Annotation automatique des gestes dans des vidéos de personnes âgées durant des conversations spontanées

par

Helmi GARRAOUI

THÈSE PRÉSENTÉE À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DU
DOCTORAT EN GÉNIE
Ph. D

MONTRÉAL, LE 29 SEPTEMBRE 2021

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

Mme Sylvie Ratté, directrice de thèse
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Luc Duong, codirecteur de thèse
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Saad Maarouf, président du jury
Département de génie électrique à l'École de technologie supérieure

M. Carlos Vasquez, membre du jury
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Pierrich Plusquellec, examinateur externe
École de psychoéducation à la Faculté des arts et des sciences

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 29 SEPTEMBRE 2021

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Mes remerciements sont destinés en premier lieu à mes Directeurs de Recherche, Sylvie Ratté et Luc Duong. Sylvie, Merci pour ta disponibilité, tes conseils, ta patience, ta générosité et ton humanité. Luc, merci pour ta rigueur scientifique et ton ardeur au travail. Tu m'as appris à toujours pousser plus loin la réflexion.

Je remercie particulièrement Saad Maarouf, Carlos Vasquez et Pierrich Plusquellec d'avoir accepté d'être des membres de jury.

Je remercie l'ensemble des étudiants du LiNCS et LIVE, pour leur aide et conseils dans mes travaux de recherche

Je dédie cet événement marquant de ma vie :

- À la mémoire de mon oncle ELHADJ Mohamed Salah disparu le 01-04-2018.

- À mon très cher père Afif, ma très chère mère Jamila. Ce modeste travail est le fruit de tous les sacrifices qu'ils ont déployés pour mon éducation et ma formation.

- À ma chère femme Asma, pour son inestimable soutien au cours de la route qui m'a mené au terme de cette recherche.

Annotation automatique des gestes dans des vidéos de personnes âgées durant des conversations spontanées

Helmi GARRAOUI

RÉSUMÉ

L'étude des gestes de l'humain en général et les mouvements de la tête et des mains en particulier constitue une partie importante de la communication non verbale. C'est pour cette raison le nombre des études qui portent sur l'analyse des gestes ne cesse pas d'augmenter Parmi ces études on trouve celles qui concernent l'analyse des vidéos afin de reconnaître et interpréter des informations spécifiques dans une conversation à travers l'annotation des gestes. Annoter les mouvements de la tête et des mains peut être une tâche difficile dans le cas des personnes âgées, car les capacités motrices et sensorielles diminuent avec l'âge. Présentement, l'annotation des gestes dans les vidéos se fait d'une façon manuelle. Ceci présente plusieurs limites. Elle est fastidieuse et imprécise vue qu'elle est soumise à la variabilité des annotateurs. De même ce type d'annotation est couteux puisque il ne peut être réalisé que par des experts. De plus les conclusions tirées ne sont pas robustes étant donné que l'annotation manuelle est souvent appliquée à des petits échantillons. Dans ce travail de recherche, nous avons proposé deux approches automatiques pour annoter les mouvements des mains et les mouvements de la tête. Ces approches sont basées sur des techniques d'apprentissage profond à savoir, le CNN et le RNN. Nous avons recours à une vérité terrain réalisée par des experts afin de valider les approches proposées.

Dans un premier temps, nous avons proposé une approche pour annoter les mouvements de la tête. La méthode développée est inspirée des standards développés par des experts dans le domaine linguistique. Le problème d'annotation a été modélisé comme un exercice de classification. Chaque geste simple, composé d'un seul mouvement, constitue une classe alors que les gestes complexes, composés de plus d'un mouvement, sont classés dans une seule classe. Pour ce faire, nous avons implémenté une méthode basée sur la technique de MTCNN, cette technique a été utilisée pour détecter le visage. Puis le LSTM est appliqué pour prédire la classe de chaque mouvement.

Dans un second temps, nous avons proposé une approche pour automatiser les phases gestuelles des mains qui existent dans la littérature depuis les années quatre-vingt. Nous nous sommes basés sur des études antérieures utilisées dans la classification des phases gestuelles. La stratégie utilisée pour annoter automatiquement les phases gestuelles est similaire à celle proposée pour annoter les mouvements des mains. En effet, le problème d'annotation est considéré comme un problème de classification. Sur un plan technique, nous avons utilisé le MobileNet pour détecter les mains et le LSTM pour prédire la phase en question.

Les approches proposées nous ont permis de réduire le coût des annotations manuelles, de minimiser le temps de travail et de fournir une annotation plus scientifique en diminuant l'impact de la subjectivité causée par l'imprécision générée par une analyse visuelle des gestes dans une vidéo. Les résultats proposent deux axes de recherche à explorer.

Premièrement, le processus d'annotation automatique des comportements non verbaux dans les vidéos est possible. Deuxièmement, les algorithmes d'apprentissage profond peuvent identifier des caractéristiques qui ne correspondent pas totalement aux observations des experts. Cette particularité établit clairement un nouveau dialogue entre les chercheurs en intelligence artificielle, les chercheurs en linguistique et les chercheurs dans les domaines liés à la communication et au vieillissement, quant à l'interprétation des résultats et des caractéristiques à partir de vidéos

Mots-clés : communication non verbale, vieillissement, annotation automatique, apprentissage profond

Automatic annotation of gestures in videos of elderly people during spontaneous conversations

Helmi GARRAOUI

ABSTRACT

The study of human gestures in general and the movements of the head and hands in particular is an important part of non-verbal communication. For this reason the number of studies related to gestures analysis follows an increasing curve. Among these studies are those which concern the videos analysis in order to recognize and interpret specific information in a conversation through gestures annotation. Noting head and hand movements of elderly people can be a difficult task, since motor and sensory skills decline with age. Currently, the annotation of gestures in videos is done manually. This has several limitations. It is tedious and imprecise since it is subject to the variability of the annotators. Likewise, this type of annotation is expensive since it must be carried out only by experts. In addition, the conclusions drawn are not robust since manual annotation is often applied to small samples. In this research work, we have proposed two automatic approaches to annotate hand movements and head movements. These approaches are based on deep learning techniques namely, CNN and RNN. We use ground truth carried out by experts in order to validate the proposed approaches.

First, we proposed an approach to annotate head movements. The developed method is inspired from standards developed by experts in the linguistic field. The annotation problem has been modeled as a classification task. Each simple gesture, consisting of a single movement, constitutes a class while complex gestures, consisting of more than one movement, are classified in a single class. To do this, we implemented a method based on the MTCNN technique, this technique was used to detect the face. Then the LSTM is applied to predict the class of each movement.

Secondly, we proposed an approach to automate the gestural phases of the hands that has existed in the literature since the Eighties. Our work is based on previous studies used in gestural phases classification. The strategy used to automatically annotate the gestural phases is similar to that proposed to annotate the movements of hands. Indeed, the annotation problem is considered a classification problem. Technically, we used MobileNet to detect hands and LSTM to predict the phase in question.

The proposed process focuses on the automatic annotation of head and hands, reducing the cost and the time of the annotation process and decreasing the impact of subjectivity caused by the imprecision generated by a visual analysis of gestures on a video. The results suggest two research focuses for further exploration. First, the automatic annotation process for nonverbal behaviors in videos is quite feasible; second, machine learning algorithms have the capacity to identify features that are not totally in sync with what humans are used to. These characteristics clearly open up a new dialog between artificial intelligence and researchers in

linguistics, and researchers in fields related to communication and aging in terms of interpreting results and features in videos

Keywords: nonverbal communication, elderly people, automatic annotation, deep learning

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

LSA	Langage des Signes Américains
ASM	Active Shape Models
MLC	Constrained Local Model
AAM	Active Appearance Model
SVM	Support Vector Machine
PCA	Principal Component Analysis
GPS	Global Positioning System
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
DL	Deep Learning
MTCNN	Multi-task Convolutional Neural Network
TCDCN	TasksConstrained Deep Convolutional Network
FCN	Fully Convolutional Networks
P-Net	Proposal Network
R-Net	Refine Network
O-Net	Output Network
SSD	Single-Shot multibox Detection
PPV	Valeur Prédictive Positive
TPR	Taux de Vrais Positifs
VP	Vrais Positifs
VN	Vrais Négatifs

FP	Faux Positifs
FN	Faux Négatifs
IS	Isotropic smoothing
DCT	Discrete Cosine Transform
DoG	Difference of Gaussian
LMC	Leap Motion Controller

INTRODUCTION

La communication joue un rôle capital dans la vie de l'être humain. Grâce à elle, l'individu peut interagir avec son entourage en échangeant des messages. En fait, la communication humaine se fait non seulement à travers la parole, mais aussi à travers les signaux non verbaux y compris les informations communiquées dans un échange social par les signes accompagnant les paroles dans une conversation. Parmi les canaux de communications, nous citons les sons, les gestes, le langage corporel, le ton, le contact visuel et le mouvement. Il est important de mentionner que seulement 5 % de nos communications sont produites par la parole, 45 % par le ton, l'inflexion et d'autres éléments de la voix et 50 % par le langage corporel, le mouvement et le contact visuel (Birdwhistell, 1977). Certains chercheurs tels que Ruesch & Kees (1969) attribuent à la communication non verbale une contribution majeure de 65 % par rapport à la transmission d'informations par voie verbale qui est égale à 35 %. D'autres chercheurs tels que Rujoiu (2009) et d'autres affirment que dans les relations interpersonnelles, l'information est transmise en proportion de 82 % par le corps et la voix, et seulement 18 % par le langage verbal.

Les comportements non verbaux tels que les mouvements de tête, les gestes de la main, la posture du corps ou l'expression du visage fournissent beaucoup d'informations sur les attitudes interpersonnelles, les intentions comportementales et les expériences émotionnelles. Ainsi, la communication non verbale joue un rôle important dans la régulation de l'interaction entre les individus. Le rôle de la communication non verbale ne se limite pas à l'échange d'information, mais elle peut aider à identifier quelques maladies de démences telles que le Parkinson et l'Alzheimer. En effet, plusieurs travaux de recherches mettent en évidence l'idée que le changement cérébral risque de se produire, plusieurs décennies, avant une première manifestation clinique chez les patients d'Alzheimer. Ces changements peuvent affecter quelques fonctions telles que la motricité (Albers et al., 2015). La détection clinique des altérations ne peut se faire que dans un stade avancé de la maladie. À ce stade-là, la consultation des médecins et des spécialistes est souvent stressante pour les patients et leurs proches. En plus, les interventions, à cette étape, ne sont pas très efficaces. Le

développement d'un outil automatique pour étudier précocement les mouvements chez les personnes âgées susceptibles d'être des victimes potentielles des maladies de démences est l'objectif principal de mon projet de recherche. L'idée est d'annoter les mouvements de la tête et des mains à partir des enregistrements vidéo des personnes âgées en utilisant des techniques d'apprentissage profond et de reconnaissance de patrons gestuels. Les résultats de ce travail de recherche peuvent être utilisés par des spécialistes de la santé, par des linguistes et par la grande communauté qui s'intéressent à l'étude de la communication non verbale chez la population âgée.

Ce document est organisé en six chapitres. Dans le premier chapitre, nous allons présenter le cadre général du projet de recherche en mettant en lumière la répercussion du vieillissement sur la société ainsi que les motivations d'aborder ce sujet de recherche. Nous allons présenter aussi la problématique de recherche ainsi que l'objectif principal et les sous objectifs associés. Le deuxième chapitre, intitulé « revue de littérature », contient les travaux qui ont une relation avec notre sujet de recherche. Nous commençons en premier lieu par expliquer la relation entre les mouvements gestuels et le langage verbal en nous basant sur quelques études. Dans un deuxième lieu nous présentons les travaux de recherches qui portent sur l'analyse des mouvements ainsi que leurs domaines d'application et puis nous détaillons les différentes techniques utilisées que se soit au niveau des suivis des mains ou au niveau de la reconnaissance d'activités humaines. Enfin nous exposons les travaux de recherche qui portent sur l'annotation des gestes. En effet, nous présentons les travaux qui se basent sur une annotation manuelle ou bien ceux qui se basent sur une annotation automatique qui concernent une partie du corps humain. Le troisième chapitre intitulé méthodologie générale présente la partie commune entre l'annotation des mouvements des mains et l'annotation des mouvements de la tête afin d'éviter la redondance dans les deux chapitres qui suivent. Nous allons mettre en lumière le processus d'annotation des gestes, le corpus utilisé et les métriques d'évaluation. Le quatrième chapitre porte sur l'approche proposée pour annoter les mouvements de la tête tandis que le cinquième chapitre met en lumière l'approche proposée pour annoter les mouvements des mains. Le sixième chapitre est une discussion générale sur les approches proposées dans notre thèse. Nous mettons en valeur les points forts et nous

présentons les points faibles ou qui nécessitent une amélioration afin d'ouvrir d'autres pistes de recherche. Nous clôturons cette thèse par une conclusion générale.

CHAPITRE 1

PROBLÉMATIQUE ET OBJECTIFS

L'évolution du nombre des personnes âgées au Canada et plus particulièrement au Québec suit une courbe exponentielle. Ceci est dû à plusieurs facteurs tels que l'amélioration de l'espérance de vie et la faible fécondité. Cette classe de population nécessite une préoccupation spéciale. En effet, certaines de ces personnes âgées subissent un déclin de leurs fonctions cognitives et motrices dues au vieillissement normal ou aux maladies cognitives telles que la maladie d'Alzheimer (MA) et les maladies de démences. La compréhension du sujet du vieillissement nous permet de mieux cerner son impact sur la communication non verbale chez les personnes âgées. Nous clarifions dans ce chapitre quelques notions qui demeurent importantes pour décortiquer le sujet de la communication non verbale en relation avec les déficits moteurs dus au vieillissement normal ou aux maladies neurodégénératives telle que la MA.

Dans ce chapitre, nous mettons en lumière la relation entre le vieillissement et l'inhibition motrice, nous présentons aussi la population québécoise par classe d'âge, nous couvrons les statistiques qui concernent le nombre des patients et les coûts engendrés par la MA. Une deuxième section porte sur les répercussions du vieillissement de la population sur la société. La troisième section est dédiée aux motivations pour étudier la communication non verbale chez les personnes âgées. Dans la quatrième section, nous présentons la problématique de recherche. Les objectifs seront présentés dans la cinquième section.

1.1 Mise en contexte

Dans la société québécoise, le nombre d'aînés (65 ans et plus) ne cesse de croître. En effet, selon l'institut national de santé publique la proportion des personnes âgées augmente très rapidement ; elle représentait 12 % de la population âgée de 65 ans et plus en 2011 et elle va atteindre 25 % en 2061 comme le montre la figure 1.

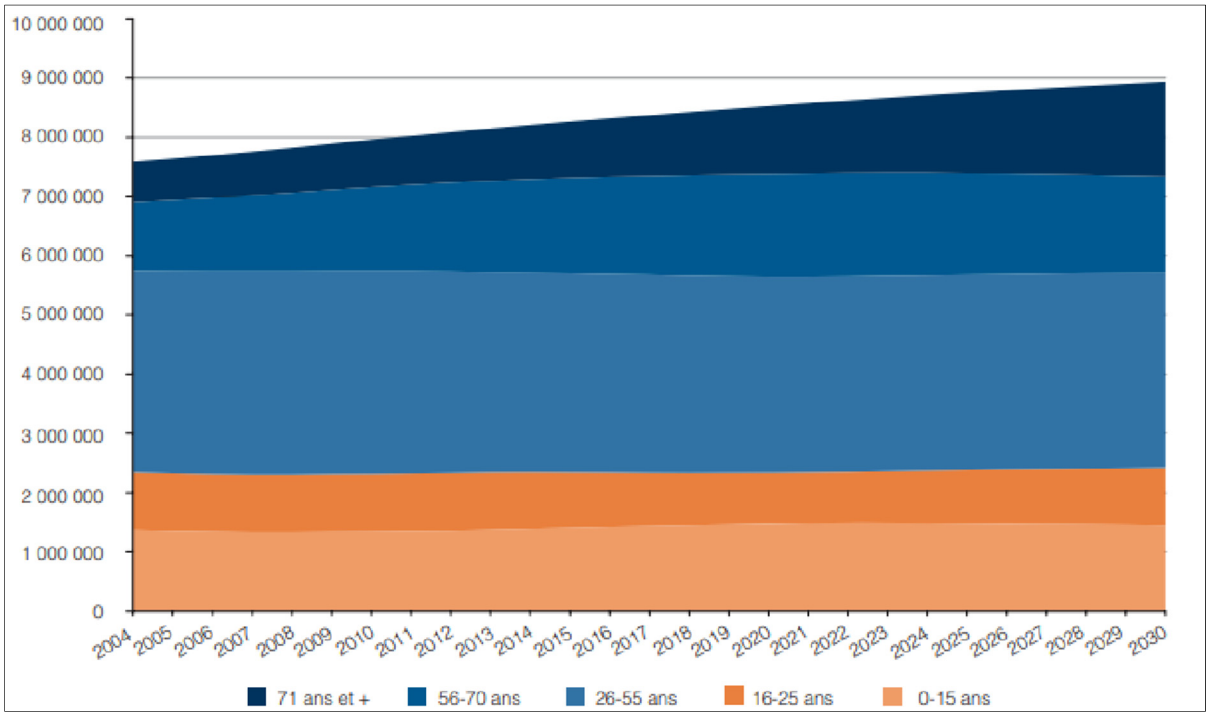


Figure 1.1 Population par classe d'âge Québec, 2004 à 2030

Ceci met en évidence l'importance des études qui peuvent contribuer à comprendre les besoins de ces personnes âgées, l'une de ces études consiste à analyser leurs communications non verbales. Il est à noter que le vieillissement est accompagné d'un processus d'inhibition motrice (Grandjean & Collette, 2011) et selon la société d'Alzheimer dans des cas, de changements cérébraux dus à la maladie d'Alzheimer.

1.2 Répercussion du vieillissement sur la société

La maladie d'Alzheimer est la cause la plus fréquente de démence chez les personnes âgées de 65 ans et plus. D'après la société d'Alzheimer, le tiers des personnes âgées souffrent de l'Alzheimer. Au Canada il y a près de 800 000 personnes qui sont âgées de 85 ans et plus. Ce nombre est en augmentation continue, il atteindra 3 000 000 en 2051 ; ce qui implique un

million de personnes potentiellement atteintes. La figure 2 montre l'évolution de la population âgée de 85 ans et plus depuis 1996 jusqu'à 2051

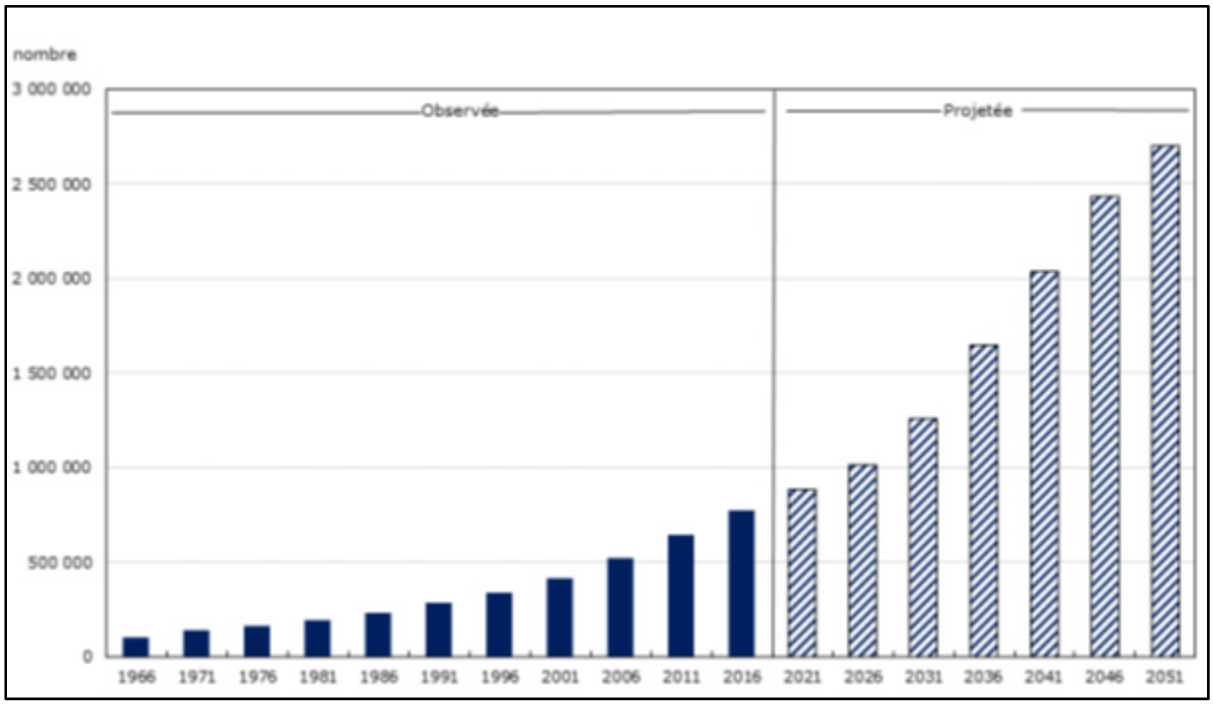


Figure 1.2 Croissance de la population âgée de 85 ans et plus entre 1996 et 2051

Les données de 2021 à 2061 sont des projections démographiques tirées du scénario de croissance moyenne des projections nationales. Les données des projections ont comme population de départ les estimations démographiques basées sur le recensement de 2011, ajustées pour tenir compte du sous-dénombrement net.

Il est à signaler que cette croissance du nombre de personnes atteintes est accompagnée d'un fardeau économique qui ne cesse d'augmenter chaque année pour atteindre 293 milliards de dollars en 2040 c'est-à-dire un fardeau économique cumulatif attribuable aux démences qui dépasse le 800 M$ comme le montre la figure 1.2. Ceci rend les maladies de démences causées par le vieillissement, un problème de santé publique et un défi pour la société.

La gravité des maladies de démences causées par le vieillissement nécessite une intervention afin de protéger la vie des patients et réduire le fardeau économique énorme. Plusieurs

études semblent indiquer qu'une intervention précoce peut être considérée comme une partie de la solution.

1.3 L'intervention précoce comme piste d'atténuation de la perte de qualité de vie chez les aînés

Nous avons mentionné dans la section précédente qu'une intervention précoce peut être considérée comme une partie de la solution puisqu'aucun traitement ne peut arrêter la maladie d'Alzheimer qui est causée principalement par le vieillissement, par contre on peut diminuer l'effet de la maladie à travers des interventions efficaces afin de maintenir plus longtemps une vie quotidienne stable pour le patient. En effet, une intervention précoce permet un gain de qualité de vie et de longévité. Le graphique de la figure 1.3 montre l'évolution de la maladie avec et sans traitement. On remarque bien sur le graphique la valeur ajoutée par le traitement, chose qui a encouragé la communauté des chercheurs à donner une importance à l'intervention précoce.

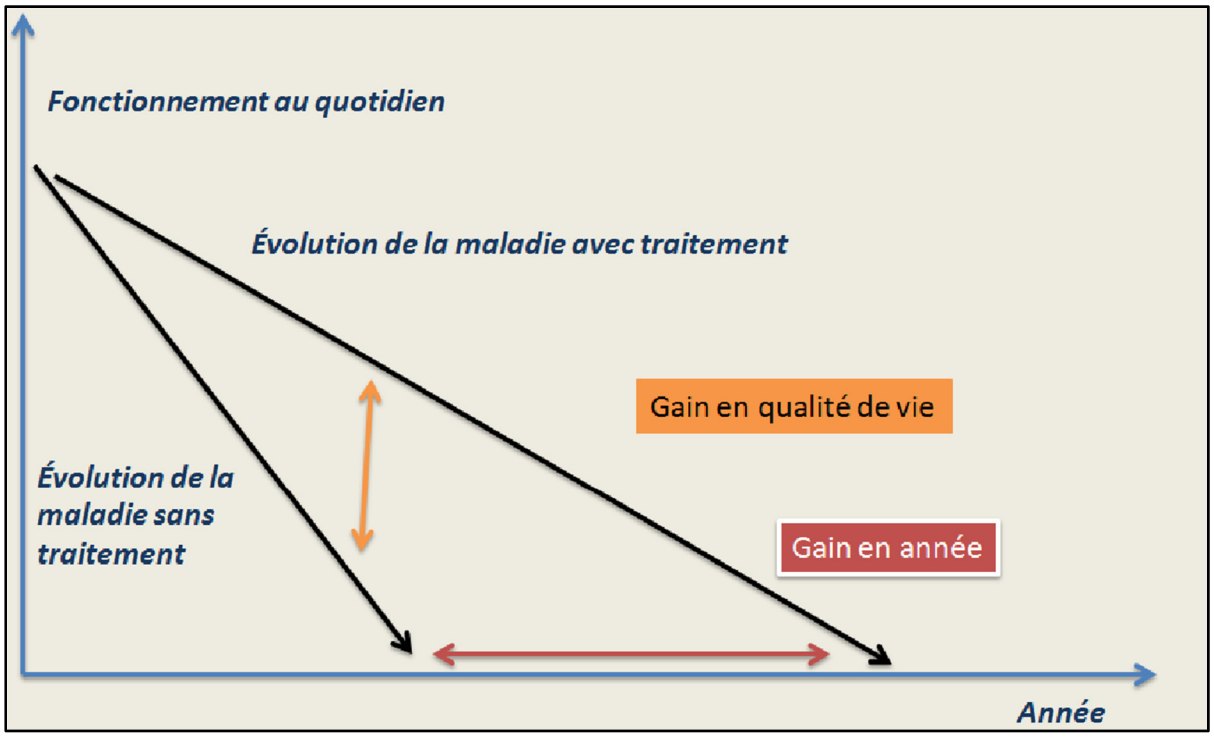


Figure 1.3 Évolution de la maladie avec et sans traitement adaptée de la présentation de Patigny (2001)

L'une de ces interventions consiste à étudier la communication non verbale. En effet, selon Sauter, McDonald, Gangi, & Messinger (2014) la communication non verbale en général et les gestes en particulier sont des indicateurs du maintien de capacités de communication et de l'évolution du déficit verbal. Cette idée a été confirmée par Przygodzki-Lionet & Schiaratura (2007). Selon cette étude, la communication non verbale facilite l'échange entre les personnes et assure la transmission des messages. Une autre étude réalisée par Di Pastena, Schiaratura, & Askevis-Leherpeux (2015) supporte les hypothèses présentées auparavant. Cette étude affirme que c'est important de mentionner que les malentendus ont des conséquences négatives sur l'état du patient.

Ce projet est axé sur l'analyse automatique des conversations de personnes âgées, dans le but de détecter des altérations dans le langage verbal et non verbal présentes chez les personnes âgées susceptibles d'être atteintes de la maladie d'Alzheimer dans un stade précoce. L'intention de ce projet est de proposer une solution économique et non invasive pour la détection des signes de la maladie d'Alzheimer et son suivi.

Ce projet propose d'analyser deux types de communications : la communication verbale et la communication non verbale. Dans le premier type, nous nous intéressons à la transcription et au signal audio. Dans le deuxième type, nous mettons l'accent sur les expressions faciales et les expressions corporelles (les gestes). L'hypothèse principale de notre étude est la possibilité d'exploiter des techniques d'apprentissage machine et d'apprentissage profond pour annoter automatiquement les gestes chez les personnes âgées.

1.4 Problématique de recherche

Plusieurs chercheurs tels que Albers et al. (2015) et P. V. Rousseau, Matton, Lecuyer, & Lahaye (2017) ont noté que le vieillissement induit des changements au niveau de la communication en général, et au niveau des gestes en particulier. Ces changements sont présents chez une majorité des personnes âgées. Ce travail de recherche vise à annoter automatiquement les gestes chez les personnes âgées en s'appuyant sur des enregistrements vidéo de conversations naturelles. À

l'heure actuelle, l'analyse des vidéos est effectuée par des experts d'une façon manuelle. Ce type d'analyse présente plusieurs limites. Il est fastidieux et imprécis vu qu'il est soumis à la variabilité des annotateurs. De même, ce type d'analyse est coûteux puisqu'il doit être réalisé que par des experts. De plus, il nécessite beaucoup de temps d'analyse. Sans oublier que les conclusions tirées ne sont pas robustes étant donné que l'analyse manuelle est souvent appliquée à des petits échantillons.

1.5 Objectifs

1.5.1 Objectif principal

L'objectif principal de notre recherche consiste à annoter automatiquement les gestes de la tête et des mains chez les personnes âgées. Plus précisément ce travail de recherche vise deux objectifs principaux qui se présentent ainsi :

- ***SO 1 : Annoter automatiquement les gestes de la tête selon les standards existants;***

- ***SO 2 : Annoter automatiquement les gestes des mains selon les standards existants.***

1.5.2 Sous objectifs

- ***SO 1 : Annoter automatiquement les gestes de la tête selon les standards existants et les analyser***;

L'objectif est d'automatiser le travail manuel fait par les experts. Pour atteindre cet objectif, nous nous basons sur une vérité terrain. Nous proposons un nouveau modèle d'annotation du langage non verbal basé sur l'application des techniques d'apprentissage profond. Grâce à l'approche proposée, nous avons été en mesure d'annoter les images d'une façon similaire aux annotations existantes dans la vérité terrain. Dans ce travail de recherche, nous allons utiliser une combinaison des techniques d'apprentissage profond composée de réseaux de neurones convolutifs et de réseaux de neurones récurrents.

- ***SO 2 : Annoter automatiquement les gestes des mains selon les standards existants***.

L'approche proposée pour annoter automatiquement les mouvements des mains est similaire à celle de l'annotation de la tête. Elle est basée sur des techniques d'apprentissage profond.

Conclusion

Dans ce chapitre, nous avons mis l'accent sur le contexte général de ce travail de recherche. Nous avons aussi mis en lumière les motivations d'aborder ce sujet de recherche. Dans une autre section, nous avons défini la problématique de recherche ainsi que les objectifs à atteindre. Dans le chapitre suivant nous allons aborder la revue de littérature, nous allons présenter les travaux de recherche qui ont une relation avec notre sujet de recherche.

CHAPITRE 2

REVUE DE LITTÉRATURE

2.1 Introduction

Ce chapitre est consacré à présenter les travaux sur lesquels notre recherche s'appuie. En effet, le sujet de cette thèse est un sujet multidisciplinaire qui englobe la communication non verbale, le vieillissement et la reconnaissance des mouvements. La combinaison de ces trois sujets rend notre sujet unique et original. Dans ce chapitre, nous présentons les travaux de recherche qui ont une relation directe ou indirecte avec le sujet. En effet, l'annotation des gestes des mains et de la tête utilise des méthodes de suivi de visage et des mains et des méthodes d'estimation de pose. Les travaux qui portent sur la reconnaissance d'activité constituent une partie importante dans notre recherche, étant donné que l'annotation des gestes est un cas particulier de la reconnaissance des gestes. Dans la première partie de ce chapitre, nous présentons quelques travaux qui ont traité la relation entre la communication non verbale et le vieillissement afin de comprendre les conséquences du vieillissement sur la communication non verbale. Nous présentons, dans une deuxième partie, les approches automatiques qui concernent le suivi du visage, le suivi des mains, la reconnaissance d'activités ainsi que l'estimation des poses et la segmentation du corps humain. Une troisième partie est consacrée à exposer les travaux de recherche qui ont mis en lumière l'annotation automatique des gestes.

2.2 La relation entre les mouvements gestuels et le langage verbal

Cette section est destinée à présenter les travaux qui ont étudié la relation entre la communication non-verbale et les mouvements gestuels. Plusieurs approches ont été présentées, nous exposons dans le tableau suivant les approches les plus citées dans les revues de littérature.

Tableau 2.1 La relation entre les mouvements gestuels et le langage verbal

Auteurs	Constatations
McNeill (McNeill, 1985)	Les mouvements gestuels et le langage verbal sont deux aspects d'un même système de représentation qui sont derrière la structure conceptuelle du message.
Feyereisen (Feyereisen, 2007)	Les gestes pourraient également faciliter l'accès au lexique ou compenser un déficit verbal.
Davis (Davis, Maclagan, & Cook, 2013)	Les gestes et les mouvements de tête ont des fonctions spécifiques : ils aident à maintenir la conversation et supportent le patient à réagir à ses auditeurs
Di Pastena (Schiaratura, Di Pastena, Askevis-Leherpeux, & Clément, 2015)	La détérioration du langage des personnes atteintes de la MA n'affecte pas le registre gestuel. Le langage et les gestes dépendent de systèmes parallèles distincts plutôt que d'une répercussion des troubles langagiers dans la communication gestuelle

La plupart des recherches supportent l'hypothèse qui confirme que les gestes et le langage non verbale sont interdépendants. En effet, les gestes comblent le déficit verbal dans certains cas et participe dans la transmission de message dans d'autres cas

2.3 Analyses des mouvements : Domaines d'application et méthodologies développées

Metaxas & Zhang (2013) mettent l'accent sur les catégories de la communication non verbale. Ils définissent deux (catégories) :

- ***une catégorie bien structurée :*** Dans cette catégorie nous trouvons essentiellement les langues des signes. En effet, plusieurs travaux de recherche ont été réalisés qui portent sur la reconnaissance des langues des signes, basées sur réseaux de neurones convolutifs. Nous citons à titre d'exemple la méthode de Huang, Zhou, Zhang, Li, & Li (2018) . Ils ont proposé une approche pour la reconnaissance des langues de signe

continu. Dans l'article de Rao, Syamala, Kishore, & Sastry (2018) les auteurs proposent une méthode pour la reconnaissance des gestes de la langue des signes indienne;

- ***une catégorie moins structurée :*** Cette catégorie inclut plusieurs domaines d'application tels que la détection de la tromperie, les expressions émotionnelles, le stress et les déficiences dans les compétences cognitives et sociales. Il est important de mentionner que ces deux catégories s'appuient sur des méthodes d'analyse de mouvement robustes telles que le suivi des mouvements et la reconnaissance des gestes. Dans la suite, nous allons présenter les méthodes d'analyse de mouvements citées dans la littérature.

2.3.1 Suivie du visage

Les mouvements du visage représentent des indices importants pour la communication non verbale. Ainsi, de nombreuses recherches en vision par ordinateur mettent l'accent sur le suivi du visage. Nous distinguons différents modèles et algorithmes de suivi de visage, nous énumérons ci-dessous les plus utilisées.

2.3.1.1 Les modèles morphables

Les modèles de visage paramétriques ont d'abord été explorés pour suivre les traits du visage. Black & Yacoob (1995) ont examiné l'usage de modèles locaux paramétriques ainsi que les mouvements d'image afin de suivre les visages humains. Pighin, Szeliski, & Salesin (1999) ont utilisé une combinaison linéaire de modèles de texture 3D; chaque modèle correspond à une expression faciale de base particulière. Ils ont utilisé une technique d'interpolation de données dispersées pour s'adapter au visage du sujet de photographies.
Les modèles de visages paramétriques doivent être conçus minutieusement au préalable en utilisant un ensemble de paramètres contrôlant les déformations entraînées par des forces élastiques ou un mouvement d'image. Vu que ces modèles ne sont pas capables de représenter les structures anatomiques des os et des muscles, des formes irréalistes peuvent être générées. Une approche alternative qui consiste à entraîner des modèles 3D à partir d'un

groupe de formes de visage et de textures (Basso, Vetter, & Blanz, 2003) (Blanz & Vetter, 1999). Ces modèles de visage 3D présentent une grande variété de visages et de mouvements faciaux. Mais, ils sont coûteux en termes de calcul et ne peuvent pas fonctionner en temps réel.

2.3.1.2 Les modèles à forme active

Les modèles à forme active connus sous le nom active shape models (ASM) ont été proposés par Cootes & Taylor (1992). Ce sont des modèles déformables vus comme une extension « intelligente » des contours actifs. Les modèles actifs de forme décrivent l'aspect physique, Ils sont utilisés pour le suivi de la tête et pour analyser le visage.

Vogler & Metaxas (1999) ont présenté une méthode de reconnaissance faciale à travers les variations de pose, allant des vues frontales aux vues de profil, et à travers une large gamme d'illuminations, y compris les ombres projetées et les réflexions spéculaires. Pour tenir compte de ces variations, l'algorithme simule le processus de formation d'image dans l'espace 3D, à l'aide des techniques d'infographie, et estime la forme et la texture 3D des visages à partir d'images uniques. L'estimation est réalisée en ajustant un modèle statistique morphable de visages 3D à des images ce qui permet un suivi robuste des visages et une estimation exacte des mouvements.

2.3.1.3 Les modèles locaux contraints

Les modèles locaux contraints (MLC) sont des extensions des modèles à forme active. Ils utilisent un ensemble indépendant de détecteurs locaux pour la détection des repères (Cristinacce & Cootes, 2006).

Comme les détecteurs de repères locaux sont acquis à partir de petites régions d'image avec des structures limitées, les réponses maximales peuvent ne pas coïncider avec les emplacements des repères corrects. Certaines méthodes ont essayé de résoudre le problème tel que celui de Y. Wang, Lucey, & Cohn (2008), en proposant une fonction quadratique convexe à partir de laquelle la moyenne et la covariance de la densité d'approximation peuvent être déduites. Saragih, Lucey, & Cohn (2009) ont utilisé la différence de carrés comme une mesure d'ajustement de repère, et ont appliqué l'approximation de Laplace pour

trouver l'estimation de covariance. Saragih, Lucey, & Cohn, (2011) ont proposé une stratégie d'optimisation dans laquelle une représentation non paramétrique des distributions de repère est maximisée dans une hiérarchie d'estimations lissées.

2.3.1.4 Les modèles actifs d'apparence

Le modèle actif d'apparence (The active appearance model) AAM est le résultat d'un ensemble d'exemples de visages constituant une base d'apprentissage, chacun de ces visages est décrit par une forme et une texture (luminance des pixels inclus dans la forme). Les formes de l'ensemble d'apprentissages sont décalées les unes par rapport aux autres et les textures sont normalisées puis alignées sur la forme moyenne des visages de l'ensemble d'apprentissage (Saragih, Lucey, & Cohn, 2011b)

Le modèle actif d'apparence a été utilisé avec succès pour le suivi du visage en temps réel. Cootes, Wheeler, Walker, & Taylor (2002) proposent des modèles actifs d'apparence basés sur la vue, qui sont une combinaison de quelques modèles 2D pour faire face aux variations de la pose. Dollár, Rabaud, Cottrell, & Belongie (2005) ont combiné le modèle actif d'apparence avec un modèle dont les paramètres globaux de mouvement de la tête obtenus à partir d'un modèle cylindrique sont utilisés comme les repères des paramètres du modèle actif d'apparence pour un ajustement ou une réinitialisation. Dans ce travail, les auteurs ont montré l'efficacité des méthodes de reconnaissance de comportement en se basant sur des caractéristiques spatio-temporelles. Ils ont montré comment l'utilisation de prototypes cuboïdes a donné naissance à un descripteur de comportement efficace et robuste. L'algorithme a été testé dans un certains domaines et a donné des bons résultats par rapport à des algorithmes bien établis

Xiao, Baker, Matthews, & Kanade (2004) et Ramnath et al. (2008) ont proposé un algorithme de suivi du visage en temps réel en combinant des modèles AAM 2D + 3D. Zhou, Liang, Sun, & Wang (2010) ont introduit des contraintes d'appariement temporelles pour renforcer la cohérence inter-trame dans le montage de modèle actif d'apparence.

2.3.1.5 Suivie de visage en utilisant les « range data »

Il existe certaines conditions environnementales qui altèrent de manière significative les caractéristiques des systèmes de suivi du visage, par exemple les ombres sur les visages et les mauvaises conditions d'éclairage. D'autre part, l'utilisation de « range data » a permis d'améliorer le suivi du visage mobile.

Yang, Shechtman, Wang, Bourdev, & Metaxas (2012) ont développé un système de suivi de visage 3D où ils ont utilisé des caméras synchronisées, de sorte qu'il était possible de capturer des images du même scénario de différents points de vue. Ces images ont été appariées à un gabarit en utilisant la régularisation de forme et l'erreur de profondeur.

Au cours des dernières années, il y a eu un intérêt considérable pour la vision par ordinateur. Les chercheurs Fanelli, Yao, Noel, Gall, & Van Gool (2010) ont développé un algorithme basé sur « Random forest » pour estimer les orientations de tête à partir des « range data ». Parmi les autres travaux nous citons celui de Zhou et al. (2010), ils ont développé une solution à maximum de vraisemblance pour suivre les formes de visage à partir des données de profondeur d'entrée bruyantes. Il y a aussi Weise, Bouaziz, Li, & Pauly (2011) ont développé un système en temps réel pour reconstruire des formes de tête 3D à partir des « range data » et les a utilisées pour générer des animations de visage.

Baltrusaitis et al. (1985) ont étendu l'approche du modèle local contraint pour le suivi des points de caractéristiques faciales.

2.3.2 Reconstruction complète du corps et estimation de la pose

Metaxas & Zhang (2013) affirment qu'en plus de la modélisation et l'analyse du visage, les mouvements et les gestes de tout le corps sont également des facteurs importants pour l'étude de la communication non verbale. Plusieurs travaux de recherche ont été réalisés pour étudier la communication non verbale, telle que la reconnaissance des langues des signes qui doivent combiner les marqueurs non manuels (par exemple, l'expression du visage) avec les mouvements du corps et les gestes pour améliorer la précision de la reconnaissance.

Une série de méthodes ont été développée pour reconstituer le geste corporel 3D à partir des séquences vidéo monoculaires. La plateforme pour la récupération de la pose en 3D à partir

de la séquence monoculaire se base sur des méthodes partielles qui utilisent l'alignement des images 2D. Il existe une vaste documentation sur les modèles partiels pour la détection et la localisation des parties du corps humain en images 2D. La plupart de ces approches se concentrent soit sur l'amélioration de l'extraction de caractéristiques pour augmenter le degré de confiance dans la détection du corps humain, soit sur l'apprentissage des classifieurs efficaces pour modéliser des configurations spatiales plausibles de pièces en 2 D. Parmi ces travaux, nous citons celui de Yang, Wang, Shechtman, Bourdev, & Metaxas (2011) qui ont amélioré la plateforme de la structure picturale en modélisant les relations contextuelles de co-occurrence entre les différentes parties de configuration. L'approche de Poselet (Bourdev & Malik, 2009) utilise l'ensemble de données 3D de pose humaine afin d'améliorer les détecteurs de différentes parties du corps et utilise la transformation de Hough dans la configuration de pose 2 D. Alors que toutes les autres approches sont basées sur l'estimation de la pose en 2D qui est très difficile à contraindre en utilisant des critères anthropométriques standards.

2.3.3 Suivi des mains

Les gestes des mains sont considérés comme un langage principal pour les personnes sourdes pour communiquer avec leurs entourages. Ils impliquent des formes, des orientations et des mouvements des mains. Ces gestes peuvent être classés en deux catégories, à savoir les gestes statiques et les gestes dynamiques. Les gestes statiques sont les gestes des mains et les poses qui sont représentés sous la forme d'images tandis que les gestes dynamiques sont définis comme l'ensemble des mouvements des mains représentés par une séquence d'images ou une vidéo (Hernandez-Rebollar, Kyriakopoulos, & Lindeman, 2004). Dans la suite, on va mettre l'accent sur la reconnaissance du geste dynamique dont les entrées sont prises à partir d'une caméra web ou d'une autre caméra.

Selon Jaliya, Thakore, & Kawdiya (2016), les étapes principales de tout système de reconnaissance de gestes sont : l'acquisition des gestes, l'analyse de fonctionnalités et enfin la classification.

Plusieurs chercheurs tels que Jaliya et al. (2016) qui s'intéressent à la reconnaissance des gestes des mains ont présenté une méthode basée sur les filtres Gabor et Machine à vecteurs de support connue sous le nom SVM. Les filtres Gabor sont utilisés pour extraire les caractéristiques gestuelles. La méthode d'analyse des composants principaux (PCA) est ensuite utilisée pour réduire la dimensionnalité de l'espace de fonctionnalité. Avec les fonctionnalités réduites de Gabor, SVM est formé et exploité pour effectuer les tâches de reconnaissance des gestes de la main.

Un algorithme de reconnaissance des gestes des mains se compose de trois étapes : Le prétraitement, l'extraction de caractéristiques et la classification. L'étape de prétraitement consiste à suivre trois sous-étapes : la segmentation qui permet de séparer la région de la main de son arrière-plan à l'aide d'un histogramme basé sur un algorithme de seuil transformé en silhouette binaire, la rotation qui utilise des capteurs pour mesurer les angles des articulations et le filtrage qui élimine efficacement le bruit de fond et le bruit d'objet de l'image binaire par une technique de filtrage morphologique (Ong & Ranganath, 2005).

Hernandez-Rebollar et al. (2004) ont proposé un prototype de système capable de reconnaître les langues des signes gestuels en temps réel. Ils ont utilisé l'espace de couleur HSV (Teinte, Saturation, Luminosité) combiné avec la détection de la peau pour supprimer l'arrière-plan complexe et créer des images segmentées. Ensuite, une détection de contour est appliquée pour localiser et sauvegarder la zone de la main. En outre, ils utilisent l'algorithme SURF pour détecter et extraire des caractéristiques clés et reconnaître les gestes des mains (les alphabets) en les comparants avec la dataset des images. Le système proposé est capable de reconnaître le signe du geste manuel et de le traduire vers les alphabets, avec un taux de reconnaissance de 63 % en moyenne.

Jaliya et al. (2016) ont présenté les différents algorithmes pour l'application de la reconnaissance des gestes. Parmi ces algorithmes nous citons :

2.3.3.1 L'algorithme de comptage de pixels

C'est un algorithme basique et simple pour la reconnaissance des gestes des mains. L'image d'entrée, obtenue à partir d'un appareil photo, est enregistrée en format RGB puis convertie en image binaire. Une fois l'image binaire obtenue, la région de la peau sera représentée par les pixels blancs et l'arrière-plan est représenté par des pixels noirs. Dans cette approche, l'algorithme compte le nombre de pixels blancs et vérifie les surfaces dans lesquelles les pixels sont noirs. En trouvant le nombre de pixels blancs, on peut identifier les doigts relevés. Comme la sortie dépend uniquement du nombre de pixels blancs, cet algorithme est invariant à la rotation de la main (Qidwai & Chen, 2009).

2.3.3.2 Détection des cercles

L'inconvénient de la méthode Comptage de pixels peut être éliminé par l'algorithme Détection des Cercles. Dans cette approche, les doigts doivent être marqués par des cercles noirs avant de les placer devant la caméra. Les images RGB sont converties en image binaire. Dans l'image convertie, la région de la peau est représentée avec des pixels blancs tandis que le fond et les cercles noirs sont représentés par les pixels noirs. À partir de cette image, le nombre d'objets circulaires est détecté.

2.3.3.3 Opérations morphologiques

Dans cet algorithme, l'image RGB est convertie en image binaire. Cette dernière est filtrée à l'aide d'un filtre médian de dimension 3. L'image binaire obtenue est érodée en utilisant un élément structurel approprié de dimension 3 à 5. Cette image est encore dilatée en utilisant un autre élément structurel de dimension 10 à 12. Le prétraitement de l'image binaire est fait pour améliorer la région de la peau (Elamaran, Narasimhan, Vijayabhaskar, & Shiva, 2013). L'image obtenue après le prétraitement est soumise à une opération d'échec ou de réussite (hit/miss). La transformation hit/miss est une opération morphologique qui reconnaît le motif donné dans une image binaire (Goth, 2011). Le motif est spécifié par deux éléments

structurants ES1 et ES2. Cette transformation reconnaît le motif du premier élément structurel en supprimant le motif du second élément structurel.

2.3.3.4 Méthodes de numérisation

C'est un algorithme robuste comparé aux algorithmes cités auparavant. Dans cette méthode, l'image RGB est convertie en image binaire. Les régions de la peau sont représentées par des pixels blancs et l'arrière-plan est représenté par des pixels noirs. Afin d'améliorer la région de la peau, l'image binaire est prétraitée par un filtre médian de dimension appropriée. L'image est divisée en deux moitiés. La moitié inférieure contient le pouce et la moitié supérieure contient les doigts. À cette étape, trois à cinq scans horizontaux linéaires sont effectués dans les deux moitiés. Afin de rendre cet algorithme à rotation invariante, la numérisation verticale se fait en dehors du balayage horizontal.

2.3.4 Reconnaissance d'activités humaines

La reconnaissance de l'activité humaine est l'un des domaines de recherche de l'apprentissage automatique. Plusieurs approches ont été développées dans ce domaine qui permet de mesurer et/ou reconnaître les activités humaines. La plupart de ces approches sont basées sur l'utilisation de capteurs. En effet, les activités peuvent être reconnues à l'aide de capteurs environnementaux tels que le Système de positionnement global, ou bien à l'aide des capteurs portables qui sont placés à différents endroits du corps. Les capteurs en construction sont des capteurs de téléphones intelligents. On peut également utiliser une combinaison de ces capteurs pour reconnaître des activités avec des résultats plus précis.
Selon Stojanova, Kocaleva, Luledjieva, & Koceski (2019), le processus de reconnaissance de l'activité humaine comprend quatre étapes principales : l'acquisition de données, le prétraitement, l'extraction de caractéristiques et la classification. Les données sont acquises à l'aide de capteurs et se dirigent vers le prétraitement ; des données prétraitées sont en outre transmises pour un processus de classification qui montre la précision de la reconnaissance.

Les travaux de recherche qui ont été réalisés dans le domaine de la reconnaissance de l'activité humaine s'intéressent surtout aux activités de la vie quotidienne (Stojanova et al., 2019). Ces travaux de recherche divisent les activités de la vie quotidienne en deux classes : des activités de bas niveau ou simples et des activités de haut niveau ou complexes :

- activités simples : consistent en une seule action répétée. Les activités simples sont également appelées activités de bas niveau ou logique. Parmi ces activités nous citons : marcher, courir, escaliers, travaux ménagers, escaliers, faire du jogging, s'asseoir, se tenir debout, etc;

- activités complexes : sont la compilation d'une série d'actions multiples ou une combinaison de différentes activités simples. Elles sont également appelées activités physiques de haut niveau. comme la cuisine, le nettoyage, l'arrosage des plantes, l'exercice physique, etc.

La plupart des travaux relatifs à la reconnaissance des activités humaines sont basés sur l'utilisation de différents types de capteurs de mouvement. L'accéléromètre est le capteur dominant dans la majorité des études. Dans la plupart des cas, il n'est utilisé avec d'autres capteurs, dans certains cas, il est combiné avec d'autres capteurs, tels que le gyroscope, et magnétomètre et le capteur des téléphones intelligents. Nous pouvons même utiliser un capteur accéléromètre en combinaison avec des capteurs environnementaux et des capteurs portables. La majorité des recherches antérieures sur la reconnaissance d'activité utilisent des capteurs portables et des systèmes embarqués (de la Concepción, Morillo, García, & González-Abril, 2017) (J. Wang, Chen, Hao, Peng, & Hu, 2019) ou des capteurs externes ou environnementaux tels que des capteurs attachés à un objet.

En plus des recherches qui s'intéressent à l'analyse des mouvements des individus, de nombreux chercheurs ont également étudié les activités de groupe utilisant des méthodes d'analyse de mouvement et / ou d'apprentissage automatique (Poppe, 2010) (Souvenir & Parrigan, 2009). Les activités de groupe de modélisation jouent un rôle important dans les systèmes de vidéosurveillance et de caméras intelligentes. En effet, il existe de nombreuses applications prometteuses comme la reconnaissance automatique et la classification des vidéos qui permettent d'analyser les vidéos d'une façon efficace. Citant l'exemple de la

recherche des tacles dans les matchs de football, les poignées de mains dans des séquences vidéo ou des mouvements de danse typique dans des vidéos de musique. Ces applications sont également importantes pour la surveillance automatique, telle que la surveillance des centres commerciaux. Ces applications permettent aussi de soutenir le vieillissement dans les endroits pour les personnes âgées dans les maisons intelligentes. Les applications d'interaction comme les interactions homme-ordinateur bénéficient également des progrès de la reconnaissance automatique de l'action humaine. Diverses activités ont été étudiées tels que la détection d'accès aux zones restreints (Bovik, 2010), comptage de voitures, la détection des personnes transportant des bagages (Kim, Choi, Yi, Choi, & Kong, 2010), les objets abandonnés (Smith, Quelhas, & Gatica-Perez, 2006), la détection d'activité de groupe, la modélisation de réseaux sociaux (Metaxas & Zhang, 2013), la surveillance des véhicules (Yang et al., 2012) et l'analyse des scènes (Swears & Hoogs, 2009).

Ces dernières années, plusieurs algorithmes ont été proposés pour améliorer la performance de l'analyse action/activité. Beaucoup d'entre eux ont mis l'accent sur la recherche d'une meilleure représentation des images et des caractéristiques extraites d'une séquence d'images. Ces algorithmes sont généralisés sur de petites variations dans l'apparence de la personne, l'arrière-plan, le point de vue et les types d'action.

Depuis plus de trois décennies, la reconnaissance des gestes de la main fait l'objet de nombreuses recherches. Un nombre important des recherches qui portent sur la reconnaissance des gestes des mains est basé sur les travaux de McNeill (1985) et de Kita, Van Gijn, & Van der Hulst (1997). Nous présentons dans cette section une brève revue des travaux publiés antérieurement liés à la reconnaissance des gestes des mains.

Plusieurs études ont mis en évidence la relation entre le vieillissement et les gestes. Werner et al. (2020) ont construit un système pour améliorer l'interaction gestuelle entre un robot d'assistance au bain et des personnes âgées à travers l'entraînement des utilisateurs sur les commandes gestuelles. Une autre étude importante réalisée par Liang, Kapetanios, Woll, & Angelopoulou (2019) pour suivre la trajectoire des mouvements des mains pour améliorer le dépistage de la démence chez les personnes sourdes vieillissantes à travers la langue des signes britannique en utilisant une vidéo 2D. Parmi les autres études qui concernent l'analyse

gestuelle, nous trouvons celle de (Leite, Duarte, Neves, De Oliveira, & Giraldi, 2017). Le système proposé tire parti de la capacité de détection infrarouge de la Kinect, qui est capable d'identifier le squelette humain. Ce système utilisait des articulations squelettiques pour réaliser une analyse gestuelle dans le but d'aider les personnes âgées à accomplir leurs tâches.

Actuellement, de nombreuses recherches et applications ont mis l'accent sur l'analyse des gestes. Ces applications ont été mises en œuvre en tenant compte de différents domaines, notamment ceux qui concernent l'interaction entre l'homme et la machine et l'interaction entre deux individus ou plus. Par conséquent, il y a eu un développement de nombreuses méthodes utilisées dans la reconnaissance des mouvements et la segmentation des gestes ; ces applications ont été mises en œuvre en utilisant différentes stratégies pour résoudre ces tâches. Parvathy & Subramaniam (2019) ont implémenté un système de reconnaissance des gestes des mains basé sur une combinaison d'un algorithme appelé Speed UpRobust Feature et d'un 2D-DiscreteWaveletsTransform. Ferstl a proposé une approche de phasage gestuel en explorant l'utilisation d'un paradigme antagoniste génératif (Ferstl, Neff, & McDonnell, 2020). Cardenas & Chavez (2020) ont proposé une approche pour la reconnaissance des gestes des mains en intégrant des caractéristiques spatio-temporelles extraites de données multimodales détectées par le capteur Kinect. Des expérimentations approfondies ont montré l'efficacité et la faisabilité de cette approche. Ameur, Khalifa, & Bouhlel (2020) ont suggéré une nouvelle méthode d'indexation chronologique des modèles pour coder les ordres temporels des gestes de la main acquises par le capteur LMC. Xing & Li (2019) ont utilisé un regroupement de scènes non supervisées et une classification audio supervisée pour extraire les caractéristiques audio/visuelles de niveau intermédiaire des vidéos de sport. La nouvelle méthode proposait un algorithme de regroupement de scènes efficace, capable de déterminer le point d'arrêt de l'algorithme sans connaissance préalable. L'annotation de vidéos est également l'une des tâches les plus connues de l'analyse gestuelle à travers l'analyse de la langue des signes. Cette tâche ne peut être effectuée que par une segmentation de phase gestuelle à l'aide d'outils manuels ou numériques citant l'exemple des travaux de Maricchiolo, Gnisci, & Bonaiuto (2012) et Madeo, Peres, & de Moraes Lima (2016).

L’apprentissage profond tel que les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents a été utilisé dans de nombreuses études grâce à son efficacité dans les tâches de reconnaissance et de segmentation des gestes des mains. Truong, Doan, Tran, Vu, & Le (2019) ont utilisé les réseaux de neurones convolutifs 3D pour la reconnaissance des gestes des mains. L’approche proposée est basée sur un CNN 3D existant utilisé pour la reconnaissance des actions humaines en appliquant la technique d’apprentissage par transfert ; ils ont obtenu un résultat très compétitif et la méthode permet de reconnaître les gestes avec précision. Furnari, Battiato, & Farinella (2018) ont suggéré une nouvelle approche pour la segmentation des vidéos égocentriques en fonction des emplacements visités par l’utilisateur. La méthode proposée fournit une sortie personnalisée et permet aux utilisateurs de spécifier les emplacements qu’ils doivent prendre en considération. D’autres études ont utilisé une combinaison entre CNN et LSTM (Tsironi, Barros, & Wermter, 2016) pour la tâche de reconnaissance dynamique gestuelle. Le modèle proposé est capable d’apprendre avec succès des gestes variant en durée et en complexité et s’avère être une base importante pour un développement ultérieur. L’approche proposée a montré des performances compétitives dans la reconnaissance des gestes.

2.4 Annotation des gestes

L’annotation des gestes est un exercice complexe. En effet, elle peut être vue, dans certain cas comme un exercice de segmentation et dans d’autres cas comme un exercice de classification et des fois comme un exercice d’étiquetage. Nous présentons dans cette section les travaux de recherche qui portent sur l’annotation.

Navarretta (2018) a étudié les types sémiotiques de gestes des mains dans des conversations naturelles ainsi que leur classification automatique. La classification automatique du type sémiotique peut contribuer à l’interprétation des gestes dans la communication homme-homme et dans les systèmes interactifs multimodaux avancés. Dans cette étude, Navaretta a annoté les gestes des mains produits par Barack Obama lors de deux discours issus des dîners annuels des correspondants de la Maison Blanche et puis il les a analysées. Dans ce travail, l’auteur a implémenté des algorithmes d'apprentissage automatique pour classer les gestes

des mains selon quatre types sémiotiques. Le score F obtenu par l'algorithme le plus performant sur la classification de quatre types sémiotiques est de 0,59. La classification distingue quatre grandes catégories de gestes Indexical Non déictique, Indexical déictique, iconiques et symboliques.

Schreer, Masneri, Lausberg, & Skomroch (2014) ont présenté un outil d'analyse et d'annotation vidéo automatique. Le système décrit permet aux chercheurs de gagner du temps en détectant automatiquement les parties du corps et en reconnaissant les mouvements des mains. L'outil peut être utilisé dans différents domaines de recherches (telles que l'analyse du comportement gestuel, l'analyse du langage des signes, l'analyse des interactions, y compris l'interaction du thérapeute avec le patient). Le résultat de l'analyse consiste en une série d'annotations représentant les mouvements des mains dans le temps et la relation spatiale entre les mains et le corps.

Alexanderson, House, & Beskow (2016) ont développé une méthode pour segmenter et étiqueter automatiquement les unités gestuelles à partir d'un flux de données de capture de mouvement 3D. Le flux gestuel est modélisé avec un modèle de Markov caché hiérarchique à 2 niveaux (HHMM) où les sous-états correspondent à des phases gestuelles. Le modèle est formé sur la base d'étiquettes d'unités gestuelles complètes et de manipulateurs auto-adaptatifs. Le modèle est testé et validé sur deux ensembles différents. Un avantage important de la méthode proposée est la possibilité de regrouper les phases gestuelles possibles de manière non supervisée.

Ferstl et al. (2020) ont exploré l'utilisation d'un paradigme d'entraînement génératif pour mapper la parole au mouvement gestuel 3D. Ils ont défini le problème de génération de gestes comme une série de sous-problèmes, y compris une dynamique de geste plausible, des configurations d'articulations réalistes et un mouvement diversifié et fluide. Chaque sous-problème est surveillé par des adversaires séparés. Pour le problème de l'application d'une dynamique gestuelle réaliste dans la sortie, ils ont entrainé trois classifieurs avec différents niveaux de détail pour détecter automatiquement les phases gestuelles. Plus de 3,8 heures de

données gestuelles ont été annotées et ont été évalués à la main à cet effet, y compris des échantillons d'un deuxième locuteur pour comparer et valider les résultats.

Di Mitri, Schneider, Klemke, Specht, & Drachsler (2019) ont introduit un outil d'inspection visuelle (VIT), qui soutient les chercheurs dans l'annotation des données multimodales et qui peut être utilisé pour des fins d'apprentissage. Alors que la plupart des solutions d'analyse d'apprentissage multimodal existantes sont conçues pour des tâches d'apprentissage et des capteurs spécifiques, l'outil traite l'annotation des données pour différents types de tâches d'apprentissage qui peuvent être capturées avec un ensemble personnalisable de capteurs de manière flexible. L'outil d'inspection visuelle soutient les chercheurs dans la triangulation de données multimodales avec des enregistrements vidéo. Il peut être utilisé pour segmenter les données multimodales en intervalles de temps et aussi pour télécharger des jeux de données annotés et l'utiliser pour l'analyse des données multimodales. Le VIT est un composant crucial qui manquait jusqu'à présent dans les outils disponibles pour la recherche dans l'analyse des données multimodales. En comblant cette lacune, les auteurs ont également identifié un flux de travail intégré qui caractérise la recherche actuelle sur l'analyse des données multimodales.

Dans l'étude de Madeo et al. (2016), le problème de la segmentation des phases de gestes est modélisé comme un problème de classification, puis une machine à vecteurs de support est employée pour concevoir un modèle capable d'apprendre les schémas de gestes inhérents à chaque phase. Cette étude présente deux points forts. La première consiste à aborder les limites de l'approche de segmentation à travers l'étude de ses performances dans différents scénarios qui représentent la complexité de l'analyse des modèles de comportement humain. Dans cette étude, les auteurs ont évoqué la possibilité d'étudier la façon dont les modèles de classification sont influencés par le comportement humain. Le deuxième point fort se réfère à la conduite d'une analyse en considérant le point de vue de spécialistes concernés par la segmentation des phases gestuelles dans le domaine de la linguistique et de la psycholinguistique.

Dans cet article I. Wang et al. (2018) ont présenté EASEL, un outil d'annotation vidéo qui utilise la segmentation et la reconnaissance des gestes pour étiqueter automatiquement les gestes dans des enregistrements Kinect. Les auteurs ont montré la façon avec laquelle l'annotation assistée peut améliorer les performances temporelles de l'annotation gestuelle. EASEL rationalise le processus d'annotation en introduisant l'annotation assistée, la segmentation et la reconnaissance automatiques des gestes pour annoter automatiquement les gestes. Pour évaluer l'efficacité de l'annotation assistée, les auteurs ont mené une étude utilisateur auprès de 24 participants et ont constaté que l'annotation assistée réduisait le temps nécessaire pour annoter les vidéos sans différence de précision par rapport à l'annotation manuelle. Les résultats confirment non seulement les avantages de l'annotation assistée pour l'annotation gestuelle dans EASEL, mais ouvrent également de nouvelles possibilités pour adapter la fonctionnalité à d'autres domaines de recherche.

Spiro, Taylor, Williams, & Bregler (2010) ont décrit une méthode d'utilisation du travail participatif pour suivre le mouvement et annoter les gestes d'humains dans la vidéo. La plateforme de déploiement choisie, Amazon Mechanical Turk, divise le travail en EIH (Exercice d'Intelligence Humain (Human Intelligence Tasks)). Vu la densité informationnelle de la vidéo, la tâche est plus complexe qu'un EIH traditionnel qui implique le traitement d'un bloc de texte ou d'une seule image. Le système décrit est une preuve de concept pour l'optimisation des annotations vidéo. Cette étude pilote montre l'effet multiplicateur obtenu en combinant l'intelligence humaine et l'intelligence artificielle dans un même pipeline. Ceci est réalisé avec une combinaison d'EIH utilisant une nouvelle interface utilisateur, combinée à des techniques automatiques telles que le suivi de modèle et le regroupement de propagation d'affinité. Les auteurs ont montré dans une étude de cas comment annoter une base de données vidéo de discours politiques avec des positions 2D et des configurations de pose des mains en 3D. Ces données sont ensuite utilisées pour certaines tâches analytiques préliminaires.

Dans cet article De Beugher, Brône, & Goedemé (2018), les auteurs ont présenté une approche semi-automatique pour l'annotation des gestes manuels. L'objectif de la méthode

proposée est de minimiser la charge de travail liée à l'annotation manuelle des gestes. La première étape consiste à extraire l'extraction des parties du corps pertinentes, y compris le visage, le torse et les mains dans chaque image à partir d'un enregistrement. Ceci est réalisé en utilisant un algorithme semi-automatique de détection des mains proposé par la même équipe de recherche, dans lequel une intervention manuelle est nécessaire au cas où la confiance d'une détection de la main est trop faible. Plusieurs expériences de validation ont révélé que la méthode proposée est capable de détecter les mains de manière très précise. L'analyse gestuelle commence par définir la position de repos de chaque main pendant tout un enregistrement. Une fois cet emplacement connu, on calcule pour chaque cadre la distance entre chaque main et sa position de repos respective. Sur la base de ce déplacement, on sera capable de segmenter un enregistrement en segments gestuels et non gestuels. De plus, les auteurs analysent automatiquement l'usage de l'espace gestuel selon la classification proposée par McNeill (1992). Une dernière analyse est faite sur la direction des mouvements des mains. Une comparaison a été faite entre l’analyse automatique des gestes proposée dans cette étude et les annotations des corpus existants, montrant que l’approche est plus précise.

Il y a eu récemment beaucoup d’études relatives à l'utilisation des annotations 2D (dans le domaine de réalité augmentée, citant l’exemple de Fakourfar, Ta, Tang, Bateman, & Tang (2016), où il y a une communication à distance entre deux utilisateurs : un local et l’autre distant. Un défi majeur avec l'utilisation des annotations 2D est l'ambiguïté de les afficher dans la réalité augmentée 3D. Nuernberger et al. (2017) se sont concentrés sur les annotations de mouvements de cercles et de flèches et ont développé des méthodes d'ancrage spécifiques pour interpréter et rendre chacune en réalité augmentée 3D. Chang et al. (2017) suivaient l’approche de Nuernberger en se concentrant sur les annotations de mouvements de cercle et de flèche; cependant, ils utilisaient des dessins gestuels à main levée en 3D en entrée, alors que l’approche de Nuernberger concerne principalement les dessins 2D en entrée. En outre, Ils examinaient plus en détail les effets de l'embellissement de l'annotation gestuelle résultante.

Bien que beaucoup de travaux ont été faits dans l'utilisation des gestes des mains 3D pour une variété de tâches en réalité virtuelle et la réalité augmentée tels que les travaux de LaViola Jr, Kruijff, McMahan, Bowman, & Poupyrev (2017) et Oda & Feiner (2012). Chang, Nuernberger, Luan, & Höllerer (2017) se sont concentrés spécifiquement sur ceux liés à la création de contenu en réalité augmentée. La plupart des travaux antérieurs se sont concentrés sur les gestes à main levée pour modéliser directement la géométrie 3D pour la visualisation (Fakourfar et al., 2016), et récemment, plusieurs applications ont adopté cette approche (par exemple, Microsoft HoloLens Skype). En réalité augmentée, le désir d'annoter des objets physiques est un cas d'utilisation important qui ne se produit pas immédiatement dans la réalité virtuelle. Plus récemment, Miksik et al. (2015) ont introduit le pinceau sémantique pour l'étiquetage des scènes du monde réel via la diffusion de rayons et la segmentation sémantique. La méthode de dessin de surface de Chang et al. (2017) peut être considérée comme une technique de pointage ou de projection de rayons basée sur l'image, similaire à la technique de lancer de rayons utilisée pour la méthode Head Crusher de Pierce (Alexanderson et al., 2016; Pierce et al., 1997)

2.5 Conclusion

La revue de littérature est une phase primordiale et importante dans n’importe quel type de recherche. Elle permet de faire une étude comparative des travaux de recherche antérieurs et de dégager les limites de ces travaux. À partir de ces limites, nous définissons notre problématique de recherche et nous nous fixons les objectifs à atteindre. Dans ce chapitre, nous avons mis en lumière la relation entre le vieillissement et la communication non verbale. Nous avons présenté, aussi, les approches qui s’intéressent à la détection ou au suivi des mouvements tels que le suivi du visage, la reconnaissance de l’expression faciale, la reconnaissance d’activité, le suivi des mains et la segmentation du corps humain. Dans le chapitre suivant, nous allons présenter la méthodologie générale pour atteindre les objectifs fixés au chapitre précédent.

CHAPITRE 3

MÉTHODOLOGIE GÉNÉRALE

3.1 Introduction

Ce chapitre est consacré à la présentation de la méthodologie que nous avons adoptée dans cette recherche. En effet, nous avons utilisé des architectures similaires dans les grandes lignes pour annoter les mouvements de la tête ainsi que les mouvements des mains. Nous allons présenter dans ce chapitre les parties communes dans le chapitre dédié à l'annotation automatique de la tête et celui de l'annotation automatique des mains pour éviter la redondance. Nous amorçons ce chapitre par le processus automatisé pour annoter les gestes de la tête et des mains. La deuxième partie est dédiée au données que nous allons utiliser pour cette recherche. Nous concluons ce chapitre par les mesures que nous avons utilisées pour évaluer les approches proposées.

3.2 L'annotation des gestes : Le processus automatisé

Le but de cette recherche est de présenter des approches pour automatiser l'annotation des mouvements de la tête ainsi que les mouvements des mains. L'objectif d'annotation des gestes a été atteint en modélisant le problème comme étant un problème de classification qui a été résolu en utilisant des architectures basées sur une utilisation combinée des réseaux de neurones convolutifs (CNN) et des réseaux de neurones récurrents (RNN). Dans les deux approches proposées, on distingue deux phases :

- phase de détection de l'objet en question (tête ou main);
- phase de prédiction de la classe.

Dans la phase de détection de l'objet en question, nous avons utilisé deux techniques différentes qui sont le Multi-task (MTCNN) pour la détection de la tête et une combinaison entre le modèle Single Shot Multibox (SSD) et le classifieur MobileNet pour la détection des mains. Dans la phase de prédiction de la classe de mouvement, nous avons eu recours au

réseau récurrent à mémoire court et long terme connu en anglais sous le nom de Long short-term memory (LSTM)

3.3 Corpus

Il est à noter que Tous les participants ont signé un formulaire d'information et de consentement autorisant l'utilisation de leur image dans des publications scientifiques.

Pour ce travail, nous allons utiliser le corpus CorpAGEst (Bolly & Boutet, 2018), il contient des conversations naturelles entre des adultes et des sujets âgés (75 ans et plus). Il s'agit d'un corpus longitudinal contenant des annotations audio et vidéo et des transcriptions synchronisées. Le but de ce corpus est d'étudier les marqueurs verbaux et gestuels et de construire un profil pragmatique des personnes âgées, en suivant leurs marqueurs verbaux et gestuels pragmatiques dans des situations réelles. Le corpus est constitué de sous-corpus transversaux et longitudinaux :

- le corpus transversal comprend 18 conversations spontanées en français avec neuf participants (8 femmes et 1 homme), dont l'âge moyen était de 85 ans. Chaque participant a tenu deux conversations. Le corpus sert à explorer les marqueurs non verbaux de posture et leur combinaison lors d'interactions verbales puisque ce sont des indicateurs représentatifs du comportement, de l'attitude et de l'état émotionnel des locuteurs;
- le corpus longitudinal a été créé dans le but de découvrir si une stratégie compensatoire peut être observée dans l'utilisation des signes pragmatiques verbaux et non verbaux chez les personnes âgées au cours du temps ; cette partie du corpus contient des conversations entre des personnes dont la langue maternelle est le français. Les conversations ont été réalisées plusieurs fois durant un an et demi. Elles concernent deux thèmes : le premier concerne des scènes en relation avec le passé et le deuxième est relié au présent afin de garantir la compatibilité des résultats entre les données transversales et les données longitudinales.

Les données étaient déséquilibrées (peu d'échantillons pour les classes de préparation et de rétraction) donc cela affectait gravement les performances du modèle, nous avons donc dupliqué les frames qui représentent ces deux classes avec des rotations (90°, 180°) et des retournements (horizontalement et verticalement).

3.4 Organisation des données

Nous avons divisé au hasard notre ensemble de données en 3 groupes :

- l'ensemble d'apprentissage est utilisé pour créer des modèles prédictifs, il représente 70 % de l'ensemble des données;
- l'ensemble de validation est utilisé pour évaluer les performances du modèle créé lors de la phase d'apprentissage. Il représente 15 % de l'ensemble des données;
- l'ensemble de tests, ou données invisibles, est utilisé pour évaluer les performances futures probables d'un modèle. Il représente 15 % de l'ensemble des données.

3.4.1 Ensemble d'apprentissage

Cet ensemble est utilisé pour ajuster le modèle. C'est l'ensemble qu'on l'utilise pour entraîner le modèle. Ce dernier voit et apprend de ces données.

3.4.2 Ensemble de validation

L'ensemble de validation est utilisé pour évaluer un modèle donné. Nous utilisons ces données pour affiner les hyper-paramètres du modèle. Nous utilisons les résultats de l'ensemble de validation pour mettre à jour les hyper-paramètres de niveau supérieur. Ainsi, l'ensemble de validation affecte le modèle indirectement. L'ensemble de validation est également appelé ensemble de développement. Cela a du sens puisque cet ensemble de données nous aide à nous améliorer pendant la phase de « développement » du modèle.

3.4.3 Ensemble de test

L'ensemble de test est utilisé une fois que le modèle est complètement entraîné (en utilisant l'ensemble d'apprentissage et l'ensemble de validation). Cet ensemble utilisé pour fournir une évaluation non biaisée d'un ajustement final du modèle sur l'ensemble de données d'entraînement.

3.5 Métriques d'évaluation des modèles proposées

Des métriques d'évaluation des modèles sont nécessaires pour quantifier les performances du modèle. Le choix des métriques d'évaluation dépend de la nature de la tâche (telle que la classification, la régression, le classement, le regroupement, la modélisation de sujets..). Certaines mesures, telles que le rappel de précision, sont utiles pour plusieurs tâches. Pour évaluer notre modèle, nous avons recours à une vérité terrain étiquetée manuellement par des experts. Puis, nous avons utilisé les mesures suivantes:

3.5.1 Précision de la classification

La précision est une mesure d'évaluation courante pour les problèmes de classification. C'est le nombre de prédictions correctes faites sous forme de rapport de toutes les prédictions faites. En utilisant la validation croisée, nous serons capables de tester notre modèle dans la phase « d'apprentissage » pour vérifier et éviter le sur-apprentissage et pour avoir une idée de la manière dont notre modèle d'apprentissage automatique se généralisera aux données indépendantes (ensemble de données de test).

3.5.2 Matrice de confusion

La matrice de confusion fournit une ventilation plus détaillée des classifications correctes et incorrectes pour chaque classe. La brève explication de la façon d'interpréter une matrice de confusion est la suivante : Les éléments diagonaux représentent le nombre de sujets (l'élément à prédire) pour lesquels l'étiquette prédite est égale à l'étiquette vraie ou la vérité

terrain, alors que tout ce qui n'est pas en diagonale a été mal étiqueté par le classifieur. Par conséquent, plus les valeurs diagonales de la matrice de confusion sont élevées, plus le modèle est précis et les prédictions sont correctes.

3.5.3 Score F

Le score F (également appelé la mesure F) est une mesure de la précision d'un test qui prend en compte à la fois la précision et le rappel du test pour calculer le score. La précision est le nombre de résultats positifs corrects divisé par le total des observations positives prévues. Le rappel, d'autre part, est le nombre de résultats positifs corrects divisé par le nombre de tous les échantillons pertinents (total des positifs réels).

3.5.4 Validation croisée

La validation croisée est une méthode statistique utilisée pour évaluer la performance des modèles d'apprentissage automatique. Elle est souvent utilisée dans l'apprentissage automatique appliqué pour comparer et sélectionner un modèle pour un problème de modélisation prédictive donné, car elle est facile à mettre en œuvre et aboutit à des estimations de compétences généralement moins biaisées que les autres méthodes.

3.5.5 Précision globale et taux d'erreurs de classification globale

Nous distinguons deux types de précision : la précision locale et la précision globale. La première concerne chaque classe et la deuxième concerne toutes les classes, il peut être considéré comme la précision moyenne de toutes les classes. Nous allons utiliser encore un autre indicateur très important pour mesurer la performance des approches proposées qui est le taux d'erreurs de classification globale qui permet de calculer le taux d'erreur moyen.

3.6 Conclusion

Dans ce chapitre, nous avons présenté la méthodologie générale qui nous a permis d'atteindre les objectifs souhaités à savoir l'annotation automatique des mouvements de la tête et les mouvements des mains. Nous avons présenté aussi la vérité terrain avec laquelle nous avons entraîné nos modèles et nous l'avons utilisé aussi pour évaluer les approches proposées. Finalement nous avons présenté les mesures que l'on va utiliser pour évaluer les résultats.

CHAPITRE 4

APPROCHE AUTOMATIQUE POUR ANNOTER LES MOUVEMENTS DE LA TÊTE CHEZ LES PERSONNES ÂGÉES

4.1 Introduction

De nos jours, l'étude de la communication non verbale est l'un des plus importants sujets qui a évolué rapidement au cours des dix dernières années. En effet, des milliers d'études empiriques ont porté sur le rôle de la communication non verbale dans la vie humaine. L'être humain a besoin de s'exprimer et de partager ses émotions au moyen de la communication verbale et non verbale. Selon des études (Birdwhistell, 1977) qui ont été faites dans le domaine de communication, les chercheurs ont constaté que les paroles participent avec, seulement, 5 % dans la compréhension du message ; le ton, l'inflexion et autres éléments vocaux participent avec 45 %. L'apport le plus important dans nos communications est celle du langage corporel, le mouvement et le contact visuel, ils représentent 50 %. Ainsi, la communication non verbale est parfois plus utile que la communication verbale pour comprendre l'intention exacte du communicateur. Certains chercheurs tels que Ruesch & Kees (1969) attribuent à la communication non verbale une contribution majeure de 65 % par rapport à la transmission d'informations par voie verbale qui est égale à 35 %. D'autres chercheurs tels que Rujoiu (2009) et d'autres affirment que dans les relations interpersonnelles, l'information est transmise en proportion de 82 % par le corps et la voix, et seulement 18 % par le langage verbal. Par conséquent, l'analyse de ses comportements peut contribuer à une meilleure compréhension de ses besoins. Le sujet de la communication non verbale devient de plus en plus important si on est dans le contexte d'une population vieillissante, étant donné que cette population a tendance à diminuer les conversations. Ceci est dû à deux facteurs, le premier est « naturel » là où on parle du vieillissement normal et l'autre est due à des maladies cognitives telles que les démences légères et sévères. Toutes ces raisons mettent l'accent sur l'importance de la communication non verbale dans la vie

des personnes âgées. Vu cette importance, plusieurs recherches ont été faites pour étudier les mouvements de la tête chez les personnes âgées. Ces études ont été faites en particulier par des linguistes. Ils ont annoté les mouvements de la tête afin de faciliter l'étude des comportements de la population vieillissante et ainsi dégager les différents paramètres et mesures qui aident les médecins et les différentes parties prenantes à suivre leurs états. Ce travail est fait par des experts-linguistes manuellement, et comme tout travail manuel demande beaucoup de temps, de même, il demande un grand budget (frais d'annotation des experts), sans oublier que les conclusions tirées ne sont pas robustes vu qu'elles sont interprétées à partir de petits échantillons. D'où l'automatisation de ce processus peut être considérée comme une contribution très importante qui peut aider les différentes parties prenantes à atteindre leurs finalités. Dans la suite de ce chapitre, nous exposons la problématique ainsi que l'objectif principal, puis nous mettons l'accent sur l'approche proposée, enparticulier le corpus utilisé dans ce travail de recherche, nous parlons aussi sur la méthodologie à suivre pour atteindre notre objectif. Nous présentons les résultats obtenus par les différentes techniques utilisées. Nous concluons le chapitre par une discussion et une conclusion.

4.2 Problématique et objectif

Plusieurs chercheurs tels que T. Rousseau (2018) et Albers et al. (2015) ont noté que le vieillissement de la population induit des changements sur tous les systèmes du corps humain. Ces changements touchent la communication, la cognition ainsi que la motricité. Ce travail de recherche vise à fournir un support scientifique aux spécialistes de la santé et aux linguistes, qui contribuent à la compréhension de la communication non verbale chez les personnes âgées. Notre contribution consiste à proposer une approche automatique qui permet d'annoter automatiquement les mouvements des mains et de la tête en s'appuyant sur des enregistrements vidéo. Présentement, l'analyse des vidéos se fait par des experts dans le domaine de la santé d'une façon manuelle. Ce type d'analyse présente plusieurs limites. Il est fastidieux et imprécis vu qu'il est soumis à la variabilité des annotateurs. De même ce type d'analyse est coûteux puisqu'il doit être réalisé uniquement par des experts. De plus, il

nécessite beaucoup de temps d'analyse. Sans oublier que les conclusions tirées ne sont pas robustes étant donné que l'analyse manuelle est souvent appliquée à des petits échantillons.
L'objectif principal de ce chapitre consiste à proposer une approche qui permet d'annoter automatiquement les mouvements de la tête dans des conversations naturelles des personnes âgées.

4.3 Approche proposée

4.3.1 Méthodologie

4.3.1.1 Stratégie d'annotation dans l'approche proposée

Comme nous l'avons mentionné dans la section 3.1, le processus d'annotation dans Corpagest présente plusieurs limites, le but de ce travail de recherche est de proposer une alternative qui permet d'automatiser le processus manuel. Pour atteindre cet objectif, nous modélisons le problème d'annotation comme un problème de classification, nous proposons une architecture basée sur une utilisation combinée de CNN et LSTM. La section suivante est consacrée à décrire la méthodologie de l'approche proposée.

L'approche proposée consiste à considérer le problème d'annotation comme étant un problème de classification. En effet, l'approche est fondée sur deux idées principales. La première vise à simplifier le modèle en créant une classe qui englobe toutes les classes composées de deux mouvements ou plus à la fois et la seconde idée, celle qui nous permettra de simuler les annotations de l'expert, est l'utilisation des techniques d'apprentissage profond. Nous avons relevé trois étapes dans l'application du processus d'apprentissage machine.

- détection de visage en utilisant MTCNN;
- construction de l'ensemble de données;
- entraînement du modèle;
- évaluation du modèle.

4.3.1.2 Algorithmes d'alignement de visage

La dernière décennie a été témoin d'une révolution dans les algorithmes de détection des objets en général et de détection de visage (la tête) en particulier. Ceci peut être expliqué par l'importance de la tête dans la communication non verbale. Ainsi, de nombreuses recherches dans le domaine de la vision par ordinateur ont mis l'accent sur le sujet de suivi du visage (face tracking). Les algorithmes les plus utilisés sont ceux de Viola-Jones (Viola & Jones, 2001). Ces algorithmes sont limités par rapport aux méthodes d'apprentissage profond telles que Tasks Constrained Deep Convolutional Network (TCDCN) (Z. Zhang, Luo, Loy, & Tang, 2014) et Multi-task CNN (MTCNN) (K. Zhang, Zhang, Li, & Qiao, 2016). En effet, ces méthodes ont montré leurs efficacités dans la détection et le suivi de visage même dans des conditions difficiles telles que l'absence de luminosité et d'éclairage suffisantes et dans les occlusions. Dans ce travail de recherche, nous utilisons la méthode de réseaux de neurones convolutifs multitâches connue sous le nom de MTCNN, cette méthode a prouvé une efficacité et une performance très élevées dans la détection des objets.

Les réseaux de neurones convolutifs multitâches sont destinés principalement à la détection et l'alignement des visages. Ils englobent trois réseaux à convolutions qui sont P-Net, R-Net et O-Net, connectés en cascade afin de sélectionner séquentiellement les meilleures fenêtres candidates à la présence de visages dans l'image à travers la reconnaissance des points de repère tels que les yeux, le nez et la bouche. On distingue trois étapes principales :

- **Étape 1 : le réseau proposé**

C'est un réseau entièrement convolutif (fully convolutional networks FCN). La différence entre un CNN et un FCN est qu'un réseau entièrement convolutif n'utilise pas de couche dense dans le cadre de l'architecture. Ce réseau est utilisé pour obtenir des fenêtres candidates et leurs vecteurs de régression de boîtes englobantes. La figure 4.1 montre l'architecture du P-Net.

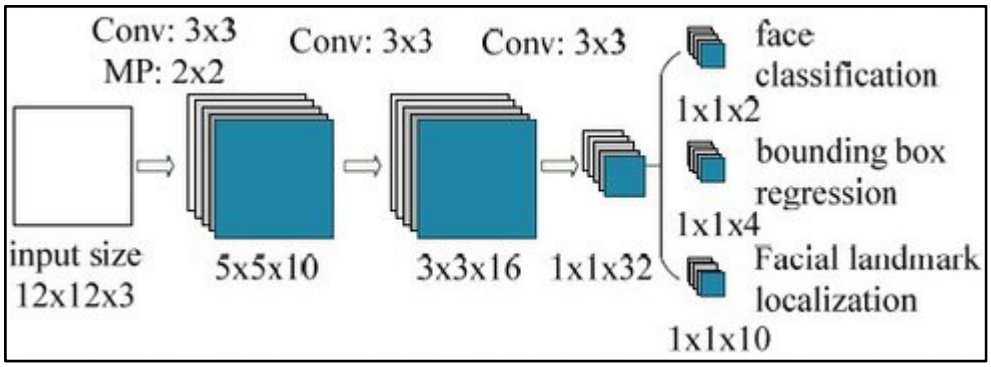


Figure 4.1 Architecture P-Net tirée de K. Zhang et al. (2016)

- **Étape 2 : le réseau Affiner**

Tous les candidats du P-Net sont intégrés dans le réseau Affiner. L'architecture R-Net présente une couche dense, c'est pour cela qu'il est considéré comme un CNN, pas un FCN. Le R-Net réduit encore le nombre de candidats en effectuant un étalonnage avec une régression par boîte englobante et utilise la suppression non maximale pour fusionner les candidats qui se chevauchent. La figure ci-dessous montre l'architecture du R-Net.

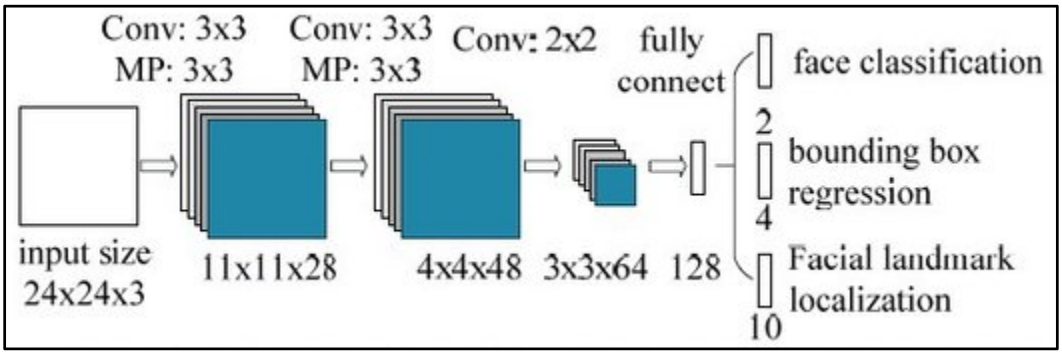


Figure 4.2 Architecture R-Net tirée de K. Zhang et al. (2016)

- **Étape 3 : le réseau de sortie**

Cette étape est similaire au R-Net, mais ce réseau de sortie vise à décrire le visage plus en détail et à afficher les positions des repères faciaux pour les yeux, le nez et la bouche.

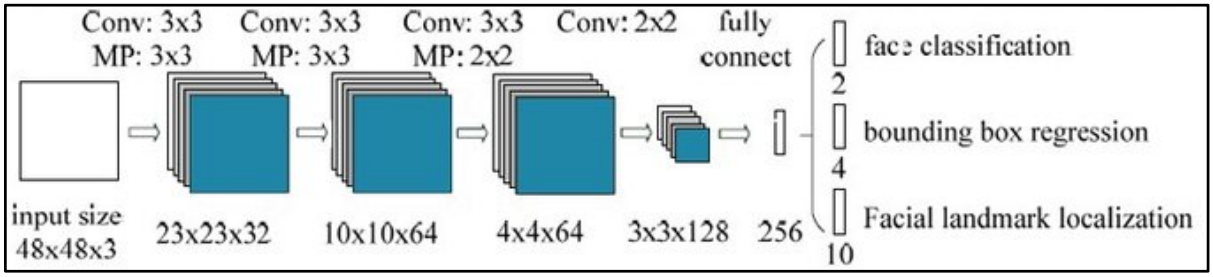


Figure 4.3 Architecture O-Net tirée de K. Zhang et al. (2016)

MTCNN est entraîné en exploitant trois tâches : la classification des visages/non-visages, la régression de la boîte englobante et la localisation des points de repère faciaux. Ci-dessous le détail des trois fonctions

- **Classification des visages :** l'objectif d'apprentissage est formulé comme étant une classification binaire basée sur la présence ou l'absence du visage. Cette classification est basée sur la perte d'entropie croisée :

$$L_i^{det} = -(y_i^{det}\ log\ log\ (p_i)\ + \left(1 - y_i^{det}\right)(1 - log\ log\ (p_i)\ )) \quad (4.1)$$

Où :

- L_i^{det} : La fonction de coût à minimiser,
- y_i^{det} : Le label de la vérité terrain (0 ou 1)
- p_i : La probabilité produite par le réseau.

- **Régression de la boîte englobante :** L'objectif est de minimiser le décalage entre les coordonnées des fenêtres candidates et les coordonnées produites par la vérité terrain.

$$L_i^{box} = \|\ddot{y}_i^{box} - y_i^{box}\|_2^2 \quad (4.2)$$

Où :

- L_i^{box} : La fonction de coût à minimiser,
- $\ddot{y}_i^{box}$: Les coordonnées prédites par le réseau.
- y_i^{box} : Les coordonnées de la vérité terrain

- **Identification des points de repère de visage à travers une régression :** Ceci est similaire à l'objectif précédent, la détection des repères faciaux est formulée comme un problème de régression et nous minimisons la distance euclidienne.

$$L_i^{Landmark} = \|\ddot{y}_i^{Landmark} - y_i^{Landmark}\|_2^2 \quad (4.3)$$

Où :

- $L_i^{Landmark}$: La fonction de coût à minimiser,
- $\ddot{y}_i^{Landmark}$: Les coordonnées de points de repère de visage prédites par le réseau.
- y_i^{box} : Les coordonnées de points de repère de visage de la vérité terrain

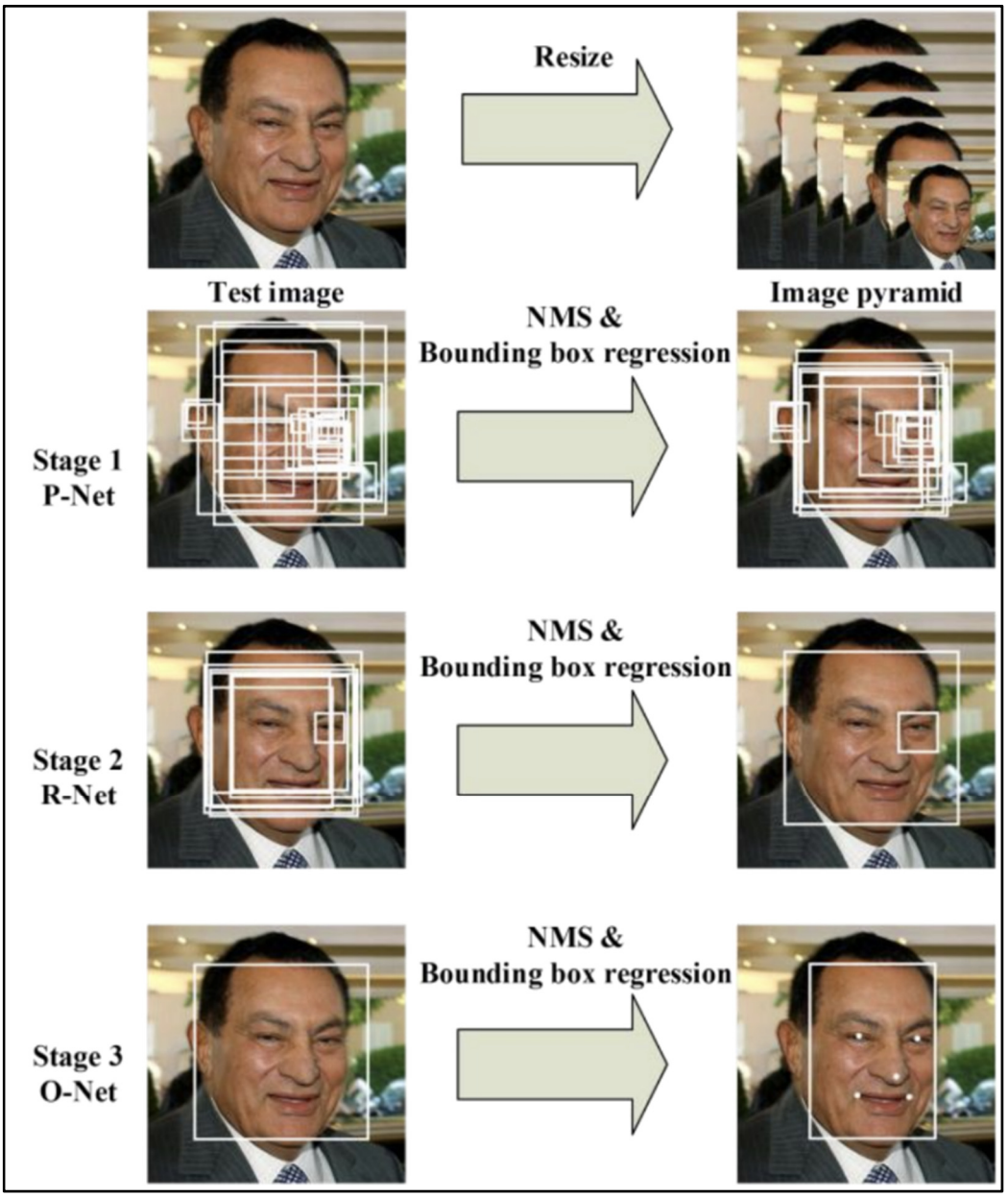


Figure 4.4 Représentation du pipeline de MTCNN
tirée de K. Zhang et al. (2016)

Le MTCNN est caractérisé par :

- **entraînements multisources :** Étant donné que le CNN est utilisé dans plusieurs tâches, il existe différents types d'entraînements utilisés dans le processus d'apprentissage tel que la classification binaire (visage/non-visage) et le visage partiellement aligné;

- **extraction d'échantillons en ligne :** Cette opération est différente de la conduite d'échantillons traditionnels. En effet, on effectue l'extraction en ligne au moment de la classification des visages pour être adaptatifs au processus d'apprentissage.

4.3.1.3 Prétraitement des données

L'une des étapes les plus importantes pour obtenir des résultats satisfaisants est le prétraitement des données. En effet, l'éclairage ou le contraste influence grandement la précision du résultat. Dans cette approche, différentes méthodes de prétraitement ont été utilisées. Nous avons procédé au nettoyage des données pour supprimer le bruit et corriger les incohérences dans les données. Nous avons recours aussi au dimensionnement des caractéristiques. En effet, nous avons dû appliquer cette technique étant donné la différence dans les tailles d'images de notre base de données, pour éviter que ces variations de dimensions entraînent de faibles performances. Les techniques de normalisation appliquées sont celles de Shin, Kim, & Kwon (2016), commençons par la normalisation basée sur la diffusion anisotrope, appelée lissage isotrope (IS), est une technique qui vise à réduire le bruit de l'image sans supprimer des parties importantes du contenu de l'image tel que les bords ou les lignes. Puis nous avons utilisé la technique de normalisation basée sur la transformée cosinus discrète (DCT) qui a été proposée par Maheshkar, Kamble, Agarwal, & Srivastava (2012) et a été couramment appliquée aux images faciales en raison de sa puissante capacité de transformation dans le traitement d'image pour éliminer les zones qui présentent une illumination forte. Enfin, la différence de Gauss (DoG) (S. Wang et al., 2012) est la technique la plus élémentaire d'amélioration des caractéristiques, qui implique la soustraction de deux images floues différemment. Habituellement, DoG est très puissant pour augmenter la visibilité des bords et pour représenter les détails de texture. La phase de prétraitement des données est clôturée par ces deux techniques :

- ***augmentation de données*** : pour avoir une bonne capacité de généralisation des modèles et éviter le sur-apprentissage dans l'apprentissage profond, un grand nombre de données est nécessaire. La récolte des données pour entraîner des modèles d'apprentissage automatique est une opération délicate. Un moyen simple pour

pallier ce problème est d'opérer l'augmentation de données, citant l'exemple des modifications sur l'image selon un ensemble de paramètres (tels que les rotations, les translations, l'augmentation du contraste, l'effet miroir, etc.);

- ***réduire le nombre de classes de mouvements de tête :*** nous avons relevé 38 classes servant à annoter les mouvements de la tête dans le corpus CorpAGEst, proposées par les experts comme indiqué au tableau 1. Parmi ces classes, 8 sont des mouvements de tête « simples » ayant une direction unique (tourner à droite, tourner à gauche, pencher à droite, pencher à gauche, incliner vers le bas, incliner vers le haut, arrière, avant). Les autres classes représentent des mouvements de tête « complexes », soit qui ont plus d'une direction (p. ex. En bas + Tourner à droite, Pencher à gauche + Tourner à droite). L'approche proposée consiste à conserver les mêmes annotations simples et à ajouter une nouvelle classe complexe qui englobe toutes les classes composées. Les annotations des experts sont listées dans le tableau suivant.

Tableau 4.1 Annotations des experts

EnAttente
PencherGaucheDroite (Waggle)
PencherGauche
PencherDroite
TournerGauche
TournerDroite
En Haut
EnBas
DeBasEnHaut
EnBas+Tourner-D
PencherGauche+TournerGauche
EnBas+TournerGauche
EnHaut+Tourner
Bas+TournerGauche
EnHaut+Tourner-D
EnArrière+Tourner
PencherDroite+EnHaut
EnBas+PencherDroite
DeBasEnHaut+TournerGauche
EnHaut+TournerGauche
TournerGauche+Tourner
TournerGauche+DeHautEnBas
TournerDroite+EnHaut
EnBas+Tourner
AvantArrière
TournerGauche+EnHaut

Tableau 4.2 Classes d'annotation proposées

EnAttente
Complexe
PencherGaucheDroite (Waggle)
PencherGauche
PencherDroite
TournerGauche
TournerDroite
EnHaut
EnBas
DeBasEnHaut

4.3.2 Le pipeline du système d'annotation

L'approche proposée est divisée en deux parties :

4.3.2.1 Phase de détection de la tête

La détection de la tête se fait en utilisant la méthode proposée par K. Zhang et al. (2016) qui consiste en un MTCNN.

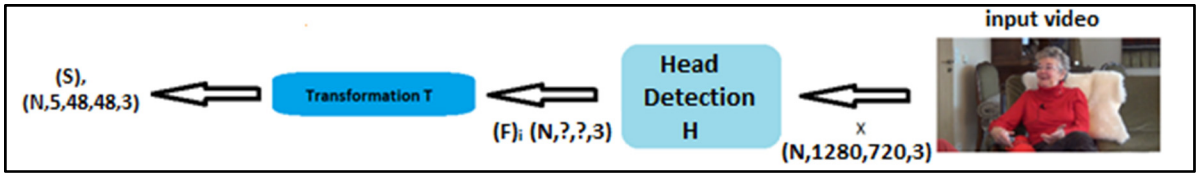


Figure 4.5 Phase de détection

La figure 4.5 illustre la phase de détection de la tête. Comme nous l'avons mentionné dans les sections précédentes, le MTCCN est utilisé pour la détection de l'objet en question.

Les points d'interrogation dépendent de la position de la tête. En effet, ils sont déterminés de telle façon que la tête soit centrée dans le cadre. Dans le cadre ci-dessous, nous expliquons les paramètres ainsi que les fonctions utilisés (pris de mon article soumis)

Tableau 4.3 Fonctions utilisées dans la phase de détection

- N : Nombre total d'images (frames)
- X : Le vidéo d'entrée
- H(X) : Fonction de détection de la tête.
- $(F)_i$: Image rognée I où i dans l'intervalle[0,N]
- T $((F)_i)$: Fonction de transformation pour redimensionner et sélectionner les 5 dernières images chaque 0,5 seconde
- $(S)_i$: Un ensemble de séquences pour la tête où S[i] est une séquence de 5 images chaque 0.5 seconde
- $(F)_{i,} = H(X)$
- $(S)_i = T((F)_I)$

4.3.2.2 Phase de prédiction

La figure 4.6 illustre l'architecture de l'approche proposée pour classifier les mouvements de la tête. Nous détaillons les différentes parties dans les sections suivantes.

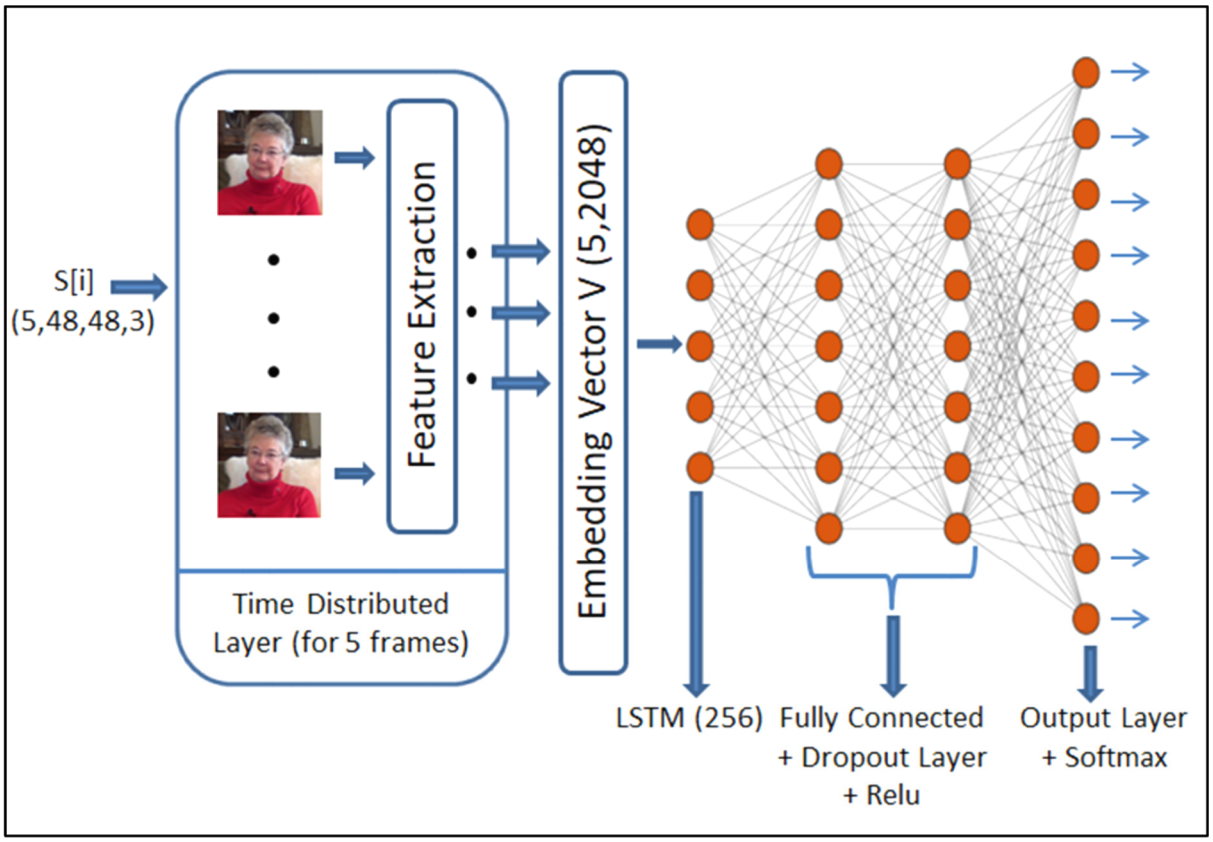


Figure 4.6 Architecture proposée

Le vecteur résultat $Y \in R^4$ peut être obtenu à partir de la séquence d'entrée $Seq \in R^{5xZxZxC}$ (Où Z et C désignent respectivement la taille des frames et le nombre des canaux) comme suit :

$V = FE\ (Seq) \in R^{5x2048}$

$T = TF\ (V) \quad \in R^{256}$

$Fc = FCL(T) \in R^4$

$Y = Softmax\ (Fc)$

Où V est le vecteur d'incorporation des caractéristiques spatiales extraites des images par la fonction FE ; T est un vecteur résultant de l'exploitation de l'évolution temporelle des caractéristiques spatiales à l'aide de la TF, une fonction basée sur le LSTM ; FCL est un représentant des 2 blocs de : couches entièrement connectées + fonction d'activation Relu + Dropout. FE s'appuie sur un bloc ResNet50 (Rezende, Ruppert, Carvalho, Ramos, & De Geus, 2017) qui permet de retracer un grand nombre d'images ayant différentes tailles de champs récepteurs, favorisant ainsi la fusion d'un grand nombre d'images avec différents champs récepteurs et améliorant la capacité de représentation multi-échelles. La figure 4.7 illustre l'architecture ResNetX utilisée

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 4.7 Architecture ResNetX

La figure 4.7 décrit l'architecture du modèle qu'on a utilisé pour atteindre notre objectif. En effet, nous présentons dans cette figure, le type du couche, le format des sorties (output Shape) ainsi que le nombre des paramètres.

Layer (type)	Output Shape	Param #
input_6 (InputLayer)	[(None, 5, 48, 48, 3)]	0
time_distributed_1 (TimeDist	(None, 5, 2048)	23587712
lstm_1 (LSTM)	(None, 256)	2360320
dense_10 (Dense)	(None, 256)	65792
dropout_2 (Dropout)	(None, 256)	0
dense_11 (Dense)	(None, 128)	32896
dropout_3 (Dropout)	(None, 128)	0
dense_12 (Dense)	(None, 10)	1290

Total params: 26,048,010
Trainable params: 25,994,890
Non-trainable params: 53,120

Figure 4.8 Résumé du modèle

4.3.2.3 Entrainement des données

L'entraînement des données a été effectué à l'aide d'un optimiseur Adam (lr = 0,000 1) et d'une taille de lot de 64 pour 14 époques.

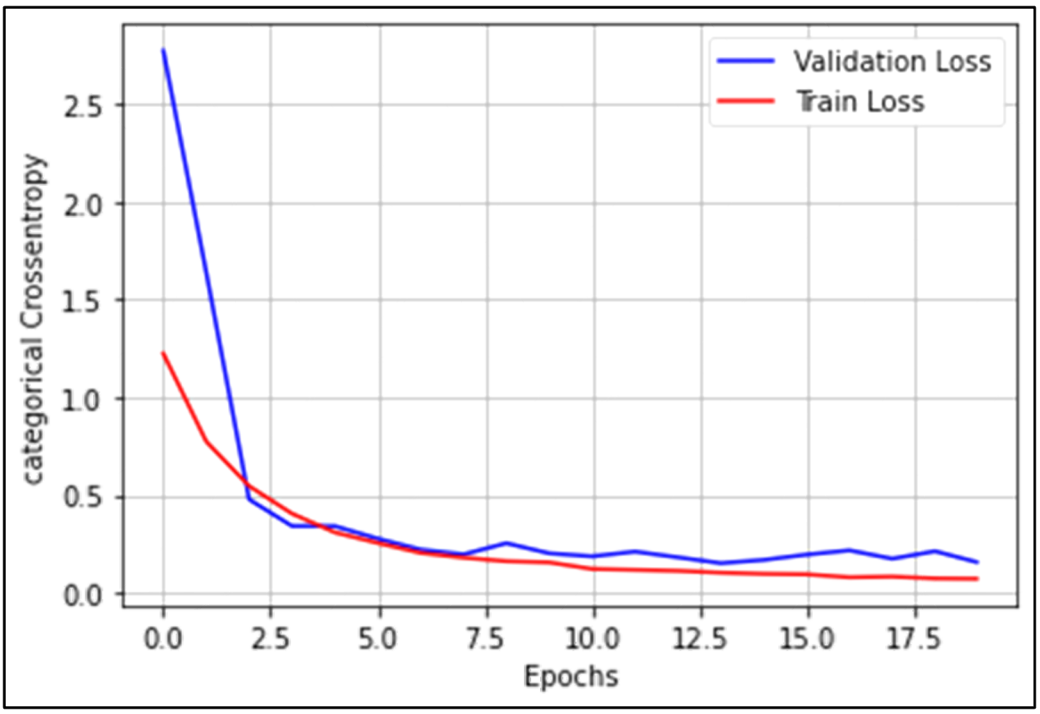


Figure 4.9 Entraînement des données

4.3.3 Résultat

Dans cette section, nous réalisons des expériences pour évaluer l'efficacité de l'approche proposée. Nous validons cette approche en utilisant une vérité terrain, annotée par des experts (linguistes). Nous avons utilisé 20 048 séquences d'entrée extraites d'une conversation spontanée.

4.3.3.1 Évaluation de l'ensemble d'apprentissage

Tableau 4.4 Performance des classes dans la phase d'apprentissage.

	Prédictions									
Classes réelles	***Stand by***	***Complex***	***Waggle***	***TiltLeft***	***TiltRight***	***TurnLeft***	***TurnRight***	***Up***	***Down***	***UpDown***
Stand by	***0.9953***	*0.0007*	*0.0000*	*0.0014*	*0.0007*	*0.0007*	*0.0014*	*0.0000*	*0.0000*	*0.0000*
Complex	*0.0019*	***0.9884***	*0.0000*	*0.0000*	*0.0045*	*0.0004*	*0.0048*	*0.0000*	*0.0000*	*0.0000*
Waggle	*0.0000*	*0.0032*	***0.9714***	*0.0000*	*0.0000*	*0.0000*	*0.0222*	*0.0000*	*0.0000*	*0.0032*
TiltLeft	*0.0008*	*0.0000*	*0.0000*	***0.9984***	*0.0008*	*0.0004*	*0.0004*	*0.0000*	*0.0000*	*0.0000*
TiltRight	*0.0000*	*0.0000*	*0.0000*	*0.0037*	***0.9963***	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*
TurnLeft	*0.0000*	*0.0000*	*0.0000*	*0.0007*	*0.0000*	***0.9989***	*0.0004*	*0.0000*	*0.0000*	*0.0000*
TurnRight	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	***1.0000***	*0.0000*	*0.0000*	*0.0000*
Up	*0.0000*	*0.0000*	*0.0000*	*0.0192*	*0.0000*	*0.0000*	*0.0000*	***0.9808***	*0.0000*	*0.0000*
Down	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	***1.0000***	*0.0000*
UpDown	*0.0000*	*0.0000*	*0.0000*	*0.0269*	*0.0049*	*0.0000*	*0.0049*	*0.0000*	*0.0000*	***0.9634***

Le tableau 4.4 est une matrice de confusion qui montre la performance de chaque classe dans la phase d'apprentissage. En effet l'exactitude dépasse 96% dans toutes les classes. Elle atteint 100 % dans certaines classes. Les résultats obtenus dans la phase d'apprentissage sont promotteurs. Ceci est un indice que les résultats dans la phase finale seront encourageants.

Tableau 4.5 Performances globales de la phase d'apprentissage

Exactitude globale	***Mauvaise classification***
0.9932	*0.0068*

Le tableau 4.5 montre les performances globales de phase d'apprentissage. En effet l'exactitude globale des classes dépasse 99% alors que le pourcentage des classes qui ne sont pas bonnes est inférieur à 1%.

Tableau 4.6 Rapport de classification de la phase d'apprentissage

Classes	***Précision***	***Rappel***	F_1 ***score***
Stand by	*0.9966*	*0.9953*	*0.9959*
Complex	*0.9992*	*0.9884*	*0.9938*
Waggle	*1.0000*	*0.9714*	*0.9855*
TiltLeft	*0.9851*	*0.9984*	*0.9917*
TiltRight	*0.9828*	*0.9963*	*0.9895*
TurnLeft	*0.9989*	*0.9989*	*0.9903*
TurnRight	*0.9868*	*1.0000*	*0.9934*
Up	*1.0000*	*0.9808*	*0.9903*
Down	*1.0000*	*1.0000*	*1.0000*
UpDown	*0.9987*	*0.9634*	*0.9807*
Moyenne macro	***0.9968***	***0.9953***	***0.9961***
Moyenne pondérée	***0.9962***	***0.9962***	***0.9962***

Le tableau 4.6 montre le rapport de classification de la phase d'apprentissage. En effet, le tableau montre les valeurs d'exactitude, de précision et mesure-F pour chaque classe. Les valeurs de chacun de ces trois mesures dépassent 96% dans chaque classe ce qui explique des moyennes macros et des moyennes pondérée pour les trois mesures qui sont proches de 100%.

4.3.3.2 Évaluation de l'ensemble de validation

Tableau 4.7 Performances des classes dans la phase de validation.

	Prédictions									
Classes réelles	***Stand by***	***Complex***	***Waggle***	***TiltLeft***	***TiltRight***	***TurnLeft***	***TurnRight***	***Up***	***Down***	***UpDown***
Stand by	***0.9022***	*0.0347*	*0.0126*	*0.0158*	*0.0032*	*0.0063*	*0.0158*	*0.0032*	*0.0000*	*0.0063*
Complex	*0.0035*	***0.9739***	*0.0017*	*0.0000*	*0.0052*	*0.0035*	*0.0122*	*0.0000*	*0.0000*	*0.0000*
Waggle	*0.0147*	*0.0147*	***0.8971***	*0.0000*	*0.0000*	*0.0147*	*0.0588*	*0.0000*	*0.0000*	*0.0000*
TiltLeft	*0.0000*	*0.0000*	*0.0000*	***0.9829***	*0.0038*	*0.0019*	*0.0019*	*0.0000*	*0.0000*	*0.0095*
TiltRight	*0.0000*	*0.0000*	*0.0000*	*0.0043*	***0.9786***	*0.0085*	*0.0000*	*0.0085*	*0.0000*	*0.0000*
TurnLeft	*0.0018*	*0.0000*	*0.0000*	*0.0159*	*0.0035*	***0.9646***	*0.0142*	*0.0000*	*0.0000*	*0.0000*
TurnRight	*0.0067*	*0.0000*	*0.0000*	*0.0067*	*0.0000*	*0.0201*	***0.9665***	*0.0000*	*0.0000*	*0.0000*
Up	*0.0000*	*0.0256*	*0.0000*	*0.0000*	*0.0256*	*0.0000*	*0.0128*	***0.9359***	*0.0000*	*0.0000*
Down	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	***1.0000***	*0.0000*
UpDown	*0.0000*	*0.0000*	*0.0000*	*0.0568*	*0.0000*	*0.0000*	*0.0170*	*0.0000*	*0.0000*	***0.9261***

Le tableau 4.7 est une matrice de confusion qui montre la performance de chaque classe dans la phase de validation. En effet l'exactitude dépasse 90% dans presque toutes les classes. Elle atteint 100 % dans certaines classes. Les résultats obtenus dans la phase de validation sont bons. Ceci est un indice que les résultats dans la phase finale seront encourageants.

Tableau 4.8 Performances globales de la phase de validation

Exactitude globale	***Mauvaise classification***
0.9601	*0.0399*

Le tableau 4.8 montre les performances globales de la phase de validation. En effet l'exactitude globale des classes dépasse 96% alors que le pourcentage des mouvements qui sont mal classées est inférieur à 4%

Tableau 4.9 Rapport de classification de la phase de validation

Classes	***Précision***	***Rappel***	$F_1\ Score$
Stand by	*0.9761*	*0.9022*	*0.9377*
Complex	*0.9756*	*0.9739*	*0.9748*
Waggle	*0.9242*	*0.8971*	*0.9104*
TiltLeft	*0.9486*	*0.9829*	*0.9655*
TiltRight	*0.9582*	*0.9786*	*0.9683*
TurnLeft	*0.9698*	*0.9646*	*0.9672*
TurnRight	*0.9372*	*0.9665*	*0.9516*
Up	*0.9605*	*0.9359*	*0.9481*
Down	*1.0000*	*1.0000*	*1.0000*
UpDown	*0.9588*	*0.9634*	*0.9422*
Moyenne macro	***0.9609***	***0.9953***	***0.9566***
Moyenne pondérée	***0.9604***	***0.9601***	***0.9600***

Le tableau 4.9 montre le rapport de classification de la phase de validation. En effet, le tableau montre les valeurs d'exactitude, de précision et mesure-F pour chaque classe. Nous remarquons que les moyennes macros sont différentes des moyennes pondérées dans le rappel et le score F_1, Ceci reviens au déséquilibre entre les frames de chaque classes dans la base de données.

4.3.3.3 Évaluation de l'ensemble de test

Tableau 4.10 Performances des classes dans la phase de test

	Prédictions									
Classes réelles	***Stand by***	***Complex***	***Waggle***	***TiltLeft***	***TiltRight***	***TurnLeft***	***TurnRight***	***Up***	***Down***	***UpDown***
Stand by	***0.9025***	*0.0346*	*0.0031*	*0.0283*	*0.0031*	*0.0063*	*0.0157*	*0.0000*	*0.0031*	*0.0031*
Complex	*0.0000*	***0.9809***	*0.0000*	*0.0000*	*0.0087*	*0.0017*	*0.0070*	*0.0000*	*0.0017*	*0.0000*
Waggle	*0.0448*	*0.0000*	***0.9552***	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*
TiltLeft	*0.0000*	*0.0000*	*0.0000*	***0.9886***	*0.0057*	*0.0019*	*0.0038*	*0.0000*	*0.0000*	*0.0000*
TiltRight	*0.0085*	*0.0000*	*0.0000*	*0.0085*	***0.9573***	*0.0043*	*0.0000*	*0.0171*	*0.0000*	*0.0043*
TurnLeft	*0.0088*	*0.0000*	*0.0000*	*0.0141*	*0.0018*	***0.9647***	*0.0106*	*0.0000*	*0.0000*	*0.0000*
TurnRight	*0.0067*	*0.0000*	*0.0000*	*0.0067*	*0.0000*	*0.0267*	***0.9599***	*0.0000*	*0.0000*	*0.0000*
Up	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0128*	*0.0000*	*0.256*	***0.9487***	*0.0000*	*0.0000*
Down	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0000*	***1.0000***	*0.0000*
UpDown	*0.0000*	*0.0000*	*0.0000*	*0.0000*	*0.0057*	*0.0000*	*0.0114*	*0.0000*	*0.0000*	***0.9086***

Le tableau 4.10 est une matrice de confusion qui montre la performance de chaque classe dans la phase de test. En effet l'exactitude dépasse 90% dans toutes les classes. Elle atteint 100 % dans certaines classes. Les résultats obtenus dans la phase de test sont bons ce qui implique l'efficacité de l'approche proposée.

Tableau 4.11 Performances globales de la phase de test

Exactitude globale	***Mauvaise classification***
0.9604	*0.0396*

Le tableau 4.11 montre les performances globales de la phase de test. En effet l'exactitude globale des classes dépasse 96% alors que le pourcentage des mouvements qui sont mal classées est inférieur à 4%

Tableau 4.12 Rapport de classification de la phase de test

Classes	***Précision***	***Rappel***	$F_1\ Score$
Stand by	*0.9567*	*0.9025*	*0.9288*
Complex	*0.9809*	*0.9809*	*0.9809*
Waggle	*0.9846*	*0.9552*	*0.9697*
TiltLeft	*0.9354*	*0.9886*	*0.9613*
TiltRight	*0.9492*	*0.9573*	*0.9532*
TurnLeft	*0.9698*	*0.9647*	*0.9672*
TurnRight	*0.9535*	*0.9599*	*0.9567*
Up	*0.9487*	*0.9487*	*0.9487*
Down	*0.9048*	*1.0000*	*0.9500*
UpDown	*0.9876*	*0.9086*	*0.9464*
Moyenne macro	*0.9571*	*0.9566*	*0.9563*
Moyenne pondérée	*0.9609*	*0.9604*	*0.9603*

Le tableau 4.12 montre le rapport de classification de la phase de test. En effet, le tableau montre les valeurs d'exactitude, de précision et mesure-F pour chaque classe. Nous remarquons que les moyennes macros ainsi que les moyennes pondérées sont bonnes, ce qui explique encore une fois l'efficacité de l'approche proposée.

4.4 Discussion et conclusion

L'approche proposée nous a permis de développer une méthode permettant d'automatiser les mouvements de la tête réduisant, ainsi, le coût des annotations manuelles et minimisant le temps de travail. Les résultats proposent deux axes de recherche à explorer. Premièrement, le processus d'annotation automatique des comportements non verbaux dans les vidéos est tout à fait faisable ; deuxièmement, les algorithmes d'apprentissage profond peuvent identifier des caractéristiques qui ne correspondent pas totalement aux observations des experts. Cette particularité établit clairement un nouveau dialogue entre les chercheurs en intelligence artificielle, les chercheurs en linguistique et les chercheurs dans les domaines liés à la communication et au vieillissement, quant à l'interprétation des résultats et des caractéristiques à partir de vidéos. La principale contribution de ce travail est le développement des techniques de reconnaissance gestuelle plus performantes adaptées au vieillissement de la population, donnant l'opportunité aux chercheurs des outils permettant d'explorer les spécificités de façon automatisée.

CHAPITRE 5

APPROCHE AUTOMATIQUE POUR ANNOTER LES MOUVEMENTS DES MAINS CHEZ LES PERSONNES ÂGÉES

5.1 Introduction

Ces dernières années, le monde a vécu des changements sociétaux et démographiques. Ces changements ont touché les pays industrialisés en particulier, ce qui explique l'augmentation du nombre de personnes âgées dans la population. Ainsi plusieurs recherches ont été faites sur le vieillissement et ils ont rendu compte que le vieillissement est associé à une diminution des capacités motrices et sensorielles (Chen, 2013). À cet effet, plusieurs études liées à l'analyse des gestes ont été réalisées. Parmi ces travaux, on trouve des recherches qui concernent l'annotation des gestes comme (Bolly & Boutet, 2018), d'autres concernent la reconnaissance de la langue des signes (Cui, Liu, & Zhang, 2017) (Hore et al., 2017) et l'analyse des gestes chez les personnes âgées (Natsumi, Saitoh, Wakita, & Nakatoh, 2018) .

Un des intérêts de l'analyse gestuelle est le problème qui émerge de la classification des phases gestuelles telle que définie dans de nombreuses études depuis les années 80 (Kendon, 1980) (Madeo et al., 2016). Les phases gestuelles sont définies comme suit : position de repos, préparation, action et rétraction. Cette logique de segmentation peut révéler des informations sur le contenu d'une conversation. Bien qu'il existe un nombre important de recherches qui portent sur les personnes âgées et l'analyse gestuelle, il existe peu d'études basées sur des approches automatiques qui concernent l'analyse gestuelle des personnes âgées

Motivé par ce défi, nous proposons une approche automatique pour classifier les mouvements de la main en nous basant sur une vérité terrain effectuée par des experts. L'approche proposée vise à imiter les annotations d'expert en linguistique réalisées dans un corpus appelé Corpagest (Bolly & Boutet, 2018). Il s'agit d'un corpus longitudinal contenant

des annotations audio et vidéo et des transcriptions synchronisées. Le but de ce corpus est d'étudier les marqueurs verbaux et gestuels et de construire un profil pragmatique des personnes âgées, en suivant leurs marqueurs verbaux et gestuels pragmatiques dans des situations réelles. Nous atteignons notre objectif d'annotation automatique en considérant le problème d'annotation comme un problème de classification. Une combinaison entre les réseaux de neurones convolutifs et les réseaux de neurones récurrents sont utilisés pour atteindre l'objectif désiré.

5.2 Problématique et objectifs

Plusieurs travaux de recherche (Kita et al., 1997) (Kendon, 1980) ont mis l'accent sur l'étude des mouvements des mains. Ils ont défini quatre phases principales : repos, préparation, action et rétraction. En effet, quel que soit le mouvement réalisé par les mains, ce mouvement doit appartenir à une phase des quatre phases citées au préalable. L'une des études les plus importantes (Bolly & Boutet, 2018) qui concernent les personnes âgées a appliqué cette classification (les phases) pour annoter manuellement les mouvements des mains enregistrés dans des vidéos afin d'étudier la communication non verbale chez les personnes âgées. Présentement l'annotation des vidéos se fait par des experts d'une façon manuelle. Ce type d'analyse présente plusieurs limites. Il est fastidieux et imprécis vu qu'il est soumis à la variabilité des annotateurs. De même ce type d'analyse est coûteux puisqu'il doit être réalisé que par des experts. De plus, il nécessite beaucoup de temps d'analyse. Sans oublier que les conclusions tirées ne sont pas robustes étant donné que l'analyse manuelle est souvent appliquée à des petits échantillons. L'automatisation de ce processus peut être considérée comme une contribution importante qui permet aux spécialistes de traiter (annoter) un grand nombre de vidéos dans un temps limité afin de tirer des conclusions robustes et établir une analyse objective. L'objectif de ce chapitre est de détailler l'approche que nous avons proposée pour atteindre notre objectif

5.3 Approche proposée

5.3.1 Méthodologie

5.3.1.1 Contexte théorique

L'objectif de ce travail de recherche est d'automatiser le processus manuel d'annotation des gestes des mains chez les personnes âgées. En effet, les classes gestuelles définies dans ce travail sont basées sur les phases gestuelles proposées par Bolly & Boutet (2018) qui se basent sur l'approche proposée par Kendon (1980). Ce dernier a défini quatre phases gestuelles (voir les illustrations ci-aprés). Les phases définies par Kendon sont les suivantes :

- Position de repos : Cette phase est caractérisée par un manque de mouvement, c'est une phase ou les mains sont inactives, techniquement est la position entre deux excursions;
- Préparation : Cette phase est le point de départ de l'exécution du mouvement. Elle se caractérise par l'augmentation de la tension des mains;
- Action : nous parlons de l'exécution du mouvement. Il se caractérise par le fait que la main reste stable. Cette phase comprend tous les mouvements, soit internes comme changement de position ou de forme de la main… soit externes comme mouvement de la main dans son ensemble, d'un endroit à un autre;
- Rétraction : cette phase est le point final de l'exécution du mouvement. Il se caractérise par la diminution de la tension de la main, c'est-à-dire le retour de la main en position de repos.

Figure 5.1 Position de repos (1/2)

Figure 5.2 Préparation (1/2)

Figure 5.3 Préparation (2/2)

Figure 5.4 Action (1/2)

Figure 5.5 Action (2/2)

Figure 5.6 Rétraction (1/2)

Figure 5.7 Rétraction (2/2)

Figure 5.8 Position de repos (2/2)

5.3.1.2 Stratégie d'annotation dans l'approche proposée

Comme nous l'avons mentionné dans les sections précédentes, le processus d'annotation dans Corpagest présente plusieurs limites, le but de ce travail de recherche est de proposer une alternative qui permet d'automatiser le processus manuel. Pour atteindre cet objectif, nous avons modélisé le problème d'annotation comme un problème de classification qui a été résolue par une combinaison de CNN et LSTM. La section suivante est consacrée à décrire la méthodologie de l'approche proposée.

5.3.2 Aperçu générale

Il existe plusieurs techniques de suivi des mains mais la plupart d'entre elles ne sont pas efficaces et peuvent se confondre dans certaines situations comme dans le cas d'un arrière-plan inhabituel ou dans le cas de changements brusques des conditions d'éclairage ou d'occlusion (Cheok, Omar, & Jaward, 2019).

Quand on essaie d'analyser la spécificité de chaque classe (Hold, Stroke, Preparation, Retraction), on peut comprendre qu'on est face à un problème de reconnaissance d'action. Par conséquent, une solution appropriée doit prendre en considération le mouvement de la main au cours du temps.

Bien que de nombreux travaux connexes ont mis l'accent sur l'analyse des mouvements des parties du corps tels que Cardenas & Chavez (2020). Ces études ne prennent en compte que la variation de position, la vitesse et l'accélération des articulations du corps. Ceci explique la confusion dans les phases de préparation et de rétraction vue que ces deux phases sont très similaires comme le montre la figure 4.2. D'où la nécessité de prendre en considération les frames qui précèdent le frame en question. Pour surmonter ce problème, nous proposons d'utiliser une architecture basée sur une combinaison de CNN-LSTM. En effet, la méthode que nous proposons utilise en entrée une séquence de 5 images pendant les 0,5 dernières secondes du mouvement de la main (0,1 seconde pour chaque image). La séquence désignée par S [i] est transmise à une couche distribuée dans le temps qui permet d'extraire les caractéristiques de chaque frame à travers une architecture ResNet50. La sortie est un vecteur d'incorporation V de forme (5, 2048) contenant les caractéristiques codées (1, 2048) des 5

images séparément. Maintenant, pour prendre en compte la variation temporelle de ces fonctionnalités, une simple couche LSTM fera cette tâche.

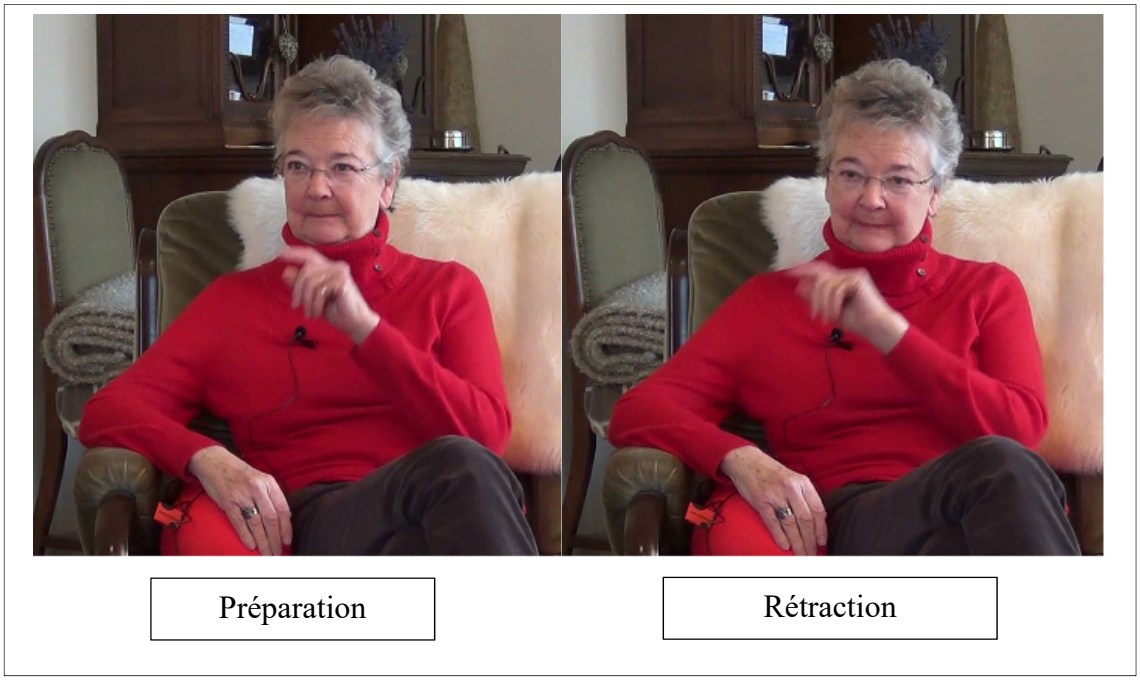


Figure 5.9 La ressemblance entre les phases de préparation et de rétraction

5.3.3 Le pipeline du système d'annotation

5.3.3.1 La phase de détection

La première étape pour atteindre notre objectif consiste à détecter les mains. Ceci peut être réalisé par une combinaison de plusieurs travaux de recherche. En effet, l'idée a été inspirée de la méthode proposée par F. Zhang et al. (2019) qui consiste en un SSD MobileNet v2. Nous avons utilisé un modèle de détection des mains déjà entraîné. Ce modèle a été entraîné sur une base de données nommée Egohands Dataset (Bambach, Lee, Crandall, & Yu, 2015).

Le modèle Single Shot Multibox (SSD) et le classifieur MobileNet ont été combinés et préentraînés sur l'ensemble de données Egohnands Dataset (Common Objects in Context).
Cet ensemble de données est très utilisé par les chercheurs en vision par ordinateur vue la richesse de cette base, elle contient des annotations de haute qualité qui dépassent 15 000 étiquettes de vérité terrain.

MobileNet a été introduit pour les applications de vision mobiles et embarquées dans (Howard et al., 2017), MobileNet est basé sur des convolutions séparables en profondeur. Fondamentalement, MobileNet est une classe de réseaux de neurones convolutifs légers qui sont beaucoup plus petits et plus rapides en performances que de nombreux autres modèles populaires. Des convolutions séparées en profondeur appliquent d'abord un seul filtre sur chaque entrée pour filtrer les données d'entrée, suivies de convolutions 1 × 1 qui combinent ces filtres en un ensemble d'entités de sortie. Ces couches séparables en profondeur imitent la fonction des couches de convolution typiques mais avec une vitesse beaucoup plus rapide et avec une légère différence. Cela minimise la taille du modèle et réduit les demandes de puissance de calcul. Toutes les couches sont suivies d'une non-linéarité batchnorm et ReLU à l'exception de la couche finale entièrement connectée qui alimente une couche softmax pour une classification sans non-linéarité. La figure ci-dessous illustre l'architecture générale de MobileNet

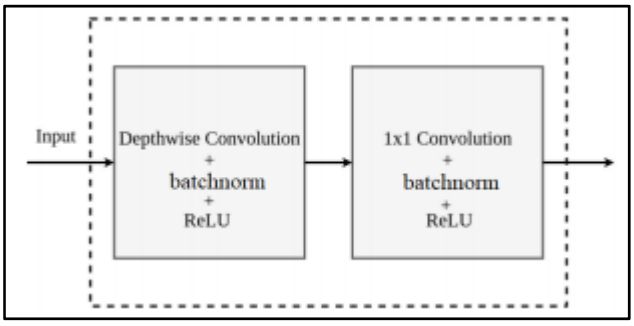


Figure 5.10 Architecture MobileNet tirée de Ghoury (2019)

SSD est proposé par Liu et al. (2016) dans le but de détecter des objets. Plusieurs réseaux de neurones ont été combinés ensemble dans SSD pour atteindre une vitesse de détection rapide

tout en gardant l'efficacité de détection. SSD utilise des couches plus superficielles avec une résolution plus élevée afin de détecter les petits objets.

Après la phase de détection, chaque vidéo se transforme en deux ensembles de séquences, un ensemble pour la main droite et l'autre pour la main gauche. La figure ci-dessous illustre la phase de prétraitement, la phase qui précède la phase de prédiction.

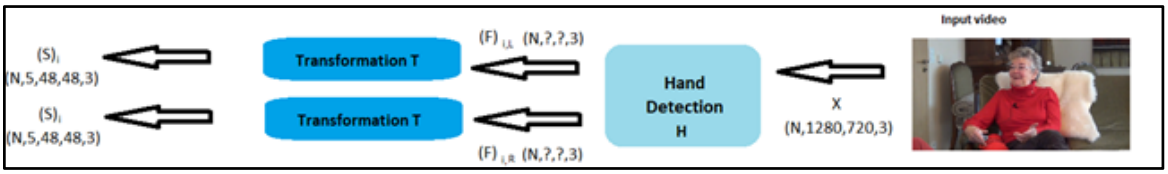


Figure 5.11 Phase de prétraitement

Tableau 5.1 Fonctions utilisées dans la phase de détection

N : Nombre total d'images
X : Vidéo d'entrée
H(X) : fonction de détection des mains
$(F)_{i,h}$: Image rongée i pour la main dont i dans l'intervalle [0,1] et h représente la main droite ou la main gauche
T $((F)_i)$: Fonction de transformation : redimensionner et sélectionner les 5 dernières images chaque 0,5 seconde
$(S)_i$: Un ensemble de séquences pour la main h où S[i] est une séquence de 5 images chaque 0.5 seconde
$(F)_{i,R}$, $(F)_{i,L}$ = H(X)
$(F)_{i,=}$ T $((F)_{i,h})$

5.3.3.2 Phase de prétraitement

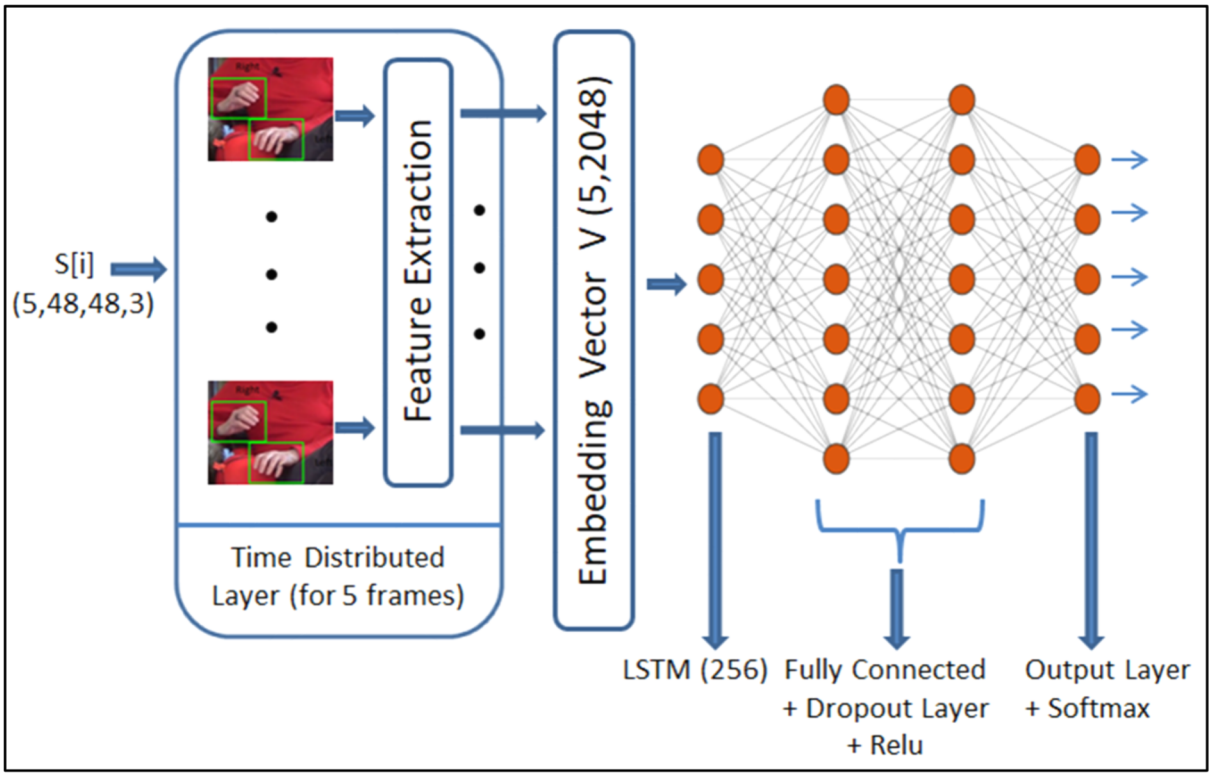


Figure 5.12 Architecture proposée

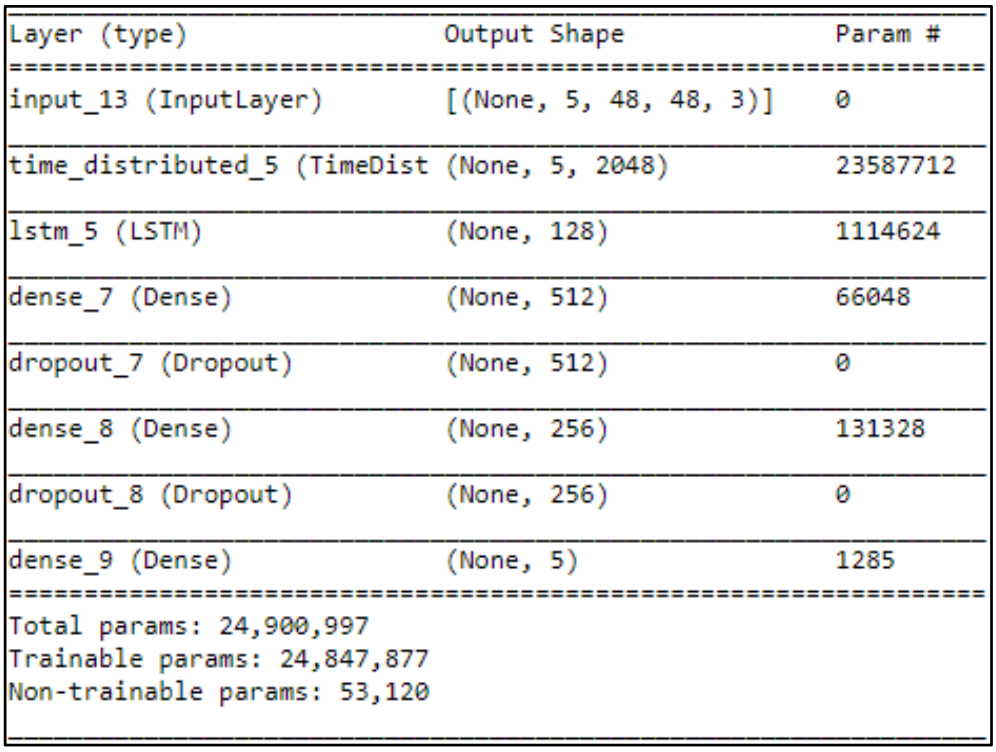
Le vecteur résultat $Y \in R^4$ peut être obtenu à partir de la séquence d'entrée $Seq \in R^{5xZxZxC}$ (Où Z et C désignent respectivement la taille des frames et le nombre des canaux) comme suit :

$V = FE\ (Seq) \in R^{5x2048}$

$T = TF\ (V) \quad \in R^{256}$

$Fc = FCL(T) \in R^4$

$Y = Softmax\ (Fc)$



Layer (type)	Output Shape	Param #
input_13 (InputLayer)	[(None, 5, 48, 48, 3)]	0
time_distributed_5 (TimeDist	(None, 5, 2048)	23587712
lstm_5 (LSTM)	(None, 128)	1114624
dense_7 (Dense)	(None, 512)	66048
dropout_7 (Dropout)	(None, 512)	0
dense_8 (Dense)	(None, 256)	131328
dropout_8 (Dropout)	(None, 256)	0
dense_9 (Dense)	(None, 5)	1285

Total params: 24,900,997
Trainable params: 24,847,877
Non-trainable params: 53,120

Figure 5.13 Résumé du modèle

Où V est le vecteur d'incorporation des caractéristiques spatiales extraites des images par la fonction FE ; T est un vecteur résultant de l'exploitation de l'évolution temporelle des caractéristiques spatiales à l'aide de la TF, une fonction basée sur le LSTM ; FCL est un représentant des 2 blocs de : couches entièrement connectées + fonction d'activation Relu + Dropout. FE s'appuie sur un bloc ResNet50 qui a été introduit par Rezende, Ruppert, Carvalho, Ramos, & De Geus (2017).

5.3.3.3 Entraînement des données

L'entraînement des données a été effectué à l'aide d'un optimiseur Adam (lr = 0,000 1) et d'une taille de lot de 64 pour 14 époques.

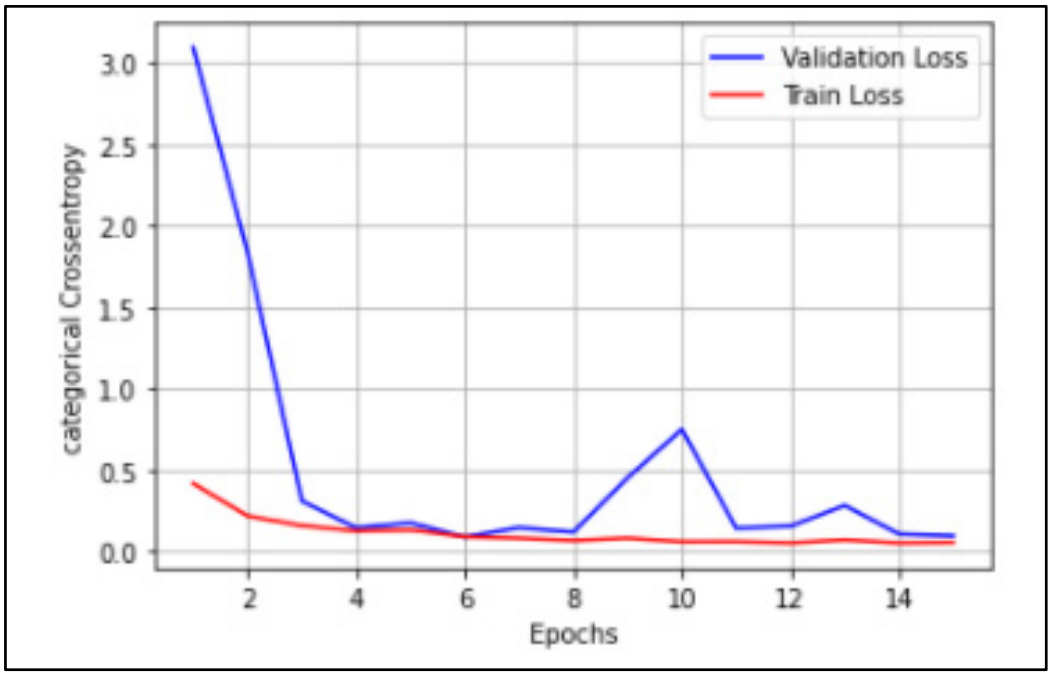


Figure 5.14 Entrainement des données

5.3.4 Résultats

Dans cette section, nous réalisons des expériences pour évaluer l'efficacité de l'approche proposée. Nous validons cette approche en utilisant une vérité terrain, annotée par des experts (linguistes). Nous avons utilisé 30 000 séquences d'entrée extraites d'une conversation spontanée. Nous avons divisé au hasard notre ensemble de données en 3 groupes :

- l'ensemble d'apprentissage est utilisé pour créer des modèles prédictifs, il représente 70 % de l'ensemble des données;
- l'ensemble de validation est utilisé pour évaluer les performances du modèle créé lors de la phase d'apprentissage. Il fournit une plateforme de test pour affiner les paramètres d'un modèle et sélectionner le modèle le plus performant. Il représente 15 % de l'ensemble des données;

- l'ensemble de tests, ou données invisibles est utilisé pour évaluer les performances futures probables d'un modèle. Si un modèle s'adapte beaucoup mieux à l'ensemble d'entraînement qu'il ne correspond à l'ensemble de test, le sur-apprentissage peut être la cause derrière ce problème. Il représente 15 % de l'ensemble des données.

5.3.4.1 Évaluation de l'ensemble d'apprentissage

Tableau 5.2 Performances des classes dans la phase d'apprentissage

Classes réelles	***Prédictions***				
	Stroke	***Preparation***	***Hold***	***Retraction***	***Transition***
Stroke	***0.9939***	*0.0000*	*0.0046*	*0.0010*	*0.0005*
Preparation	*0.0015*	***0.9984***	*0.0001*	*0.0000*	*0.0000*
Hold	*0.0001*	*0.0001*	***0.9989***	*0.0000*	*0.0009*
Retraction	*0.0000*	*0.0015*	*0.0000*	***0.9985***	*0.0000*
Transition	*0.0016*	*0.0005*	*0.0111*	*0.0000*	***0.9868***

Le tableau 5.2 est une matrice de confusion qui montre la performance de chaque classe dans l'ensemble d'apprentissage. En effet l'exactitude dépasse 98% dans toutes les classes. Elle est proche de 100 % dans certaines classes. Les résultats obtenus dans la phase d'apprentissage sont bons.

Tableau 5.3 Performances globales de la phase d'apprentissage

Exactitude globale	***Mauvaise classification***
0.9962	***0.0038***

Le tableau 5.3 montre les performances globales de la phase d'apprentissage. En effet l'exactitude globale des classes dépasse 99% alors que le pourcentage des mouvements qui sont mal classées est inférieur à 1%

Tableau 5.4 Rapport de classification de la phase d'apprentissage

	Précision	***Rappel***	$F_1\ score$
Stroke	*0.9954*	*0.9939*	*0.9947*
Preparation	*0.9990*	*0.9984*	*0.9987*
Hold	*0.9939*	*0.9989*	*0.9964*
Retraction	*0.9985*	*0.9985*	*0.9985*
Transition	*0.9973*	*0.9868*	*0.9920*
Moyenne macro	***0.9968***	***0.9953***	***0.9961***
Moyenne pondérée	***0.9962***	***0.9962***	***0.9962***

Le tableau 5.4 montre le rapport de classification de la phase d'apprentissage. En effet, le tableau montre les valeurs d'exactitude, de précision et mesure-F pour chaque classe. Nous remarquons que les moyennes macros ainsi que les moyennes pondérées des trois mesures sont bonnes, ce qui aide à obtenir des résultats promoteurs dans phase de test.

5.3.4.2 Évaluation de l'ensemble de validation

Tableau 5.5 Performances des classes dans la phase de validation

Classes réelles	***Prédictions***				
	Stroke	***Preparation***	***Hold***	***Retraction***	***Transition***
Stroke	***0.9822***	*0.0036*	*0.0047*	*0.0047*	*0.0047*
Preparation	*0.0028*	***0.9958***	*0.0007*	*0.0000*	*0.0007*
Hold	*0.0013*	*0.0017*	***0.9932***	*0.0000*	*0.0038*
Retraction	*0.0000*	*0.0000*	*0.0000*	***1.0000***	*0.0000*
Transition	*0.0053*	*0.0011*	*0.0317*	*0.0011*	***0.9608***

Le tableau 5.5 est une matrice de confusion qui montre la performance de chaque classe dans l'ensemble de validation. En effet l'exactitude dépasse 96% dans toutes les classes. Elle est égale à 100 % dans certaines classes. Les résultats obtenus dans la phase de validation sont bons.

Tableau 5.6 Performances globales de la phase de validation

Exactitude globale	***Mauvaise classification***
0.9880	*0.0120*

Le tableau 5.6 montre les performances globales de la phase de validation. En effet l'exactitude globale des classes dépasse 98% alors que le pourcentage des mouvements qui sont mal classées est inférieur à 2%

Tableau 5.7 Rapport de classification de la phase de validation

	Précision	***Rappel***	$F_1\ Score$
Stroke	*0.9857*	*0.9822*	*0.9840*
Preparation	*0.9945*	*0.9958*	*0.9952*
Hold	*0.9853*	*0.9932*	*0.9893*
Retraction	*0.9915*	*1.0000*	*0.9957*
Transition	*0.9848*	*0.9608*	*0.9727*
Moyenne macro	***0.9884***	***0.9864***	***0.9874***
Moyenne pondérée	***0.9880***	***0.9880***	***0.9880***

Le tableau 5.7 montre le rapport de classification de la phase de validation. En effet, le tableau montre les valeurs d'exactitude, de précision et mesure-F pour chaque classe. Nous remarquons que les moyennes macros ainsi que les moyennes pondérées des trois mesures sont bonnes, elles dépassent 98%, ce qui aide à obtenir des résultats promoteurs dans phase de test.

5.3.4.3 Évaluation de l'ensemble de test

Tableau 5.8 Performances des classes dans la phase de test

	Prédictions				
Classes réelles	***Stroke***	***Preparation***	***Hold***	***Retraction***	***Transition***
Stroke	***0.9763***	*0.0036*	*0.0154*	*0.0012*	*0.0036*
Preparation	*0.0007*	***0.9986***	*0.0000*	*0.0000*	*0.0007*
Hold	*0.0008*	*0.0004*	***0.9924***	*0.0004*	*0.0059*
Retraction	*0.0051*	*0.0000*	*0.0000*	***0.9949***	*0.0000*
Transition	*0.0095*	*0.0032*	*0.0307*	*0.0000*	***0.9566***

Le tableau 5.5 est une matrice de confusion qui montre la performance de chaque classe dans l'ensemble de test. En effet l'exactitude dépasse 95% dans toutes les classes. Elle est proche de 100 % dans certaines classes. Les résultats obtenus dans la phase de test sont bons ce qui explique l'efficacité de l'approche proposée.

Tableau 5.9 Performance globale de la phase de test

Exactitude globale	***Mauvaise classification***
0.9864	*0.0136*

Le tableau 5.9 montre les performances globales de la phase de test. En effet l'exactitude globale des classes dépasse 98% alors que le pourcentage des mouvements qui sont mal classées est inférieur à 2%

Tableau 5.10 Rapport de classification de la phase de test

	Précision	***Rappel***	$F_1\ score$
Stroke	*0.9821*	*0.9763*	*0.9792*
Preparation	*0.9952*	*0.9986*	*0.9969*
Hold	*0.9824*	*0.9924*	*0.9874*
Retraction	*0.9966*	*0.9949*	*0.9957*
Transition	*0.9805*	*0.9566*	*0.9684*
Moyenne macro	***0.9874***	***0.9838***	***0.9855***
Moyenne pondérée	***0.9864***	***0.9864***	***0.9864***

Le tableau 5.10 montre le rapport de classification de la phase de test. En effet, le tableau montre les valeurs d'exactitude, de précision et mesure-F pour chaque classe. Nous remarquons que les moyennes macros ainsi que les moyennes pondérées des trois mesures sont bonnes, elles dépassent 98%, ce qui explique encore une fois l'efficacité de l'approche proposée

5.4 Discussion et conclusion

Dans ce chapitre, nous avons présenté une approche automatique pour l'annotation des mouvements des mains chez les personnes âgées. En effet, nous avons commencé le chapitre par une introduction qui porte sur le contexte général de ce travail de recherche. Puis nous avons mentionné les principaux travaux de recherche, qui sont en lien avec cette étude, dans une section nommée revue de littérature. Dans une troisième partie, nous avons détaillé la méthodologie qui nous a permis d'atteindre l'objectif fixé au préalable (annotation automatique des mouvements des mains selon les standards). La méthode proposée va permettre de réduire le coût du processus d'annotation. Cette étude a illustré comment des techniques automatisées peuvent être utilisées pour annoter des vidéos selon des normes proposées par des experts. Les résultats obtenus nous ont permis de conclure que l'automatisation du processus d'annotation est possible. Cependant, plusieurs questions s'imposent ; ces questions sont liées à l'interprétation des résultats d'annotation. En effet, dans certains cas il y a une différence entre le résultat donné par l'approche proposée et la vérité terrain. Ceci nous confirme l'idée que malgré le progrès scientifique et la révolution technologique nous sommes toujours incapables d'inventer une technologie capable de simuler le raisonnement de l'être humain. L'approche proposée peut donner des résultats plus précis si on augmente la base de données sur laquelle nous avons travaillé. Ceci va être un futur projet, nous allons enrichir la base de données par d'autres enregistrements afin que le modèle proposé puisse donner des meilleurs résultats.

CHAPITRE 6

DISCUSSION GÉNÉRALE

Dans cette thèse, nous avons proposé des approches pour automatiser l'annotation des mouvements de la tête et des mains. Les deux approches proposées considèrent le problème d'annotation comme un problème de classification. Elles sont basées sur l'utilisation des techniques d'apprentissage profond qui sont essentiellement CNN et RNN. Les expériences ont donné des résultats prometteurs. Les modèles ont atteint un score F variant de 0,928 8 à 0,996 9, en fonction de la complexité des conditions et de la phase gestuelle elle-même.

Nous pouvons affirmer que cette étude apporte deux contributions significatives : une contribution technique qui consister à mixer plusieurs techniques existantes ensemble. En effet, le mariage entre les techniques de détection (MTCNN et le SSD MobileNet) avec les LSTM est une idée originale. La deuxième contribution consiste à automatiser un processus manuelle qui va aider les différentes parties prenantes dans leurs recherches. Notre approche a pu analyser la complexité imposée par des données dépendant des variations du comportement humain. Les résultats obtenus montrent que notre approche est suffisamment robuste pour supporter la mise en place d'un système spécialisé dans la segmentation de phase gestuelle et des outils d'annotation pour soutenir la prise de décision des spécialistes.

L'automatisation de l'annotation des phases gestuelles a pu imposer plus d'efficacité et de précision au processus d'annotation, réduisant l'impact de la subjectivité généralement causée par deux facteurs : la fatigue causée par l'annotation manuelle et l'imprécision générée par l'analyse visuelle des mouvements dans une séquence d'images.

Les deux approches proposées dans cette thèse sont suffisamment robustes pour être intégrées dans plusieurs applications qui doivent identifier les phase significatives des gestes, telles que les applications utilisées pour : identifier les structures prosodiques (Eisenstein, Barzilay, & Davis, 2008) (Kettebekov, Yeasin, & Sharma, 2005) ; identifier les perturbations du mouvement ou même les perturbations liées à la construction du discours (Quek, 2004).

Il est à noter que dans certains cas les annotations automatiques se diffèrent des annotations manuelles. Autrement dit les annotations détectées par la machine ne sont pas détectées par l'être humain et inversement. Ceci peut être expliqué par la subjectivité des annotations manuelles d'une part et par la complexité des mouvements d'autre pat. Notre approche présente, cependant, une limitation liée aux moments de début et de fin des phases, ceci peut être expliqué par l'hésitation ou la rapidité des mouvements de la personne qui gesticule. En effet, parfois, Nous trouvons une difficulté à déchiffrer la phase gestuelle soit parce que la phase est composée de plusieurs phases inachevées ou parce qu'une phase est combinée avec une autre de telle manière nous trouvons une difficulté à identifier le début et la fin de chaque phase.

Ces recherches ouvrent également des opportunités de recherches complémentaires telles que la relation entre la main gauche et la main droite dans un premier niveau et la relation entre les mouvements de la tête et les mouvements des mains dans un second lieu. Nous pouvons aussi mettre l'accent sur la relation des gestes de chaque main avec le discours ; et enfin, analyser différents contextes de gesticulation naturelle tels que les entretiens, les débats, les conférences et la gesticulation dans le contexte des langues des signes.

CONCLUSION

Dans cette thèse, nous avons abordé un sujet très important, le sujet de l'annotation des gestes chez les personnes âgées. En effet, l'évolution du nombre de cette population suit une courbe exponentielle, la raison pour laquelle les études qui concernent la population vieillissante ne cessent pas de croître afin de l'aider à vivre confortablement d'une part et pour minimiser les fardeaux économiques d'autre part. Ce travail de recherche s'inscrit dans le cadre des efforts pour fournir un support scientifique aux différentes parties prenantes concernées par le vieillissement afin de mieux comprendre ou de faciliter l'étude de leurs comportements. Le point de départ dans ce travail de recherche était que les études qui concernent la population vieillissante ne sont pas consacrées à une personne ou une communauté et que chacun peut contribuer à sa façon. En tant que chercheur dans le domaine de technologie de l'information, la meilleure façon de contribuer à l'étude de la communication non verbale chez les personnes âgées est de fournir une approche automatique pour annoter les mouvements chez cette population afin d'aider les spécialistes et les différentes parties prenantes à étudier, diagnostiquer et traiter l'anomalie. Nous avons développé deux approches : une approche automatique pour annoter les mouvements de la tête. La deuxième approche concerne l'annotation automatique des mouvements des mains. Dans les deux approches, nous avons recours à une vérité terrain réalisée par des experts-linguistes. Nous avons utilisé des techniques d'apprentissage profond pour atteindre notre objectif. Une combinaison entre les réseaux de neurones convolutifs et les réseaux de neurones récurrents a été mise en place pour annoter les mouvements de la tête ainsi que les mouvements des mains. Dans la première approche, nous avons utilisé une technique, qui s'appelle MTCNN, pour détecter le visage et le LSTM pour prédire la classe de chaque mouvement. Dans la deuxième approche, nous nous sommes basés sur des études antérieures pour classifier les mouvements des mains. Techniquement parlant, nous avons utilisé le MobileNet pour détecter les mains et le LSTM pour prédire la classe. Les approches proposées nous ont permis de réduire le coût des annotations manuelles et minimiser le temps de travail. Les résultats proposent deux axes de recherche à explorer. Premièrement, le processus d'annotation automatique des comportements non verbaux dans les vidéos est tout

à fait faisable ; deuxièmement, les algorithmes d'apprentissage approfondi peuvent identifier des caractéristiques qui ne correspondent pas totalement aux observations des experts. Cette particularité établit clairement un nouveau dialogue entre les chercheurs en intelligence artificielle, les chercheurs en linguistique et les chercheurs dans les domaines liés à la communication et au vieillissement, quant à l'interprétation des résultats et des caractéristiques à partir de vidéos. La principale contribution de ce travail est le développement des techniques de reconnaissance gestuelle plus performantes adaptées au vieillissement de la population, donnant l'opportunité aux chercheurs d'explorer les spécificités de façon automatisée.

Malgré l'importance des résultats obtenus, notre travail de recherche, comme tout travail de recherche, présente certaines limites. En effet, dans ce genre de travaux la taille de base de données qui représente la vérité terrain est très importante pour que l'approche soit applicable sur différents vidéos avec les mêmes taux. Ceci était un défi dans notre recherche car le nombre des vidéos ainsi que le nombre des personnes qui ont participé dans les conversations étaient limités. Nous suggérons comme un futur travail d'augmenter le nombre des conversations filmées en tenant compte de l'aspect longitudinal pour que nous puissions suivre l'évolution de l'état des participants.

ANNEXE I

DIFFUSION SCIENTIFIQUE

a. Garraoui, H. Ratté, S. (29 Mai, 2019). "*Approche automatisée pour l'étude de la communication non verbale chez les personnes âgées*". 87e Congrès de l'"Acfas. Outaouais, Canada. Conférence.

b. Garraoui, H. Ratté, S. Duong, L. (30 Juillet 2020). "*Automatic annotations of hand movements among elderly people susceptible to Alzheimer's diseas* ". The Journal of the Alzheimer's Association. Volume 16, Issue S7: Dementia Care and Psychological Factors. Poster.

c. Garraoui, H. Ratté, S. Duong, L. (Mai 2021). "*Dynamic approach for hand gesture phases recognition using CNN-LSTM*". Article soumis au journal « Journal of Visual Communication and Image Representation »

d. Garraoui, H. Ratté, S. Duong, L. "*Dynamic approach for head gestures annotations* " Article prêt pour soumission.

e. Garraoui, H. Ratté, S. Duong, L. (Décembre 2021). "*Computer-based annotation of hand gestures among elderly people susceptible to Alzheimer's disease*". 14th World Congress on Dementia and Alzheimer's disease. présentation orale.

BIBLIOGRAPHIE

Albers, M. W., Gilmore, G. C., Kaye, J., Murphy, C., Wingfield, A., Bennett, D. A., . . . Devanand, D. P. (2015). At the interface of sensory and motor dysfunctions and Alzheimer's disease. *Alzheimer's & dementia, 11*(1), 70-98.

Alexanderson, S., House, D., & Beskow, J. (2016). Automatic annotation of gestural units in spontaneous face-to-face interaction. Dans *Proceedings of the Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction* (pp. 15-19).

Ameur, S., Khalifa, A. B., & Bouhlel, M. S. (2020). Chronological pattern indexing: An efficient feature extraction method for hand gesture recognition with Leap Motion. *Journal of Visual Communication and Image Representation, 70*, 102842.

Baltrusaitis, R., Cady, R., Cassiday, G., Cooperv, R., Elbert, J., Gerhardy, P., . . . Steck, D. (1985). The Utah Fly's eye detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 240*(2), 410-428.

Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. Dans *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1949-1957).

Basso, C., Vetter, T., & Blanz, V. (2003). Regularized 3D morphable models. Dans *First IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis, 2003. HLK 2003.* (pp. 3-10). IEEE.

Birdwhistell, R. (1977). Some discussion of ethnography, theory and method. *About Bateson: Essays on Gregory Bateson*, 103-144.

Black, M. J., & Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. Dans *Proceedings of IEEE international conference on computer vision* (pp. 374-381). IEEE.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. Dans *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 187-194).

Bolly, C. T., & Boutet, D. (2018). The multimodal CorpAGEst corpus: Keeping an eye on pragmatic competence in later life. *Corpora, 13*(3), 279-317.

Bourdev, L., & Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. Dans *2009 IEEE 12th International Conference on Computer Vision* (pp. 1365-1372). IEEE.

Bovik, A. C. (2010). *Handbook of image and video processing*. Academic press.

Cardenas, E. J. E., & Chavez, G. C. (2020). Multimodal hand gesture recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes. *Journal of Visual Communication and Image Representation, 71*, 102772.

Chang, Y. S., Nuernberger, B., Luan, B., & Höllerer, T. (2017). Evaluating gesture-based augmented reality annotation. Dans *2017 IEEE Symposium on 3D User Interfaces (3DUI)* (pp. 182-185). IEEE.

Chen, W. (2013). Gesture-based applications for elderly people. Dans *International Conference on Human-Computer Interaction* (pp. 186-195). Springer.

Cheok, M. J., Omar, Z., & Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics, 10*(1), 131-153.

Cootes, T. F., & Taylor, C. J. (1992). Active shape models—'smart snakes'. Dans *BMVC92* (pp. 266-275). Springer.

Cootes, T. F., Wheeler, G. V., Walker, K. N., & Taylor, C. J. (2002). View-based active appearance models. *Image and Vision Computing, 20*(9-10), 657-664.

Cristinacce, D., & Cootes, T. F. (2006). Feature detection and tracking with constrained local models. Dans *Bmvc* (Vol. 1, pp. 3). Citeseer.

Cui, R., Liu, H., & Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. Dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7361-7369).

Davis, B. H., Maclagan, M., & Cook, J. (2013). 'Aw, so, how's your day going?': Ways that persons with dementia keep their conversational partner involved. *Pragmatics in dementia discourse*, 83-116.

De Beugher, S., Brône, G., & Goedemé, T. (2018). A semi-automatic annotation tool for unobtrusive gesture analysis. *Language Resources and Evaluation, 52*(2), 433-460.

de la Concepción, M. Á. Á., Morillo, L. M. S., García, J. A. Á., & González-Abril, L. (2017). Mobile activity recognition and fall detection system for elderly people using Ameva algorithm. *Pervasive and Mobile Computing, 34*, 3-13.

Di Mitri, D., Schneider, J., Klemke, R., Specht, M., & Drachsler, H. (2019). Read between the lines: An annotation tool for multimodal data for learning. Dans *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 51-60).

Di Pastena, A., Schiaratura, L. T., & Askevis-Leherpeux, F. (2015). Joindre le geste à la parole: les liens entre la parole et les gestes co-verbaux. *LAnnee psychologique, 115*(3), 463-493.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. Dans *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (pp. 65-72). IEEE.

Eisenstein, J., Barzilay, R., & Davis, R. (2008). Gestural cohesion for topic segmentation. Dans *Proceedings of ACL-08: HLT* (pp. 852-860).

Elamaran, V., Narasimhan, K., Vijayabhaskar, P., & Shiva, G. (2013). Comparison of wavelet filters in hybrid domain watermarking. *Research Journal of Information Technology, 5*(3), 393-401.

Fakourfar, O., Ta, K., Tang, R., Bateman, S., & Tang, A. (2016). Stabilized annotations for mobile remote assistance. Dans *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 1548-1560).

Fanelli, G., Yao, A., Noel, P.-L., Gall, J., & Van Gool, L. (2010). Hough forest-based facial expression recognition from video sequences. Dans *European Conference on Computer Vision* (pp. 195-206). Springer.

Ferstl, Y., Neff, M., & McDonnell, R. (2020). Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics, 89*, 117-130.

Feyereisen, P. (2007). How do gesture and speech production synchronise? *Current psychology letters. Behaviour, brain & cognition,* (22, Vol. 2, 2007).

Furnari, A., Battiato, S., & Farinella, G. M. (2018). Personal-location-based temporal segmentation of egocentric videos for lifelogging applications. *Journal of Visual Communication and Image Representation, 52*, 1-12.

Goth, G. (2011). Brave NUI world. *Communications of the ACM, 54*(12), 14-16.

Grandjean, J., & Collette, F. (2011). Capacités d'inhibition et vieillissement normal. *Le vieillissement cognitif normal. Maintenir l'autonomie de la personne âgée.*, 65-76.

Hernandez-Rebollar, J. L., Kyriakopoulos, N., & Lindeman, R. W. (2004). A new instrumented approach for translating American Sign Language into sound and text. Dans *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* (pp. 547-552). IEEE.

Hore, S., Chatterjee, S., Santhi, V., Dey, N., Ashour, A. S., Balas, V. E., & Shi, F. (2017). Indian sign language recognition using optimized neural networks. Dans *Information Technology and Intelligent Transportation Systems* (pp. 553-563). Springer.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. Dans *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32).

Jaliya, U., Thakore, D., & Kawdiya, D. (2016). A survey on hand gesture recognition. *IRJET*.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the. *The relationship of verbal and nonverbal communication,* (25), 207.

Kettebekov, S., Yeasin, M., & Sharma, R. (2005). Prosody based audiovisual coanalysis for coverbal gesture recognition. *IEEE transactions on multimedia, 7*(2), 234-242.

Kim, I. S., Choi, H. S., Yi, K. M., Choi, J. Y., & Kong, S. G. (2010). Intelligent visual surveillance—a survey. *International Journal of Control, Automation and Systems, 8*(5), 926-939.

Kita, S., Van Gijn, I., & Van der Hulst, H. (1997). Movement phases in signs and co-speech gestures, and their transcription by human coders. Dans *International Gesture Workshop* (pp. 23-35). Springer.

LaViola Jr, J. J., Kruijff, E., McMahan, R. P., Bowman, D., & Poupyrev, I. P. (2017). *3D user interfaces: theory and practice*. Addison-Wesley Professional.

Leite, D. Q., Duarte, J. C., Neves, L. P., De Oliveira, J. C., & Giraldi, G. A. (2017). Hand gesture recognition from depth and infrared Kinect data for CAVE applications interaction. *Multimedia Tools and Applications, 76*(20), 20423-20455.

Liang, X., Kapetanios, E., Woll, B., & Angelopoulou, A. (2019). Real time hand movement trajectory tracking for enhancing dementia screening in ageing deaf signers of British

sign language. Dans *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 377-394). Springer.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. Dans *European conference on computer vision* (pp. 21-37). Springer.

Madeo, R. C. B., Peres, S. M., & de Moraes Lima, C. A. (2016). Gesture phase segmentation using support vector machines. *Expert Systems with Applications, 56*, 100-115.

Maheshkar, V., Kamble, S., Agarwal, S., & Srivastava, V. K. (2012). DCT-based reduced face for face recognition. *International Journal of Information Technology and Knowledge Management, 5*(1), 97-100.

Maricchiolo, F., Gnisci, A., & Bonaiuto, M. (2012). Coding hand gestures: a reliable taxonomy and a multi-media support. Dans *Cognitive behavioural systems* (pp. 405-416). Springer.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological review, 92*(3), 350.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Metaxas, D., & Zhang, S. (2013). A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing, 31*(6-7), 421-433.

Miksik, O., Vineet, V., Lidegaard, M., Prasaath, R., Nießner, M., Golodetz, S., . . . Torr, P. H. (2015). The semantic paintbrush: Interactive 3d mapping and recognition in large outdoor spaces. Dans *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3317-3326).

Natsumi, K., Saitoh, Y., Wakita, Y., & Nakatoh, Y. (2018). Changes in fundamental frequency and gesture of response corresponding to the understanding level of partner's talk. Dans *2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)* (pp. 1-4). IEEE.

Navarretta, C. (2018). The automatic annotation of the semiotic type of hand gestures in Obama's humorous speeches. Dans *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Oda, O., & Feiner, S. (2012). 3D referencing techniques for physical objects in shared augmented reality. Dans *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 207-215). IEEE.

Ong, S. C., & Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Computer Architecture Letters, 27*(06), 873-891.

Parvathy, D. P., & Subramaniam, K. (2019). Rapid speedup segment analysis based feature extraction for hand gesture recognition. *Multimedia Tools and Applications*, 1-16.

Pierce, J. S., Forsberg, A. S., Conway, M. J., Hong, S., Zeleznik, R. C., & Mine, M. R. (1997). Image plane interaction techniques in 3D immersive environments. Dans *Proceedings of the 1997 symposium on Interactive 3D graphics* (pp. 39-ff.).

Pighin, F., Szeliski, R., & Salesin, D. H. (1999). Resynthesizing facial animation through 3d model-based tracking. Dans *Proceedings of the Seventh IEEE International Conference on Computer Vision* (Vol. 1, pp. 143-150). IEEE.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing, 28*(6), 976-990.

Przygodzki-Lionet, N., & Schiaratura, L. (2007). Communication dans le domaine judiciaire: une situation de négociation. Dans *Actes du congrès* (pp. 263-270).

Qidwai, U., & Chen, C.-h. (2009). *Digital image processing: an algorithmic approach with MATLAB*. CRC press.

Quek, F. (2004). The catchment feature model: A device for multimodal fusion and a bridge between signal and sense. *EURASIP Journal on Advances in Signal Processing, 2004*(11), 1-18.

Ramnath, K., Koterba, S., Xiao, J., Hu, C., Matthews, I., Baker, S., . . . Kanade, T. (2008). Multi-view AAM fitting and construction. *International journal of computer vision, 76*(2), 183-204.

Rao, G. A., Syamala, K., Kishore, P., & Sastry, A. (2018). Deep convolutional neural networks for sign language recognition. Dans *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)* (pp. 194-197). IEEE.

Rezende, E., Ruppert, G., Carvalho, T., Ramos, F., & De Geus, P. (2017). Malicious software classification using transfer learning of resnet-50 deep neural network. Dans *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1011-1014). IEEE.

Rousseau, P. V., Matton, F., Lecuyer, R., & Lahaye, W. (2017). The Moro reaction: More than a reflex, a ritualized behavior of nonverbal communication. *Infant Behavior and Development, 46*, 169-177.

Rousseau, T. (2018). *Maladie d'Alzheimer et troubles de la communication*. Elsevier Health Sciences.

Ruesch, J., & Kees, W. (1969). *Nonverbal communications: Notes on the visual perception of human relations*. University of California Press.

Rujoiu, O. (2009). Academic dishonesty: Copy-paste method. Shame and guilt among Romanian high school students. *Revista Romana de Comunicare si Relatii Publice, 11*(2).

Saragih, J. M., Lucey, S., & Cohn, J. F. (2009). Face alignment through subspace constrained mean-shifts. Dans *2009 IEEE 12th International Conference on Computer Vision* (pp. 1034-1041). Ieee.

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011a). Deformable model fitting by regularized landmark mean-shift. *International journal of computer vision, 91*(2), 200-215.

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011b). Real-time avatar animation from a single image. Dans *Face and Gesture 2011* (pp. 117-124). IEEE.

Sauter, D. A., McDonald, N. M., Gangi, D. N., & Messinger, D. S. (2014). Nonverbal expressions of positive emotions.

Schiaratura, L. T., Di Pastena, A., Askevis-Leherpeux, F., & Clément, S. (2015). Expression verbale et gestualité dans la maladie d'Alzheimer: une étude en situation d'interaction sociale. *Gériatrie et Psychologie Neuropsychiatrie du Vieillissement, 13*(1), 97-105.

Schreer, O., Masneri, S., Lausberg, H., & Skomroch, H. (2014). Coding hand movement behavior and gesture with NEUROGES supported by automatic video analysis. Dans *9th International Conference on Methods and Techniques in Behavioral Research (Measuring Behavior 2014), Wageningen, The Netherlands*.

Shin, M., Kim, M., & Kwon, D.-S. (2016). Baseline CNN structure analysis for facial expression recognition. Dans *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 724-729). IEEE.

Smith, K. C., Quelhas, P., & Gatica-Perez, D. (2006). *Detecting abandoned luggage items in a public space*. IDIAP.

Souvenir, R., & Parrigan, K. (2009). Viewpoint manifolds for action recognition. *EURASIP Journal on Image and Video Processing, 2009*, 1-13.

Spiro, I., Taylor, G., Williams, G., & Bregler, C. (2010). Hands by hand: Crowd-sourced motion tracking for gesture annotation. Dans *2010 IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition-Workshops (pp. 17-24). IEEE.

Stojanova, A., Kocaleva, M., Luledjieva, M., & Koceski, S. (2019). High level activity recognition using android smart phone sensors-Review. *Balkan Journal of Applied Mathematics and Informatics, 2*(2), 27-36.

Swears, E., & Hoogs, A. (2009). Functional scene element recognition for video scene analysis. Dans *2009 Workshop on Motion and Video Computing (WMVC)* (pp. 1-8). IEEE.

Truong, D.-M., Doan, H.-G., Tran, T.-H., Vu, H., & Le, T.-L. (2019). Robustness analysis of 3D convolutional neural network for human hand gesture recognition. *International Journal of Machine Learning and Computing, 9*(2), 135-142.

Tsironi, E., Barros, P. V., & Wermter, S. (2016). Gesture Recognition with a Convolutional Long Short-Term Memory Recurrent Neural Network. Dans *ESANN*.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Dans *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 1, pp. I-I). IEEE.

Vogler, C., & Metaxas, D. (1999). Parallel hidden markov models for american sign language recognition. Dans *Proceedings of the Seventh IEEE International Conference on Computer Vision* (Vol. 1, pp. 116-122). IEEE.

Wang, I., Narayana, P., Smith, J., Draper, B., Beveridge, R., & Ruiz, J. (2018). EASEL: Easy Automatic Segmentation Event Labeler. Dans *23rd International Conference on Intelligent User Interfaces* (pp. 595-599).

Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters, 119*, 3-11.

Wang, S., Li, W., Wang, Y., Jiang, Y., Jiang, S., & Zhao, R. (2012). An Improved Difference of Gaussian Filter in Face Recognition. *Journal of Multimedia, 7*(6), 429-433.

Wang, Y., Lucey, S., & Cohn, J. F. (2008). Enforcing convexity for improved alignment with constrained local models. Dans *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.

Weise, T., Bouaziz, S., Li, H., & Pauly, M. (2011). Realtime performance-based facial animation. *ACM transactions on graphics (TOG), 30*(4), 1-10.

Werner, C., Kardaris, N., Koutras, P., Zlatintsi, A., Maragos, P., Bauer, J. M., & Hauer, K. (2020). Improving gesture-based interaction between an assistive bathing robot and

older adults via user training on the gestural commands. *Archives of Gerontology and Geriatrics, 87*, 103996.

Xiao, J., Baker, S., Matthews, I., & Kanade, T. (2004). Real-time combined 2D+ 3D active appearance models. Dans *CVPR (2)* (pp. 535-542).

Xing, J., & Li, X. (2019). Feature Extraction Algorithm of Audio and Video based on Clustering in Sports Video Analysis. *Journal of Visual Communication and Image Representation*, 102694.

Yang, F., Shechtman, E., Wang, J., Bourdev, L. D., & Metaxas, D. N. (2012). Face morphing using 3D-aware appearance optimization. Dans *Graphics Interface* (pp. 93-99). Citeseer.

Yang, F., Wang, J., Shechtman, E., Bourdev, L., & Metaxas, D. (2011). Expression flow for 3D-aware face component transfer. Dans *ACM SIGGRAPH 2011 papers* (pp. 1-10).

Zhang, F., Li, Q., Ren, Y., Xu, H., Song, Y., & Liu, S. (2019). An Expression Recognition Method on Robots Based on Mobilenet V2-SSD. Dans *2019 6th International Conference on Systems and Informatics (ICSAI)* (pp. 118-122). IEEE.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499-1503.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2014). Facial landmark detection by deep multi-task learning. Dans *European conference on computer vision* (pp. 94-108). Springer.

Zhou, M., Liang, L., Sun, J., & Wang, Y. (2010). AAM based face tracking with temporal matching and face segmentation. Dans *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 701-708). IEEE.