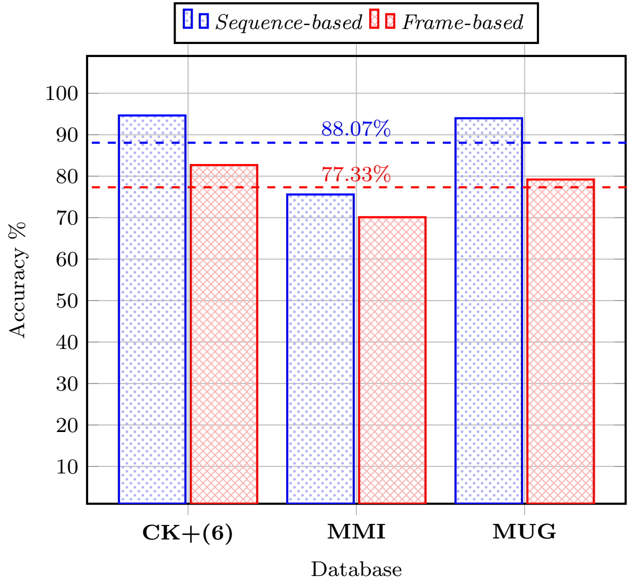


Table 10 represents a more thorough evaluation using five different metrics that might be summarized as follows: *F1_Score*, *Recall*, *Precision*, *Accuracy* and $\sum$ *Attributes* (number of attributes).

Table 10 Comparison between *sequence-based* and *frame-based* using various metrics.

Dataset	**Approach**	*F1_Score*	*Recall*	*Precision*	*Accuracy*	$\sum$ *Attributes*
CK+(6)	*Sequence-based*	**94.27**	**94.65**	**95.84**	**94.65**	**1114**
	Frame-based	82.53	82.68	85.63	82.68	1708
MMI	*Sequence-based*	**74.40**	**75.59**	**77.18**	**75.59**	**806**
	Frame-based	68.32	70.11	73.42	70.11	2050
MUG	*Sequence-based*	**94.12**	**94.19**	**94.49**	**94.19**	**724**
	Frame-based	79.71	79.20	84.54	79.20	2278
CK+(7)	*Sequence-based*	**92.30**	**92.68**	**93.73**	**92.68**	**1988**
	Frame-based	78.76	79.00	81.72	79.00	1139

Table 11 Computational time for the *sequence-based* and *frame-based* versions.

Dataset	**Classifier**	**Approach**	**Computational time** (ms)	
			Training	*Predicting*
CK+(6)	SVM	*Sequence-based*	**6.74**	**0.02**
		Frame-based	1145.25	1.14
	k-NN	*Sequence-based*	**0.31**	**0.16**
		Frame-based	0.65	5.40
MMI	SVM	*Sequence-based*	**6.28**	**0.02**
		Frame-based	2519.00	8.06
	k-NN	*Sequence-based*	**0.31**	**0.17**
		Frame-based	1.48	22.78
MUG	SVM	*Sequence-based*	**8.33**	**0.02**
		Frame-based	2985.30	8.34
	k-NN	*Sequence-based*	**0.31**	**0.17**
		Frame-based	10.00	27.00
CK+(7)	SVM	*Sequence-based*	**10.15**	**0.02**
		Frame-based	1361.30	1.10
	k-NN	*Sequence-based*	**0.29**	**0.14**
		Frame-based	0.57	5.34

In Table 11 is presented an evaluation in terms of *computational time* during training and predicting phases. The measure of the training time is done using six samples (image sequences) for the **CK+** (*six emotions*), **MMI** and **MUG** datasets and seven samples for the **CK+** (*seven emotions*) dataset. The predicting time is measured using a single sample.

7.7 Cross-dataset Evaluation

In order to estimate the generalization capability of both versions of the proposed approach, we performed a *cross-dataset* evaluation. It consists in generating a *model* (after the *training phase*) with a specific dataset and evaluating its performance in terms of accuracy using the other datasets. In Table 12 is represented the obtained results for each versions of the proposed approach.

Table 12 *Cross-dataset* evaluation for both *sequence-based* and *frame-based* versions.

		Training		
Approach	***Dataset***	**CK+**	**MMI**	**MUG**
Sequence-based	**CK+**	**99.68**%	74.11%	76.38%
	MMI	53.77%	**99.50**%	56.28%
	MUG	78.73%	70.57%	**99.89**%
Frame-based	**CK+**	**91.26**%	71.20%	71.52%
	MMI	61.81%	**86.93**%	61.81%
	MUG	63.59%	64.66%	**85.18**%

8 Discussion

The first evaluation highlighted the performance of the different employed classification techniques. Indeed, we notice from the Fig. 6 that the best results were achieved by the two proposed classifiers (namely SVM and k-NN). Moreover, the SVM classifier allowed to reach the highest accuracy: 94.65%, 75.59% and 93.98% with the **CK+**, **MMI** and **MUG** datasets, respectively. Thus, it confirms the choice of two classifiers, especially the SVM for its robustness and generalization capability.

The main objective of this work remains the recognition of the *six basic emotions* (Ekman and Friesen 1971) through facial expressions. From Tables 2, 4, 6 and 8 that represent the *confusion matrix*, we notice that the proposed approach performs better when recognizing the **SU**, **HA** and **DI** emotions with a peak accuracy of 98.80%, 100% and 99.35%, respectively. However, it has issues when identifying the **SA**, **FE** and **CO** emotions with lowest accuracy of 71.88%, 42.86% and 77.78%, respectively. It might be explained by the fact that we have used only a single type of descriptors that provides relevant representations for **SU**, **HA** and **DI** emotions. Moreover, the highest *overall* value is attributed to the **MUG** dataset (93.65%) while the lowest is attributed to the **MMI** (73.56%).

We also compared both versions of our approach to other existing methods in terms of accuracy. From Tables 3, 5, 7 and 9, we notice that the *sequence-based* version outperforms the existing ones with the following accuracy: 94.64%, 75.59%, 94.19% and 92.68% when evaluated with the **CK+** (*six emotions*), **MMI**, **MUG** and **CK+** (*seven emotions*) datasets, respectively. Thus, it confirms the efficiency of the *spatio-temporal* representation combined with the SVM classifier.

The next evaluation is dedicated to the comparison between the two versions of the proposed *dynamic* AFER approach. Thus, from the Fig. 7 and Table 10, we clearly notice that the *sequence-based* version achieve the best performance in terms of accuracy. Moreover, the average accuracy over the three benchmark datasets is estimated to 88.07% and 77.33% for the *sequence-based* and *frame-based* version, respectively. In almost all the cases, the *sequence-based* version considers fewer attributes and it might be seen as an asset since it reduces the processing time and computational load. One of the most important criteria when designing a *dynamic* AFER remains the processing time. From Table 11, we notice that the *sequence-based* version takes less time for both *training* and *predicting* phases than the *frame-based* one. Indeed, the training duration of the *sequence-based* and *frame-based* versions using an SVM classifier is ≈ 7.87 (ms) and ≈ 2002.63 (ms), respectively. Moreover, we notice that the k-NN classifier is faster during the training phase since it does not generate a *model*. However, the SVM classifier performs better during the predicting phase and takes ≈ 0.02 (ms) while the k-NN classifier takes ≈ 0.16 (ms). In summary, the best performance in terms of computational time is achieved using the *sequence-based* version with the *spatio-temporal* representation combined with the SVM classifier.

Finally, we measure the generalization capability of both versions of the proposed approach by performing a *cross-dataset* evaluation. We notice from the Table 12 that the *sequence-based* version performs better in terms of accuracy and takes less computing time. However, the *frame-based* version reached a higher accuracy when training with both **CK+** and **MUG** datasets and evaluating the generated model with the **MMI** dataset.

9 Conclusion

In this paper, we introduced a new *dynamic* AFER approach. Unlike *static* AFER methods, the *dynamic* ones process a larger amount of information and are more realistic since a facial expression includes a temporal aspect through three different *transitional states* (namely *onset*, *apex* and *offset*). Based on the previous works, we proposed two different versions. The *sequence-based* version generates a *spatio-temporal* representation while the *frame-based* one process each frame individually. We have evaluated both versions with three different benchmark datasets (namely **CK+**, **MMI** and **MUG**). The evaluation was performed following different phases and each one provided useful information about the performance of the proposed approach.

Indeed, the first evaluation highlighted the efficiency of the two employed machine learning techniques namely SVM and k-NN. Then, we compared both versions of our approach to existing methods and it has been found that the *sequence-based* version combined with an SVM classifier outperforms the studied methods in terms of accuracy. The next part of the evaluation was dedicated to the comparison between the *sequence-based* and *frame-based* versions. It has been clearly confirmed that the *sequence-based* version is the most effective since it yields the best accuracy while using less attributes. The last evaluation highlighted the time efficiency of the *sequence-based* version in both training and predicting phases.

Nevertheless, the proposed *dynamic* AFER approach has some flaws that need to be overcome. Indeed, it supports only image sequences that contain a *frontal* view of the face. Therefore, we have to improve our approach in order to be able to perform the recognition from other angles. Another issue that has to be investigated consists in the low obtained accuracy with two of the six basic emotions namely sadness and fear. In the context of this work, we employed *geometric-based* features and the inclusion of other types of features might improve the efficiency. Finally, the performance of the proposed *dynamic* AFER approach is closely related to the feature extraction stage that depends on the facial fiducial points detection technique. Moreover, the chosen technique might fail to extract the features which cause the interruption of the recognition process. In summary, several improvements might be provided to the proposed approach and we are planning to achieve them.

Acknowledgements The present work is financially supported by Canadian Foundation for Innovation (CFI) and Université du Québec à Rimouski (UQAR). We also would acknowledge every person and organization for providing the various benchmark facial expression datasets (**CK+** (Kanade et al. 2000), **MMI** (Pantic et al. 2005) and **MUG** (Aifanti et al. 2010)) used for the validation of our approach.

References

Agarwal S, Santra B, Mukherjee DP (2016) Anubhav: recognizing emotions through facial expression. The Visual Computer DOI 10.1007/s00371-016-1323-z

Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Machine learning 6(1):37–66

Aifanti N, Papachristou C, Delopoulos A (2010) The mug facial expression database. In: 11th international Workshop on Image analysis for Multimedia Interactive Services, IEEE, pp 1–4

Castellano G, Kessous L, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech, Springer, pp 92–103. DOI 10.1007/978-3-540-85099-1_8

Cohen I, Sebe N, Garg A, Chen LS, Huang TS (2003) Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and Image Understanding 91(1):160–187, DOI 10.1016/S1077-3142(03)00081-X

Cootes T, Edwards G, Taylor C (2001) Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6):681–685, DOI 10.1109/34.927467

Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models-their training and application. Computer vision and image understanding 61(1):38–59, DOI 10.1006/cviu.1995.1004

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20(3):273–297, DOI 10.1007/BF00994018

Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. Journal of personality and social psychology 17(2):124, DOI 10.1037/h0030377

Ekman P, Rosenberg EL (2005) What the Face Reveals Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, DOI 10.1093/acprof:oso/9780195179644.001.0001

Fasel B, Luettin J (2003) Automatic facial expression analysis: A survey. Pattern Recognition 36(1):259–275, DOI 10.1016/S0031-3203(02)00052-3

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Machine Learning 63(1):3–42, DOI 10.1007/s10994-006-6226-1

Ghimire D, Jeong S, Lee J, Park SH (2017) Facial expression recognition based on local region specific features and support vector machines. Multimedia Tools and Applications 76(6):7803–7821, DOI 10.1007/s11042-016-3418-y

Guo M, Hou X, Ma Y, Wu X (2017) Facial expression recognition using elbp based on covariance matrix transform in klt. Multimedia Tools and Applications 76(2):2995–3010, DOI 10.1007/s11042-016-3282-9

Gupta O, Raviv D, Raskar R (2018) Illumination invariants in deep video expression recognition. Pattern Recogn 76(C):25–35, DOI 10.1016/j.patcog.2017.10.017

Kamarol SKA, Jaward MH, Klviinen H, Parkkinen J, Parthiban R (2017) Joint facial expression recognition and intensity estimation based on weighted votes of image sequences. Pattern Recogn Lett 92(C):25–32, DOI 10.1016/j.patrec.2017.04.003

Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, pp 46–53, DOI 10.1109/AFGR.2000.840611

Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1867–1874, DOI 10.1109/CVPR.2014.241

Konar A, Halder A, Chakraborty A (2015) Introduction to Emotion Recognition, John Wiley & Sons, Inc., pp 1–45. DOI 10.1002/9781118910566.ch1

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444, DOI 10.1038/nature14539

Lee SH, Baddar WJ, Ro YM (2016) Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos. Pattern Recogn 54(C):52–67, DOI 10.1016/j.patcog.2015.12.016

Lei Z, Zengcai W, Guoxin Z (2017) Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and gabor multiorientation fusion histogram. Mathematical Problems in Engineering 2017:1–12, DOI 10.1155/2017/7206041

Liu M, Wang R, Li S, Huang Z, Shan S, Chen X (2016) Video modeling and learning on riemannian manifold for emotion recognition in the wild. Journal on Multimodal User Interfaces 10(2):113–124, DOI 10.1007/s12193-015-0204-5

Long F, Bartlett MS (2016) Video-based facial expression recognition using learned spatiotemporal pyramid sparse coding features. Neurocomput 173(P3):2049–2054, DOI 10.1016/j.neucom.2015.09.049

Louppe G, Wehenkel L, Sutera A, Geurts P (2013) Understanding variable importances in forests of randomized trees. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems 26, Curran Associates, Inc., pp 431–439

Mehrabian A (1968) Communication without words, 2nd edn, pp 51–52

Niki A, Anastasios D (2014) Linear subspaces for facial expression recognition. Signal Processing: Image Communication 29(1):177 – 188, DOI doi.org/10.1016/j.image.2013.10.004

Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: IEEE International Conference on Multimedia and Expo, IEEE, pp 5–pp, DOI 10.1109/ICME.2005.1521424

Peng Y, Yin H (2018) Facial expression analysis and expression-invariant face recognition by manifold-based synthesis. Machine Vision and Applications 29(2):263–284, DOI 10.1007/s00138-017-0895-6

Shalev-Shwartz S, Ben-David S (2014) Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, DOI 10.1017/CBO9781107298019

Maximiano da Silva FA, Pedrini H (2015) Effects of cultural characteristics on building an emotion classifier through facial expression analysis. Journal of Electronic Imaging 24:24 – 24 – 9, DOI 10.1117/1.JEI.24.2.023015

Maximiano da Silva FA, Pedrini H (2016) Geometrical features and active appearance model applied to facial expression recognition. International Journal of Image and Graphics 16(04):1650019, DOI 10.1142/S0219467816500194

Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, vol 1, pp I–I, DOI 10.1109/CVPR.2001.990517

Wang L, He DC (1990) Texture classification using texture spectrum. Pattern Recognition 23(8):905–910, DOI 10.1016/0031-3203(90)90135-8

Wei Z, Youmei Z, Lin M, Jingwei G, Shijie G (2015) Multimodal learning for facial expression recognition. Pattern Recognition 48(10):3191–3202, DOI 10.1016/j.patcog.2015.04.012

Xijian F, Tardi T (2015) A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. Pattern Recognition 48(11):3407–3416, DOI 10.1016/j.patcog.2015.04.025

Xijian F, Tardi T (2017) A dynamic framework based on local zernike moment and motion history image for facial expression recognition. Pattern Recognition 64:399–406, DOI 10.1016/j.patcog.2016.12.002

Yaddaden Y, Bouzouane A, Adda M, Bouchard B (2016) A new approach of facial expression recognition for ambient assisted living. In: Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, ACM, p 14, DOI 10.1145/2910674.2910703

Yaddaden Y, Adda M, Bouzouane A, Gaboury S, Bouchard B (2017) Facial expression recognition from video using geometric features. In: Proceedings of the 8th International Conference on Pattern Recognition Systems, IET, pp 1–6, DOI 10.1049/cp.2017.0133

Yaddaden Y, Adda M, Bouzouane A, Gaboury S, Bouchard B (2018) User action and facial expression recognition for error detection system in an ambient assisted environment. Expert Systems with Applications 112:173–189, DOI 10.1016/j.eswa.2018.06.033

Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: A literature survey. ACM Comput Surv 35(4):399–458, DOI 10.1145/954339.954342

Chapitre 6

Titre :

Reconnaissance Automatique des Actions Utilisateurs et des Expressions Faciales pour la Détection d'erreur dans un Environnement d'assistance Ambiante

Résumé - *Ce chapitre décrit un système d'assistance qui s'intègre dans un environnement intelligent. Ainsi, il sera constitué de deux principaux composants pour la reconnaissance des actions utilisateur durant l'activité et la détection automatique d'éventuelles erreurs. La première opération se fera à l'aide de tags* ***RFID*** *disposés sur les objets qui seront manipulés durant les expérimentations. À partir de la position estimée de chaque objet, les différentes actions de l'activité réalisée sont identifiées. Sur la base de l'analyse de ces informations, il est possible de détecter la présence d'erreurs ou d'anomalies. Afin d'améliorer les performances en termes de détection d'erreurs, nous proposons d'utiliser les expressions faciales. Pour ce faire, nous proposons une architecture de réseau de neurones à convolution inspirée de* ***LeNet-5*** *proposée par LeCun et al. [39]. L'architecture proposée est évaluée avec cinq différentes bases de données :* ***JAFFE****,* ***RaFD****,* ***KDEF****,* ***MMI*** *et* ***CK+****. Les résultats obtenus attestent de l'efficacité de notre réseau en termes de taux de reconnaissance. Nous l'avons donc utilisé pour améliorer la détection d'erreurs à travers les expressions faciales et il s'est avéré qu'il permet de réduire considérablement le taux de fausses détections.*

Mots clés :

Environnement Intelligent, Reconnaissance d'activité, Identification par Radiofréquence, Reconnaissance d'expressions Faciales, Réseau de Neurones à Convolution, Détection d'erreur

Contributions associées :

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "User action and facial expression recognition for error detection system in an ambient assisted environment," in *Journal of Expert Systems with Applications*, Elsevier, vol. 112, 2018, pp. 173-189. (Statut : Publié)

Y. Yaddaden, M. Adda, A. Bouzouane, S. Gaboury, B. Bouchard, "Facial expressions based error detection for smart environment using deep learning," in *Ubiquitous Intelligence Computing (UIC)*, IEEE, 2017, pp. 1-7. (Statut : Présenté)

Expert Systems With Applications 112 (2018) 173–189

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

User action and facial expression recognition for error detection system in an ambient assisted environment

Yacine Yaddaden[a,*], Mehdi Adda[a,b], Abdenour Bouzouane[a], Sébastien Gaboury[a], Bruno Bouchard[a]

[a] *Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA), Département d'informatique et de mathématique, Université du Québec à Chicoutimi (UQAC), 555 Boulevard de l'Université, Chicoutimi, Québec G7H 2B1, Canada*

[b] *Département de mathématiques, d'informatique et de génie, Université du Québec à Rimouski (UQAR), 300 Allée des Ursulines, Rimouski., Québec G5L 3A1, Canada*

ARTICLE INFO

Article history:
Received 27 January 2018
Revised 1 May 2018
Accepted 14 June 2018
Available online 15 June 2018

Keywords:
Smart environment
Activity recognition
Radio frequency IDentification
Facial expression recognition
Convolutional neural network
Error detection

ABSTRACT

Emotion recognition through facial expressions represents a relevant way to understand and even predict the human behavior. Thus, it has been used in various fields such as human-robot interaction and ambient assistance. Nevertheless, it remains a challenging task since expressed emotions might be affected by different parameters such as ethnic origins, age and so on. In this paper, we introduce an efficient facial expression recognition approach based on a Convolutional Neural Network architecture. Carried experimentation on five benchmark facial expression datasets confirms the efficiency of the proposed approach with recognition rates higher than **95%**. In the context of ambient assistance, we introduced an error detection module using a user action recognition from Radio Frequency IDentification tags placed on various objects of daily living. The experiments performed in a smart environment show a consistent improvement of the error detection module when including facial expression recognition. Indeed, the false positive detection rate is significantly reduced by over **20%**.

1. Introduction

Since the dawn of time, human being's interaction and communication were achieved through emotions and until nowadays, they play a crucial role in daily interactions (Ekman & Friesen, 1971). With the advent of Information and Communications Technology (ICT), automatic emotion recognition capability becomes necessary when designing devices used by human beings. Indeed, emotions allow a more intuitive Human-Computer Interaction (HCI) that enhances significantly the user's experience and the effectiveness of the designed devices. Thus, automatic emotion recognition found several applications in various fields such as entertainment, surveillance, e-learning (Ammar, Neji, Alimi, & Gouardères, 2010), healthcare (Tang, Yusuf, Botzheim, Kubota, & Chan, 2015), human-robot interaction (Zhang, Jiang, Farid, & Hossain, 2013). Emotions might be represented by various *modalities* such body gait and gesture, speech and facial expressions (Castellano, Kessous, & Caridakis, 2008). Accordingly, we distinguish two types of automatic emotion recognition systems. *Unimodal* employ a unique source of information while *multimodal* systems combine several ones to improve the performance in terms of accuracy (Perez-Gaspar, Caballero-Morales, & Trujillo-Romero, 2016; Zhang et al., 2013). Moreover, in Mehrabian (1968), it is stated that during an interaction, 55% of the information is transmitted through facial expressions while the verbal and vocal parts contribute with 7% and 38%, respectively. This study highlights the relevancy of using facial expressions to recognize the emotional state.

Abbreviations: **ICT**, Information and Communications Technology; **HCI**, Human-Computer Interaction; **SH**, Smart Home; **AmI**, Ambient Intelligence; **ADL**, Activity of Daily Living; **AFER**, Automatic Facial Expression Recognition; **DL**, Deep Learning; **CNN**, Convolutional Neural Network; **ARM**, Action Recognition Module; **EDM**, Error Detection Module; **RFID**, Radio Frequency IDentification; **PCA**, Principal Component Analysis; ***k*** **-NN**, *k*-Nearest Neighbors; **SVM**, Support Vector Machine; **HOG**, Histogram of Oriented Gradient; **PHOG**, Pyramid Histogram of Gradients; **CASAS**, Center for Advanced Studies in Adaptive Systems; **COACH**, Cognitive Orthosis for Assisting aCtivities in the Home; **POMDP**, Partially Observable Markov Decision Process; **TEBRA**, TEeth BRushing Assistance; **IN**, Intensity Normalization; **FE**, FEar; **SU**, SUrprise; **HA**, HAppiness; **DI**, DIsgust; **AN**, ANger; **SA**, SAdness; **NE**, NEutral *state*; **HE**, Histogram Equalization; **AAL**, Ambient Assisted Living; **DT**, Decision Tree; **TP**, True Positive *detection*; **FP**, False Positive *detection*; **JAFFE**, JApanese Female Facial Expression; **RaFD**, Radboud Faces Database; **KDEF**, Karolinska Directed Emotional Faces; **MMI**, Man-Machine Interaction; **CK+**, Cohn-Kanade Extended; **LIARA**, Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités; **DOMUS**, labotatoire de DOMotique et informatique mobile Ã l'Université de Sherbrooke.

[*] Corresponding author.

E-mail addresses: yacine.yaddaden1@uqac.ca (Y. Yaddaden), mehdi_adda@uqar.ca (M. Adda), abdenour_bouzouane@uqac.ca (A. Bouzouane), sebastien_gaboury@uqac.ca (S. Gaboury), bruno_bouchard@uqac.ca (B. Bouchard).

https://doi.org/10.1016/j.eswa.2018.06.033
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

During the last decades, several initiatives have been undertaken to design ICT-based ambient assistance systems. The main purpose is allowing the elderly to keep their independence and stay at home while ensuring their well-being. Moreover, one of the most common ways to provide assistance for the aging population remains the Smart Home (SH). It might be seen as a common house equipped with Ambient intelligence (AmI) technological devices. Basically, we distinguish three main components. The *sensors* collect raw data directly from the smart environment, the *actuators* provide responses according to the needs of the user and the *processing unit* represents the computation component that performs both activity recognition and error detection. Usually, indoor assistance aims to furnish, when it is necessary, adequate assistance to the elderly when performing their Activities of Daily Living (ADL). Furthermore, automatic emotion recognition has been integrated into SH in order to achieve various purposes such as monitoring mood and mental state, stress level, pain presence. However, designing automatic emotion recognition systems remains a challenging task. Indeed, emotions might vary depending on the used modality, how the person expresses them, acquisition device and so on. Designing an intuitive ambient assisting system might also be challenging. Even if the existing systems offer a relatively good error detection rate, they still suffer from a high false positive detection rate. Which means that the systems tend to trigger assistance when it is not necessary which in turn leads to a lack of intuitiveness.

In this paper, we introduce a new robust Automatic Facial Expression Recognition (AFER) approach. It employs a new trend of supervised machine learning technique derived from Deep Learning (DL). Indeed, we propose an *optimized* Convolutional Neural Network (CNN) architecture inspired from the popular **LeNet-5** and introduced by LeCun, Bottou, Bengio, and Haffner (1998). It allows identification of expressed emotions through facial expressions from frontal face images. The proposed *static* AFER approach also includes a preprocessing stage that enhances significantly its performance in terms of accuracy. In the context of ambient assistance, we propose a system that includes a user Action Recognition Module (ARM) and an Error Detection Module (EDM). Its main purpose consists in recognizing the user actions during the realization of an ADL, it also detects the presence of any anomaly or error. The raw data is retrieved from Radio Frequency IDentification (RFID) tags placed on various objects of daily living and the presence of potential errors is detected after analysis of the recognized user actions (Belley, Gaboury, Bouchard, & Bouzouane, 2015). Moreover, we investigate the potential contribution of the proposed *static* AFER approach to improve the EDM performance. Our work presented in this paper includes several contributions that might be summarized as follows:

- A robust and accurate CNN architecture optimized for *static* AFER,
- An efficient user ARM based on the RFID technology,
- An EDM enhanced by the proposed *static* AFER approach.

The evaluation of the presented work is achieved through two sets of experiments. The first one aims to evaluate the performance of the proposed *static* AFER approach. Indeed, we performed experimentation involving the use of five distinct benchmark facial expression datasets. Each one includes frontal face images of participants expressing the *six basic emotions* namely fear, surprise, happiness, disgust, anger and sadness in addition to the neutral state. Moreover, the evaluation does not only consider the recognition rate aspect but also the enhanced performance of the proposed CNN architecture when compared to **LeNet-5** in terms of convergence and accuracy. Carried experimentation also highlights the effect of adding preprocessing steps and image enhancement techniques. The main purpose consists in measuring the generalization capability of the proposed *static* AFER approach before using it in the context of ambient assistance. The second set of experiments is conducted in a testing smart environment following a proposed experimentation protocol. The main purpose is to collect raw data (objects position and frontal face images) when the participants perform ADL following four different predefined scenarios. The collected datasets are exploited to evaluate the different component of the proposed ambient assisting system namely the user ARM and EDM. Moreover, it allows highlighting the contribution of the proposed *static* AFER approach in enhancing the system's performance.

This paper is organized as follows: In Section 2, we introduce basic knowledge related to assisting systems deployed in smart environments and fundamentals about AFER systems. We also present and discuss existing methods and related works. In Section 3 and 4, we introduce an overview of the proposed *static* AFER approach and how it contributes to enhance the performance of an ambient assisting system. In Section 5, we describe and detail the validation process of the proposed *static* AFER approach and the experimental protocol for the conducted experiments in a testing smart environment. In Section 6, we present and analyze obtained results before providing discussions and interpretations in Section 7. Finally, in Section 8 we present concluding remarks and further works we are planning to achieve.

2. Background

In this section, we introduce fundamentals about AFER and related works in this field. Since the proposed *static* AFER approach is mainly based on a DL technique, we presents basis about this kind of learning algorithms. Moreover, we introduce briefly generalities about ambient assisting systems integrated in smart environments.

2.1. Automatic facial expressions recognition

Affective computing is an active field of study since it enables the recognition of human's emotional state through specific *modalities* such as speech, body gestures and facial expressions. The last one provides an important amount of information and allows a relatively good emphasis on the emotional state (Mehrabian, 1968). Each year, several AFER methods are proposed aiming to enhance the existing ones. Most of them use the same system architecture which is composed of the same building blocks as a common pattern recognition system (Konar, Halder, & Chakraborty, 2015).

Before going further and present existing AFER methods, we briefly detail the different blocks as shown in Fig. 1. In the context of AFER, the system might handles two categories of data input: (1) *static* that might be defined as either 2D or 3D images, (2) *dynamic* which are basically videos or image sequences that take into account the temporal aspect (Khan, Xu, Chan, & Yan, 2017). Depending on the input, the *feature extraction* stage provides a better representation of the data input by extracting pertinent descriptors. Various types of features might be used to design an AFER system, they are divided into three distinct categories (Konar et al., 2015): (1) *geometric-based* features rely on facial fiducial points to compute specific distances and use them as a feature vector, (2) *appearance-based* descriptors exploit the entire image to generate a more pertinent textural representation through various transformation techniques, (3) *hybrid-based* combine the two previous features type in order to enhance the system's performance, it might be achieved either in features extraction or classification level. The size of the extracted feature vectors might be significant. They might also include redundant and noisy information that affects the system's performance. Therefore, another but not mandatory stage might be applied and consists in *feature selection* or *features reduction*. Its purpose is to select the more relevant and

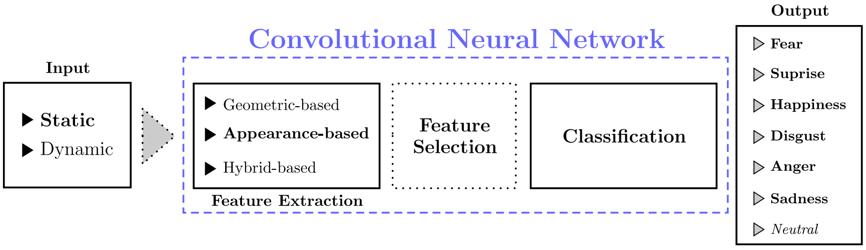


Fig. 1. Overview of a common **AFER** system.

representative attributes while removing the noisy and redundant ones. Moreover, this operation does not just enhance the accuracy but it also increases the system's processing speed. The most common existing technique remains the Principal Component Analysis (PCA). The last stage consists in *classification* and relies on a *supervised* machine learning technique and allows recognizing unlabeled data. However, a model has to be built and requires a training set of labeled data. The final system's *output* consists in the different emotions or facial expressions that might be identified. In fact, most of the existing are based on Ekman's works who considers that human beings can express *six basic emotions*: fear (FE), surprise (SU), happiness (HA), disgust (DI), anger (AN) and sadness (SA) (Ekman & Friesen, 1971). In Fig. 1 is shown as outputs the six basic emotions with the *neutral* state (NE).

Among the existing AFER methods, we have selected some recent ones to review. Thus, Uçar, Demir, and Güzeliş (2016) proposed an approach based on *Discrete Curvelet Transform* that allows computing coefficients after extracting the face region with a technique introduced in Viola and Jones (2001) and applying *Histogram Equalization*. From the extracted features, the authors compute statistical measures such as entropy, standard deviation and mean values before feeding a classifier composed of *Spherical Clustering* and *Online Sequential Extreme Learning Machine*. (Yaddaden, Adda, Bouzouane, Gaboury, & Bouchard, 2017; Yaddaden, Bouzouane, Adda, & Bouchard, 2016) introduced two different methods for either *static* and *dynamic* input. In both cases, the feature vector consists of all the possible Euclidean distances computed using extracted facial fiducial points. The authors have also introduced a method of feature selection based on *variance*. For the classification part, *k*-Nearest Neighbors (*k*-NN) and Support Vector Machine (SVM) have been exploited for the first and second proposed method, respectively. Ali, Zhuang, and Ibrahim (2017) presented a method which has proven its efficiency when applied to two benchmark facial expression datasets. The authors employed the Histogram of Oriented Gradient (HOG) in order to construct a feature vector that feeds a *Sparse Representation* based classifier allowing the recognition of various emotions. Fan and Tjahjadi (2015) proposed an AFER system to recognize emotions through facial expression image sequences. It detects the facial fiducial points in order to perform a face alignment. Then, it extracts two dynamic features that consist in *Pyramid Histogram of Gradients* (PHOG) in *Three Orthogonal Planes* and *Dense Optical Flow*. The two resultant feature vectors are concatenated before feeding an SVM classifier. Similarly, the method proposed by Fang et al. (2014) processes image sequences, but the authors preferred to employ geometric-based features. To achieve this, the method extracts the face region using the Viola et Jones technique then, it detects and tracks the different facial fiducial points. Based on their positions, several features are extracted such as *Fourier Coefficients, Polynomial Fitting*, PCA and so on. For the recognition part, the authors employed different techniques but the best performance was achieved using the *Fuzzy-Rough based Nearest Neighbor algorithm*. In Jiang and Jia (2016), the authors employed the *k*-NN to select the testing or validation set of samples. Moreover, their AFER method is based on the use of *Two-Dimensional Local Discriminative Component Analysis* algorithm for both feature extraction and classification. In Zhang et al. (2013), the authors proposed a multimodal system for HCI including speech and facial expressions. The proposed AFER method extracts seventeen *Action Units* using a built upper and lower facial action analyzers based on supervised neural networks. Using the resultant *Action Units*, a neural network classifier is trained to recognize the six basic emotions. The method proposed in Samara, Galway, Bond, and Wang (2017) extracts forty-nine facial fiducial points before computing all the possible Euclidean distances. The resultant feature vector feeds a multi-class SVM classifier for recognition. In Alphonse and Dharma (2017), the authors proposed a method that begins by detecting and extracting the face region from the input images. Then, the extract *Enhanced Gabor* and PHOG features before combining them to generate a unique descriptor vector. A dimension reduction is performed using the *Pearson General Kernel-based Discriminant Analysis* technique before classification employing *Pearson General Kernel-based Extreme Learning Machine*. In Zavaschi, Britto, Oliveira, and Koerich (2013) is proposed a method that employs two appearance-based features namely *Local Binary Patterns* and *Gabor filters*. Then, an *Multi-Objective Genetic Algorithm* is used for classification. It search for the best ensemble using accuracy and ensemble's size as objective functions. In Owusu, Zhan, and Mao (2014), the authors introduced a method that beings by detecting and extracting the face region using the Viola et Jones technique before reducing the dimension of the input by applying the *Bessel Transform*. Then, it extract the *Gabor* related features and applying the *AdaBoost* feature reduction technique. The resultant feature vectors descriptors feed an supervised neural network for the recognition. In Mlakar, Fister, Brest, and Potočnik (2017) is presented a method that employs HOG features before applying a *Multi-Objective Differential Evolution* oriented feature selection technique. The classification part of the method is achieved through an SVM classifier.

Based on this review, improving the existing systems requires to optimize each one of the different building blocks. It might be achieved by choosing the adequate combination of feature type, selection and classification technique. However, it might not be enough since the performance varies from a dataset to another. Therefore, new approaches have to be proposed in order to enhance the generalization aspect through the different datasets.

2.2. Deep learning

It might be noticed from the previous section that building an efficient pattern recognition system requires careful engineering in order to design a descriptor extractor allowing to transform raw data into more relevant representations. They might be seen as feature vectors that feed a classifier in order to detect or recognize specific patterns. Thus, systems with *hand-crafted feature engineering* capability include a considerable limitation as it is required to design the feature extractor according to the type of data inputs and patterns targeted for recognition. Recently, a new trend of machine learning techniques emerged namely *Deep Learning* allowing to automatically discover the adequate and relevant representations from raw data such as images. Indeed, they enable the extraction of several representations levels beginning from the lower level input to higher and more abstract one. LeCun, Bengio, and Hinton (2015) demonstrate how DL oriented techniques integrate several layers and how each one extracts specific level of representation. In the case of an image, the first layer of representation defines the presence or absence of edges at specific orientations and locations. The next one allows detection of motifs by spotting particular arrangements of edges. The third layer allows to combine the detected motifs in order to spot parts of the object to detect. The last layers might merge the detected parts to match the entire object.

Over the years, several DL related techniques have been introduced such as *Deep Belief Network, Recurrent Neural Network*, CNN, *Stacked Auto Encoder* and other derived techniques (Bengio et al., 2009). In the context of computer vision, CNN are the most solicited as they allow to resolve complex problems of object detection and recognition from images or videos (LeCun, Kavukcuoglu, & Farabet, 2010). A common CNN classifier is composed of three distinct components, each one perform a specific task (Aghdam & Heravi, 2017; Bengio et al., 2009). The first one consists in *convolutional layer* and allows to extract the relevant descriptors from the input image. It might be described by the Eq. (1).

$$S(i,j) = \sum_{m}\sum_{n} I(i-m, j-n)K(m,n) \tag{1}$$

This operation enables extracting the *feature map* represented by *S* from the input image *I*. It is achieved using the *kernel function* represented by *K*. It relies on two different parameters *m* and *n* which represent the height and width of the *kernel function* window, respectively. *i* and *j* define the location of the pixel in the input image *I* and output feature map *S*. The other component is *detector layer* and consists in applying a *nonlinear activation function*. Its main aim is to ensure that the representation in the input space is mapped to a different space in the output. Several functions have been proposed and used such as *Sigmoid, Rectified Linear Unit* and *Hyperbolic Tangent*. The last component is called *pooling layer* and it enables replacing a group of close pixels by a statistical summary. Various methods are available such as *average-pooling* and *max-pooling*. In the context of our work, we chose to use the last one since it yields the best performance (see Eq. (2)).

$$y_{i,j,k} = \max_{pq}(x_{i,j+p,k+q}) \tag{2}$$

The value of the statistical summary is represented by *y*, *i* represents the *i*th *feature map* and *j, k* are the co-ordinates. There are two important parameters *p, q* to set and represent the neighborhood to which is applied the *max-pooling* operation. The main advantage of applying such operation lies on ensuring that the output representation become *invariant* to small translations of the input. An additional component that consists in *fully-connected hidden layer* is required to provide the adequate output format. It allows to map the different *feature maps* into a one-dimensional vector. Even if the CNN has been widely used, it still has some limitations and the most critical one remains the *overfitting*. Basically, it happens when the built model fits too well to the training set. Then, it becomes difficult for the model to generalize to new examples that were not in the training set. In order to overcome such issue, several techniques have been proposed and the most common remain: *data augmentation* that consists in increasing the dataset's size by generating synthetic images from original ones and *dropout layers* (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) allowing to ignore a certain number of hidden units based on a specific probability.

Several works have been conducted to proposed AFER methods using DL techniques. Thereby, Shan, Guo, You, Lu, and Bie (2017) proposed an AFER method based on a Deep CNN. It begins by detecting and extracting the face region from the input image using the Viola et Jones technique. Then, it applies the *histogram equalization* as image enhancement technique before feeding a CNN. Lopes, de Aguiar, De Souza, and Oliveira-Santos (2017) introduced a AFER method that consists of two stages. The first one is preprocessing and lies on the application of operations on face images: detection, alignment, extraction and resizing. It also includes image enhancement through the application of *intensity normalization*. To overcome the *overfitting* issue, the authors have used *data augmentation* by generating synthetic images with different orientations. For the classification part, the authors proposed to exploit two different architectures: 1) a CNN with seven different output corresponding to the six basic emotions in addition to the neutral state, 2) a combination of binary CNN, each one to identify a specific emotion. The last architecture achieved the best performance while rendering the model more complex. In Chen, Yang, Wang, and Zou (2017) is proposed another CNN architecture that includes *Batch Normalization* layer to improved its performance. (Mollahosseini, Chan, & Mahoor, 2016) presented a CNN based AFER method using pre-trained models: *GoogleNet* and *AlexNet*. In Zavarez, Berriel, and Oliveira-Santos (2017), the proposed method takes as input the face image and the eyes position for face alignment before grayscale conversion. Moreover, the method includes data augmentation to reduce the *overfitting* effect. In terms of classification, the authors fine-tuned pre-trained CNN models using face images and called VGG-Face. Similarly, Mavani, Raman, and Miyapuram (2017) fine-tuned an existing and pre-trained CNN model for AFER recognition. Moreover, they introduced optimizations by tuning the hyper-parameters. Sun, Zhao, and Jin (2017) introduced a new AFER method that begins by pre-processing steps such as face detection, alignment and extraction before applying *illumination normalization*. Then, it extracts the *Multi-scale Dense Local Binary Pattern* as descriptors. The resultant feature vector feeds a *Stacked Binary Auto-Encoder* for *unsupervised* feature learning. The classification is achieved using a deep *Binarized Neural Network*. In Liu, Li, Shan, and Chen (2013), the authors presented an *AU-aware Deep Networks* composed of three main components. An *Over-complete Representation* that exploits a convolutional and a pooling layer, *AU-aware Receptive Fields* that select the relevant features to keep, *Hierarchical Feature Learning* that uses *Restricted Boltzmann Machine*. The recognition is achieved using a linear SVM for each feature vector. In Mengyi, Shaoxin, Shiguang, Ruiping, and Xilin (2014), the authors proposed to use a 3D CNN to recognize facial expressions from image sequences. In Ruiz-Garcia, Elshaw, Altahhan, and Palade (2017) is proposed a *Stacked Convolutional Auto Encoder* that basically combines two DL oriented techniques to form a more efficient AFER method. In order to enhance their system, the authors have used a *Batch Normalization* technique. In Shin, Kim, and Kwon (2016), the authors evaluated the performance of four different CNN architectures with five benchmark facial expression datasets. Moreover, they investigated the effect of preprocessing steps and image enhancement techniques in improving the global accuracy.

In the above review of existing AFER methods based on DL oriented techniques, most authors focused on achieving the highest recognition rates. However, in the context of our work, we do not only focus in reaching the highest accuracy but also *lighten* the CNN architecture in order to ease the implementation in embedded systems with limited hardware resources. To achieve this objective, we designed our own *light* CNN architecture inspired from **LeNet-5** proposed in Zhang et al. (2013). We have improved its efficiency by tuning the different parameters. A detailed description of

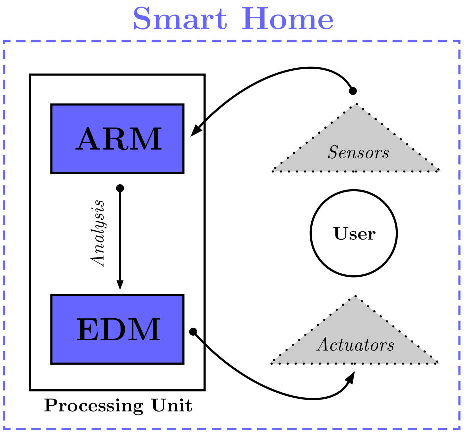


Fig. 2. Components of a basic *ambient assisting* system.

the proposed CNN architecture will be presented in the next section.

2.3. Smart environment & ambient assistance

One of the most common and popular technological-based way to provide assistance to the elderly remains the SH. Nowadays, several experimental SH have been implemented in different laboratories as testing environments (Tomita, Russ, Sridhar et al., 2010). Their main purpose consists in the validation of designed ambient assisting systems using collected datasets related to the interaction of users with the testing smart environment. Laboratoire d'Intelligence Ambiante pour la Reconnaissance d'Activités (LIARA) (Fortin-Simard et al., 2015) and labotatoire de DOMotique et informatique mobile Ã l'Université de Sherbrooke (DOMUS) (Bouchard, Giroux, & Bouzouane, 2006) represent two experimental SH which share nearly the same architecture and technology. Indeed, DOMUS might be described as a common apartment with a bathroom, living room, kitchen and dining room. It contains more than two hundred sensors including twenty RFID readers, various infrared sensors and motion detectors. DOMUS has been designed to ease the development of approaches aiming to provide pervasive cognitive assistance to the elderly and people suffering from dementia and related diseases. On the other hand, LIARA has the same rooms where are placed more than one hundred sensors hidden as much as possible such that the user feel that they are in a standard apartment. In this experimental SH, the most used sensors remain the passive RFID technologies (tags & antennas) but it also includes electromagnetic sensors, accelerometers, load cells, ultrasonic sensors, and much more. In order to provide prompt assisting services to the resident whenever needed, LIARA lab includes several actuators such as few screens to show video guidance and also IP speakers installed in every corner.

Smart environments alone offer a relatively low level of assistance. Therefore, more complex *ambient assisting* systems have to be developed and integrated to these existing infrastructures in order to furnish higher level assistance such as activity recognition, error detection and even predicting the user's behavior based on his habits. In the following, we present some existing systems that provide assistance during the performance of various ADL. Fig. 2 represents the mains components of a common ambient assisting system integrated to a smart environment.

In Belley et al. (2015), the authors introduced a *nonintrusive* ambient assisting system integrated to the LIARA lab. It main purpose is to provided assistance and guidance to the elderly and people with cognitive disabilities during the performance of ADL. Moreover, the proposed system relies on RFID technology to collect datasets while the actuators that provide cue and guidance consists in lights, screens and speakers. The carried experimentation might be described by two scenarios related to the preparation of breakfast. Thus, the proposed ambient assisting system begins by identifying details about the performed activity. After analysis, it detect the presence of erratic behaviors and provides specific guidance depending on the user's profile.

Mihailidis, Boger, Craig, and Hoey (2008) introduced a system called Cognitive Orthosis for Assisting aCtivities in the Home (COACH) whose main purpose is to assist the elderly and people suffering at various levels of cognitive impairment during the activity of hand-washing. The system performs three different actions: (1) *tracking* allowing the system to detect the hand and towel positions from captured video data, (2) *decision-making* based on Partially Observable Markov Decision Process (POMDP) and enables decision making in conditions of uncertainty, (3) *prompting* that provides guidance to the individual through audio and audio-video prompts with three levels of assistance (minimal, maximal and maximal with video demonstration). The system has been tested with six older adults suffering from moderate-to-severe dementia. The efficiency of the system has been proven by the increase of independence when performing hand-washing.

Another example of assisting system for ADL in bathroom is TEeth BRushing Assistance (TEBRA) proposed by Peters, Hermann, and Wachsmuth (2013). The system exploits various sensors including two 2D cameras observing the scene from an overhead and a frontal perspective, a toothbrush equipped with sensors (accelerometer, gyroscope and magnetometer) and a flow sensor is installed in the water pipe. From these sensors, ten different features are extracted and discretized into an intermediate representation of state space variables. For the recognition part, the authors proposed the use of *Bayesian Network* to classify the user behaviors based on the computed variables. Assistance should only be given to the user when necessary in order to foster his independence. Therefore, the system uses a *Finite State Machine* to model the timing behavior of the TEBRA system and detect any anomaly. The system provides two types of visual prompts (pictograms and teal-time videos) including verbal command. The authors validated their system with thirteen regular users and the obtained results were promising.

Bouchard, Giroux, and Bouzouane (2007) introduced a keyhole plan recognition model in order to assist aging people suffering from Alzheimer's disease at early-intermediate stages. The proposed approach allows performing error detection during ADLs executed in a kitchen. The presented approach is based on the work of Baum and Edwards (1993). Indeed, they state that a patient suffering from Alzheimer's disease may commit six different categories of errors during an ADL namely initiation, organization, realization, sequence, judgment and completion. The proposed ambient assisting system is based on a plan recognition process allowing to interpret a set of observed actions. The system is also able to predict the errors before they happen. In order to validate their system, several experiments have been conducted in the DOMUS lab where various sensors are exploited to collect low-level data.

Jean-Baptiste, Rotshtein, and Russell (2016) proposed an automatic prompting system to guide stroke survivors during ADL called CogWatch. The proposed ambient assisting system consists of various modules: (1) *monitoring module* aiming to collect data provided by sensors placed in the different involved objects and by the camera which tracks the user hands coordinates, (2) *action recognition system* that exploits *Hidden Markov Models* to infer the user's current action, (3) *task manager* based on POMDP, it compromises two main modules that are *action policy* allowing to select the best next action based on the action recognition system output and error recognition enabling to detect and identify the committed errors by the user, (4) *cue selector* that might be either vo-

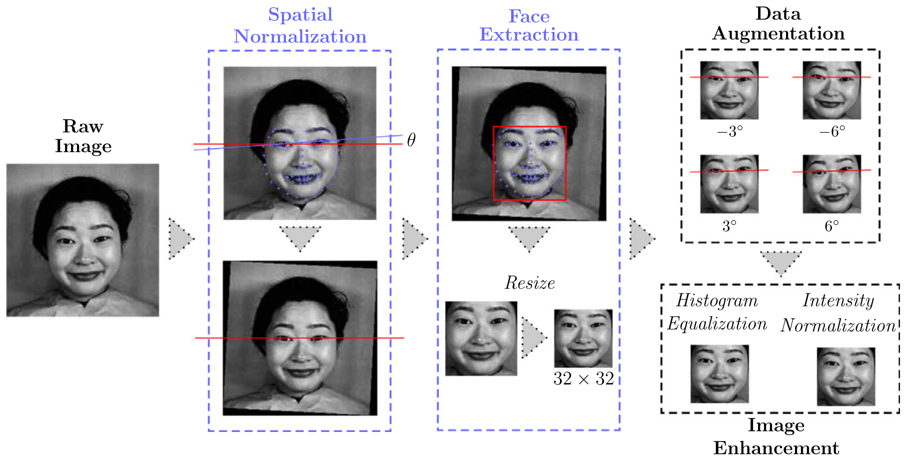


Fig. 3. Preprocessing workflow on an image from **JAFFE** dataset.

cal commands, written messages, pictures giving a hint about the next step to take or a video showing an actor performing the next action to perform.

Based on this brief review, we notice that the efficiency of an ambient assisting system is closely related to the EDM and ARM (see Fig. 2). Indeed, the ARM allows providing useful information about the performed ADL. The analysis of these information allows detecting the potential presence of error or abnormal behavior. However, some issues related to the EDM might be encountered such as a high false positive detection rate. Thus, the user receives guidance when it is not necessary and might be annoying. Therefore, enhancing the system's accuracy is not enough, the false positive detection rate has also to be reduced.

3. Automatic facial expression recognition

Most of the different existing AFER approaches based on DL techniques include *preprocessing stages* aiming to format and enhancing the image quality. These operations might ease the recognition process and also increase the accuracy. In the context of our work, we applied various operations to the input image as it is shown in Fig. 3.

In our defined preprocessing workflow, the first operation consists in *spatial normalization* that basically performs face alignment. As shown in the Algorithm 1, we used a technique introduced by Kazemi and Sullivan (2014). It allows extracting sixty-eight facial fiducial points L from a face image I_{in} in a relatively efficient way. L is a vector containing the $N = 68$ facial fiducial points and each one is represented by a Cartesian coordinates $(L[i].x, L[i].y)$ where i corresponds to the i^{th} facial fiducial points. In order to compute the positions of eyes centers $P_{eye1}(X_{eye1}, Y_{eye1})$ and $P_{eye2}(X_{eye2}, Y_{eye2})$, we need eight different landmarks $i = \{36, 37, 39, 40, 42, 44, 45, 47\}$. Then, two Euclidean distances $line_d$ and $line_h$ are computed. The first one links both eyes centers and the other one is a perfect horizontal line. Finally, the value of the correction angle θ is computed and the image is aligned as I_{out}.

Algorithm 1: Spatial Normalization.

```
Data: Input image $I_{in}$
Result: Output aligned image $I_{out}$
begin
    $L$ = FaceFiducialPointsExtraction($I_{in}$);
    $X_{eye1} = ((L[39].x - L[36].x)/2) + L[36].x$;
    $Y_{eye1} = ((L[40].y - L[37].y)/2) + L[37].y$;
    $X_{eye2} = ((L[45].x - L[42].x)/2) + L[42].x$;
    $Y_{eye2} = ((L[47].y - L[44].y)/2) + L[44].y$;
    $line_h = \sqrt{(X_{eye1} - X_{eye2})^2}$;
    $line_d = \sqrt{(X_{eye1} - X_{eye2})^2 + (Y_{eye1} - Y_{eye2})^2}$;
    if $Y_{eye1} > Y_{eye2}$ then
        $\theta = -\arccos(line_h/line_d)$;
    else
        $\theta = \arccos(line_h/line_d)$;
    end
    $I_{out}$ = ImageRotate($I_{in}$, $\theta$);
end
```

The next preprocessing operation consists in *face extraction*. The most common technique to detect the face region is proposed by Viola et Jones but in our case, we exploited the computed facial fiducial points using the Kazemi et Sullivan technique to extract the face region. Moreover, the CNN architecture that we use takes as input an image with a specific size. Therefore, we resize the face image to 32×32 as shown in Fig. 3. Finally, we perform some *image enhancement* using two common and popular techniques. The main objective of applying such techniques is reducing the effects of brightness and illumination. The first used technique is *linear intensity normalization* and consists in applying the Eq. (3) in a pixel level.

$$I_{out} = (I_{in} - Min)\frac{Max_{new} - Min_{new}}{Max - Min} + Min_{new} \tag{3}$$

Basically, from the input image I_{in} is computed the maximum and minimum pixel intensity values represented by Max and Min, respectively. Then, new values of the maximum and minimum are set namely $Min_{new} = 0$ and $Max_{new} = 255$. Finally, for each pixel in the input image is computed a new value stored in the output image I_{out}. The second used technique is *histogram equalization* allowing to adjust the image's histogram in order to have a uniform distribution of pixel intensity values. Thus, the lower contrast pixel will be increased and vice versa.

After preprocessing all the images of the dataset, the next step is employing the proposed *supervised* DL technique in order to recognize facial expressions from unlabeled images. We propose a CNN architecture inspired from **LeNet-5** (LeCun et al., 1998). As illustrated in Fig. 4, we distinguish two types of stages. The first one is *automatic features extraction* and it consists of *Convolutional layer* ($Conv_i$) and *Max-Pooling layer* ($Pool_i$). The main objective of this stage is to generate *features maps* C_j and S_j corresponding to the output of $Conv_i$ and $Pool_i$, respectively. They might be seen as descriptors of the input image. Moreover, another operation is applied to the *features map* C_j and consists in a *nonlinear activation function*, that is in our case the *Hyperbolic Tangent*. After executing

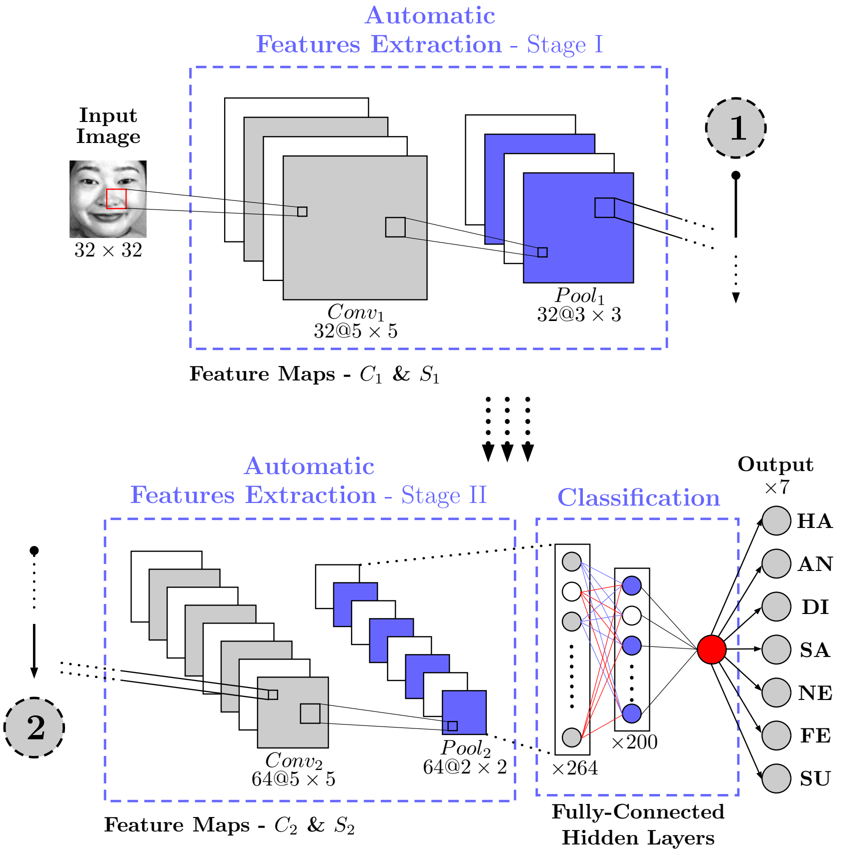


Fig. 4. The proposed CNN architecture *optimized* for **AFER**.

the two different stages of features extraction, the generated *features maps* are reshaped to a one-dimensional vector using a *fully-connected hidden layer*. Finally, a *LogSoftMax*() function is applied in order to generate seven different probabilities corresponding to the six different emotions in addition of *neutral* state. The highest value corresponds to the identified facial expression. In addition to this components, we added *dropout layers* in order to overcome and reduce the *overfitting* effect during the training phase. The probability parameter of this layers is set to $P_{dropout} = 0.5$.

The implementation of the proposed *static* AFER approach has been achieved using different libraries. The image preprocessing operations have been main performed using the popular **OpenCV** (Bradski, 2000) interfaced with Python. For the CNN part, even if different technologies might be employed, we chose **Torch7** (Collobert, Kavukcuoglu, & Farabet, 2011) interfaced with Lua.

4. Proposed *ambient assisting system*

Based on the different existing works and studies related to ambient assistance in an Ambient Assisted Living (AAL), we design our own *ambient assisting* system. One of the most main novelty consists in using facial expressions to enhance the system's performance. Indeed, we investigate by conducted experimentation in an experimental smart environment how including AFER might contribute to improving the performance of the proposed ambient assisting system.

As shown in Fig. 5, several components are involved to provide an effective assistance. In the context of our work, we focus mainly in the design of the ARM and EDM. In the following, we detail each one of the two components.

4.1. Action recognition module

The proposed ambient assisting system enables the recognition of each individual action or step during an ADL through the variation of the object positions. As illustrated in Fig. 3, the positions of the different objects are computed using an approach proposed in (Bilodeau, Bouzouane, Bouchard, & Gaboury, 2017). It considers the signal strength retrieved from the different *antennas* when objects with passive RFID tags are near to them. The approach returns for each object the Cartesian coordinates representing its current position. For each object, there are two different vectors $X_{obj} = [x_1, x_2, \ldots, x_{n_{obj}}]$ and $Y_{obj} = [y_1, y_2, \ldots, y_{n_{obj}}]$ that contain the variation of each coordinate. Nevertheless, these vectors are considered as raw data and in order to perform a relatively good recognition, some preprocessing steps have to be applied. Thus, from each vector is computed three different *statistical measures*.

Eqs. (4)–(6) represent the statistical measures namely *Kurtosis, Skewness* and *Variance*. n_{obj} represents the total number of the object position records or in other words, the input vector length. $\overline{X}_{obj}$ represents the mean value of the input vector X_{obj}. For each object are computed six different features $F_{obj} = [F_X, F_Y]$ where $F_X = [\gamma_2(X_{obj}), \gamma_1(X_{obj}), \sigma^2(X_{obj})]$, $F_Y = [\gamma_2(Y_{obj}), \gamma_1(Y_{obj}), \sigma^2(Y_{obj})]$.

$$Variance(X_{obj}) = \sigma^2(X_{obj}) = \frac{\sum_{i=1}^{n_{obj}} (x_i - \overline{X}_{obj})^2}{n_{obj}} \tag{4}$$

$$Skewness(X_{obj}) = \gamma_1(X_{obj}) = \frac{\sum_{i=1}^{n_{obj}} \frac{(x_i - \overline{X}_{obj})^3}{n_{obj}}}{\sigma^3(X_{obj})} \tag{5}$$

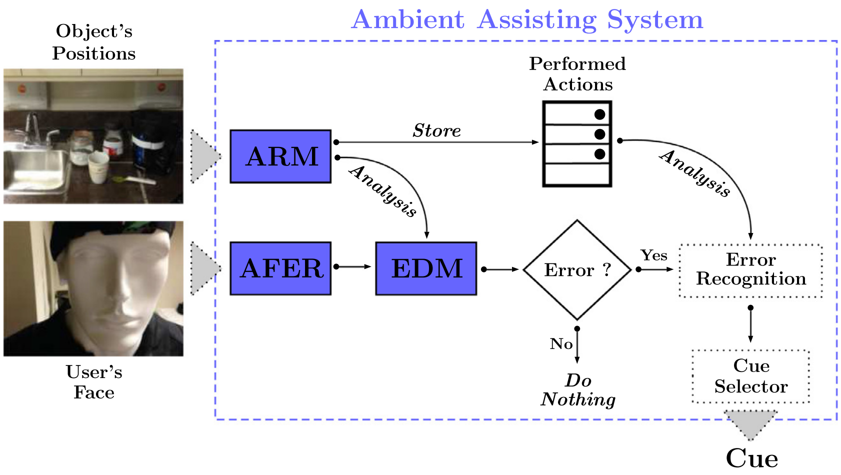


Fig. 5. Overview of the proposed *ambient assisting* system.

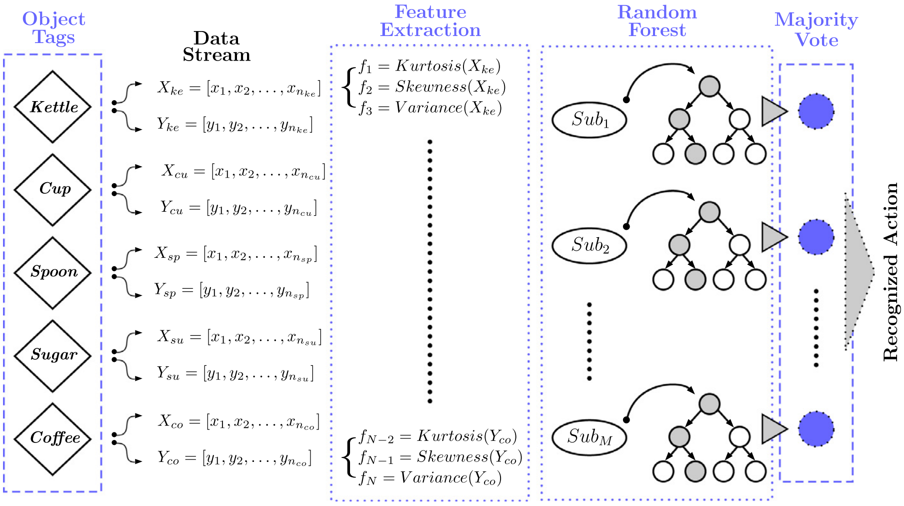

Object Tags
Data Stream
Feature Extraction
Random Forest
Majority Vote
Kettle
Cup
Spoon
Sugar
Coffee
$X_{ke} = [x_1, x_2, \dots, x_{n_{ke}}]$
$Y_{ke} = [y_1, y_2, \dots, y_{n_{ke}}]$
$X_{cu} = [x_1, x_2, \dots, x_{n_{cu}}]$
$Y_{cu} = [y_1, y_2, \dots, y_{n_{cu}}]$
$X_{sp} = [x_1, x_2, \dots, x_{n_{sp}}]$
$Y_{sp} = [y_1, y_2, \dots, y_{n_{sp}}]$
$X_{su} = [x_1, x_2, \dots, x_{n_{su}}]$
$Y_{su} = [y_1, y_2, \dots, y_{n_{su}}]$
$X_{co} = [x_1, x_2, \dots, x_{n_{co}}]$
$Y_{co} = [y_1, y_2, \dots, y_{n_{co}}]$
$f_1 = Kurtosis(X_{ke})$
$f_2 = Skewness(X_{ke})$
$f_3 = Variance(X_{ke})$
$f_{N-2} = Kurtosis(Y_{co})$
$f_{N-1} = Skewness(Y_{co})$
$f_N = Variance(Y_{co})$
Sub_1
Sub_2
Sub_M
Recognized Action

Fig. 6. Description of the proposed user **ARM**.

$$Kurtosis(X_{obj}) = \gamma_2(X_{obj}) = \frac{\sum_{i=1}^{n_{obj}} \frac{(x_i - \overline{X}_{obj})^4}{n_{obj}}}{\sigma^4(X_{obj})} \tag{6}$$

The different feature vectors corresponding to each object are concatenated to form a single descriptor vector that feeds a *supervised* machine learning technique. We have tested different techniques such as SVM, *k*-NN and Decision Tree (DT), but the best performance in terms of accuracy was achieved using a *Random Forest* classifier. This technique belongs to the category of *ensemble learning* methods. Basically, it consists in dividing the training set in M random subsets namely $Sub_1, Sub_2, \dots, Sub_M$ as shown in Fig. 6 and each subset is used to train a specific DT. When performing the recognition of an unlabeled feature vector, each DT generates a specific output. The final decision is generated using a *majority vote* function. In our case, we have used specific parameters that allow to achieve a relatively good performance. The number of trees or estimators has been set to $M = 15$ and the *entropy* as impurity metric. The implementation of the ARM has been achieved using the machine learning library **Scikit-Learn** (Pedregosa et al., 2011) interfaced with Python.

4.2. Error detection module

One of the most critical components of an ambient assisting system is the EDM. Indeed, its main purpose is to detect the potential presence of errors or anomalies when performing an ADL. When the EDM detects an error, it triggers the error recognition component before providing the adequate guidance. Moreover, the efficiency of an EDM might be defined by its capability to trigger assistance only when it is necessary. Usually, potential errors during an ADL are detected based on the analysis of recognized user actions. However, a major issue is raised when relying solely on this information. The performance of the EDM is then directly dependent on the ARM performance. In case the ARM misidentifies the user actions, the EDM will be automatically affected since it triggers assistance even if it is not required. Thus, it leads to an increase of the false positive detection rate and it might be annoying to the user since he receives guidance when he does not need one. As shown in Fig. 7, the proposed EDM includes two different sources of information from the ARM and AFER components.

Similarly to existing systems, the proposed EDM processes information provided by the ARM in order to detect the presence of potential anomalies during an ADL (Bouchard et al., 2007). As illustrated in Fig. 7, the ARM recognizes the different user actions. Following the performed activity, the analysis of the recognized user actions might leads to the detection of various anomalies. In-

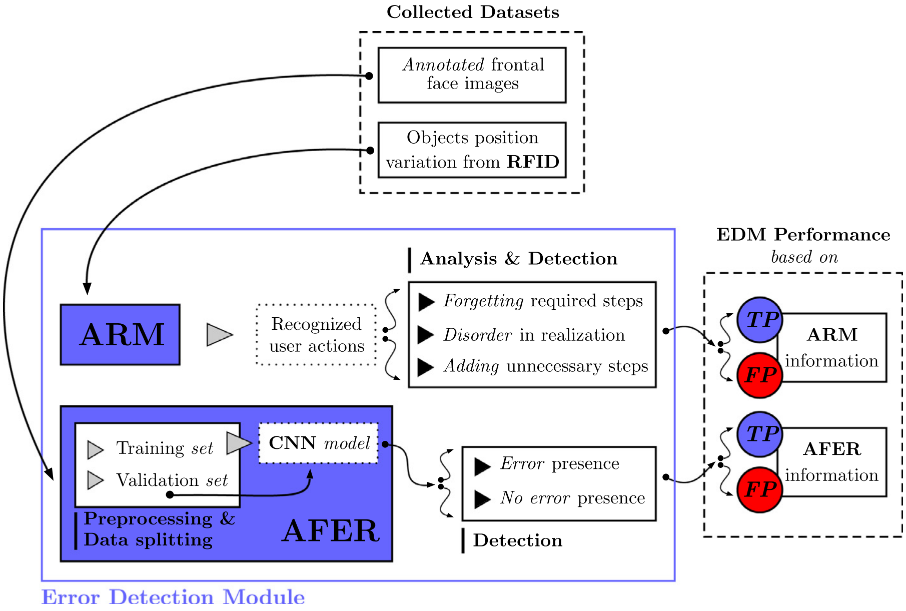


Fig. 7. Description of the proposed **EDM**.

deed, the user might have forgotten required steps, performed the activity in an inadequate way or added unnecessary steps. Moreover, the ARM does only allows the EDM to detect the presence of potential committed errors, it also provides useful information to identify the type of errors. Nevertheless, even if it represents an essential component, it remains insufficient to the EDM. Indeed, using only the ARM for error detection might raise a dependency issue and lead to an increase of the false positive detection rate. Thus, we have to add another source of information in order to improve the EDM performance and at the same time enhancing the ambient assistance system.

In this paper, the main contribution remains the *static* AFER based on a DL technique. Due to the large amount of information they provide, facial expressions might be employed to improve the performance of the EDM in terms of detection. To the best of our knowledge, no previous work has been conducted to investigate the presence of a potential relationship between facial expressions and committed errors during the realization of an ADL. Therefore, we have to construct our own dataset of face images from participants performing an ADL that consists in *coffee making*. Since there are no universal or specific facial expressions corresponding to committing error during an ADL, we have to *manually* label the collected face images. As shown in Fig. 7, a CNN model is built using the labeled face images from the training set. The labels are either *error* or *no error* presence. Furthermore, we have been motivated by the use of a DL technique for its capacity to find the most relevant patterns autonomously.

The EDM evaluation is performed using two different criteria TP (True Positive *detection*) and FP (False Positive *detection*) as shown in Fig. 7. The first one describes the detection of anomalies when they are actually present during the ADL while the second one represents the detection of anomalies when there are actually none. Based on these criteria, we investigate the potential contribution of the facial expressions to enhance the EDM performance.

5. Experimentation & validation

In this section, we detail the validation process for the proposed *static* AFER approach and the different components of the proposed ambient assisting system namely the ARM and EDM. Moreover, we present the used benchmark facial expression datasets in addition to the conducted experiments aiming to collect a new dataset from the LIARA lab.

5.1. Validation of the static AFER approach

In this paper, one of the main contributions remains the *static* AFER approach based on a CNN architecture. In order to evaluate its performance and compare it with existing methods, we employed several benchmark facial expression datasets. In Table 1 is detailed the five different datasets which we used. Each one consists of the six basic emotions in addition to the *neutral* state. The number of images per emotions and their resolution are different from a dataset to another. The JApanese Female Facial Expression (**JAFFE**) (Lyons, Akamatsu, Kamachi, & Gyoba, 1998) include only Japanese female participants. However, the other datatsets namely Radboud Faces Database (**RaFD**) (Langner et al., 2010), Karolinska Directed Emotional Faces (**KDEF**) (Lundqvist, Flykt, & Öhman, 1998), Man-Machine Interaction (**MMI**) (Pantic, Valstar, Rademaker, & Maat, 2005) and Cohn–Kanade Extended (**CK+**) (Kanade, Cohn, & Tian, 2000), include participants with different ages and ethnicities.

In order to evaluate the performance of the proposed *static* AFER approach, we followed a specific validation workflow presented in Fig. 8. The different image preprocessing operations described previously are applied to each image of the current used benchmark facial expression dataset. Then, the entire dataset is divided into ten different subsets $S = \{S_1, S_2, \ldots, S_{10}\}$ according to the *ten-folds cross-validation* technique. Thus, ten different iterations will be executed. During each one, a CNN model is trained using nine different subsets and the validation is achieved using the last one. In order to overcome the issue of randomized initialization when training a CNN model, each iteration corresponding to a specific fold is performed ten times and the average value is attributed to the fold. Finally, the average value is computed from the generated recognition rates corresponding to the ten folds. Finally, results are generated and represent the evaluation of the proposed *static* AFER approach when tested with a specific dataset.

5.2. Validation of the ARM

The evaluation of the ARM is achieved using our own collected dataset. Indeed, eight different participants and members of the

Table 1
Benchmark facial expression datasets.

Dataset		JAFFE	RaFD	KDEF	MMI	CK+
Emotions	HA	29	67	140	2136	691
	AN	30	67	140	1376	431
	DI	28	67	140	1225	486
	SA	30	67	140	1927	234
	NE	30	67	140	3891	327
	FE	30	67	140	287	240
	SU	30	67	140	1642	686
	Total	**207**	**469**	**980**	**12484**	**3095**
Resolution		256 × 256	681 × 1024	562 × 762	768 × 576	640 × 490

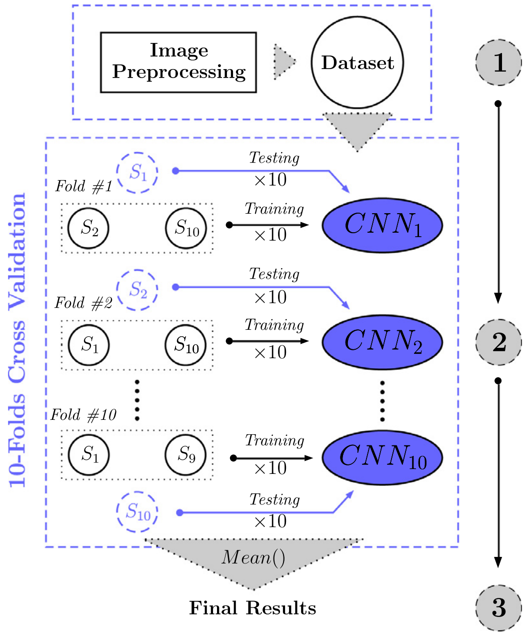


Fig. 8. Validation process of the proposed *static* **AFER** approach.

LIARA lab accepted to perform the various basic actions of the chosen ADL that consists in *coffee making*. It consists of five different basic steps: (1) *fill_kettle*: the participant takes the kettle and fill it with water from the sink, (2) *fill_cup*: it consists in filling the cup with water from the kettle, (3) *put_sugar*: the participant uses the spoon to put some sugar in the cup, (4) *put_coffee*: it consists in using the spoon to put some coffee in the cup, (5) *stir*: it is considered as the final step and consists in stirring the content of the cup.

These basic actions rely on the use of different objects such as kettle, sugar container, cup, spoon, and coffee container. As illustrated in Fig. 9, these objects have *passive RFID tags* attached to them and placed in the LIARA lab kitchen. There are four different *antennas* in the LIARA lab allowing to receive the signal from the different objects and transmitting it to the server. The approach proposed by Bilodeau et al. (2017) is exploited to compute the current positions of the different object in the kitchen using the signal strength. We developed an application in order to receive the data stream from the server, label (manually by a human expert) and store it in different files. The generated files represent our dataset. Eight different participants, that are students and members of the LIARA lab, and each one performs the basic actions five times. After collecting the raw data, we got for each basic action forty records. The validation of the ARM is achieved using the *one-leave-out* validation strategy. Thus, the performance of the system is evaluated for each participant using the data from the other participants for training.

5.3. Experimental protocol

One of the main purpose of this work consists in the investigation of how the facial expressions might enhance the error detection during an ADL. In order to achieve that, we define an experimental protocol that we followed to collect the needed dataset. Moreover, all the details related to the experiment such as *subjects, working place, collected data* and *experiment steps* are presented in the following:

- **Subjects:** in the context of the experiments, *seven* different participants, that are students and members of the LIARA lab, are prompted to perform *coffee making*. The subjects are males and females of 22 → 30 years old. All the needed instructions have been given to them in order to ensure that the experiments will be correctly performed.
- **Working place:** the experiments are performed in the LIARA lab and more precisely in the kitchen room. During the ADL, the participants used the same objects that we described previously for the ARM validation. In addition to these objects, we have used a device allowing to record the subject's face during the ADL and called *Facial Recorder System*.
- **Collected data:** in order to validate the proposed ambient assisting system, we need two types of dataset. The first one is related to the object's position variation used for the ARM and received from the LIARA lab server using our developed application. The other type of dataset consists in face image records and it is collected through a device that we have made. This type of data is used for the proposed *static* AFER approach. Both types are labeled by a human expert during the ADL and two distinct possibilities might be distinguished: *presence* and *absence* of committed errors or abnormal behaviors.
- **Experiment steps:** in order to push the participants to commit errors while recording their facial expressions, we define four different scenarios in collaboration with a *neuroscientist*. These different scenarios might be described as follows: (1) *normal:* during this first scenario, the participant is asked to perform *coffee making* in his own way or how he usually does it, (2) *guided:* in this case, the participant must perform the different basic actions of the ADL in a specific order and all the details are given to him through a paper note where are written all the steps, (3) *anomalies:* in order to induce the subjects to express specific facial expressions, we have changed the location of certain objects. The subject is asked to search by himself where the objects are located, 4) *distraction:* during this last scenario, the subject is asked to perform the ADL in the same way as when he was guided but this time, he will not be able to use the paper note with the different steps. We also say to him four different words that he must remember until the completion of the ADL.

At the end of the experiments, we get the two different types of dataset. In summary, we collected 28 labeled facial videos and 32 files containing the variation of object's position during the per-

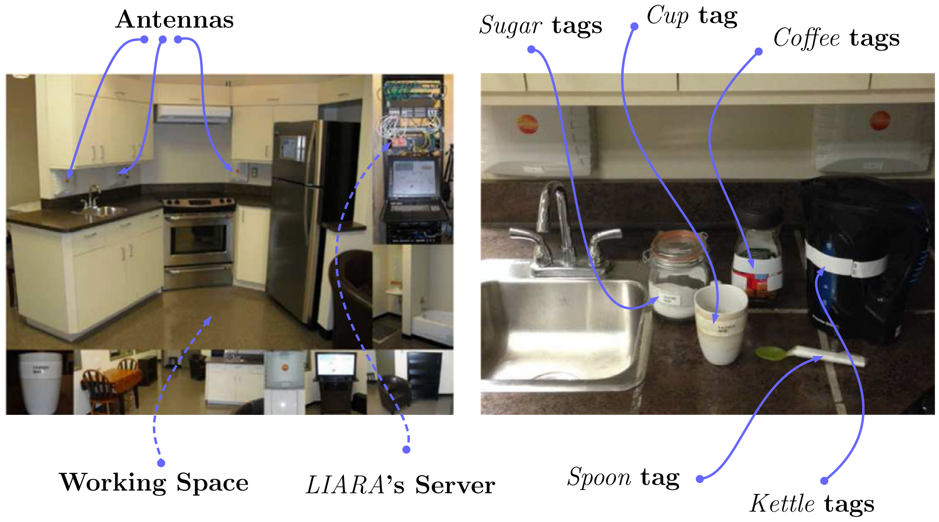


Fig. 9. The LIARA lab & used objects with RFID tags.

Table 2
AFER approach accuracy comparatively to state-of-the-art methods (**JAFFE**).

Dataset	Architecture	Approach	Σ Emotions	Accuracy
JAFFE	*Shallow*	(Mlakar et al., 2017)	Seven	92.75%
		(Yaddaden et al., 2016)		92.29%
		(Uçar et al., 2016)		94.65%
		(Owusu et al., 2014)		96.81%
	Deep	(Shan et al., 2017)	Six	76.74%
		(Lopes et al., 2017)	Seven	86.74%
		(Chen et al., 2017)		87.73%
		Proposed		**95.30% ± 1.70%**

formed ADL. The evaluation of the two components using the collected dataset is achieved in a separate way. Thereby, we employ the proposed *static* AFER approach to evaluate the presence of potential errors during the activity using the facial video records. The proposed ARM is evaluated for detecting basic user actions during the ADL in addition to the presence of potential errors.

Table 3
Confusion matrix using the **JAFFE** dataset.

	FE	SU	HA	DI	AN	SA	NE
FE	**95.17**	4.50	0.00	0.50	0.00	3.17	0.00
SU	0.00	**94.83**	5.83	0.00	0.00	0.00	3.00
HA	0.68	0.00	**99.32**	0.68	0.00	0.68	0.00
DI	9.11	0.00	0.00	**93.04**	4.64	0.00	0.00
AN	0.00	0.00	0.00	7.67	**91.83**	2.50	1.00
SA	1.00	0.00	3.33	0.00	0.00	**94.83**	4.50
NE	0.00	0.00	0.00	0.00	0.00	3.30	**99.00**
			Overall = **95.43%**				

Table 4
AFER approach accuracy comparatively to state-of-the-art methods (**RaFD**).

Dataset	Architecture	Approach	Σ Emotions	Accuracy
RaFD	*Shallow*	(Ali et al., 2017)	Six	85.00%
		(Jiang & Jia, 2016)	Seven	94.52%
	Deep	(Mavani et al., 2017)	Seven	95.71%
		(Sun et al., 2017)		96.93%
		(Zavarez et al., 2017)		85.97%
		Proposed		**97.57% ± 1.33%**

6. Results and analysis

In this section, we present the obtained results and evaluated performance of the proposed *static* AFER approach in addition to those of the proposed ambient assisting system.

6.1. Performance of the *static* AFER approach

The evaluation using each dataset is detailed following two distinct tables. The first one represents the global recognition rate compared with those of existing methods in either *shallow* or *deep* based architectures. The other table presents the confusion matrix and details the accuracy for each one of the six basic facial expressions in addition to the *neutral* state.

6.1.1. *JAFFE* dataset

In Table 2 is presented the obtained accuracy when using the **JAFFE** dataset. As shown in Table 2, the proposed *static* AFER approach outperforms the other methods. We observe a considerable difference in terms of accuracy when compared to the best *deep* based method proposed by Chen et al. (2017) and it is evaluated to 7.57%.

From Table 3, we notice that the best performance in terms of accuracy is achieved when recognizing *happiness* and *neutral* state with values greater than 99.00%. The lowest values are attributed to *disgust* and *anger* and might be explained by the fact that they are similar and hard to distinguish.

6.1.2. *RaFD* dataset

RaFD is the next datatset we used for evaluation. We obtained the highest recognition rate when compared to the other datasets. Moreover, as shown in Table 4, the proposed *static* AFER approach outperforms existing methods with a considerable difference when compared with the best *shallow* based architecture proposed by Jiang and Jia (2016) that is evaluated to 3.05%.

Table 5
Confusion matrix using **RaFD** dataset.

	FE	SU	HA	DI	AN	SA	NE
FE	**95.82**	3.51	0.00	0.00	0.00	0.00	3.06
SU	1.12	**99.63**	0.00	0.00	0.00	0.00	0.00
HA	0.00	0.00	**99.78**	0.00	0.00	0.00	0.66
DI	0.00	0.00	0.00	**99.18**	0.22	0.00	1.19
AN	0.00	0.00	0.00	0.00	**97.91**	4.48	1.79
SA	1.57	0.00	0.00	0.00	2.84	**95.45**	3.28
NE	0.45	0.90	0.00	0.67	7.61	2.91	**95.82**
			Overall = **97.66%**				

Table 6
AFER approach accuracy comparatively to state-of-the-art methods (**KDEF**).

Dataset	Architecture	Approach	Σ Emotions	Accuracy
KDEF	*Shallow*	(Ali et al., 2017)	Six	78.00%
		(Samara et al., 2017)	Seven	81.84%
		(Yaddaden et al., 2016)		79.69%
	Deep	(Zavarez et al., 2017)	Seven	72.55%
		(Ruiz-Garcia et al., 2017)		86.73%
		(Shin et al., 2016)		59.15%
		Proposed		**90.62% ± 1.60%**

Table 7
Confusion matrix using **KDEF** dataset.

	FE	SU	HA	DI	AN	SA	NE
FE	**83.63**	9.07	2.63	0.90	4.73	3.80	3.44
SU	6.29	**92.86**	0.64	0.00	0.21	0.64	1.03
HA	0.60	0.330	**98.70**	0.60	0.20	0.40	0.33
DI	2.04	0.32	1.11	**89.61**	3.82	5.21	0.96
AN	5.14	0.11	1.11	8.00	**85.04**	4.86	2.11
SA	4.82	0.11	0.00	3.36	1.50	**89.86**	3.93
NE	0.00	0.64	0.00	0.11	2.50	4.57	**94.54**
			Overall = **90.61%**				

We notice that among all the different facial expressions, the lowest accuracy value is greater than 95.00% (see Table 5). However, the best values are attributed to *surprise, happiness, disgust* with accuracy values greater than 99.00%.

6.1.3. ***KDEF*** *dataset*

When using **KDEF** dataset, the obtained accuracy also outperforms the other methods (see Table 6). Indeed, a considerable difference is noticed when compared with the best *shallow* based method proposed by Samara et al. (2017) and it is estimated to 8.78%.

In Table 7, we note that the best accuracy is achieved when recognizing *happiness* and *neutral* state with values greater than 94.00%. The lowest values are attributed to *fear* and *anger* with recognition rate values greater than 80.00%.

6.1.4. ***MMI*** *dataset*

Even if the obtained recognition rate is lower compared to the other tested dataset, the accuracy attributed to the **MMI** dataset outperforms the other existing methods for both architectures (see Table 8). We notice an important difference when compared with the best *deep* based approach proposed by Mollahosseini et al. (2016) and it is evaluated to 7.94%.

In Table 9, we notice that the best recognition rates are achieved when recognizing *surprise* and *happiness* with values greater than 90.00%. The lowest value is attributed to *fear* and it might be explained by the fact that unlike the other datasets, the face images in the **MMI** dataset include various objects such as glasses and hats. Moreover, it is well-known that DL techniques perform better when processing large size datasets and as shown in Table 1, *fear* has the lowest number of samples.

Table 8
AFER approach accuracy comparatively to state-of-the-art methods (**MMI**).

Dataset	Architecture	Approach	Σ Emotions	Accuracy
MMI	*Shallow*	(Alphonse & Dharma, 2017)	Seven	82.10%
		(Mlakar et al., 2017)		84.07%
		(Fan & Tjahjadi, 2015)		74.30%
		(Fang et al., 2014)	Six	75.96%
	Deep	(Zavarez et al., 2017)	Seven	67.03%
		(Mollahosseini et al., 2016)		77.90%
		(Mengyi et al., 2014)		63.40%
		(Liu et al., 2013)		74.76%
		Proposed		**85.84% ± 0.86%**

Table 9
Confusion matrix using **MMI** dataset.

	FE	SU	HA	DI	AN	SA	NE
FE	**29.90**	19.64	4.64	2.85	3.14	6.07	41.70
SU	2.19	**92.43**	0.72	0.06	0.33	1.18	4.26
HA	0.63	0.39	**94.27**	0.37	0.05	0.36	4.99
DI	0.67	0.67	1.55	**82.03**	4.91	0.75	11.69
AN	0.64	0.62	0.31	1.91	**85.87**	2.04	14.81
SA	1.17	0.62	0.07	0.32	0.46	**89.11**	10.64
NE	1.07	2.98	2.85	1.72	2.81	7.36	**82.21**
			Overall = **79.40%**				

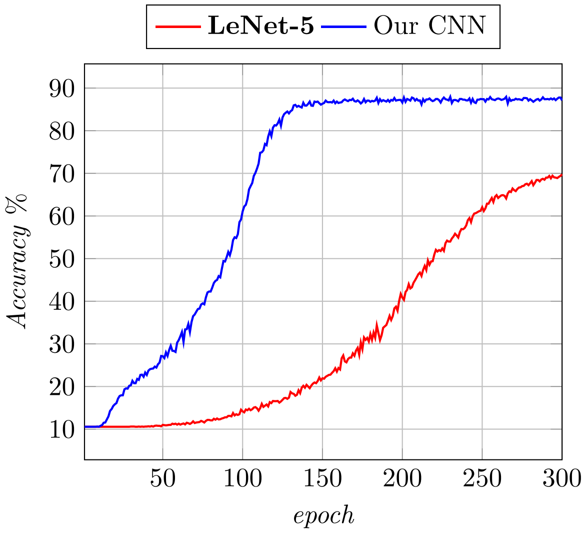


Fig. 10. Comparison between the proposed CNN and **LeNet-5**.

6.1.5. ***CK+*** *dataset*

In Table 10 is presented the comparison, in terms of accuracy, between our proposed *static* AFER approach and existing methods when testing with **CK+** dataset. We notice a considerable difference when compared with the best *deep* based method proposed by Mollahosseini et al. (2016) and it is evaluated to 3.17%.

The confusion matrix of this dataset is presented in Table 11 and shows that the best accuracy is reached when recognizing *surprise* and *happiness* with values greater than 99%. Moreover, the lowest value is attributed to *neutral* state with an accuracy greater than 89.00%.

6.1.6. *Other evaluations*

The proposed *static* AFER approach is mainly based on an optimized CNN architecture. Moreover, this architecture is inspired from an original one called **LeNet-5** proposed by LeCun et al. (1998). In Fig. 10 is shown the convergence and average accuracy of our proposed CNN and **LeNet-5** when using **JAFFE** dataset. Our architecture allows a faster convergence than **LeNet-5** and higher accuracy. Indeed, our proposed CNN reaches conver-

Table 10
AFER approach accuracy comparatively to state-of-the-art methods (**CK+**).

Dataset	Architecture	Approach	Σ Emotions	Accuracy
CK+	*Shallow*	(Yaddaden et al., 2017)	Six	92.54%
		(Uçar et al., 2016)	Seven	95.17%
		(Zavaschi et al., 2013)		88.90%
		(Zhang et al., 2013)	Six	71.30%
	Deep	(Zavarez et al., 2017)	Seven	88.58%
		(Shan et al., 2017)	Six	80.30%
		(Mollahosseini et al., 2016)	Seven	93.20%
		Proposed		**96.37% ± 0.80%**

Table 11
Confusion matrix using **CK+** dataset.

	FE	SU	HA	DI	AN	SA	NE
FE	**96.72**	0.61	1.00	0.00	0.37	0.31	3.52
SU	0.33	**99.18**	0.15	0.15	0.04	0.02	0.49
HA	0.04	0.02	**99.88**	0.04	0.00	0.00	0.10
DI	0.06	0.31	1.08	**96.26**	0.37	0.00	2.32
AN	0.67	0.00	0.00	1.08	**93.78**	3.40	3.34
SA	0.00	0.00	0.00	0.31	3.69	**92.53**	3.76
NE	0.70	0.32	0.87	1.62	9.92	5.83	**89.30**
			Overall = **95.38%**				

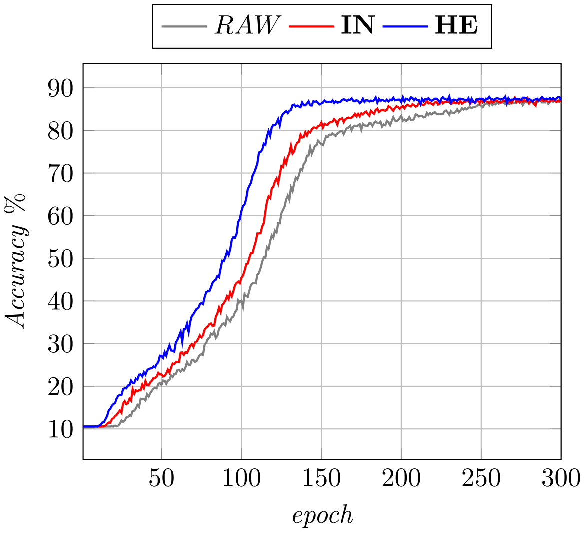


Fig. 11. Affect of various image enhancement techniques.

gence after 125 *epoch* while **LeNet-5** reaches it after 300 *epoch*. There is also a considerable difference between the two architectures in terms of accuracy and estimated to ≈ 18.00%.

In Fig. 11 is illustrated the effect of the different image enhancement techniques namely RAW (no enhancement), Intensity Normalization (IN) and Histogram Equalization (HE). From the graph, we notice that the convergence is faster when applying image enhancement techniques.

Similarly, in Fig. 12 we note that the best reached performance in terms of recognition rate for the different datasets is achieved using an image enhancement technique.

6.2. ARM performance

The ARM is another main component of the proposed ambient assisting system. Therefore, we evaluated it using our own collected dataset in the LIARA lab. We begin by evaluating the ARM performance when recognizing only samples of basic actions. As illustrated in Fig. 13, the obtained performance is relatively good and we achieved 100.00% for subjects 3 and 8. The value of the resulting accuracy in the subject independent way is estimated to 93.90% and represented by the red dashed line.

In Table 12 is represented the confusion matrix and we notice that each basic action is relatively well recognized with estimated values greater than 90.00%. The worst recognition rate is attributed to *no_action* and estimated to 63.00% and might be explained by the presence of signal noise.

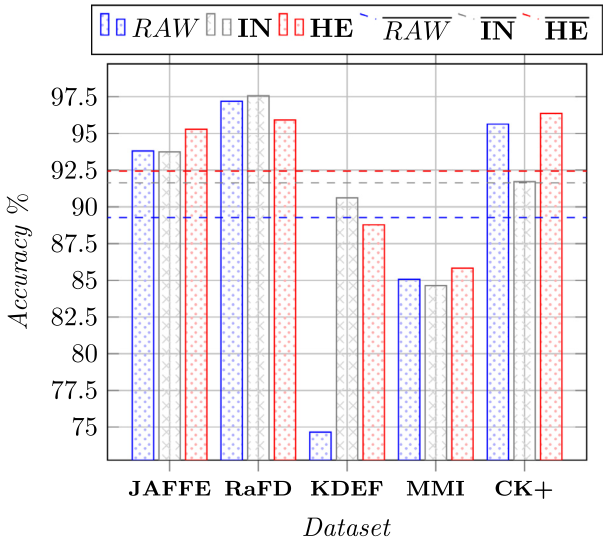


Fig. 12. Accuracy using the five datasets with various image enhancement techniques.

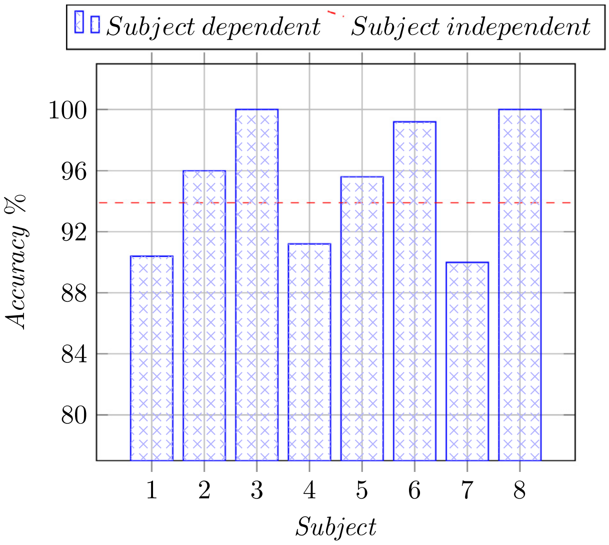


Fig. 13. Basic action recognition accuracy per subject.

Table 12
Confusion matrix for action recognition in a *subject independent* way.

	no_action	stir	put_sugar	fill_cup	fill_kettle	put_coffee
no_action	**63.00**	28.00	4.00	0.00	2.00	3.00
stir	0.00	**95.00**	0.00	4.00	0.00	1.00
put_sugar	0.00	3.00	**92.00**	0.00	0.00	5.00
fill_cup	0.00	0.00	1.00	**99.00**	0.00	0.00
fill_kettle	0.00	0.00	0.00	5.00	**95.00**	0.00
put_coffee	0.00	0.00	4.00	0.00	0.00	**96.00**
			Overall = **90.00%**			

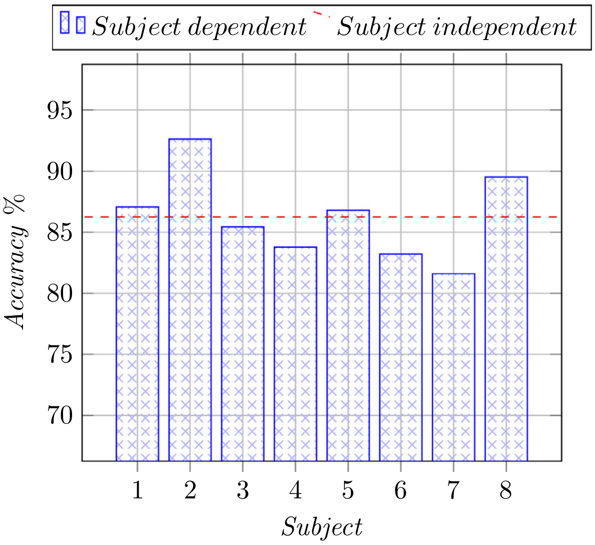


Fig. 14. Basic action recognition accuracy during *coffee making*.

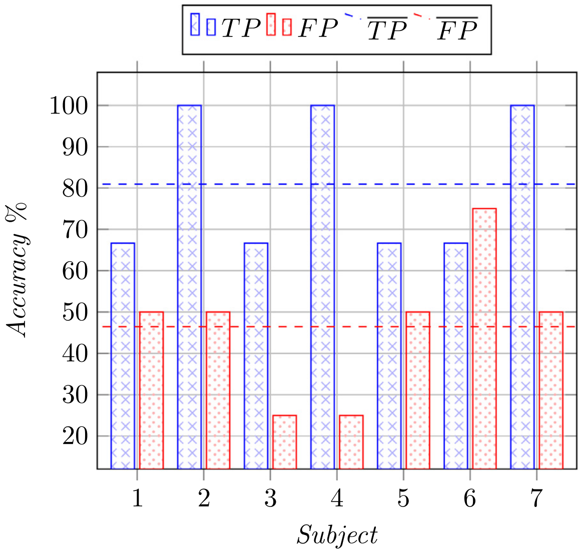


Fig. 15. **EDM** evaluation using **ARM** information.

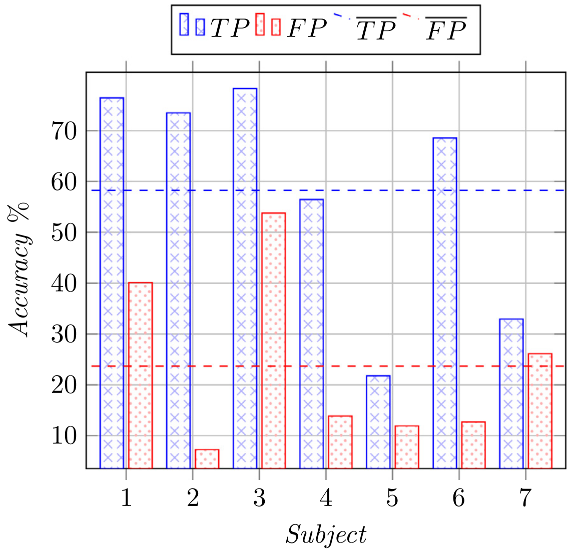


Fig. 16. **EDM** evaluation using **AFER** information.

We have also evaluated the proposed ARM when recognizing the basic actions during *coffee making*. The Fig. 14 illustrates the obtained accuracy for each subject. We notice that even if the performance is lower than when recognizing only basic actions, it remains relatively good since the global accuracy is equal to 86.25%.

6.3. EDM performance

The EDM allows the ambient assisting system to detect the presence of potential anomalies during an ADL. In this paper we evaluated the performance of the proposed EDM using two distinct sources of information.

In Fig. 15 is represented the performance of the EDM when using information provided by the ARM. As shown in Fig. 7, the recognized user actions are analyzed to detect potential errors during *coffee making*. The evaluation is represented by *TP* and *FP* rates in addition to their average values namely $\overline{TP}$ and $\overline{FP}$.

Fig. 16 describes the EDM performance through the *TP*, *FP*, $\overline{TP}$, $\overline{FP}$ criteria. The used source of information consists in facial expressions from the different participants when performing *coffee making*.

7. Discussion

In the section, we discuss and provide interpretations about the presented results in the previous section. The results for each component of the system are explained independently.

7.1. Static AFER approach

The proposed *static* AFER approach is mainly based on a *optimized* CNN architecture inspired from **LeNet-5** proposed by LeCun et al. (1998). As shown in Fig. 10, the proposed CNN shows good performance not only in terms of accuracy but also in terms of convergence. Indeed, we notice an improvement in terms of accuracy higher than 18% and it took half the number of epoch to converge. Moreover, the proposed *static* AFER approach includes several image preprocessing steps as shown in Fig. 3. During the evaluation, we notice that image enhancement techniques contributes significantly to improve accuracy (see Fig. 12) and convergence (see Fig. 11). In summary, the high performance of the proposed *static* AFER approach is related to the *optimized* CNN architecture in addition to the inclusion of different preprocessing steps.

We compared the proposed *static* AFER approach with other existing methods using five benchmark facial expression datasets. In all cases, our approach outperforms the existing methods in terms of accuracy. Indeed, the estimated recognition rate is higher than 95% when using **JAFFE, RaFD** and **CK+** datasets (see Tables 2, 4 and 10). However, we notice a relatively low performance for the **KDEF** and **MMI** datasets with 90.62% and 85.84% (see Tables 6 and 8). Furthermore, the evaluation through confusion matrix highlights the performance of the proposed *static* AFER approach when recognizing the six basic emotions in addition to the *neutral* state. We notice that obtained accuracy is relatively good when recognizing the different emotions except for *fear* in **MMI** dataset (see Table 9). It might be due to the lowest number of samples representing this emotion and the fact that DL techniques perform better with large datasets.

7.2. ARM

In the context of ambient assistance, one of the most important component remains the ARM that allows recognizing the user actions during an ADL. We tested the proposed ARM using the collected dataset provided through the experiments conducted in the LIARA lab. We begin the evaluation by recognizing only basic user actions. As shown in Fig. 13, the proposed ARM yields relatively good performance with an average accuracy of 93.90%. Motivated by the obtained results, we extended the proposed ARM to recognize the basic user actions during the *coffee making*. Even if it provides lower performance, it remains efficient since it yields 86.25% accuracy (see Fig. 14). As illustrated in Table 12, the different basic user actions from the activity of *coffee making* are relatively well recognized with an accuracy higher than 90% except for *no_action* that yields 63.00% and it is due to the presence of signal noise.

The conducted experimentation highlights the good performance of the proposed ARM. However, several issues might be raised such as the dependency on the approach that allows computing the object's position from RFID tags (Bilodeau et al., 2017). The improvement of this approach might improve the performance of the proposed ARM. Moreover, the inclusion of more sensors, that provide a larger amount of information, might improve the ARM performance.

7.3. EDM

Another main component of an ambient assisting system consists in the EDM that allows detecting potential committed errors or abnormal behaviors during an ADL. Similarly to existing systems, we employs information provided by the ARM. The EDM performs analysis in order to detect the potential errors. Thus, in Fig. 15 shows that using information from the ARM allows yielding a relatively high performance in terms of *TP* with an average of $\overline{TP} = 80.95\%$. However, we note a high rate in terms of *FP* with an average of $\overline{FP} = 46.43\%$. In order to improve the performance of the EDM, we have to reduce the value of *FP*.

In this paper, we investigate the contribution of the proposed *static* AFER approach to enhance the EDM performance. From the Fig. 16, we notice that using facial expressions reduce significantly the value of *FP* by over 20%. Thus, the two source of information (ARM and AFER) are complementary since each one improve the performance of the EDM in terms of *TP* and *FP*. Even if there are no universal or specific facial expressions that correspond to the fact of committing errors during an ADL, the use of a DL techniques allows finding the most relevant representation and highlighting the needed patterns that describe such facial expressions. Moreover, we have to combine the two sources of information (ARM and AFER) in order to improve the performance of the EDM and at the same time the proposed ambient assisting system.

8. Conclusion and further works

In this paper, we introduced a new *static* AFER approach based on the use of a DL technique. Indeed, we proposed a CNN architecture inspired from **LeNet-5** and *optimized* for AFER from images. We also proposed a preprocessing stage that enhances significantly the performance of the *static* AFER approach in terms of accuracy and convergence. We evaluated it with five different benchmark facial expression datasets and for each one, our approach outperformed the existing methods. Moreover, our work is mainly oriented to ambient assistance. Therefore, we proposed an ambient assisting system where we focused on two main components that consist in ARM and EDM. We conducted experiments in the LIARA lab in order to collect datasets during the realization of an ADL by different participants. We validated the ARM that provided relatively good results when recognizing the basic user actions during the ADL. Then, we evaluated the EDM using information provided by the analysis of the recognized user actions during the activity. However, we noticed that even if the ARM information allows obtaining relatively good performance in detecting potential errors, the EDM still suffer from a high false positive detection rate. Therefore, we investigated the potential contribution of AFER to overcome this issue. Indeed, the obtained results confirm that using facial expressions information contributes significantly to reducing the false positive detection rate by over 20%.

Nevertheless, the presented work is only the first step of many that are necessary to get a complete ambient assisting system. Next, we are planning to extend the CNN architecture to process image sequences. This might be achieved by adding *recurrent neural cells* to consider the temporal aspect. In the ARM side, we employed information retrieved from RFID tags only and its performance might be enhanced by considering information from other sensors available in the LIARA lab. Moreover, the proposed ambient assisting system might be seen as a prototype with the purpose of validating the ARM and EDM components. Since we are ultimately targeting elderly in loss of autonomy and people with cognitive disabilities, we are extending our experiments to those populations. It will be interesting to see whether the results obtained in this work hold or not with this targeted population and what are the eventual necessary adjustments. Furthermore, we have used a designed device in order to collect face images from the participant but it might be difficult to use for real applications. Therefore, we are planning to exploits images collected from cameras placed in the smart environment. The realization of these enhancements might lead to provide an efficient ambient assisting system that might be implemented in real smart environments.

Acknowledgments

The present work is financially supported by various organizations such as Natural Sciences and Engineering Research Council of Canada (NSERC), Canadian Foundation for Innovation (CFI) and Université du Québec à Rimouski (UQAR). Moreover, we would like to acknowledge every member of the LIARA lab who was involved in the experiments. We also would thank every person and organization that provided the various benchmark facial expression datasets (**JAFFE, RaFD, KDEF, MMI** and **CK+**) used for the evaluation and validation of the proposed *static* AFER approach.

References

Aghdam, H. H., & Heravi, E. J. (2017). *Guide to convolutional neural networks: A practical application to traffic-sign detection and classification*. Springer. doi:10.1007/978-3-319-57550-6.

Ali, A. M., Zhuang, H., & Ibrahim, A. K. (2017). An approach for facial expression classification. *International Journal of Biometrics, 9*(2), 96–112. doi:10.1504/IJBM.2017.085665.

Alphonse, A. S., & Dharma, D. (2017). Enhanced Gabor (e-gabor), hypersphere-based normalization and Pearson general kernel-based discriminant analysis for dimension reduction and classification of facial emotions. *Expert Systems with Applications, 90*, 127–145. doi:10.1016/j.eswa.2017.08.013.

Ammar, M. B., Neji, M., Alimi, A. M., & Gouardères, G. (2010). The affective tutoring system. *Expert Systems with Applications, 37*(4), 3013–3023. doi:10.1016/j.eswa.2009.09.031.

Baum, C., & Edwards, D. F. (1993). Cognitive performance in senile dementia of the alzheimers type: The kitchen task assessment. *American Journal of Occupational Therapy, 47*(5), 431–436. doi:10.5014/ajot.47.5.431.

Belley, C., Gaboury, S., Bouchard, B., & Bouzouane, A. (2015). Nonintrusive system for assistance and guidance in smart homes based on electrical devices identification. *Expert Systems with Applications, 42*(19), 6552–6577. doi:10.1016/j.eswa.2015.04.024.

Bengio, Y., et al. (2009). Learning deep architectures for ai. *Foundations and Trends® in Machine Learning, 2*(1), 1–127. doi:10.1561/2200000006.

Bilodeau, J.-S., Bouzouane, A., Bouchard, B., & Gaboury, S. (2017). An experimental comparative study of RSSI-based positioning algorithms for passive RFID localization in smart environments. *Journal of Ambient Intelligence and Humanized Computing*, 1–17. doi:10.1007/s12652-017-0531-3.

Bouchard, B., Giroux, S., & Bouzouane, A. (2006). A smart home agent for plan recognition of cognitively-impaired patients. *Journal of Computers, 1*(5), 53–62. doi:10.4304/jcp.1.5.53-62.

Bouchard, B., Giroux, S., & Bouzouane, A. (2007). A keyhole plan recognition model for alzheimer's patients: First results. *Applied Artificial Intelligence, 21*(7), 623–658. doi:10.1080/08839510701492579.

Bradski, G. (2000). The opencv library.. *Dr. Dobb's Journal: Software Tools for the Professional Programmer, 25*(11), 120–123.

Castellano, G., Kessous, L., & Caridakis, G. (2008). *Emotion recognition through multiple modalities: face, body gesture, speech* (pp. 92–103). Springer.

Chen, X., Yang, X., Wang, M., & Zou, J. (2017). Convolution neural network for automatic facial expression recognition. In *Proceedings of the international conference on applied system innovation (ICASI)* (pp. 814–817). IEEE. doi:10.1109/ICASI.2017.7988558.

Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). Torch7: A matlab-like environment for machine learning. *Biglearn, Nips workshop*.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion.. *Journal of Personality and Social Psychology, 17*(2), 124. doi:10.1037/h0030377.

Fan, X., & Tjahjadi, T. (2015). A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognition, 48*(11), 3407–3416. doi:10.1016/j.patcog.2015.04.025.

Fang, H., Mac Parthaláin, N., Aubrey, A. J., Tam, G. K., Borgo, R., Rosin, P. L., et al. (2014). Facial expression recognition in dynamic sequences: An integrated approach. *Pattern Recognition, 47*(3), 1271–1281. doi:10.1016/j.patcog.2013.09.023.

Fortin-Simard, D., Bilodeau, J.-S., Bouchard, K., Gaboury, S., Bouchard, B., & Bouzouane, A. (2015). Exploiting passive RFID technology for activity recognition in smart homes. *IEEE Intelligent Systems, 30*(4), 7–15. doi:10.1109/MIS.2015.18.

Jean-Baptiste, E. M., Rotshtein, P., & Russell, M. (2016). Cogwatch: Automatic prompting system for stroke survivors during activities of daily living. *Journal of Innovation in Digital Ecosystems, 3*(2), 48–56. doi:10.1016/j.jides.2016.10.003.

Jiang, B., & Jia, K. (2016). Robust facial expression recognition algorithm based on local metric learning. *Journal of Electronic Imaging, 25*(1), 1–8. doi:10.1117/1.JEI.25.1.013022.

Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE international conference on automatic face and gesture recognition* (pp. 46–53). IEEE. doi:10.1109/AFGR.2000.840611.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1867–1874). doi:10.1109/CVPR.2014.241.

Khan, S., Xu, G., Chan, R., & Yan, H. (2017). An online spatio-temporal tensor learning model for visual tracking and its applications to facial expression recognition. *Expert Systems with Applications, 90*, 427–438. doi:10.1016/j.eswa.2017.08.039.

Konar, A., Halder, A., & Chakraborty, A. (2015). *Introduction to emotion recognition* (pp. 1–45)). John Wiley & Sons, Inc..

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and emotion, 24*(8), 1377–1388. doi:10.1080/02699930903485076.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. doi:10.1038/nature14539.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. doi:10.1109/5.726791.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. In *Proceedings of the IEEE international symposium on circuits and systems* (pp. 253–256). IEEE. doi:10.1109/ISCAS.2010.5537907.

Liu, M., Li, S., Shan, S., & Chen, X. (2013). Au-aware deep networks for facial expression recognition. In *Proceedings of the 10th IEEE international conference and workshops on automatic face and gesture recognition* (pp. 1–6). IEEE. doi:10.1109/FG.2013.6553734.

Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition, 61*, 610–628. doi:10.1016/j.patcog.2016.07.026.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The karolinska directed emotional faces (kdef) - cd rom from department of clinical neuroscience, psychology section, karolinska institutet.

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *Proceedings of the 3rd IEEE international conference on automatic face and gesture recognition (fg)* (pp. 200–205). IEEE. doi:10.1109/AFGR.1998.670949.

Mavani, V., Raman, S., & Miyapuram, K. P. (2017). Facial expression recognition using visual saliency and deep learning. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 2783–2788). doi:10.1109/ICCVW.2017.327.

Mehrabian, A. (1968). *Communication without words* (2nd, pp. 51–52)).

Mengyi, L., Shaoxin, L., Shiguang, S., Ruiping, W., & Xilin, C. (2014). Deeply learning deformable facial action parts model for dynamic expression analysis. In *Proceedings of the Asian conference on computer vision* (pp. 143–157). Springer. doi:10.1007/978-3-319-16817-3_10.

Mihailidis, A., Boger, J. N., Craig, T., & Hoey, J. (2008). The coach prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC Geriatrics, 8*(1), 28. doi:10.1186/1471-2318-8-28.

Mlakar, U., Fister, I., Brest, J., & Potočnik, B. (2017). Multi-objective differential evolution for feature selection in facial expression recognition systems. *Expert Systems with Applications, 89*, 129–137. doi:10.1016/j.eswa.2017.07.037.

Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *Proceedings of the IEEE winter conference on applications of computer vision (wacv)* (pp. 1–10). IEEE. doi:10.1109/WACV.2016.7477450.

Owusu, E., Zhan, Y., & Mao, Q. R. (2014). A neural-adaboost based facial expression recognition system. *Expert Systems with Applications, 41*(7), 3383–3390. doi:10.1016/j.eswa.2013.11.041.

Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. In *Proceedings of the IEEE international conference on multimedia and expo (icme)*. IEEE. doi:10.1109/ICME.2005.1521424. 5–pp

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Perez-Gaspar, L.-A., Caballero-Morales, S.-O., & Trujillo-Romero, F. (2016). Multimodal emotion recognition with evolutionary computation for human-robot interaction. *Expert Systems with Applications, 66*, 42–61. doi:10.1016/j.eswa.2016.08.047.

Peters, C., Hermann, T., & Wachsmuth, S. (2013). TEBRA - An automatic prompting system for persons with cognitive disabilities in brushing teeth. In *Proceeding of the 6th international conference on health informatics (healthinf)* (pp. 12–23). doi:10.5220/0004193800120023.

Ruiz-Garcia, A., Elshaw, M., Altahhan, A., & Palade, V. (2017). Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. In *Proceedings of the international joint conference on neural networks* (pp. 1586–1593). IEEE. doi:10.1109/IJCNN.2017.7966040.

Samara, A., Galway, L., Bond, R., & Wang, H. (2017). Affective state detection via facial expression analysis within a human–computer interaction context. *Journal of Ambient Intelligence and Humanized Computing*, 1–10. doi:10.1007/s12652-017-0636-8.

Shan, K., Guo, J., You, W., Lu, D., & Bie, R. (2017). Automatic facial expression recognition based on a deep convolutional-neural-network structure. In *Proceedings of the IEEE 15th international conference on software engineering research, management and applications* (pp. 123–128). IEEE. doi:10.1109/SERA.2017.7965717.

Shin, M., Kim, M., & Kwon, D.-S. (2016). Baseline cnn structure analysis for facial expression recognition. In *Proceedings of the 25th IEEE international symposium onrobot and human interactive communication (ro-man)* (pp. 724–729). IEEE. doi:10.1109/ROMAN.2016.7745199.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958.

Sun, W., Zhao, H., & Jin, Z. (2017). An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. *Neurocomputing, 267*, 385–395. doi:10.1016/j.neucom.2017.06.050.

Tang, D., Yusuf, B., Botzheim, J., Kubota, N., & Chan, C. S. (2015). A novel multimodal communication framework using robot partner for aging population. *Expert Systems with Applications, 42*(9), 4540–4555. doi:10.1016/j.eswa.2015.01.016.

Tomita, M. R., Russ, L. S., Sridhar, R., et al. (2010). Smart home with healthcare technologies for community-dwelling older adults. *Smart home systems*. InTech. doi:10.5772/8411.

Uçar, A., Demir, Y., & Güzeliş, C. (2016). A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering. *Neural Computing and Applications, 27*(1), 131–142. doi:10.1007/s00521-014-1569-1.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition: 1* (pp. 511–518). IEEE. doi:10.1109/CVPR.2001.990517.

Yaddaden, Y., Adda, M., Bouzouane, A., Gaboury, S., & Bouchard, B. (2017). Facial expression recognition from video using geometric features. In *Proceedings of the 8th international conference on pattern recognition systems* (pp. 4:1–4:6). IET. doi:10.1049/cp.2017.0133.

Yaddaden, Y., Bouzouane, A., Adda, M., & Bouchard, B. (2016). A new approach of facial expression recognition for ambient assisted living. In *Proceedings of the 9th ACM international conference on pervasive technologies related to assistive environments* (pp. 14:1–14:8). ACM. doi:10.1145/2910674.2910703.

Zavarez, M. V., Berriel, R. F., & Oliveira-Santos, T. (2017). Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *Proceedings of the 30th sibgrapi conference on graphics, patterns and images* (pp. 405–412). IEEE. doi:10.1109/SIBGRAPI.2017.60.

Zavaschi, T. H., Britto, A. S., Oliveira, L. E., & Koerich, A. L. (2013). Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications, 40*(2), 646–655. doi:10.1016/j.eswa.2012.07.074.

Zhang, L., Jiang, M., Farid, D., & Hossain, M. (2013). Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications, 40*(13), 5160–5168. doi:10.1016/j.eswa.2013.03.016.

Chapitre 7

Conclusion Générale

Tout au long de ce travail de recherche qui s'est concrétisé par la présente thèse, nous nous sommes focalisés sur l'assistance aux personnes âgées souffrant de déficience cognitive. En effet, cette partie de la population, lorsqu'elle atteint un certain âge, devient incapable de réaliser les tâches les plus basiques de la vie quotidienne. Par conséquent, elle a besoin d'une assistance quasi-permanente fournie par des infrastructures et personnel spécialisés. Cependant, la mise en place de ces solutions classiques représente un fardeau pour l'économie en raison des énormes coûts qui y sont liés. Par conséquent, le défi actuel est de trouver des solutions qui soient à la fois efficaces et accessibles financièrement.

Dans le cadre de ce travail de recherche, nous nous sommes intéressés à l'utilisation des technologies de l'information et de la communication afin de développer un système permettant de fournir une assistance adéquate aux personnes souffrant de déficience cognitive durant la réalisation des activités de la vie quotidienne. Tout comme la majorité des systèmes d'*assistance ambiante*, celui que nous avons proposé est composé de deux modules principaux. Celui qui est dédié à la reconnaissance automatique des activités est basé sur la technologie **RFID** [4] dont dispose le **LIARA** [20] où se sont déroulées les expérimentations. Le second module permet d'analyser les actions reconnues à l'aide du précédent module afin de détecter, de façon automatique, la présence d'erreurs ou d'anomalies lors de l'exécution de l'activité. Notre principale contribution réside dans l'implication des expressions faciales afin d'améliorer les performances en termes de détection d'erreurs [68].

Avant d'aborder la partie dédiée à l'*assistance ambiante*, nous avons d'abord travaillé sur la reconnaissance automatique des émotions à travers les expressions faciales. L'objectif étant le développement de différentes méthodes ayant un taux de reconnaissance élevé et pouvant être implémenté dans des environnements avec des ressources matérielles limitées (tels que les systèmes embarqués). Nous avons commencé par comparer les performances lors de l'utilisation des trois type de descripteurs : *géométriques* [61, 67], *apparences* [65] et *hybrides* [66]. Chacun d'eux apporte son lot d'avantages et de limitations. Ainsi, la représentation *géométrique* offre un bon taux de reconnaissance mais elle est généralement plus adaptée aux image frontales. Ce problème est résolu par la représentation d'*apparence* qui opère sur l'intégralité de l'image peu importe l'angle choisi. La combinaison des deux qui est achevée par la représentation *hybride* offre les meilleures performances, mais elle est plus complexe que les deux méthodes précédentes. Avec l'engouement suscité par l'apparition des nouvelles techniques d'*apprentissage profond*, nous avons proposé une architecture de *réseau de neurones à convolution* [64] inspirée de **LeNet-5** proposée par LeCun et al. [38]. Le principal avantage de ce genre de technique réside dans la génération *automatique* de représentations pertinentes. L'architecture proposée offre les meilleures performances en termes de reconnaissance et elle est *optimisée* pour les expressions faciales ainsi que pour une implémentation en environnement à ressources matérielles limitées tel que le **Raspberry Pi** et le robot **NAO**.

La majeur partie des méthodes proposées sont *statiques* afin de réduire la complexité de traitement et surtout faciliter l'implémentation dans des environnements à ressources matérielles limitées. Cependant, nous avons également travaillé sur des méthodes *dynamiques* opérant sur des séquences d'images. Le principal intérêt de ce genre de méthodes est la quantité d'information qui est prise en considération. En effet, elles permettent de caractériser les différentes phases de transition de l'expression faciale (*début*, *sommet* et *fin*), mais elles sont généralement plus complexes. Dans le cadre de cette thèse, nous avons proposé une méthode basée sur la représentation *spatio-temporelle* [**23**, **66**] ayant à la fois de bonnes performances et réduisant la complexité du traitement pour une implémentation en environnement à ressources matérielles limitées.

Après avoir évalué les différentes méthodes en utilisant des bases de données *benchmark*, nous avons été en mesure de définir la méthode la plus adéquate pour l'amélioration de la détection automatique des erreurs. En se basant sur les résultats obtenus, il s'est avéré que la méthode basée sur l'*apprentissage profond* était la plus adéquate. Nous avons ainsi mené des expérimentations au niveau de l'environnement intelligent **LIARA** afin de collecter les données nécessaires pour la validation du système d'assistance proposé. Les données collectées sont de deux types distincts à savoir **RFID** et vidéos. Lors de la validation du système d'assistance, nous avons constaté qu'effectivement les expressions faciales contribuent à l'amélioration des performances. Plus précisément, le taux de fausses détection dans le module de détection automatique d'erreurs s'est vu réduit de plus de 20%.

7.1 Objectifs accomplis

Dans le cadre de cette thèse, le principal objectif consistait à développer un système d'assistance fournissant une aide adéquate aux personnes âgées et souffrant de déficience cognitive. Cependant, la contribution la plus significative de ce travail réside dans l'utilisation des expressions faciales afin d'améliorer les performances du système d'assistance proposé. Ainsi, la majeure partie de la thèse est dédiée à la reconnaissance automatique des expressions faciales (voir Chapitre 2, Chapitre 3, Chapitre 4 et Chapitre 5). Nous avons pu travailler sur des méthodes *statiques* et *dynamiques* en utilisant les trois différents types de descripteurs. Nous avons fait appel à diverses techniques d'apprentissage automatique classiques dont nous avons pu comparer les performances. Nous avons également travaillé avec une des nombreuses techniques d'apprentissage profond à savoir le *réseau de neurones à convolution*. Parmi les méthodes proposées, nous avons choisi la plus performante afin de l'intégrer au système d'assistance développé et qui a été validé au sein de l'environnement intelligent **LIARA** (voir Chapitre 6).

Dans le Chapitre 2, nous avons travaillé avec une représentation *géométrique*.

Pour la partie reconnaissance, nous avons utilisé une architecture *multi-classes* basée sur la combinaison de classifieurs *mono-classe*. Malgré l'existence de différents classifieurs *mono-classe*, notre choix s'est porté sur les *machines à vecteurs de support*. Les résultats obtenus démontrent l'efficacité de l'architecture proposée en termes de taux de reconnaissance. De plus, nous avons remarqué qu'elle était plus adéquate lorsque la base de données d'entrée est d'une taille réduite. En comparant ses performances à celles obtenues par la combinaison de classifieurs *bi-classes*, nous avons constaté qu'elle offrait des résultats très satisfaisants avec un nombre réduit d'attributs. C'est très avantageux lorsque ces méthodes développées sont destinées à être implémentées dans des environnements limités en termes de ressources matérielles.

Nous nous sommes également intéressés aux caractéristiques d'*apparence*. Ainsi, dans le Chapitre 3, nous avons proposé une méthode utilisant une représentation basée sur la technique *étendue* des *motifs binaires locaux*. Contrairement aux méthodes existantes, nous avons appliqué cette technique sur des sous-régions spécifiques du visage (les yeux, le nez et la bouche). Cette délimitation des régions a permis à la fois de réduire la complexité du traitement et de se débarrasser des données redondantes et inutiles. Afin d'avoir une représentation plus affinée, nous avons utilisé deux méthodes de *sélection d'attributs* basées sur la *transformation* dont nous avons comparé les performances. Les résultats ont mis en évidence l'efficacité de l'*analyse en composantes principales* même si l'utilisation de l'*analyse en composantes indépendantes* a fourni des résultats prometteurs. Comme les *motifs binaires locaux* sont sensibles aux effets de contraste et de luminosité, nous avons appliqué certains pré-traitements afin de remédier à cette limitation. De façon globale, les résultats obtenus sont prometteurs et confirment l'efficacité de ce descripteur.

Avec les résultats prometteurs obtenus à l'aide des deux descripteurs *géométrique* et d'*apparence*, nous nous sommes penchés sur leur combinaison en proposant une méthode *hybride*. Dans le Chapitre 4, nous avons proposé l'utilisation de la représentation *géométrique* qui a fait ses preuves dans [61, 67] en combinaison avec une représentation d'*apparence* basée sur les coefficients issus de la *transformée en ondelettes discrète*. La fusion de ces deux représentations est réalisée en *aval* au niveau du bloc de classification à l'aide d'une architecture *multi-classe* basée sur les *machines à vecteurs de support bi-classes*. Les résultats attestent de l'efficacité de la combinaison des deux descripteurs en atteignant plus de 99% de taux de reconnaissance avec la base de données **RaFD** [35]. Même si la méthode proposée est plus complexe, elle a permis d'atteindre de bien meilleures performances que chacune des deux représentations utilisées individuellement.

Dans les premiers chapitres, nous nous sommes penchés principalement sur les approches *statiques* qui opèrent sur de simples images. Cependant, ça ne reflète pas la réalité, car les expressions faciales incluent l'aspect temporel qui apporte une information supplémentaire concernant la progression et l'intensité de l'émotion. Ainsi, dans le Chapitre 5 est proposée une méthode *dynamique* basée sur une représentation *spatio-temporelle* efficace. Elle utilise des descripteurs *géométriques* combinés à des *mesures de dispersion* pour la représentation temporelle. Dans la première version proposée [63], il était nécessaire d'appliquer une *normalisation* au niveau des séquences afin d'avoir le même nombre d'images par séquence. C'est une limitation et pour cette raison, nous avons proposé une deuxième version [62] offrant de meilleures performances sans recourir à la *normalisation*. En raison de l'importante quantité d'informations considérée, les méthodes *dynamiques* sont généralement plus complexes que les *statiques*. Cependant, la représentation *spatio-temporelle* a permis de condenser les informations afin de réduire la taille ainsi que la complexité.

Dans toutes les méthodes que nous avons proposés, nous avons utilisé des techniques d'apprentissage classiques dont l'efficacité est reflétée à travers les résultats obtenus. Néanmoins, elles présentent une limitation qui réside dans le fait que les descripteurs doivent être *manuellement* définis au préalable. Pour cette raison, nous nous sommes intéressés à l'*apprentissage profond*. En effet, dans le Chapitre 6 est proposée une nouvelle architecture de *réseau de neurones à convolution* qui est *optimisée* pour la reconnaissance d'expressions faciales. Elle est basée sur l'architecture **LeNet-5** proposée par LeCun et al. [38] mais la notre a permis d'obtenir de meilleures performances en termes de taux de reconnaissance ainsi que de rapidité de convergence.

L'intérêt de ce genre de techniques réside dans la génération *automatique* de représentation pertinente à partir des données d'entrée.

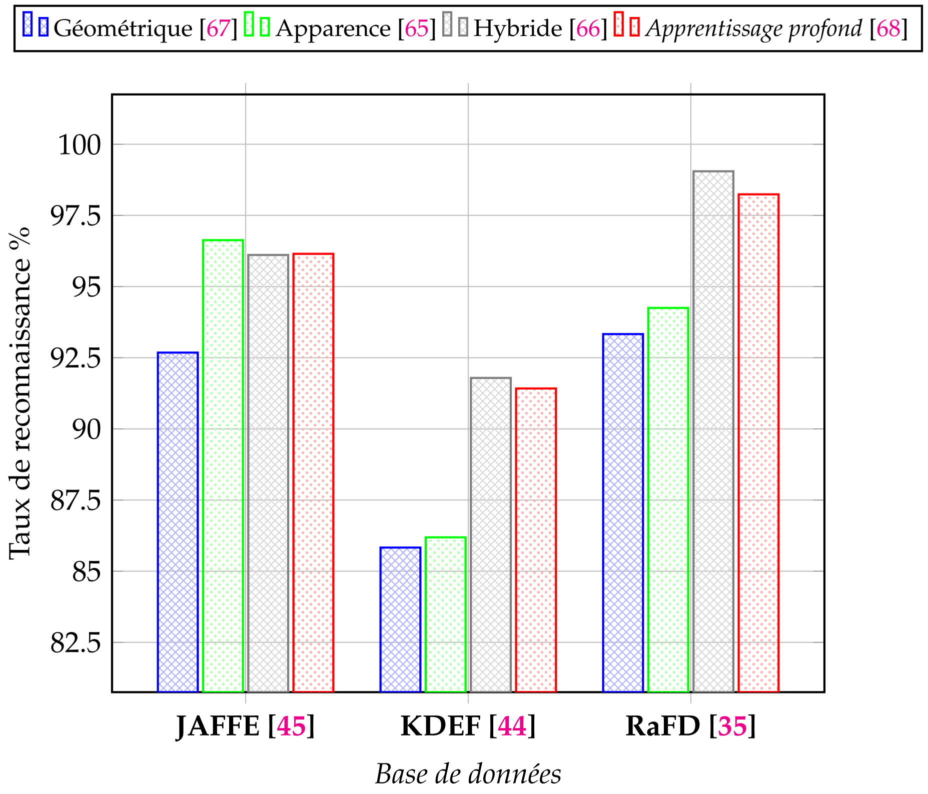
Le Chapitre 6 a été entièrement consacré au développement d'un système permettant de fournir une assistance adéquate lors de l'exécution d'une activité de la vie quotidienne (la *préparation du café*). Le système est composé de deux parties distinctes, la première est destinée à la reconnaissance des actions de base de l'activité. Elle est basée sur l'utilisation de la technologie **RFID** [4]. Le taux de reconnaissance enregistré est prometteur et est estimé à 90%. La seconde partie, quant à elle, est destinée à la détection d'erreurs ou d'anomalies durant l'exécution de l'activité. Il a été possible d'utiliser les informations issues de l'analyse de différentes actions de bases reconnues, mais les performances obtenues étaient insuffisantes. C'est là qu'interviennent les expressions faciales. En se basant sur les performances des méthodes de reconnaissance automatique des expressions faciales, il s'est avéré que la plus adéquate était celle qui est basée sur l'*apprentissage profond*. Nous l'avons intégré au module de détection automatique d'erreurs du système d'assistance proposé. Les résultats obtenus lors de la validation du système ont confirmés l'intérêt d'utiliser les expressions faciales. En effet, le taux de fausse détection a été réduit de plus de 20%, ce qui permet de déclencher l'assistance uniquement lorsqu'elle est nécessaire.

7.2 Comparaison des approches développées

Comme décrit dans l'Introduction Générale, l'ensemble des méthodes de reconnaissance automatique des expressions faciales ont été évaluées à l'aide de *six* bases de données *benchmark*. Trois d'entre elles sont *statiques* et les autres *dynamiques*. Dans ce qui suit, nous présenterons une comparaison entre les différentes méthodes de reconnaissance automatique des expressions faciales proposées en termes de taux de reconnaissance.

Dans la Figure 7.1 sont représentées les performances des différentes méthodes proposées qui ont été évaluées avec des bases de données *statiques*. Nous pouvons aisément remarquer que les meilleures performances sont atteintes par la méthode

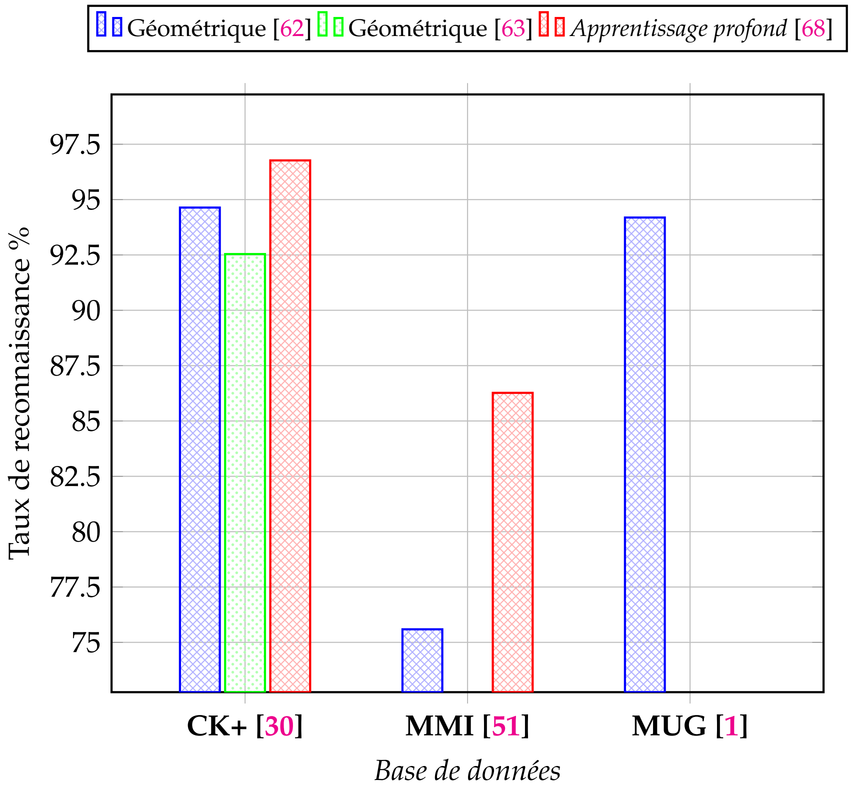
FIGURE 7.1 – Comparaison avec base de données *statique*.



basée sur la représentation *hybride* et celle utilisant la technique d'*apprentissage profond*. Elles ont permis d'atteindre un taux de reconnaissance moyen (sur les trois bases de données benchamrk) de plus de 95%. Les deux autres méthodes restantes utilisant un seul type de descripteur (*géométrique* ou d'*apparence*) restent néanmoins efficaces puisqu'elles ont permis d'atteindre un taux de reconnaissance moyen de plus de 90%.

Nous avons également comparé les performances des méthodes évaluées avec des bases de données *dynamiques*. Ainsi, dans la Figure 7.2 nous remarquons que la seconde version de la méthode basée sur la représentation *spatio-temporelle* [62] produit un meilleur taux de reconnaissance que la première version [63] avec une différence de plus de 2% (voir base de données **CK+**). Même si elle n'est pas *dynamique*, la méthode basée sur l'*apprentissage profond* fournit les meilleures performances.

FIGURE 7.2 – Comparaison avec base de données *dynamique*.



7.3 Limitations connues

Malgré les bonnes performances qu'elles ont permis d'atteindre lors de l'évaluation, les différentes méthodes proposées possèdent néanmoins des limitations dont nous avons connaissance. La plus critique est liée au fait qu'elles opèrent uniquement sur des images *frontales*. Plus précisément, ce sont celles qui utilisent des descripteurs *géométriques* qui sont les plus concernées. Ceci peut-être problématique si les dispositifs d'acquisition ne capturent qu'une partie du visage ou bien sous un angle où il est plus délicat d'appliquer les méthodes proposées. De plus, la plupart des méthodes proposées incluent des *pré-traitements* dont le plus important est la détection et extraction de la zone du visage. La méthode utilisée est celle proposée par Viola et Jones [58]. Par conséquent, leur bon fonctionnement est étroitement lié aux performances de ce *pré-traitement*.

Au niveau du système d'assistance, il y a également différentes limitations qu'il est nécessaire de prendre en considération à des fins d'améliorations futures. La première limitation réside dans l'obligation d'utiliser un dispositif spécifique pour enregistrer les images *frontales* du visage de l'individu lors des expérimentations. Le dispositif est intrusif et n'est pas très adéquat pour une utilisation dans un environnement intelligent. Les expérimentations ont été conduites avec des personnes *neurotypique* alors que le système cible des personnes souffrant de déficience cognitive. Dans ce travail de recherche, nous nous sommes focalisés uniquement sur les deux modules de reconnaissance d'activités et de détection d'erreurs en négligeant la partie réponse. Cette dernière est censée indiquer aux individus s'ils ont commis une erreurs et comment y remédier.

7.4 Perspectives et travaux futurs

Comme décrit dans la section précédente, les méthodes proposées, malgré les résultats prometteurs, ont besoin d'être améliorées. Pour cette raison, nous prévoyons diverses améliorations afin de remédier aux limitations citées précédemment.

La première amélioration que nous prévoyons d'apporter est de permettre aux différentes méthodes proposées de faire de la reconnaissance automatique d'expressions faciales sous différents angles. Ensuite, nous prévoyons d'étendre l'architecture du *réseau de neurones à convolution* afin qu'elle puisse opérer sur des données *dynamiques*. Pour cela, nous envisageons d'utiliser une autre technique d'*apprentissage profond* à savoir le *réseau de neurones récurrent* [22].

Cependant, le plus gros travail que nous prévoyons d'effectuer est directement lié à l'*assistance ambiante*. Plus précisément, il consiste à exploiter des images acquises à partir de caméras murales qui seront disposées dans l'environnement intelligent. La méthode que nous avons utilisée pour la reconnaissance automatique des expressions faciales doit être modifiée de telle sorte qu'elle puisse opérer sur différents angles et surtout traiter des séquences d'images. Nous avons utilisé la technologie **RFID** pour la partie reconnaissance des actions utilisateur durant l'activité, mais nous prévoyons de travailler avec des données vidéos afin de n'avoir qu'une seule source d'information. Finalement, après avoir testé et validé le nouveau système, nous envisageons de mener de nouvelles expérimentations, mais cette fois-ci avec des sujets souffrants de déficience cognitive.

Dans le cadre de ce travail de recherche, il a été question de système d'assistance destiné à être intégré dans un environnement intelligent. Cependant, il est possible d'y ajouter un agent intelligent sous forme d'un robot humanoïde qui permettra d'améliorer grandement l'interaction. En faisant le lien entre l'individu et son environnement, le robot augmentera l'efficacité de l'assistance fournie. De plus, le robot est généralement mobile, ce qui lui permet de suivre l'individu plus facilement. Parmi les technologies existantes, nous pouvons citer le **NAO**[1] qui dispose de nombreux outils facilitant le développement de méthodes d'assistance et d'interaction. Cependant, il dispose de ressources matérielles limitées et pour cette raison, les méthodes à y intégrer ne doivent pas être trop complexes. Certaines des méthodes que nous avons développées pour la reconnaissance automatique des expressions faciales ont été testées avec succès sur le **Raspberry Pi**[2] qui est un système embarqué à ressources matérielles limitées. Ainsi, comme perspective, nous projetons d'implémenter les méthodes que nous avons développé sur le **NAO** et faire des expérimentations avancées avec.

La population que nous avons ciblée par ce travail de recherche est âgée et souffre de déficience cognitive. Au vue de l'importante quantité d'informations relative à l'individu et fournie par les expressions faciales, une autre application serait de pouvoir détecter de façon automatique le niveau et la progression de la déficience cognitive. En effet, l'assistance fournie à chaque individu doit être adaptée à son besoin, c'est-à-dire que toutes les personnes souffrant de déficience cognitive n'ont pas forcément besoin du même niveau d'assistance. Pour cette raison, une des perspectives futures de ce présent travail de recherche est de faire en sorte que l'assistance s'adapte au besoin de l'individu suivant son niveau de déficience cognitive. C'est l'information fournie par les expressions faciales lorsque l'individu commet une erreur durant l'exécution d'une activité de la vie quotidienne qui servira de base pour

[1] https://www.softbankrobotics.com/emea/fr/nao

[2] https://www.raspberrypi.org/products/raspberry-pi-3-model-b/

suivre la progression du niveau de déficience cognitive.

Bibliographie

[1] Niki AIFANTI, Christos PAPACHRISTOU et Anastasios DELOPOULOS. « The MUG facial expression database ». In : *11th international Workshop on Image analysis for Multimedia Interactive Services*. IEEE. 2010, p. 1–4.

[2] Corinne BELLEY et al. « Nonintrusive system for assistance and guidance in smart homes based on electrical devices identification ». In : *Expert Systems with Applications* 42.19 (2015), p. 6552–6577. DOI : 10.1016/j.eswa.2015.04.024.

[3] Yoshua BENGIO et al. « Learning deep architectures for AI ». In : *Foundations and trends® in Machine Learning* 2.1 (2009), p. 1–127. DOI : 10.1561/2200000006.

[4] Jean-Sébastien BILODEAU et al. « An experimental comparative study of RSSI-based positioning algorithms for passive RFID localization in smart environments ». In : *Journal of Ambient Intelligence and Humanized Computing* (2017), p. 1–17. DOI : 10.1007/s12652-017-0531-3.

[5] Bruno BOUCHARD, Sylvain GIROUX et Abdenour BOUZOUANE. « A smart home agent for plan recognition of cognitively-impaired patients. » In : *JCP* 1.5 (2006), p. 53–62. DOI : 10.4304/jcp.1.5.53-62.

[6] Walter B CANNON. « The James-Lange theory of emotions : A critical examination and an alternative theory ». In : *The American journal of psychology* 39.1/4 (1927), p. 106–124. DOI : 10.2307/1415404.

[7] Jean-Marc COLLETTA et Anna TCHERKASSOF. *Les émotions : cognition, langage et développement*. T. 247. Editions Mardaga, 2003.

[8] J.M. COLLETTA et A. TCHERKASSOF. *Les émotions : cognition, langage et développement*. Psychologie et sciences humaines. Mardaga, 2003.

[9] Diane J Cook et al. « CASAS : A smart home in a box ». In : *Computer* 46.7 (2013), p. 62–69. DOI : 10.1109/MC.2012.328.

[10] Timothy F Cootes, Gareth J Edwards et Christopher J Taylor. « Active appearance models ». In : *IEEE Transactions on Pattern Analysis & Machine Intelligence* 6 (2001), p. 681–685.

[11] Timothy F Cootes et al. « Active shape models-their training and application ». In : *Computer vision and image understanding* 61.1 (1995), p. 38–59. DOI : 10.1006/cviu.1995.1004.

[12] Roddy Cowie et al. « 'FEELTRACE' : An instrument for recording perceived emotion in real time ». In : *ISCA tutorial and research workshop (ITRW) on speech and emotion*. 2000.

[13] Navneet Dalal et Bill Triggs. « Histograms of oriented gradients for human detection ». In : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. T. 1. IEEE. 2005, p. 886–893. DOI : 10.1109/CVPR.2005.177.

[14] Charles Darwin et Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[15] Gérard Dreyfus. *Apprentissage statistique*. Editions Eyrolles, 2008.

[16] Paul Ekman. « Are there basic emotions ? » In : 99.3 (1992), p. 550–553. DOI : 10.1037/0033-295X.99.3.550.

[17] Paul Ekman et Wallace V Friesen. « Constants across cultures in the face and emotion. » In : *Journal of personality and social psychology* 17.2 (1971), p. 124. DOI : 10.1037/h0030377.

[18] Paul Ekman et Wallace V. Friesen. « Measuring facial movement ». In : *Environmental psychology and nonverbal behavior* 1.1 (1976), p. 56–75. DOI : 10.1007/BF01115465.

[19] Terrence Fong, Illah Nourbakhsh et Kerstin Dautenhahn. « A survey of socially interactive robots ». In : *Robotics and Autonomous Systems* 42.3 (2003), p. 143–166. DOI : 10.1016/S0921-8890(02)00372-X.

[20] Dany FORTIN-SIMARD et al. « Exploiting passive RFID technology for activity recognition in smart homes. » In : *IEEE Intelligent Systems* 30.4 (2015), p. 7–15. DOI : 10.1109/MIS.2015.18.

[21] Shaogang GONG, Chen Change LOY et Tao XIANG. « Security and Surveillance ». In : *Visual Analysis of Humans : Looking at People*. Sous la dir. de Thomas B. MOESLUND et al. London : Springer London, 2011, p. 455–472. DOI : 10.1007/978-0-85729-997-0_23.

[22] Ian GOODFELLOW, Yoshua BENGIO et Aaron COURVILLE. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[23] Alfred HAAR. « Zur Theorie der orthogonalen Funktionensysteme ». In : *Mathematische Annalen* 69.3 (1910), p. 331–371. DOI : 10.1007/BF01456326.

[24] Don H HOCKENBURY et Sandra E HOCKENBURY. *Discovering psychology*. Macmillan, 2010.

[25] Eva HUDLICKA. « Affective Game Engines : Motivation and Requirements ». In : *Proceedings of the 4th International Conference on Foundations of Digital Games*. FDG '09. Orlando, Florida : ACM, p. 299–306. DOI : 10.1145/1536513.1536565.

[26] Farzad HUSAIN, Babette DELLEN et Carme TORRAS. « Scene Understanding Using Deep Learning ». In : *Handbook of Neural Computation*. Elsevier, 2017, p. 373–382. DOI : 10.1016/B978-0-12-811318-9.00020-X.

[27] Carroll E. IZARD. *Human Emotions*. Springer US, 1977. DOI : 10.1007/978-1-4899-2209-0.

[28] Alejandro JAIMES et Nicu SEBE. « Multimodal human–computer interaction : A survey ». In : *Computer Vision and Image Understanding* 108.1 (2007), p. 116–134. DOI : 10.1016/j.cviu.2006.10.019.

[29] William JAMES. « What is an Emotion ? » In : *Mind* 9.34 (1884), p. 188–205.

[30] Takeo KANADE, Jeffrey F COHN et Yingli TIAN. « Comprehensive database for facial expression analysis ». In : *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 2000, p. 46–53. DOI : 10.1109/AFGR.2000.840611.

[31] Vahid KAZEMI et Josephine SULLIVAN. « One millisecond face alignment with an ensemble of regression trees ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, p. 1867–1874. DOI : 10.1109/CVPR.2014.241.

[32] Faiza KHALFI. « Reconnaissance automatique des émotions par données multimodales : expressions faciales et des signaux physiologiques ». Thèse de doct. Université Paul Verlaine-Metz, 2010.

[33] T. KO. « A survey on behavior analysis in video surveillance for homeland security applications ». In : *2008 37th IEEE Applied Imagery Pattern Recognition Workshop(AIPR)*. 2008, p. 1–8. DOI : 10.1109/AIPR.2008.4906450.

[34] Amit KONAR, Anisha HALDER et Aruna CHAKRABORTY. « Introduction to Emotion Recognition ». In : *Emotion Recognition*. John Wiley & Sons, Inc., 2015, p. 1–45. DOI : 10.1002/9781118910566.ch1.

[35] Oliver LANGNER et al. « Presentation and validation of the Radboud Faces Database ». In : *Cognition and emotion* 24.8 (2010), p. 1377–1388. DOI : 10.1080/02699930903485076.

[36] Yann LECUN, Yoshua BENGIO et Geoffrey HINTON. « Deep learning ». In : *Nature* 521.7553 (2015), p. 436–444. DOI : 10.1038/nature14539.

[37] Yann LECUN, Koray KAVUKCUOGLU et Clément FARABET. « Convolutional networks and applications in vision ». In : *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2010, p. 253–256. DOI : 10.1109/ISCAS.2010.5537907.

[38] Yann LECUN et al. « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11 (1998), p. 2278–2324. DOI : 10.1109/5.726791.

[39] Yann LECUN et al. « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11 (1998), p. 2278–2324. DOI : 10.1109/5.726791.

[40] M. LEO et al. « Computer vision for assistive technologies ». In : *Computer Vision and Image Understanding* 154 (2017), p. 1–15. DOI : 10.1016/j.cviu.2016.09.001.

[41] G.C. LITTLEWORT et al. « Towards Social Robots : Automatic Evaluation of Human-Robot Interaction by Facial Expression Classification ». In : *Advances in Neural Information Processing Systems 16*. Sous la dir. de S. THRUN, L. K. SAUL et B. SCHÖLKOPF. MIT Press, 2004, p. 1563–1570.

[42] André Teixeira LOPES et al. « Facial expression recognition with convolutional neural networks : coping with few data and the training sample order ». In : *Pattern Recognition* 61 (2017), p. 610–628. DOI : 10.1016/j.patcog.2016.07.026.

[43] Elena LOZANO-MONASOR et al. « Facial expression recognition in ageing adults : from lab to ambient assisted living ». In : *Journal of Ambient Intelligence and Humanized Computing* 8.4 (2017), p. 567–578. DOI : 10.1007/s12652-017-0464-x.

[44] Daniel LUNDQVIST, Anders FLYKT et Arne ÖHMAN. *The Karolinska directed emotional faces (KDEF) - CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*. 1998.

[45] Michael LYONS et al. « Coding facial expressions with gabor wavelets ». In : *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE. 1998, p. 200–205. DOI : 10.1109/AFGR.1998.670949.

[46] Albert MEHRABIAN. « Communication without words ». In : *Psychology Today*. 2^{e} éd. 1968, p. 51–52.

[47] Alex MIHAILIDIS et al. « The COACH prompting system to assist older adults with dementia through handwashing : An efficacy study ». In : *BMC geriatrics* 8.1 (2008), p. 28. DOI : 10.1186/1471-2318-8-28.

[48] Qin NI, Ana Belén GARCÍA HERNANDO et Iván Pau de la CRUZ. « The elderly's independent living in smart homes : A characterization of activities and sensing infrastructure survey to facilitate services development ». In : *Sensors* 15.5 (2015), p. 11312–11362. DOI : 10.3390/s150511312.

[49] Armelle NUGIER. « Histoire et grands courants de recherche sur les émotions ». In : *Revue électronique de psychologie sociale* 4.4 (2009), p. 8–14.

[50] Timo OJALA, Matti PIETIKÄINEN et David HARWOOD. « A comparative study of texture measures with classification based on featured distributions ». In : *Pattern recognition* 29.1 (1996), p. 51–59. DOI : 10.1016/0031-3203(95)00067-4.

[51] Maja PANTIC et al. « Web-based database for facial expression analysis ». In : *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2005, 5–pp. DOI : 10.1109/ICME.2005.1521424.

[52] Christian PETERS, Thomas HERMANN et Sven WACHSMUTH. « TEBRA - An automatic prompting system for persons with cognitive disabilities in brushing teeth ». In : *Proceeding of the 6th International Conference on Health Informatics (HealthInf)*. Barcelona, Spain, 2013, p. 12–23. DOI : 10.5220/0004193800120023.

[53] Robert PLUTCHIK. « A psychoevolutionary theory of emotions ». In : *Social Science Information* 21.4-5 (1982), p. 529–553. DOI : 10.1177/053901882021004003.

[54] James A RUSSELL. « A circumplex model of affect. » In : *Journal of personality and social psychology* 39.6 (1980), p. 1161. DOI : 10.1037/h0077714.

[55] Nicu SEBE et al. « Multimodal approaches for emotion recognition : a survey ». In : *Internet Imaging VI*. T. 5670. International Society for Optics et Photonics. 2005, p. 56–68. DOI : 10.1117/12.600746.

[56] Anthony TEOLIS. « Discrete Wavelet Transform ». In : *Computational Signal Processing with Wavelets*. Boston, MA : Birkhäuser Boston, 1998, p. 89–126. DOI : 10.1007/978-1-4612-4142-3_5.

[57] Silvan S TOMKINS. « Affect theory ». In : *Approaches to emotion* 163.163–195 (1984).

[58] Paul VIOLA et Michael JONES. « Rapid object detection using a boosted cascade of simple features ». In : *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. T. 1. IEEE. 2001, p. 511–518. DOI : 10.1109/CVPR.2001.990517.

[59] J. WHITEHILL, M. BARTLETT et J. MOVELLAN. « Automatic facial expression recognition for intelligent tutoring systems ». In : *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2008, p. 1–6. DOI : 10.1109/CVPRW.2008.4563182.

[60] Ian H WITTEN et al. *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[61] Yacine YADDADEN et al. « A New Approach of Facial Expression Recognition for Ambient Assisted Living ». In : *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*. ACM. 2016, 14 :1–14 :8. DOI : 10.1145/2910674.2910703.

[62] Yacine YADDADEN et al. « An Automatic Facial Expression Recognition Approach using an Efficient Spatio-Temporal Representation ». In : *Journal of Ambient Intelligence and Humanized Computing* (2019).

[63] Yacine YADDADEN et al. « Facial Expression Recognition from Video using Geometric Features ». In : *Proceedings of the 8th IET International Conference on Pattern Recognition Systems*. Institution of Engineering et Technology. 2017, 4(6 .)–4(6 .)(1). DOI : 10.1049/cp.2017.0133.

[64] Yacine YADDADEN et al. « Facial expressions based error detection for smart environment using deep learning ». In : *Proceedings of the 14th International Conference on Ubiquitous Intelligence Computing*. IEEE. 2017, p. 1–7. DOI : 10.1109/UIC-ATC.2017.8397536.

[65] Yacine YADDADEN et al. « Facial Sub-region for automatic Emotion Recognition using Local Binary Patterns ». In : *Proceedings of the 4th International Conference on Signal, Image, Vision and their Applications*. IEEE. 2018, p. 1–6.

[66] Yacine YADDADEN et al. « Hybrid-based Facial Expression Recognition Approach for Human-Computer Interaction ». In : *Proceedings of the 20th International Workshop on Multimedia Signal Processing*. IEEE. 2018, p. 1–6. DOI : 10.1109/MMSP.2018.8547081.

[67] Yacine YADDADEN et al. « One-Class and Bi-Class SVM Classifier Comparison for Automatic Facial Expression Recognition ». In : *Proceedings of the International Conference on Applied Smart Systems*. IEEE. 2018, p. 1–6.

[68] Yacine YADDADEN et al. « User action and facial expression recognition for error detection system in an ambient assisted environment ». In : *Expert Systems with Applications* 112 (2018), p. 173–189. DOI : 10.1016/j.eswa.2018.06.033.