Analyse de la posture par apprentissage automatique profond

Kévin Reby

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

MENTION INFORMATIQUE

Par **Kévin RÉBY**

Analyse de la posture par apprentissage automatique profond

Sous la direction de : **Marie BEURTON AIMAR**

Soutenue le 14 décembre 2022

Membres du jury :

M. Olivier ALATA	Professeur/HDR	Université de Saint-Étienne	Rapporteur
M. Pascal BALLET	Maître de conférence/HDR	Université de Brest	Invité
Mme Marie BEURTON AIMAR	Maître de conférence/HDR	Université de Bordeaux	Directrice de thèse
M. Pascal BOURDON	Maître de conférence	Université de Poitiers	Examinateur
M. David HELBERT	Maître de conférence/HDR	Université de Poitiers	Rapporteur
M. Serge MARCHAND	Professeur	Université de Sherbrooke	Examinateur
M. Akka ZEMMARI	Professeur/HDR	Université de Bordeaux	Président du jury

Table des matières

Table des figures

Résumé

La convergence des secteurs de la santé et du numérique a conduit au développement de thérapeutiques numériques (DTx). Ce travail a été réalisé dans le cadre du développement de DTx pour la douleur. En effet, la douleur est un problème important auquel sont confrontés les individus, les familles, les prestataires de soins et la société dans son ensemble. La douleur est généralement mesurée par l'auto-évaluation du patient ou les impressions du clinicien, soit par des entretiens cliniques, soit par des échelles d'évaluation visuelle et numérique de la douleur. Bien qu'utile, l'auto-déclaration de la douleur ne peut être utilisée avec les jeunes enfants, les patients présentant certaines déficiences neurologiques ou psychiatriques, ou avec de nombreux patients en soins postopératoires ou dans des états de conscience transitoires. Les progrès récents de la vision par ordinateur et de l'apprentissage automatique pour l'analyse et la modélisation automatiques du comportement humain pourraient jouer un rôle essentiel pour surmonter certaines limitations dans un contexte clinique. Par conséquent, l'évaluation automatique et objective des troubles de la douleur à partir de signaux comportementaux présente un intérêt croissant pour les cliniciens et les informaticiens.

Le développement et le succès des méthodes basées sur le Deep Learning dans le domaine de la vision par ordinateur ont permis des avancées significatives dans le domaine de la reconnaissance automatique de la douleur. La plupart des études se sont concentrées sur la détection faciale de la douleur, alors que peu de recherches ont été menées sur la pose humaine.

Dans ce travail, nous proposons d'utiliser un modèle de réseau convolutif spatio-temporel (ST-GCN) à deux flux et des mécanismes d'attention spatiale et temporelle, pour évaluer le comportement d'une personne à partir de sa posture. Nous avons testé notre modèle sur UI-PRMD, un ensemble de données de référence qui fournit des données sur le squelette à l'aide de systèmes de capture de mouvement. Nos résultats montrent que nos modèles ST-GCN basés sur l'attention sont plus performants que les méthodes de pointe pour la prédiction du score de qualité et la classification binaire.

Abstract

The convergence of the health and digital sectors has led to the development of digital therapeutics (DTx). This work was done in the context of the development of a DTx for pain management. Pain is an important issue facing individuals, families, healthcare providers and society as a whole. Pain is usually measured by patient self-report or clinician impressions, either through clinical interviews or through visual and numerical pain scales. Although helpful, self-reported pain cannot be utilized on infants, people with specific neurological or mental conditions, those recovering from surgery, or people in temporary states of consciousness. Recent advances in computer vision and machine learning for automatic analysis and modelling of human behaviour could play a key role in overcoming some of the limitations in a clinical context. Therefore, the automatic and objective assessment of pain disorders from behavioural signals is of increasing interest to clinicians and computer scientists.

The development and success of Deep Learning-based methods in the field of computer vision has led to significant advances in the field of automatic pain recognition. Most studies have focused on facial pain detection, while little research has been conducted on human pose.

In this work, we propose to use a two-stream spatio-temporal convolutional network (ST-GCN) model and spatial and temporal attention mechanisms to evaluate a person's behaviour based on their posture. We tested our model on UI-PRMD, a reference dataset that provides skeletal data using motion capture systems. Our results show that our attention-based ST-GCN models outperform state-of-the-art methods for quality score prediction and binary classification.

Glossaire

— DTx : Digital Therapeutics, thérapies numériques
— Active Learning : Apprentissage actif
— ANN : Artificial Neural Network
— AU : Action Unit
— Convolutionnal Neural Network (CNN) : Réseau de neurones convolutifs
— Deep learning : Apprentissage profond
— ECG : Électrocardiogramme
— EEG : Électroencéphalogramme
— FACS : Facial action coding system
— FDA : Food and Drug Administration, l'administration américaine des denrées alimentaires et des médicaments
— fMRI : functional magnetic resonance imaging, Imagerie par résonance magnétique fonctionnelle
— fNIRS : functional near infrared spectroscopy, spectroscopie proche infrarouge fonctionnelle
— GAN : generative adversarial network, réseaux antagonistes génératifs
— GCN : Graph Convolutional Network
— GRU : Gated Recurrent Unit
— HAS : Haute Autorité de Santé
— HAR : Human Action Recognition
— IA : Intelligence Artificielle
— IASP : Association internationale pour l'étude de la douleur
— Information retrieval ou IR : recherche d'information
— k Nearest Neighbors ou kNN : méthode des k plus proches voisins
— LSTM : Long Short-Term Memory
— Machine Learning : Apprentissage automatique
— Natural Language Processing (NLP) : Traitement du Automatique du Langage Naturel, ou TAL
— Recurrent Neural Network (RNN) : Réseau de neurones récurrents
— RGPD : Règlement Général sur la Protection des Données
— Seq2seq : Sequence to sequence
— SDK : Software Development Kit, kit de développement logiciel
— ST-GCN : Spatial Temporal Graph Convolutional Network
— Support Vector Machines (SVM) : Machines à vecteurs de support

Remerciements

Je souhaiterais commencer par remercier toutes les personnes qui m'ont permis de faire cette thèse malgré mon CV « atypique », en particulier ma directrice de thèse, Marie Beurton Aimar, pour m'avoir fait découvrir le deep learning en stage de master, ainsi que les fondateurs de l'entreprise Lucine, Maryne Cotty Esslous et Aymeric Esperance pour m'avoir fait confiance pour mener ce projet de recherche.

Je voudrais également adresser mes remerciements à tous les membres de mon jury de thèse pour leur intérêt et pour avoir accepté d'évaluer mes travaux.

Je remercie mes ami·e·s et collègues du Labri : Myriam, Gala, Claire, Elsa, Cécil, Luc et Manh Tu, ainsi que celles et ceux de Lucine : Charlotte, Linh, Laurie, Adrien et Cécilia pour leur soutien, ainsi que mes amis Yann et Odile pour m'avoir initié à la Boxe Française.

Je remercie également les membres de ma famille pour m'avoir soutenue lorsque j'ai décidé de reprendre mes études, puis de faire une thèse.

Enfin, je remercie Bastien, pour avoir partagé chaque instant de cette thèse. Et je n'oublie pas non plus mon chat Hécate, pour son soutien psychologique et la ronronthérapie. Grâce à eux, le confinement ne fut pas si difficile.

Introduction

Contexte et Motivation

La convergence et le développement des technologies de la santé et du numérique a conduit au développement de la santé numérique, une vaste catégorie de technologies numériques de santé visant à améliorer la santé et le bien-être des personnes, à optimiser la qualité et la sécurité des soins, à accroître l'accès aux traitements, à rendre les services de santé plus efficaces et à réduire les coûts globaux des soins de santé [235]. C'est dans ce contexte que se sont développé les **thérapies numériques** (*Digital therapeutics*). Aujourd'hui, des DTx sont sur le marché ou en cours de développement pour un large éventail de pathologies physiques, psychiques et comportementales, principalement chroniques, telles que : le diabète, la gestion des traitements oncologiques, l'anxiété, la dépression, l'insomnie, le Trouble Déficit de l'Attention avec ou sans Hyperactivité (TDAH), la toxicomanie, ou encore la douleur [136]. Ce travail de thèse s'est effectué dans le cadre du développement de DTx pour la douleur.

En effet, malgré l'amélioration des connaissances, la **douleur** est encore souvent mal prise en charge. La douleur chronique est un problème de santé publique. En Europe, 150 millions de personnes sont atteintes de douleurs chroniques. En France, 70% des patients douloureux ne reçoivent pas de traitement approprié et seulement 3% d'entre eux reçoivent un soin personnalisé [218]. COmme l'indique le Ministère de la Santé et de la Prévention, la loi relative aux droits des malades et à la qualité du système de santé du 4 mars 2002 reconnaît le soulagement de la douleur comme un droit fondamental de toute personne. La lutte contre la douleur est également une priorité de santé publique inscrite dans la loi de santé publique de 2004. [1]

Pour pouvoir traiter la douleur, il faut d'abord la détecter et l'évaluer [101]. Chez l'homme, la douleur est exprimée par la voix, par les expressions du visage, mais aussi à travers la posture et des comportements particuliers. La **posture** peut se définir simplement par la position des différents segments corporels à un instant donnée [214]. Un mouvement humain met en jeu la coordination d'un grand nombre d'articulations et de muscles entraînant une modification de la posture. Il peut être volontaire ou réflexe. En cas de pathologies ou de troubles, la douleur peut altérer l'équilibre du corps, et donc perturber la posture et la motricité. Ainsi, de nombreux symptômes douloureux sont à l'origine de troubles de la posture.

1. solidarites-sante.gouv.fr/soins-et-maladies/prises-en-charge-specialisees/douleur/article/la-douleur

En vision par ordinateur, la **reconnaissance de gestes et d'actions** est un problème complexe en raison de la complexité du corps humain (le nombre de degrés de liberté des articulations, la variabilité d'apparences entre les différentes personnes) et des ambiguïtés visuelles telles que l'auto-occultation ou encore la perte d'information sur la profondeur). L'analyse posturale est un sujet de recherche aux nombreuses applications pratiques comme la réalité virtuelle et augmentée, l'interface homme machine, ou encore l'analyse du geste sportif, principalement étudié en biomécanique [115].

Dans cette thèse, nous souhaitons étudier comment détecter et évaluer la douleur à partir de l'analyse des postures, par l'utilisation des techniques avancées en **apprentissage automatique profond** (*Deep Learning*). La détection automatique de la douleur à partir de la posture est un sujet d'actualité présentant de nombreux attraits pour les professions médicales et paramédicales pour l'amélioration de la prise en charge des patients et la communauté scientifique s'est emparée du sujet. La littérature montre qu'il s'agit d'une approche prometteuse et les résultats s'améliorent régulièrement en termes de précision [318].

Cadre de la thèse

Les travaux de cette thèse dans le cadre du dispositif national **Cifre** (Conventions Industrielles de Formation par la REcherche) qui subventionne toute entreprise de droit français qui embauche un doctorant pour le placer au cœur d'une collaboration de recherche avec un laboratoire public [2]. Les Cifre sont intégralement financées par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation qui en a confié la mise en œuvre à l'ANRT (Association Nationale de la Recherche et de la Technologie) [3]. L'entreprise d'accueil de ce doctorat est **Lucine** [4], une start-up spécialisée dans les DTx pour traiter et soulager la douleur, fondée en 2017 par Maryne Cotty-Eslous et Aymeric Esperance.

Organisation du manuscrit

Dans le chapitre 1, nous définissons la thérapie numérique et ce qu'est la douleur, comment la classifier et la caractériser, ainsi que des problématiques importantes pour son évaluation liée au nombreux biais possible. Nous décrivons aussi les principaux marqueurs physiologiques observables et mesurables dont les signes posturaux afin de permettre une détection automatique de la douleur.

Le chapitre 2 décrit les différentes méthodes pour la capture et l'analyse de mouvements grâce à la reconnaissance d'action à partir de squelettes.

Dans le chapitre 3 nous décrivons tout d'abord les concepts fondateurs de l'apprentissage automatique et de l'apprentissage profond, notamment des CNNs et des RNNs. Nous abordons par la suite les notions de mécanismes d'attention et

2. https ://www.enseignementsup-recherche.gouv.fr/fr/les-cifre-46510

3. https ://www.anrt.asso.fr/fr/le-dispositif-cifre-7844

4. https ://lucine.fr/

de réseaux à graphes qui sont des concepts importants pour la méthode que nous proposons.

Dans le chapitre 4 nous détaillons les différentes méthodes existantes pour l'apprentissage automatique profond pour l'analyse du comportement et la reconnaissance d'actions.

Le chapitre 5 décrit nos travaux sur la détection des émotions à partir de la posture. À partir de la base de données BoLD, nous avons extrait des caractéristiques géométriques tels que les angles entre différents segments du corps afin de prédire une séquence d'émotions grâce à des réseaux de neurones de type LSTM et une architecture de type Encoder-Decoder.

Enfin, le Chapitre 6 décrit les protocoles d'apprentissage et d'évaluation que nous proposons pour notre tâche d'évaluation du comportement douloureux. Nous discutons des deux bases de données de rééducation physique sur lesquelles nous avons mené nos expériences, leurs caractéristiques et les éventuels pré-traitements qui ont été nécessaires à leurs utilisations. Ensuite, nous appliquons la méthodologie que nous avons mise en place sur la base de données UI-PRDMD, avec un modèle basé sur les mécanismes d'attention issus du Transformer et une architecture inspirée des ST-GCNs.

Première partie

Douleur et Analyse de mouvements

Chapitre 1

Contexte

"Of pain you could wish only one thing : that it should stop. Nothing in the world was so bad as physical pain. In the face of pain there are no heroes."

— George Orwell, 1984

Sommaire

La **douleur** constitue le premier motif de consultation médicale, et elle soulève d'importants défis concernant sa prise en charge. L'objectif de l'entreprise **Lucine** est de créer de nouvelles solutions de **thérapies numériques** qui soulagent la douleur chronique, appelées *Digital Therapeutics* (DTx). La douleur est une expérience multidimensionnelle complexe, avec des facteurs physiques, psychologiques et sociaux, impliquant des composantes sensorielles et affectives (émotionnelles). Ces facteurs peuvent modifier la façon dont la douleur est perçue, notamment pour la douleur chronique. Son évaluation de manière objective et fiable est nécessaire en médecine pour établir un diagnostic différentiel, choisir et adapter le traitement adéquat, etc. Actuellement, la pratique courante consiste à se fier à l'**auto-évaluation** du patient. Cependant, pour certains patients, comme ceux souffrant de troubles mentaux ou les nouveau-nés, cette évaluation n'est ni fiable ni valable. Il existe plusieurs marqueurs caractéristiques de la douleur, dont certaines sont indépendantes de l'origine culturelle, de la personnalité ou du sexe du patient. Il s'agit notamment de changements spécifiques dans l'expression du visage, la posture et les paramètres biologiques tels que la fréquence cardiaque, la conductance de la peau ou l'activité électrique des muscles squelettiques. Ces informations peuvent être utilisées pour évaluer la douleur de façon automatique, c'est-à-dire sans que le patient la signale.

Au cours de la dernière décennie, grâce aux progrès conjoints en vision par ordinateur et en apprentissage automatique, la **reconnaissance automatique** de la douleur est devenue un sujet de recherche d'un intérêt considérable. À l'heure actuelle, les systèmes de reconnaissance de la douleur sont principalement basés sur la reconnaissance faciale. Cette reconnaissance faciale se base sur différentes techniques dont l'utilisation des **FACS** (*Facial Action Coding System*) et des **AUs** (*Action Units*) [77]. Parallèlement, afin d'automatiser ces processus d'analyse d'images et de vidéos, les techniques d'apprentissage automatique profond (*deep learning*) se sont largement développées depuis la dernière décennie. Le **deep learning** constitue désormais l'état de l'art en analyse d'image et en vision par ordinateur, y compris dans la reconnaissance faciale, la reconnaissance de la posture, de gestes ou d'actions.

1.1 Digital Therapeutics

En 2015, le terme de **thérapie numérique** apparaît pour la première fois. Il est alors défini comme « des traitements comportementaux fondés sur des données probantes et délivrés en ligne qui peuvent accroître l'accessibilité et l'efficacité des soins de santé » [258, 35]. Depuis, la *Digital Therapeutics Alliance* (DTA), association regroupant les principaux acteurs du domaine, a défini les DTx comme suit : « la délivrance aux patients d'interventions thérapeutiques fondées sur des données probantes et pilotées par des logiciels pour prévenir, gérer ou traiter un trouble médical ou une maladie. Ils sont utilisés indépendamment ou de concert avec des médicaments, des dispositifs ou d'autres thérapies pour optimiser les soins aux patients et les résultats en matière de santé » [73]. De plus, la DTA précise que les DTx intègrent une technologie avancée avec de meilleures pratiques en matière de conception, de soutien clinique, de convivialité et de sécurité des données. Ces produits sont examinés et approuvés par les autorités de réglementation pour compléter les allégations relatives aux risques, à l'efficacité et à l'utilisation prévue. Les DTx ont également pour objectif de permettre à toutes les parties prenantes (les patients,

les prestataires de soins de santé, les assurances et mutuelles), de disposer d'outils intelligents et accessibles pour aborder une variété de conditions par le biais d'interventions basées sur les données qui sont de haute qualité, sûres et efficaces [65].

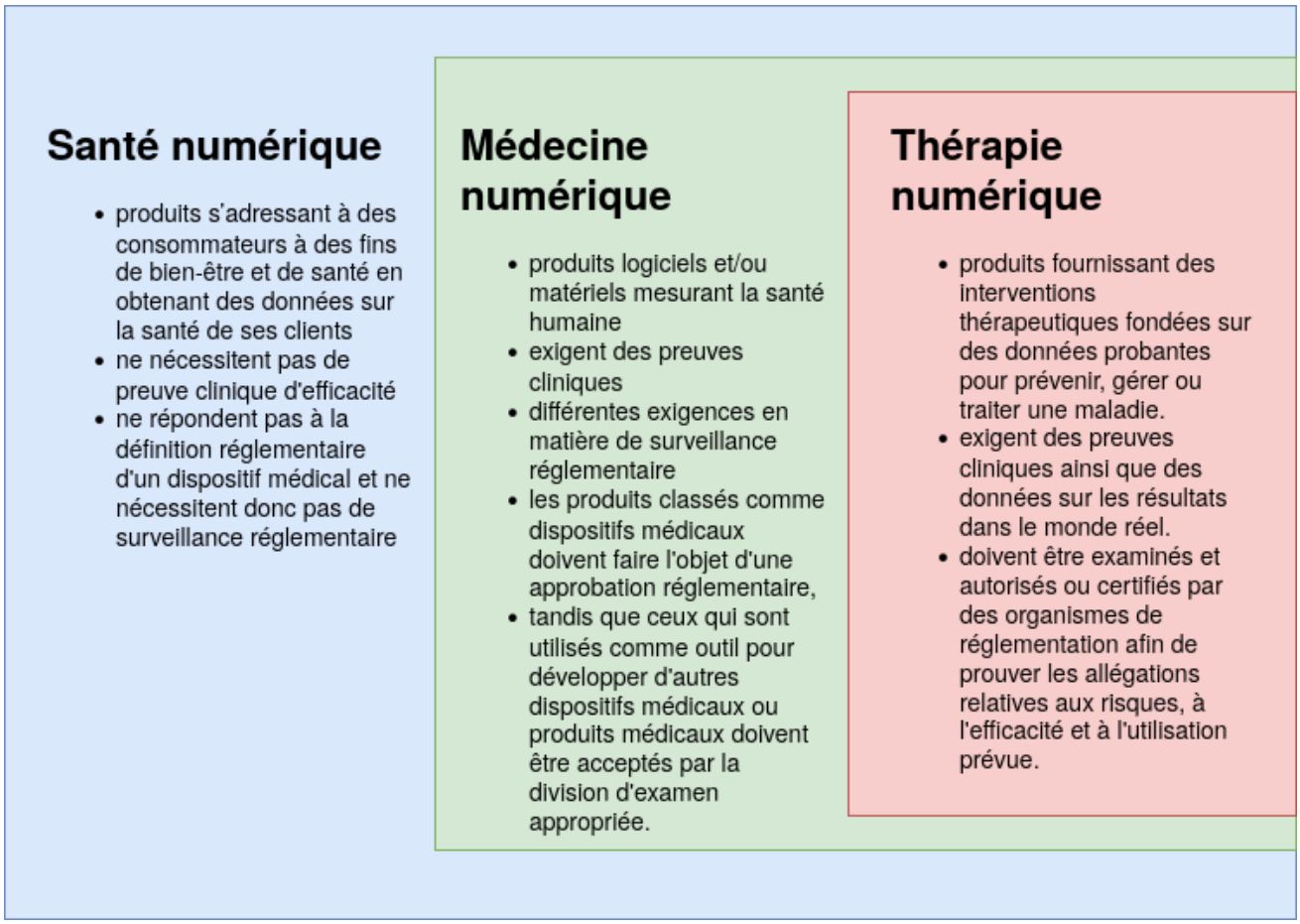


FIGURE 1.1 – Différences entre la santé numérique, la médecine numérique et la thérapie numérique

Comme le montre la figure 1.1, les *Digital Therapeutics* ou DTx constituent une sous-catégorie de la **santé numérique**, qui représente un ensemble de technologies, de produits et de services dans le domaine de la santé et du bien-être (voir le tableau 6.10 en annexe). Il faut donc bien distinguer la santé numérique (regroupant des applications et autres produits qui ne nécessitent pas de preuves cliniques), la médecine numérique (qui s'appuie sur des preuves cliniques, mais sans nécessairement requérir une approbation réglementaire) et les thérapies numériques (qui eux nécessitent une approbation réglementaire). Les DTx se distinguent donc des applications de bien-être, car ils reçoivent une homologation de la part d'une autorité de régulation telle que le FDA (*Food and Drug Administration*) ou l'ANSM (Agence nationale de sécurité du médicament et des produits de santé), permettant ainsi sa commercialisation et son remboursement.

Les DTx ont pour objectifs de fournir aux patients de nouvelles solutions thérapeutiques pour prévenir, gérer ou traiter un large éventail de symptômes et pathologies physiques, mentales et comportementales. Les DTx ont pour objectifs de donner

aux patients et aux prestataires des soins de santé numériques grâce à des outils intelligents et accessibles. Ces outils traitent un large éventail de conditions par le biais d'interventions de haute qualité, sûres et efficaces, reposant sur des données collectées et analysées [136].

Les DTx veulent être considérées comme une classe émergente de médicaments qui fournissent des thérapies fondées sur des preuves. Comme les médicaments, les **produits thérapeutiques numériques** se composent de principes actifs et d'excipients. Alors que le « principe actif numérique » est principalement responsable du résultat clinique, les « excipients numériques » (assistant virtuel, rappels, systèmes de récompense, etc.) sont nécessaires pour garantir la meilleure expérience utilisateur au patient et pour permettre l'utilisation prolongée de la thérapie. Afin de permettre l'interaction avec les patients, les thérapies numériques peuvent prendre différentes formes numériques destinées aux patients, comme des applications pour smartphones, des applications pour téléphones mobiles, des applications pour téléphones portables, etc.

Pour permettre l'interaction avec les patients, la thérapeutique numérique peut prendre différentes formes numériques adaptées aux patients, comme des applications pour smartphones, des jeux vidéo, des programmes de réalité virtuelle ou encore de réalité augmentée. Le processus de recherche et développement de DTx consiste en un développement de logiciel, un développement pilote et un développement clinique complet. Les essais cliniques randomisés et contrôlés de confirmation sont essentiels pour générer des preuves de bénéfices pour l'approbation réglementaire, le remboursement et la prescription. Les thérapies numériques ont le potentiel de transformer la gestion des maladies chroniques et de représenter la première option thérapeutique proposée par chaque médecin à chacun de ses patients souffrant de maladies chroniques et de dépendance. Comme ces attentes peuvent ou non être satisfaites et que les avantages potentiels peuvent s'accompagner d'effets indésirables et/ou non voulus, l'introduction, la mise en œuvre, l'utilisation et le financement de la thérapeutique numérique doivent être soigneusement évalués et suivis [235].

Pour être considéré comme un DTx, il est donc nécessaire de remplir trois critères majeurs :

— Être un dispositif médical
— Pouvoir être prescrit par un professionnel de santé
— Prétendre à un remboursement par les institutions réglementaires

En 2017, dans le contexte de la crise des opioïdes (voir section 1.2) la FDA a pour la première fois homologué un Dtx : *reSET* de Pear Therapeutics. La valeur du marché mondial du DTx est estimée à 1,8 milliard $ en 2018, et devrait atteindre 7,1 milliards $ d'ici à 2025. Un rapport récent estime que les plus grandes applications du DTx seront prochainement le diabète et la perte de poids, d'autres applications étant susceptibles d'être observées dans des conditions telles que la bronchopneumopathie chronique obstructive (BPCO), les troubles du développement, ou encore le trouble du stress post-traumatique (TSPT), avec l'utilisation de la réalité virtuelle (VR) [65].

Exemples

Les DTx fonctionnent tous ou presque via une application et traitent essentiellement des maladies chroniques. On peut classer les DTx en trois grandes catégories :

— Premièrement, les DTx basés sur la gamification, sous la forme de jeux vidéos, comme « Somryst », approuvé par la FDA, qui traite l'insomnie chronique chez des patients adultes.
— Deuxièmement, les médicaments numériques basés sur la mobilisation des sens ou thérapies numériques sensitives. Lucine développe des thérapies numériques personnalisées de ce type pour soulager la douleur chronique.
— Troisièmement, les DTx en tant qu'applications de suivi comme « Diabeloop », une application d'autonomisation des suivis pour les personnes diabétiques. Cette thérapie est approuvée par les autorités françaises et peut être remboursée par certains types d'assurances.

Les DTx peuvent aussi être déployées en complémentarité avec des thérapies classiques existantes. En effet, on relève deux stratégies différentes déployées par les laboratoires de DTx :

— une approche en « standalone » (DTx indépendante), où l'entreprise crée une thérapie numérique entièrement nouvelle pour traiter une maladie,
— ou une approche « around-the-pill » (DTx en soutien à un médicament existant), où le laboratoire va plutôt chercher à développer un traitement numérique complémentaire à l'usage d'un médicament existant.

Par exemple, le DTx « DBLG1 » développé par Diabeloop est associé à un capteur de glucose en continu (CGM) et une pompe à insuline [1]. Toutes les cinq minutes, un résultat de glycémie est envoyé à un terminal via la technologie Bluetooth. L'algorithme de DBLG1 analyse les données en temps réel et calcule la juste dose d'insuline à administrer selon les paramètres biologiques personnalisés du patient (âge, poids, vitesse d'élimination) ainsi que les informations renseignées (repas, activité physique). Le corps est ainsi calculé, numérisé. Et les données générées lui reviennent sous forme d'injection d'insuline. Cette approche en soutien à un médicament existant est un exemple d'un DTx qui ne remplace pas un médicament ordinaire, mais fonctionne en parallèle pour traiter un aspect de la maladie que le médicament seul ne peut résoudre.

La douleur est un phénomène complexe, difficile à traiter. Une des solutions pour améliorer sa prise en charge est le développement de **thérapies numériques**. Le recours à ces thérapies numériques vient en complément des traitements traditionnels, car elles ont pour but de réduire fortement la consommation de molécules aux effets secondaires non négligeables. Pour pouvoir traiter la douleur avec des DTx, il faut d'abord la comprendre, l'évaluer, pour pouvoir ensuite proposer l'accompagnement thérapeutique adéquat, ce qui est l'objectif de l'entreprise **Lucine**.

La suite de ce chapitre sera consacrée à l'étude de la douleur et des marqueurs physiologiques observables et mesurables pour la détecter et l'évaluer de façon **automatique**.

1. https ://www.diabeloop.fr/produits

1.2 La Douleur

La douleur est une réponse naturelle dont le but est de protéger notre corps contre une blessure réelle ou potentielle. Cependant, quand la douleur est trop forte ou lorsqu'elle persiste, elle peut entraîner des souffrances aussi bien physiques que psychiques, ainsi que des dépenses médicales et économiques importantes, tant au niveau personnel que pour la collectivité. La douleur est une **expérience individuelle**, c'est-à-dire unique et personnelle, ce qui rend difficile son appréciation objective extérieure [230]. Chez les patients non communicants (tels que les nouveaux nés, certains autistes, ou encore des patients présentant des troubles mentaux, etc), les experts utilisent un ensemble de marqueurs tels que les vocalisations (cris, pleurs, gémissements), les expressions faciales, mais aussi des gestes et des modifications de comportement (positions antalgiques, interactions avec l'entourage...). La douleur est un phénomène complexe et peut-être sujet à différentes interprétations. En effet, si la douleur repose sur un ensemble de processus physiologiques, elle reste également intrinsèquement subjective. Ainsi, bien qu'à première vue la douleur soit un phénomène purement physique, elle peut également être ressentie en l'absence de toute activation sensorielle nocive. Elle peut être également être créée, modifiée, renforcée ou supprimée par d'autres expériences émotionnelles et psychologiques, comme la peur, le stress et la mémoire (par le souvenir d'autres expériences douloureuses passées par exemple) [41, 237, 156].

Il existe plusieurs définitions de la douleur dans la littérature. Ainsi, en 1968, l'infirmière américaine Margo McCaffery, pionnière dans le domaine des soins pour la gestion de la douleur, a défini la douleur comme « ce que la personne qui l'éprouve dit être, et qui existe quand elle le dit » comme l'explique Pasero [215]. Cette définition est intéressante, car elle souligne que la douleur est considérée comme une **expérience subjective**, sans mesure objective. Elle affirme également que c'est le patient, et non le clinicien, qui fait autorité en matière de douleur. En 1979, l'Association internationale pour l'étude de la douleur (*International Association for the study of pain*, ou IASP) [2] a introduit la définition de la douleur qui est devenue la plus courante : une « expérience sensorielle et émotionnelle désagréable associée à un dommage tissulaire réel ou potentiel, ou décrite en termes d'un tel dommage » [276, 40].

Cette définition de la douleur fut ensuite révisée par l'IASP en 2020 en soulignant davantage son aspect complexe et **multidimensionnelle**. La douleur est alors décrite comme :

> « une expérience sensorielle et émotionnelle désagréable associée à une lésion tissulaire réelle ou potentielle, ou ressemblant à une telle lésion, et est complétée par l'ajout de six notes clés et de l'étymologie du mot « douleur » pour un contexte supplémentaire précieux :
> — La douleur est toujours une expérience personnelle qui est influencée à divers degrés par des facteurs biologiques, psychologiques et sociaux.
> — La douleur et la nociception sont des phénomènes différents. La douleur ne peut pas être déduite uniquement de l'activité des neurones

2. www.iasp-pain.org/terminology

sensoriels.
— Au travers de leurs expériences de vie, les individus apprennent le concept de la douleur. Il convient de respecter le fait qu'une personne qualifie une expérience de douloureuse. Bien que la douleur joue généralement un rôle adaptatif, elle peut avoir des effets négatifs sur la fonction et le bien-être social et psychologique.
— **La description verbale n'est qu'un comportement parmi d'autres pour exprimer la douleur** ; l'incapacité à communiquer n'exclut pas la possibilité qu'un humain ou un animal non humain éprouve de la douleur. » [3]

Un problème de Santé Public

La douleur est un problème majeur de santé publique auquel sont confrontés à la fois les individus, les familles, les professionnels de santé, et donc la société dans son ensemble, comme l'a démontré la récente **crise des opioïdes** [6]. La douleur aiguë est la raison la plus courante pour laquelle les patients consultent un médecin. Le mot « douleur » a toujours été un des mots clés les plus recherchés dans le moteur de recherche Google [4], quel que soit le pays, comme le montre la Figure 1.2 (les valeurs sont calculées sur une échelle de 0 à 100, où 100 correspond à la position la plus populaire par rapport au nombre total de recherches, une valeur de 50 indique une position moins de deux fois plus fréquente).

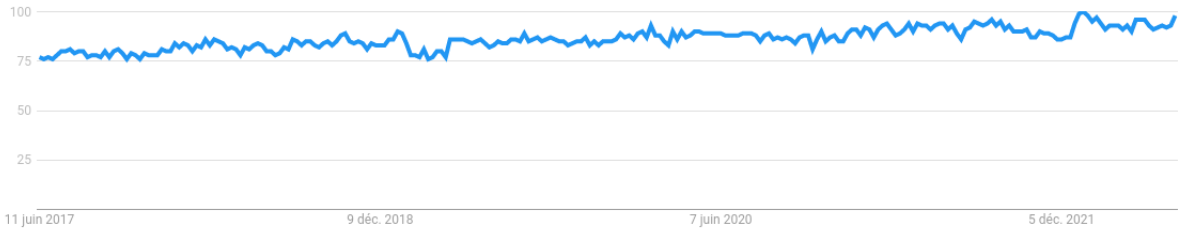


FIGURE 1.2 – Google Trends pour le mot clé "douleur".

Les opioïdes sont des substances composées d'extraits de la graine de pavot, ou de composés synthétiques aux propriétés similaires, et sont couramment utilisés pour le traitement de la douleur, notamment grâce à l'utilisation de médicaments tels que la morphine, le fentanyl et le tramadol. [5]. Cependant, leur utilisation prolongée ou en dehors de toute surveillance médicale peut entraîner une dépendance aux opioïdes et générer d'autres problèmes de santé (notamment des difficultés respiratoires). Ainsi, dans le monde, environ 0,5 million de décès sont dus à la consommation de drogues et plus de 70 % de ces décès sont attribuables aux opioïdes, et plus de 30 % de ces décès étant dus à une surdose. En 2000, le Congrès des États-Unis déclarait les dix années suivantes « Décennie de la lutte contre la douleur et de la recherche ».

3. https ://www.iasp-pain.org/publications/iasp-news/iasp-announces-revised-definition-of-pain/

4. https ://trends.google.fr/

5. https ://www.who.int/news-room/fact-sheets/detail/opioid-overdose

La douleur est alors devenue le « cinquième signe vital » (les quatre autres étant tension artérielle, le pouls, la fréquence respiratoire et la température), et l'évaluation numérique de la douleur est devenue une caractéristique standard des dossiers médicaux. L'évaluation de la douleur est donc devenue une caractéristique standard de la pratique médicale. En conséquence, les médecins se sont retrouvés confrontés à une épidémie de douleur auparavant non signalée auparavant et ont donc commencé à distribuer massivement des opioïdes comme l'**OxyContin** pour soulager les patients. Ainsi, entre 1997 et 2010, la prescription d'Oxycontin chaque année a considérablement augmenté pour atteindre 6,2 millions. Ce médicament, commercialisé par le laboratoire Purdue, a été accusé d'être le principal responsable de la crise des opioïdes qui a ravagé les États-Unis, faisant plus de 450 000 morts. Purdue a été incriminée pour sa stratégie marketing auprès des médecins pour pousser à la prescription de son médicament, tout en sachant qu'il pouvait provoquer une forte accoutumance. Les patients ont ainsi eu tendance à surévaluer leur douleur pour recevoir de l'Oxycontin. Ce mauvais usage de l'Oxycontin serait à lui seul à l'origine de plus de 300 000 morts par overdose depuis les débuts de sa commercialisation sur le marché américain en 1996. Le laboratoire Purdue est depuis en cessation de paiement depuis septembre 2019 et a plaidé coupable pour fraude et corruption en lien avec sa promotion agressive de l'OxyContin. D'autres grands laboratoires ayant vendu des opioïdes, ont été poursuivis, ainsi que des distributeurs, pharmacies, et certains médecins prescripteurs. Le laboratoire Johnson Johnson, les distributeurs McKesson, Cardinal Health et AmerisourceBergen ont ainsi accepté de verser 26 milliards de dollars en dédommagement aux victimes [6].

Cette crise des opioïdes a démontré l'intérêt de rechercher des solutions thérapeutiques alternatives en compléments des traitements classiques par antalgiques, tel que les DTx.

Problème de l'auto-évaluation

Toutes les définitions précédentes montrent que la douleur est prise en compte uniquement à travers la personne qui la ressent. La tolérance face à la douleur varie d'un individu à l'autre selon son vécu, ses croyances, sa personnalité (voir section 1.3). L'**auto-évaluation** du patient est donc pour l'instant considéré comme l'indicateur le plus fiable de la douleur, facile à obtenir grâce aux échelles de mesure (voir section 1.2.3), et est donc devenu la référence pour quantifier la douleur.

Cependant, cette méthode peut être critiquée. De façon générale, toutes les échelles dépendent presque entièrement de l'évaluation verbale des patients, que ce soit via des entretiens cliniques ou des questionnaires [55]. Ainsi, l'auto-évaluation n'est pas possible dans le cas de patients non communicants (nouveaux-nés, patients inconscients, certains patients atteints de démence, etc). De plus, la fiabilité de cette technique peut être remise en question précisément à cause de sa subjectivité [61]. En effet, la douleur est une réponse qui peut être orientée vers un objectif, tel que recevoir des traitements antalgiques, ou encore être affectée par des biais de déclarations et des problèmes de mémoire [284].

6. https ://www.sciencesetavenir.fr/sante/opiaces-la-justice-valide-la-faillite-du-laboratoire-purdue-immunite-partielle-pour-la-famille-sackler$_1$57222

Il est donc nécessaire de chercher à développer d'autres méthodes de mesures et d'évaluation objectives de la douleur afin d'incorporer les observations comportementales qui sont de forts indicateurs de la douleur et peuvent se produire sans que l'individu en ait conscience, et sont donc moins biaisés. Ces indicateurs de la douleur qui permettent de l'évaluer de façon plus objective seront développés dans les sections 1.4 et 6.

1.2.1 Douleur et Nociception

Comme l'a indiqué l'IASP dans sa définition de 2020, il est essentiel de souligner la différence entre la nociception et la douleur. La **nociception** fait référence au processus physiologique par lequel l'information sur des lésions tissulaires est transmise au système nerveux central (voir Fig. 1.3). Autrement dit, la nociception correspond aux processus neuronaux d'encodage des stimuli nociceptifs, et est donc un phénomène objectif, contrairement à la douleur [244].

Nociception

De la périphérie au cortex, on parle donc de nociception. Tout comme dans d'autres modalités sensorielles, telles que la vision ou l'audition, la nociception génère des signaux biologiques qui assurent la médiation entre les événements du monde extérieur et le milieu interne d'un organisme. Un comportement réflexe en aval peut alors être généré pour protéger l'organisme et augmenter ainsi sa capacité d'évolution. Chez la plupart des animaux, y compris l'homme, la réponse à la nociception et sa modulation peuvent se produire avant et probablement sans perception, ce qui la rend très conservée entre les espèces.

Douleur

La douleur en revanche est une **interprétation**, c'est un processus qui résulte de la perception d'informations nociceptives, que la source de ces informations soit externe (par exemple, une décharge électrique ou une brûlure) ou interne (comme un déchirement musculaire par exemple). Lorsque le cerveau traite ces informations, nous pouvons consciemment ressentir le stimulus comme douloureux et réfléchir à sa localisation, son intensité, ses qualités sensorielles et émotionnelles, ou tout autre élément caractéristique. C'est alors que des expressions comportementales complexes (faciales, corporelles, verbales) peuvent se manifester. Les différences individuelles doivent également être prises en compte.

La douleur est donc une perception liée aux expériences personnelles et subjectives, qui survient dans le cerveau. La douleur n'est pas seulement un phénomène sensoriel, mais a aussi une composante affective et émotionnelle, et cognitive : la douleur est ainsi caractérisée par sa gravité, son emplacement, sa durée, elle est influencée par l'évaluation du patient de la gravité de sa blessure, et de ses valeurs culturelles.

Les cognitions telles que les pensées (telles que « à quel point ça fait mal », « quand est-ce que ça va s'arrêter »), les croyances (« la douleur reflète forcément une lésion tissulaire », « l'exercice physique l'aggrave »), les attentes (liées aux effets

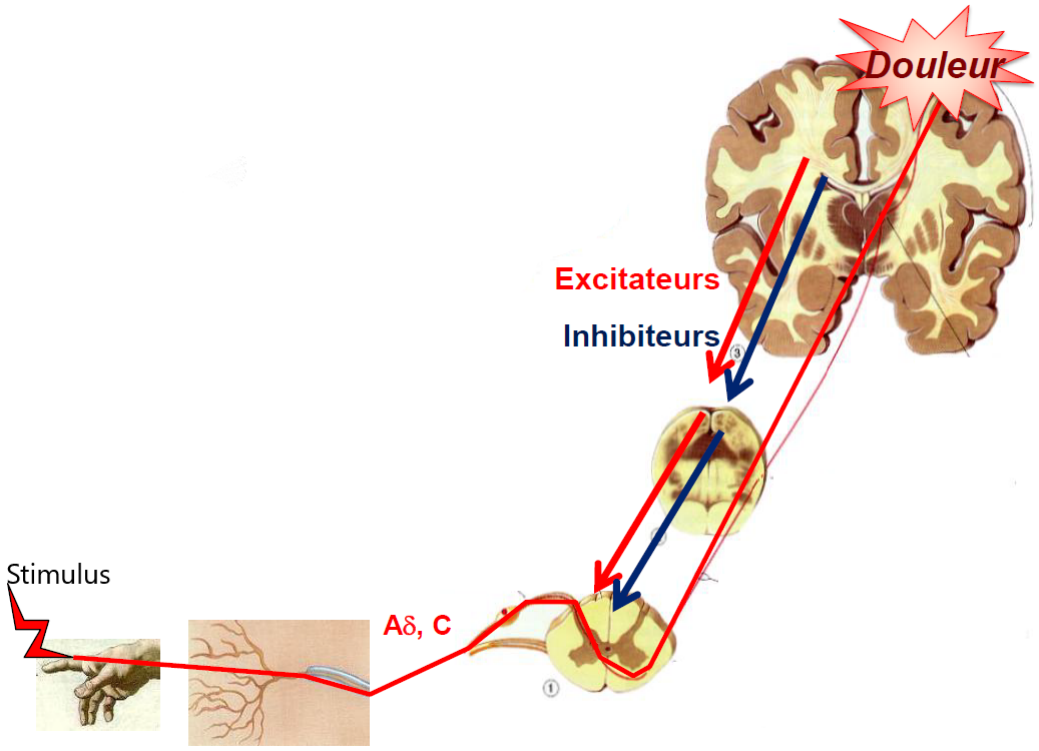


FIGURE 1.3 – De la nociception à la douleur

placebo et nocebo par exemple), ou encore les émotions (peur, joie) sont d'autres modulateurs importants de la douleur. La nociception ne conduit donc pas forcément à l'expérience de douleur, et inversement la douleur peut survenir en l'absence de nociception[244].

Réponses à la douleur

L'expérience de la douleur doit être distinguée de la cause de la douleur (comme les lésions tissulaires avec nociception), de la réponse à la douleur (via la communication verbale et non verbale) et de l'évaluation de la douleur (par exemple par un soignant) [112]. Ceci explique que pour un même stimulus nociceptif, les personnes peuvent ressentir la douleur différemment en fonction de divers facteurs socioculturels, comme la présence ou non d'un proche dans la pièce, ou le fait d'avoir été continuellement exposé à des soins de santé discriminatoires.

De multiples zones du système nerveux, du système nerveux périphérique au cortex cérébral, participent au processus de la douleur et entraînent :

- des **réponses physiologiques** : conductance cutanée, rythme cardiaque ECG, pression artérielle, dilatation de la pupille, EEG, fMRI, fNIRS
- des **réponses comportementales** : expressions faciales (FACS), mouvements corporels (réflexes de protection, torsion, frottements...), vocalisations (cris, pleurs, soupirs, grognements etc).

Ces différentes réponses peuvent être mesurées à l'aide capteurs spécifiques pour

les réponses physiologiques et de caméras pour les réponses comportementales pour être ensuite utilisée pour évaluer de façon automatique la douleur à l'aide d'algorithme d'apprentissage profond.

1.2.2 Classification

Dans sa forme la plus simple, la douleur est d'abord classée comme aiguë ou chronique. Néanmoins, la distinction entre douleur aiguë et douleur chronique n'est pas claire, bien que l'on ait traditionnellement utilisé un intervalle de temps arbitraire entre le début de la douleur : ainsi une douleur de plus de 3 mois peut être considérée comme chronique.

La **douleur aiguë** est généralement associée à des lésions tissulaires ou à la menace de lésions tissulaires, et a pour but essentiel de modifier rapidement le comportement de la personne qui la ressent afin d'éviter ou de minimiser les lésions et d'optimiser les conditions dans lesquelles la guérison peut avoir lieu, en s'arrêtant lorsque celle-ci est terminée. La gravité de la douleur aiguë varie de légère à modérée à intense. Elle est évoquée par une maladie ou une blessure spécifique, elle sert un but biologique pendant la guérison et elle est auto-limitative (afin d'éviter son aggravation). Parmi les exemples de douleur aiguë, on peut citer les coupures, les brûlures, mais aussi les douleurs liées à une maladie aiguë, comme une appendicite aiguë. En revanche, la **douleur chronique** persiste au-delà de l'évolution prévue d'une maladie aiguë, elle n'a ni but biologique ni d'évaluation claire. Elle peut avoir un impact important sur la psychologie de la personne atteinte, en plus d'avoir un effet sur leur bien-être physique. [7]

Au-delà de cette simple classification binaire, on peut définir différents types de douleur qui peuvent être liées à différents types de pathologies, et peuvent donc être classées selon différentes terminologies et ontologies [231, 241]. La figure 6.12 montre ainsi une classification selon les signes cliniques observés, la figure 6.13 montre la classification ICD-11 définie par l'Organisation Mondiale de la Santé (OMS) [8], et la figure 6.14 montre celle de l'IASP (toutes ces figures sont en Annexe).

Douleur chronique

La douleur chronique est généralement décrite en médecine humaine comme une douleur qui persiste au-delà du temps normal de guérison, ou comme une douleur persistante causée par des conditions où la guérison n'a pas eu lieu ou qui disparaît puis réapparaît. Une douleur chronique est définie par la Haute Autorité de Santé (HAS) comme un syndrome multidimensionnel exprimé par la personne qui en est atteinte.

Selon l'HAS, il y a douleur chronique, quelles que soient sa topographie et son intensité, lorsque la douleur présente plusieurs des caractéristiques suivantes :

— persistance ou récurrence, qui dure au-delà de ce qui est habituel pour la cause initiale présumée, notamment si la douleur évolue depuis plus de 3 mois

7. https ://onlinelibrary.wiley.com/doi/full/10.1111/jsap.12200

8. https ://icd.who.int/fr

— réponse insuffisante au traitement
— détérioration significative et progressive du fait de la douleur, des capacités fonctionnelles et relationnelles du patient dans ses activités de la vie journalière, au domicile comme à l'école ou au travail.

Comme indiqué dans [170], quand elle devient chronique, la douleur perd alors son rôle de signal d'alarme et devient une maladie en tant que telle, quelle que soit son origine. La douleur chronique est souvent associée à d'autres facteurs qui participe à son entretien comme [9] :

— des manifestations psychopathologiques ;
— une demande insistante par le patient de recours à des médicaments ou à des procédures médicales souvent invasives, alors qu'il déclare leur inefficacité à soulager ;
— une difficulté du patient à s'adapter à la situation.

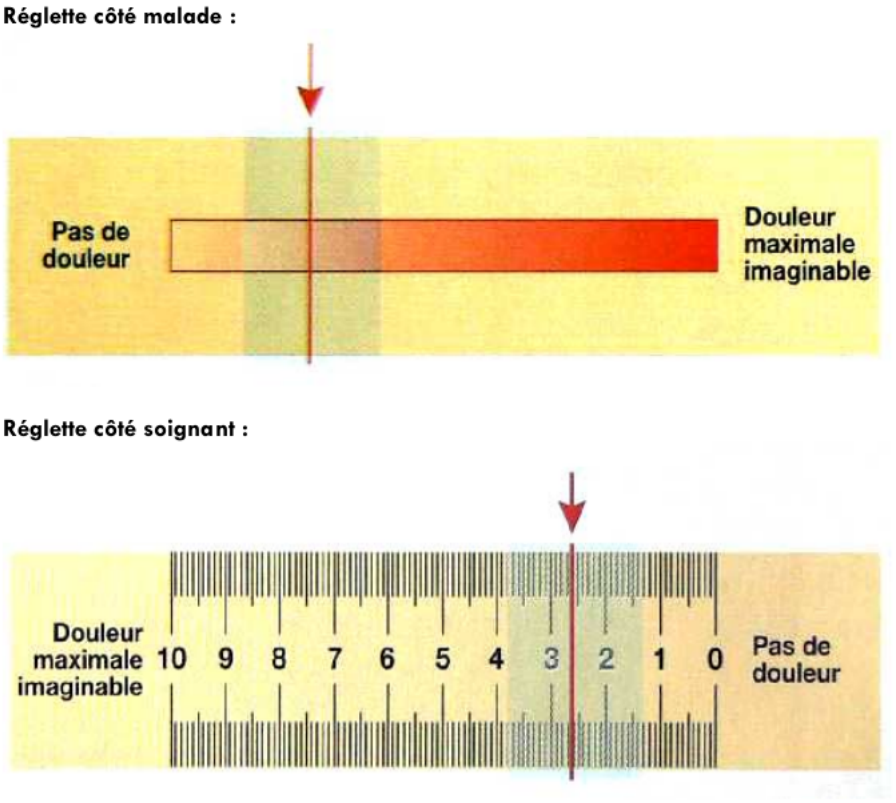
Ainsi, la douleur aiguë et la douleur chronique sont des entités cliniques différentes, et la douleur chronique peut être considérée comme un état pathologique. De plus, la douleur étant une expérience multidimensionnelle complexe impliquant des composantes à la fois sensorielles et affectives (émotionnelles), les facteurs émotionnels peuvent modifier la façon dont la douleur est perçue, et ceci est particulièrement vrai pour la douleur chronique. Les approches thérapeutiques de la prise en charge de la douleur doivent refléter ces différents profils. Le traitement de la douleur aiguë vise à traiter la cause sous-jacente et à interrompre les signaux nociceptifs à différents niveaux du système nerveux, tandis que les approches thérapeutiques de la douleur chronique doivent reposer sur une approche multidisciplinaire et une gestion holistique (global) de la qualité de vie du patient.

Dans le cadre de la détection automatique de la douleur, les bases de données académiques se sont concentrées sur la douleur aiguë, facile à reproduire et à contrôler en laboratoire (voir 6.15 dans la section 1.5).

1.2.3 Évaluation de la douleur

Comme nous l'avons vu précédemment, la douleur est un phénomène complexe, car en partie subjectif, puisqu'il s'agit d'une expérience individuelle unique. Il est donc difficile de la mesurer. Chez les patients (et les animaux), les professionnels de la santé utilisent donc un ensemble de marqueurs dont notamment les **signes comportementaux** pour faire un diagnostic et orienter la prise en charge en conséquence. L'outil de mesure de la douleur le plus utilisé aujourd'hui est l'**échelle numérique** de 0 à 10. La première version fut introduite en 1948 par Kenneth Keele, un cardiologue britannique, qui a demandé à ses patients d'évaluer leur douleur avec un score compris entre 0 (aucune douleur) et 3 (douleur sévère). Au fil des ans, cette échelle de la douleur s'est allongée jusqu'à 10 afin de s'adapter à plus de situations en offrant plus de gradations. Dans certains contextes, les patients, plutôt que de choisir un chiffre, positionnent une marque (à l'aide de visages souriants ou grimaçants, ou encore à l'aide de code couleur) sur une ligne de dix centimètres (voir figure 1.4).

9. Source : http ://www.sfetd-douleur.org/la-douleur-chronique



- **Correspondance entre les outils d'auto-évaluation et l'intensité de la douleur**

Source : Méthodologie des essais cliniques dans le domaine de la douleur. Institut UPSA de la Douleur.

Type de douleur	EVA
Douleur « légère »	1 à 3 cm
Douleur « modérée »	3 à 5 cm
Douleur « intense »	5 à 7 cm
Douleur « très intense »	> 7cm
Seuil d'intervention thérapeutique*	3 / 10

* seuil au-delà duquel la mise en route d'un traitement est indispensable, en deçà l'intervention thérapeutique reste à l'appréciation du patient et du soignant.

FIGURE 1.4 – Échelle d'évaluation de la douleur (Source : CHU de Bordeaux)

Il existe donc différentes échelles pour l'évaluation par les professionnels de santé et l'auto-évaluation de la douleur (par les patients), selon le type de patients, afin de proposer la meilleure estimation possible selon l'âge et/ou la pathologie du patient. En voici par exemple une liste non exhaustive :

- échelle numérique
- échelle verbale simple
- échelle visuelle analogique
- échelle EOC (Échelle Observation Comportementale)
- échelle ECPA (Échelle Comportementale de la douleur chez la personne âgée)
- échelle DOLOPLUS
- échelle ALGOPLUS
- échelle comportementale Behavioral Pain Scale (BPS)
- questionnaire DN4
- échelle de douleur et d'inconfort du nouveau-né (EDIN)
- échelle d'évaluation de la Douleur Aigüe du Nouveau-né (DAN)

— grille PIPP (Premature Infant Pain Profile)
— échelle Neo Facial Coding System (NFCS)
— échelle Objective Pain Scale (OPS)
— échelle Postopérative Pain Mesure for Parents (PPMP)
— échelle Évaluation Enfant Douleur (EVENDOL)
— grille HEDEN
— échelle Douleur Enfant San Salvadour

1.3 Complexité de la douleur

Biais et stéréotypes

Toutes ces échelles numériques sont loin d'être satisfaisantes. En effet, la nature auto-déclarée des scores de douleur peut conduire à remettre en question leur exactitude. Ce doute ouvre la porte aux **stéréotypes** et aux **préjugés**. De plus, l'expérience de la douleur se produit souvent dans un contexte social complexe et lorsqu'elle est communiquée à d'autres personnes, la douleur est un message éloquent qui a plusieurs fonctions, comme signaler un danger dans l'environnement ou motiver les autres à aider la personne qui souffre. Plusieurs étapes sont nécessaires pour que la douleur exprimée par la personne en souffrance soit perçue par des observateurs externes et chacune de ces étapes peut être influencée par de nombreux facteurs personnels, relationnels et contextuels.

Influences culturelles

Les réponses culturelles des patients à la douleur peuvent être divisées en deux catégories : stoïque ou émotive. Les patients stoïques expriment moins leur douleur et ont tendance à se retirer socialement. En revanche, les patients émotifs sont plus enclins à verbaliser l'expression de leur douleur, préfèrent être entourés et attendent des autres qu'ils réagissent à leur douleur afin de valider leur malaise. Les différences culturelles dans la réaction à la douleur aggravent les difficultés inhérentes à la communication. Bien que presque tout le monde ressente la douleur de la même manière, plusieurs études ont montré qu'il existe des différences importantes dans la façon dont les gens expriment leur douleur et s'attendent à ce que les autres réagissent à leur malaise [334, 7, 217, 178, 275]. Il existe également des attitudes fondées sur la culture concernant l'utilisation des analgésiques. Il est donc important de comprendre l'impact de la culture sur l'expérience de la douleur pour bien évaluer la douleur en conséquence, et assurer des soins efficaces et adaptés à la culture des patients.

Ainsi, si on prend l'exemple des États-Unis, l'édition 2014 du manuel de soins infirmiers américain « Nursing : A Concept-Based Approach to Learning » indiquait encore que les Amérindiens « peuvent choisir un chiffre sacré lorsqu'on leur demande d'évaluer la douleur » et que la validité des auto-évaluations sera probablement affectée par le fait que les Juifs « croient que la douleur doit être partagée » et les Noirs « croient que la souffrance et la douleur sont inévitables ». Depuis, l'éditeur du manuel a supprimé le passage offensant des nouvelles éditions. Mais les préjugés demeurent courants, et de nombreuses études ont montré des disparités dans le traitement de la douleur. Hoffman et al. ont publié un article indiquant que les

patients noirs sont beaucoup moins susceptibles que les patients blancs de se faire prescrire des médicaments pour le même niveau de douleur déclaré, et qu'ils reçoivent des doses plus faibles [133]. D'autres articles montrent que les femmes sont jusqu'à vingt-cinq pour cent moins susceptibles que les hommes de recevoir des traitements opiacés contre la douleur [171, 42]. En France, on parle encore de **syndrome méditerranéen** chez les populations du pourtour méditerranéen (Espagne, Portugal, Italie et Maghreb) pour désigner un comportement d'exagération des symptômes de la part d'un patient du fait de ses origines et de sa culture [259]. Il s'agit d'une généralisation en pensant que les patients expressifs sont souvent d'origine hispanique, moyen-orientale et méditerranéenne, tandis que les patients stoïques sont souvent d'origine nord-européenne et asiatique. Si nous utilisons de telles généralisations pour aider à comprendre le comportement humain, nous devons cependant toujours garder à l'esprit que si la culture est un cadre qui oriente le comportement humain, tout le monde dans chaque culture ne se conforme pas à un ensemble de comportements ou de croyances attendus. L'utilisation rigide de généralisations conduit à des stéréotypes culturels qui, à leur tour, peuvent conduire à de graves inexactitudes [78].

Les réactions des personnes à la douleur sont donc intéressantes, car elles reflètent aussi les différentes **perceptions et attentes culturelles** à l'égard des soins médicaux et des traitements en général. Parfois les piqûres sont considérées comme plus puissant qu'une pilule, un gros comprimé sera plus efficace qu'un petit, un médicament à mauvais goût sera plus actif qu'un médicament qui a bon goût, selon le contexte socio-culturel du patient. Dans certaines cultures, les personnes ont tendance à penser que plus une procédure est invasive, plus elle est efficace. Dans les pays où les injections sont courantes, un analgésique par voie intraveineuse sera préféré à des comprimés analgésiques narcotiques, même si ces derniers sont efficaces, et un patient pourrait croire que sans injection, le traitement est inadéquat. D'autres peuvent rejeter complètement les analgésiques par crainte des effets nocifs et des risques de dépendance. De plus, des raisons culturelles et/ou religieuses peuvent aussi empêcher une personne de demander des analgésiques. Il est donc nécessaire pour les médecins et les infirmières d'anticiper les besoins d'un patient en matière de douleur et d'entamer des discussions pour expliquer les raisons de l'utilisation d'analgésiques afin de déterminer le traitement qui leur le plus adapté.

Douleur et émotions

La douleur est traditionnellement étudiée et traitée séparément des émotions. Mais si la plupart du temps la douleur est considérée comme purement physique, la définition de l'IASP indique bien que la douleur est en fait une expérience subjective avec une **composante sensorielle et émotionnelle**. En outre, des travaux ont montré qu'il existe des expressions faciales spécifiques de la douleur et qu'elles peuvent être distinguées des expressions des émotions de base. Cependant, comme nous pouvons le voir dans la (figure 1.5, la douleur est souvent associée à des émotions négatives (dégoût, peur, colère, tristesse) et/ou peut être surprenante, ce qui peut conduire à des expressions modifiées ou mélangées, on parle d'émotions composées. De plus, il existe une grande variabilité entre les individus d'une même population quant à leur expressivité faciale. Chaque personne commencent à réagir à des intensités de douleur différentes. Une réactivité accrue à la douleur a été observée chez les

personnes atteintes de démence et les personnes âgées souffrant de troubles cognitifs, qui ont une capacité limitée ou nulle à signaler la douleur. L'expression de la douleur peut également être simulée ou exagérée, mais les expressions faciales simulées sont contrôlées par d'autres voies corticales, diffèrent des expressions authentiques et peuvent être distinguées, dans certains cas de manière plus fiable, par la vision par ordinateur que par les humains [60].

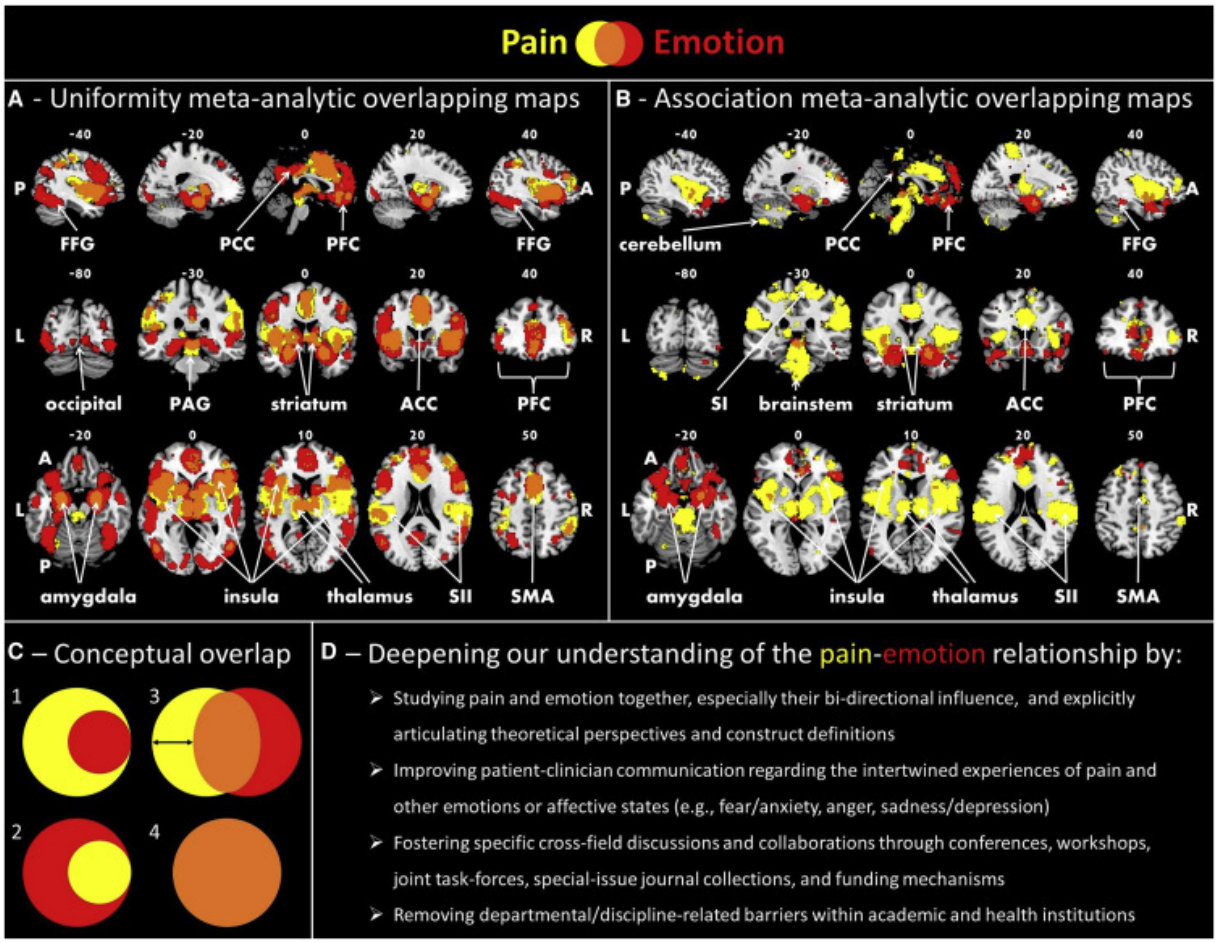


FIGURE 1.5 – Visualisation de la relation entre douleur et émotion [112]

Douleur et personnalité

Bien qu'il n'existe pas de personnalité unique sujette à la douleur, les attributs de la personnalité tels que l'introversion ou l'extraversion, l'optimisme et le pessimisme, et des troubles de la personnalité affectent la capacité des patients à faire face à la douleur. Les patients souffrant de troubles de la personnalité ont tendance à réagir de manière excessive au d'émotions négatives comme le stress, ou encore à la maladie, selon le schéma caractéristique de leur trouble de la personnalité. Les caractéristiques du type de personnalité sont accentuées par la condition médicale, comme la douleur chronique. Par conséquent, ces caractéristiques provoquent des problèmes prévisibles de gestion du comportement dans le contexte de la douleur [57, 19].

1.4 Marqueurs physiologiques de la douleur

Il est donc maintenant bien établi que la douleur est un phénomène multidimensionnel, faisant intervenir un grand de variables comportementales, psychologiques et sociales. La douleur peut être exprimée d'une multitude de façons, en fonction de la culture de la personne qui l'exprime, de sa personnalité et de son état émotionnel. Le contexte social et environnemental immédiat a une grande influence sur la perception de la douleur, tout comme l'expérience passée et la culture. Par conséquent, une cause standard de douleur (comme une intervention chirurgicale par exemple) peut générer d'énormes différences individuelles dans la perception de la douleur. L'expérience de la douleur chez tout individu se manifestera donc par des réponses émotionnelles et comportementales influencées par sa culture, son histoire personnelle et ses perceptions uniques, entraînant une grande variabilité des réponses. Bien qu'utiles, les échelles de mesures basées sur l'auto-évaluation de la douleur présentée précédemment ne sont donc pas totalement satisfaisantes, car biaisées. Les praticiens manquent de moyens objectifs, systématiques et efficaces pour mesurer la douleur. Or, comme nous pouvons le voir dans le tableau 1.1 il existe un certain nombre de **marqueurs physiologiques** de la douleur, observables et quantifiables.

Vidéos et images RGB	visage
	tête
	corps
	voix
Imagerie médicale	IRM
	TEP
	imagerie spectroscopique proche infrarouge (fNIRS)
Signaux physiologiques	EMG
	ECG
	EEG
	EDA
	pupillométrie
Marqueurs biologiques	inflammatory markers
	immuno
	hormones
Omiques	génomique
	protéomique
	métabolomique
	transcriptomique
Psychologie	personnalité
	stress
	émotions
Dossiers de santé électroniques (EHR)	antécédents médicaux, chirurgicaux, familiaux
	âge, genre, poids, situation professionnelle
	tests et questionnaires

Tableau 1.1 – Marqueurs de la douleur

Un marqueur est défini comme un élément « utilisé pour détecter ou confirmer la présence d'une maladie ou d'une condition d'intérêt ou pour identifier les indivi-

dus présentant un sous-type de la maladie »[121]. Si la douleur est une expérience sensorielle subjective le plus souvent déclarée par le patient, il existe néanmoins un ensemble de marqueurs susceptibles de produire des indicateurs objectifs liés à la douleur, notamment lorsqu'ils sont utilisés en combinaison (voir figure 6.11), car il est évident que la complexité de la douleur ne peut être saisie par un seul marqueur. L'utilisation d'outils analytiques avancés de science des données, telles que les algorithmes d'apprentissage automatique et les réseaux neuronaux, pour combiner plusieurs marqueurs objectifs en marqueurs composites de la douleur, a beaucoup plus de chances de réussir [290, 318].

La réponse comportementale à la douleur remplit différentes fonctions. Pour protéger son propre corps, la douleur attire l'attention, interrompt le comportement associé et incite à agir pour l'atténuer, comme le retrait réflexe de la main d'une surface chaude. Certains comportements servent à en montrer notre besoin d'aide à nos alliés (c'est-à-dire les personnes susceptibles de nous aider), d'autres au contraire servent à dissimuler notre faiblesse/vulnérabilité, un comportement qui s'est probablement développé, car il permettait d'augmenter les chances de survie et de reproduction. Les réponses comportementales à la douleur comprennent les expressions faciales, les mouvements du corps et les vocalisations. La douleur chronique entraîne souvent des changements permanents dans le comportement quotidien et les interactions sociales.

1.5 Détection automatique de la douleur

L'un des objectifs de l'**informatique affective**, telle que définie par Rosalind Picard [223], est l'étude et le développement d'algorithmes permettant de reconnaître les états internes de l'être humain tels que les émotions, le stress, la dépression et la douleur, à l'aide de différents signaux (imagerie médicale, interactions sociales, expressions faciales). L'informatique affective s'est beaucoup développé depuis les années 2000 et a de nombreux domaines d'application, tels que l'interaction homme-machine ou la médecine[250, 123, 102]. Les progrès récents de la vision par ordinateur et de l'apprentissage automatique pour l'analyse et la modélisation automatiques du comportement humain pourraient jouer un rôle essentiel pour surmonter certaines limitations dans le contexte clinique. Les méthodes actuelles de dépistage et d'évaluation de la douleur dépendent presque entièrement des symptômes rapportés verbalement par les patients lors d'entretiens cliniques.

Parmi l'ensemble des marqueurs physiologiques de la douleur développé précédemment, tous ne sont pas facilement observables et mesurables ou nécessitent un équipement conséquent. Plusieurs études ont donc étudié la faisabilité d'une évaluation automatique de la douleur à partir de l'analyse des expressions faciales, des vocalisations, de la posture et des changements de paramètres physiologiques [318]. Par conséquent, l'évaluation automatique et objective des troubles de la douleur à partir de signaux comportementaux présente un intérêt croissant pour les cliniciens et les informaticiens.

1.5.1 Reconnaissance faciale

Le visage est l'un des canaux les plus puissants de communication non verbale. Plusieurs projets de recherche en cours explorent l'utilisation du comportement facial comme indicateur des états internes du patient, du comportement social, de la biométrie et de la psychopathologie. Le visage transmet des informations sur l'âge, le sexe, les antécédents et l'identité d'une personne, ce qu'elle ressent ou pense. L'expression faciale fournit des indices sur les émotions [74, 5], l'intention, la vigilance, la douleur, la personnalité, régule le comportement interpersonnel et communique le statut psychiatrique et biomédical, entre autres fonctions. Le visage a donc suscité un vif intérêt chez les spécialistes du comportement. Ainsi, il existe plusieurs études sur la détection automatique de la douleur [260, 285], dont certaines chez les enfants [245] et les personnes atteintes de démence [194].

La détection de la douleur est également fondamentale pour la protection de la santé des animaux. Un système de détection automatique permettrait d'améliorer la façon dont nous surveillons et soignons nos animaux. L'évaluation des niveaux de douleur chez les animaux est un processus crucial, mais chronophage. Si des méthodes "générales" sont utilisées quotidiennement en milieu vétérinaire, chaque espèce exprime sa douleur différemment. Afin de compléter les traditionnelles **échelles d'observation** (comportementales, physiologiques) de la douleur animale, des efforts récents ont été faits vers la mise en place d'échelles se basant sur les expressions du visage. Appelées *Grimace Scales*, elles s'appuient sur des distances entre différents points du visage de l'animal pour détecter des expressions douloureuses. Inspirés par le système des FACS utilisé pour caractériser les expressions faciales chez l'humain, ces échelles ont été développées pour de nombreuses espèces : souris [85, 63], rats [278], moutons [122, 192], lapins [155], bétail [238], porcs [298], chats [134], cheval [8, 64]. Le projet « EquineML », mené par deux équipes de recherche suédoises, explore actuellement ce même type d'application chez le cheval [114]. En parallèle, certains travaux essayent d'automatiser la pose des marqueurs faciaux chez le mouton [233, 129]. Chez les souris, un modèle a été proposé dans l'objectif de détecter automatiquement la douleur chez une souris en observant ses expressions faciales [85] : pour la tâche de la détection binaire de la douleur (douleur/non-douleur), les auteurs affichent 94% de réussite. Il est aussi intéressant de noter que le système FACS a été développé pour d'autres espèces, sans que le lien à la douleur soit étudié (par exemple l'orang-outang [32]).

Deep Pain

Un exemple de reconnaissance faciale de la douleur par Deep learning est celui de Deep Pain [240]. Ce modèle est basé sur la base de données UNBC-McMaster constituée de 200 séquences vidéo de 25 patients souffrant de douleurs à l'épaule. Les images ont été étiquetées à l'aide de la métrique de Prkachin et Solomon (PSPI) basée sur le système de codage des FACS, qui code différents mouvements des muscles du visage avec différents niveaux d'intensité (voir figures 1.7 et 1.6).

Le réseau CNN permet donc d'extraire des caractéristiques indépendamment de chaque image, puis de regrouper leurs prédictions sur l'ensemble de la vidéo. Cependant, ce type d'architecture ignore la dimension temporelle. Il faut donc ajouter une

FIGURE 1.6 – pipeline de prétraitement des images de Deep Pain

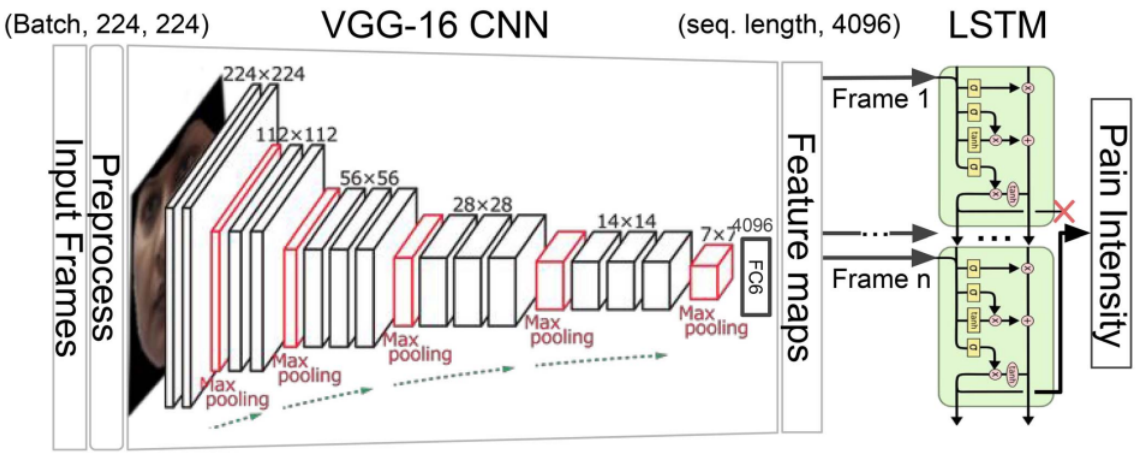


FIGURE 1.7 – Deep Pain

couche récurrente (LSTM) pour encoder et capturer l'ordre temporel de la séquence. L'architecture est de type end to end avec une entrée d'images RGB. La prédiction finale au niveau vidéo est la moyenne des prédictions de chaque clip. L'inconvénient de l'utilisation des LSTMs ici est qu'ils utilisent les caractéristiques des dernières couches issues du CNN, et ne peuvent donc pas être en mesure de capturer les mouvements fins à bas niveau. CNN et LSTM seront détaillés dans la section 3.

1.5.2 Reconnaissance posturale

L'étude de la **posture douloureuse** est un sujet moins exploré que pour la reconnaissance faciale. La plupart des études se sont concentrées sur la détection faciale de la douleur, alors que peu de recherches ont été menées sur la détection à partir de la posture [318]. Les travaux de Beatrice de Gelder [68] et de Nadia Bianchi-Berthouze [161] ont montré l'intérêt d'étudier les expressions corporelles des émotions. Bien que moins étudié que l'expression faciale, il existe néanmoins plusieurs études sur l'expression des émotions par le corps [161, 82, 69] et des études se focalisant sur un type de pathologie ou un type de douleur [191]. Walsh et al. ont montré que la douleur peut s'exprimer par la posture [301]. Boyi Hu et al. ont utilisé le Deep Learning pour reconnaître les douleurs lombaires [26]. Les travaux de Philipp Werner [221, 219, 220] se sont limités aux mouvements de la tête et au haut du corps. Dans cette thèse, nous nous intéresserons au corps entier.

1.5.3 Bases de données académiques

Il existe plusieurs bases de données sur la douleur. Werner et al. ont résumé ces bases de données dans le tableau 6.15 en annexe [318]. Ces données sont majoritairement des images provenant de vidéos, avec des images centrées sur le visage et parfois d'autres marqueurs (comme les vocalisations ou des marqueurs physiologiques tels que l'ECG). Ces bases de données utilisent principalement le système FACS ou ses dérivés comme l'échelle de douleur *PSPSI* (Prkachin and Solomon Pain Intensity) [228] pour coder l'intensité de la douleur. La grande diversité des expressions liées à la douleur à l'intérieur de chaque modalité suggère que la classification de l'intensité de la douleur devrait être abordée comme un problème de reconnaissance multimodale.

Une approche combinant 3 modalités : audio, vidéo et physiologies a donc été proposé [216]. Au lieu d'utiliser les informations fournies par une modalité unique, une approche de fusion bien conçue devrait être capable de combiner de façon appropriée des informations complémentaires provenant de plusieurs sources afin d'améliorer à la fois la robustesse d'un système de classification et sa performance. La base de donnée *SenseEmotion* utilisé par cette équipe a ainsi été constitué à partir de 45 individus sains. La douleur a été provoquée artificiellement (par élévation locale de la température). Plusieurs enregistrements ont ainsi été effectués :

— rythme respiratoire,
— électrocardiogramme,
— électromyogramme,
— activité électrodermique
— vidéos, pour extraire des points de repères ainsi que la posture de la tête,
— sons.

Chaque modalité est caractérisée par des propriétés spécifiques qui fournissent des renseignements sur la douleur induite artificiellement. Un système de classification amélioré en combinant de manière appropriée les informations fournies par plusieurs modalités a été créé. Les systèmes à classificateurs multiples permettent de profiter de la diversité ainsi que de la complémentarité des informations extraites. En plus, les systèmes de classification à modalité unique sont moins robustes en raison de leur dépendance à l'égard d'une modalité unique, en particulier en cas de données manquantes. Pour faire la fusion des modalités, plusieurs architectures de Forets Aléatoires ont été évalués. Cette approche permet d'atteindre des taux de classification de respectivement 83,39 %, 59,53 % et 43,89 % pour des tâches de classification de l'intensité de la douleur à 2 classes, 3 classes et 4 classes. Pour améliorer ce système, des données plus réalistes sont nécessaires. La fusion des modalités par deep learning est une piste qui reste encore à explorer. Ces travaux constituent donc un premier pas vers la mise en place d'échelles douleur spécifiques.

En ce qui concerne la posture douloureuse, il n'existe que peu de base de données disponibles : parmi elles, *BioVid* contient des données centrées uniquement sur la tête et le haut du corps, seule EmoPain contient des données sur l'ensemble du corps. En effet, EmoPain [201] (*The Emotion and Pain Project*[10]) est un projet ayant pour objectif de concevoir et de développer un système intelligent qui permet-

10. http ://www.emo-pain.ac.uk/

tra une surveillance et une évaluation omniprésentes de l'humeur et des mouvements des patients liés à la douleur à l'intérieur. Par rapport aux expressions faciales, les mouvements posturaux liés à la douleur ne sont pas aussi bien établis. Par conséquent, un codage a été déterminé au moyen d'un processus itératif par quatre kinésithérapeutes et des psychologues. Les experts ont été utilisés pour l'étiquetage du comportement du corps compte tenu de la difficulté de cette tâche et de la connaissance de l'environnement que cela exige. Cette base de données est étudiée en détails dans la section 6.

Problématique et objectifs

La douleur est également un problème aux proportions épidémiques, avec des conséquences physiologiques et psychologiques graves, et a également un impact important sur la qualité de vie. Ainsi, les patients dont la douleur n'est pas correctement prise en charge peuvent souffrir d'anxiété, de peur, de colère, de dépression ou de dysfonctionnement cognitif. Pour bien traiter la douleur, il faut donc la comprendre, l'évaluer, pour pouvoir ensuite proposer l'accompagnement thérapeutique adéquat, ce qui est l'objectif de l'entreprise Lucine avec des thérapies numériques.

L'auto-évaluation n'est pas suffisante dans un certain nombre de cas et n'est pas entièrement fiable. Il est donc nécessaire de développer d'autres méthodologies pour mesurer la douleur de façon objective à l'aide de marqueurs tels des mouvements particuliers et des modifications de la posture. Ces modifications peuvent être capturées à l'aide de caméras pour être ensuite analysées de façon automatique par des algorithmes d'apprentissage automatique profond.

L'objectif de cette thèse est donc d'examiner l'expression de la douleur par la posture et ses variations, et de construire un modèle d'apprentissage automatique capable de détecter cette expression de la douleur.

Chapitre 2

Analyse de mouvements

Sommaire

L'évolution de la **vision par ordinateur** a permis de développer des systèmes automatiques efficaces pour le traitement des données posturales et l'analyse des mouvements et du comportement des personnes. En étudiant les mouvements du corps humain, il est possible de reconnaître les actions et les gestes, notamment ceux utilisés par les gens pour communiquer des informations de manière non verbale. Ces tâches sont essentielles dans de nombreuses applications de vision par ordinateur, comme la reconnaissance d'événements particuliers et la vidéosurveillance. L'analyse des mouvements peut également être utilisée dans des dispositifs de surveillance et de monitoring médical.

La **reconnaissance d'actions** et la **reconnaissance de gestes** sont des domaines clés pour la compréhension du comportement humain. Si ces deux termes sont souvent confondus, au sens strict, on peut distinguer : la reconnaissance d'actions qui est axée sur la reconnaissance d'actions humaines génériques (telles que « marcher », « manger », « répondre au téléphone ») ou sportives (« course », « passe », « saut ») effectuées par une personne, la reconnaissance d'activités qui est la reconnaissance d'une séquence complexe d'actions réalisées par plusieurs humains pouvant interagir les uns avec les autres de manière conjointe, et enfin la reconnaissance des gestes qui est axée sur la reconnaissance de mouvements corporels effectués par un utilisateur avec une signification particulière dans un certain contexte (« viens », « bonjour »). Bien que toutes ces tâches aient des applications différentes, elles sont reliées, car toutes deux basés sur l'analyse de la posture et du mouvement du corps, par exemple à travers des séquences vidéos, [59] et peuvent être regroupées ensemble sous le terme générique d'**analyse de mouvements** ou plus simplement sous le terme de reconnaissance d'actions *Human Action Recognition, HAR* [137].

2.1 Squelettisation

La nature des données d'entrée est une question importante. Les données **RGBD** pour l'analyse du mouvement humain comprennent trois modalités : les couleurs RGB (*Red Green Blue*), et la profondeur (*Depth*) et le squelette. La principale caractéristique des données RGB est leur forme, leur couleur et leur texture qui apporte les avantages de l'extraction des points intéressants et du flux optique [224]. Par rapport aux données RGB, la profondeur est insensible aux variations d'illumination, invariante aux changements de couleur et de texture, fiable pour estimer la silhouette du corps et le squelette, et fournit de riches informations structurelles en 3D sur la scène. Enfin, le squelette contient les positions des articulations humaines et est une caractéristique de haut niveau pour la reconnaissance de mouvement.

Il existe plusieurs approches :
- *top down* où l'on cherche d'abord à détecter un humain à l'aide d'une *bounding box*, puis un algorithme d'estimation du squelette est utilisé pour localiser les points clés (les articulations).
- À l'inverse, dans l'approche *bottom up*, on localise d'abord les articulations, puis on les assemble en instances humaines.

Pour construire un squelette, il faut d'abord localiser dans l'espace les points

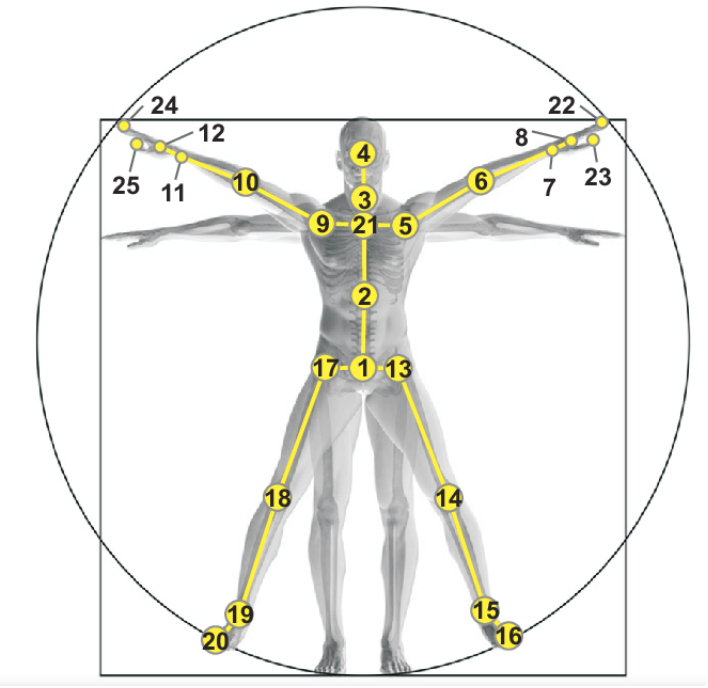


FIGURE 2.1 – Exemple de squelettisation [261]

d'intérêts correspondants majoritairement aux articulations du corps humain. Il faut donc d'abord extraire des données de l'environnement en utilisant différents capteurs [59]. Ces données sont ensuite traitées pour en extraire des informations utiles selon l'objectif, c'est-à-dire ici la reconnaissance des gestes. Un système d'apprentissage pourra ensuite être utilisé pour pouvoir classer de nouvelles données.

Un système de reconnaissance de gestes est donc caractérisé par plusieurs paramètres :

— le capteur utilisé : les caméras 2D, les caméras 3D, les systèmes de stéréovision, les marqueurs optiques, les capteurs inertiels,
— le choix des caractéristiques : les points d'intérêts 2D ou 3D, les angles d'Euler ou les quaternions de ces points, les trajectoires, les silhouettes
— le choix du classifieur : les sacs de mots, les SVM, les k-Plus Proches Voisins (*k-Nearest Neighbors, k-NN*), les CNN, les DTW (déformation temporelle dynamique, ou *Dynamic Time Warping*) ou les modèles de Markov cachés, ou *Hidden Markov Model, HMM*).

La squelettisation n'est pas la seule méthode de modélisation du corps. Comme expliqué dans [59], Bourdev et Malik ont introduit des parties de postures qui ont une faible variabilité entre elles appelées *poselets* [25]. La détection de ces poselets permet de retrouver la posture d'une personne dans une image. Dantone et al. ont proposé une approche utilisant deux couches successives de forêts d'arbres aléatoires pour trouver la localisation des articulations d'une personne dans une image [67]. Enfin, Renna et al. ont développé un filtrage particulaire pour définir un modèle 3D représentant le haut du corps [236].

2.1.1 Localisation de points d'intérêts

À partir d'une image RGB, on peut estimer les **coordonnées** dans l'espace des points de repères (2D ou 3D). Les recherches sur la reconnaissance de la posture basée sur la squelettisation se sont beaucoup développées ces dernières années. En tant que représentations de haut niveau, les squelettes sont à la fois robustes aux apparences, aux environnements et aux variations de vue de caméra. En effet, lorsqu'on s'intéresse à la description des mouvements du corps humain, une solution classique est de suivre le mouvement du squelette de la personne filmée.

La squelettisation peut également être obtenu directement par des capteurs de profondeur comme les caméras Kinect ou Realsense d'Intel, ou par des systèmes de capture optique de mouvements (mocap) comme Vicon. Des algorithmes d'estimation de pose comme OpenPose [33] peuvent également être utilisés pour la détection automatique de points de repère.

(a) Caméra Kinect (b) Caméra Vicon

FIGURE 2.2 – Exemple de caméras RGBD

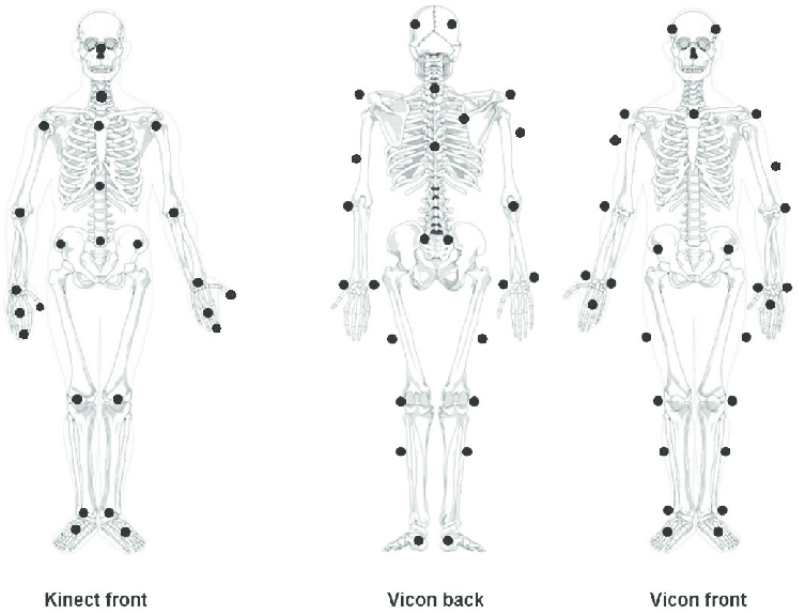


FIGURE 2.3 – Configuration des points d'intérêts de Kinect et Vicon [213]

2.1.2 Caméras de profondeur

Les caméras de profondeur donnent des informations de distance entre la scène filmée et la caméra [59]. L'image obtenue est appelée carte de profondeur (*depth map*). Ces caméras ont plusieurs avantages : on peut directement extraire des cartes de profondeur la géométrie de la scène, et ces caméras sont moins sensibles aux changements de luminosité que les caméras 2D [150, 97].

Il existe différentes techniques d'acquisition d'images en 3D (voir Fig. 2.4) [111]. Ici, nous nous intéresserons uniquement aux techniques d'acquisition les plus utilisées en reconnaissance d'actions actuellement : les caméras basées sur une technologie de temps-de-vol *Time-of-Flight* et les caméras projetant des lumières structurées *Strucured-Light* [249]. Ces méthodes ont été utilisées pour obtenir les données présentées dans les sections 4 et 6.2.

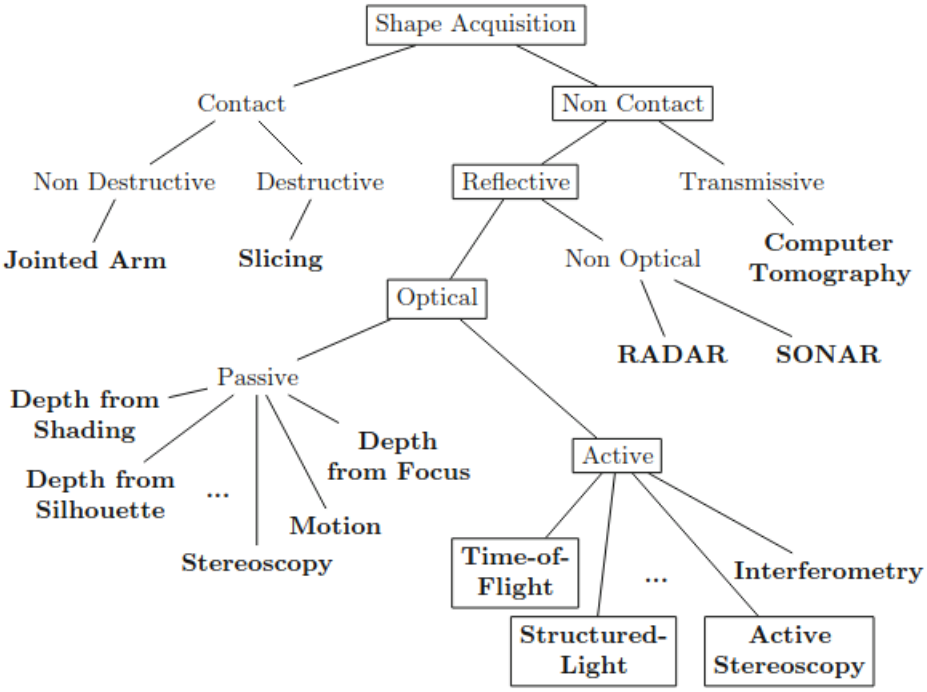


FIGURE 2.4 – Taxonomie des différentes techniques d'acquisition de données 3D [111]

La méthode de la lumière structurée est une technique de stéréovision active. Les Caméras à lumière structurée projettent un motif laser sur la cible et estiment la profondeur par triangulation [333]. Une séquence de motifs connus est projetée sur un objet, qui est déformé par la forme géométrique de l'objet. Cet objet est ensuite observé par une caméra depuis une direction différente. En analysant la déformation du motif observé, c'est-à-dire la différence par rapport au motif projeté d'origine, on peut extraire des informations sur la profondeur. C'est le cas de la Microsoft Kinect 1 sortie en 2010.

D'autres caméras fonctionnent sur le principe du temps-de-vol (*Time of Flight*, TOF) [175]. C'est le cas de la Kinect 2 [205] sortie en 2012. La technologie TOF est

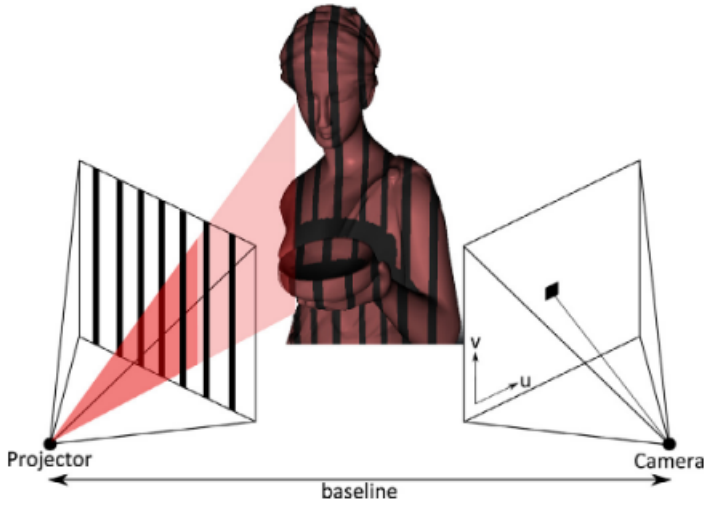


FIGURE 2.5 – Caméra à lumière structurée [249]

basée sur la mesure du temps que la lumière émise par une source pour atteindre un objet et revenir vers le réseau de capteurs. Généralement, la lumière provient d'un laser ou d'une LED fonctionnant dans le domaine proche infrarouge (850 nm), donc invisible pour l'œil humain. Un capteur d'image reçoit la lumière et convertit l'énergie photonique en courant électrique. En mesurant le retard de phase de la lumière réfléchie sur l'objet filmé, on peut calculer la distance parcourue par la lumière. Dans les capteurs TOF, la distance est mesurée pour chaque pixel, ce qui donne une carte de profondeur, c'est-à-dire une matrice de points en 3D où chaque point est un voxel. Une carte de profondeur peut être rendue dans un espace tridimensionnel sous la forme d'un nuage de points. Ces points 3D peuvent être connectés pour former un maillage.

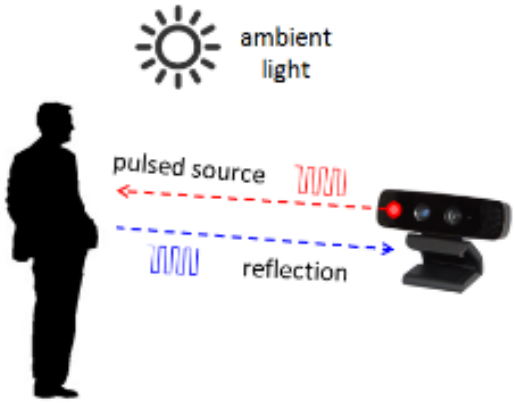


FIGURE 2.6 – Principe de la caméra TOF [175]

Grâce à l'arrivée des caméras RGB-Depth (RGBD) les recherches sur la squelettisation humaine se sont multipliées. Une des squelettisations les plus connues est celle fournie par le (*Software Development Kit, SDK* (kit de développement logiciel) de la Kinect, un capteur RGBD commercialisé par Microsoft (voir Fig. 2.2). Grace au développement de ces capteurs de profondeurs, la reconnaissance de mouvement

s'est beaucoup développée, car cette dimension supplémentaire (la profondeur) est insensible aux changements d'éclairage et inclut de nouvelles informations 3D de la scène. De plus, les positions 3D de points de repères du corps (notamment les articulations) peuvent être estimées à partir de cartes de profondeur. De nombreuses méthodes basées sur des données RGBD ont donc été proposées pour l'analyse du mouvement humain, comme développé dans [92]. On peut ainsi citer Plagemann et al. qui ont proposé une méthode de squelettisation basée sur des images issues de caméras de profondeur utilisant la technologie du temps de vol [226].

2.1.3 Captures de mouvements IMU

Dans bon nombre de travaux, les activités et les comportements sont analysés à partir de données fournies par des capteurs corporels comme les *Inertial Measurement Units, IMUs* (unité de mesure inertielle). Un IMU est un dispositif électronique qui peut intégrer plusieurs capteurs comme un accéléromètre, un gyroscope, un magnétomètre, un capteur de température ou encore un capteur de pression, ce qui permet de récupérer plusieurs types de données comme l'ensemble des coordonnées de points d'intérêts et différents types de caractéristiques tel que l'accélération ou la rotation avec les angles d'Euler ou les quaternions [330, 159]. Les IMU présentent plusieurs avantages. Ils sont fixés directement sur le corps de l'utilisateur, ils obtiennent donc des données relativement précises. De plus, ils ne sont pas affectés par les conditions d'éclairage. Enfin, ils peuvent être couplés avec d'autres capteurs comme des électrocardiogrammes (ECG) ou des électromyogrammes (EMG) [201] ce qui permet de rajouter d'autres types d'informations aux modèles. Ils ont été utilisés pour la base de données EmoPain présentée en section 1.5.3.

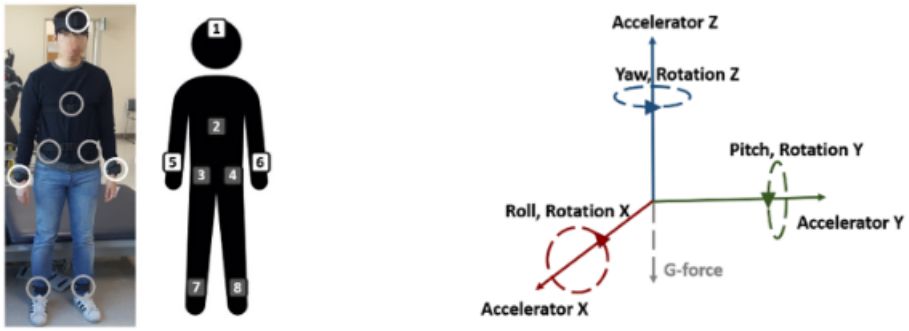


FIGURE 2.7 – Exemple de capteurs IMU [159]

2.1.4 Autres techniques

Marqueurs optiques

Certaines technologies comme celle des caméras Vicon [1] utilisent des marqueurs optiques pour faire de la capture de mouvements. Le concept général de la capture de mouvements (*motion capture* ou mocap) est de filmer un mouvement réalisé par une personne (ou un animal) réelle et de le numériser pour le transformer en animation 3D. L'acteur filmé pendant la capture de mouvement est recouvert d'une

1. https ://www.vicon.com/

série de marqueurs. Le mouvement de ces marqueurs est analysé afin d'extraire leur position dans l'espace. En sachant à quels points d'intérêts ou articulations d'un squelette d'animation est associé chacun des marqueurs, il est possible d'animer le squelette en fonction du mouvement de ces derniers. On génère ainsi l'animation 3D. Cette méthode est utilisée notamment en animation, dans les jeux vidéos ou encore le cinéma. Le matériel de capture est habituellement composé d'un élément de référence (caméra, émetteur, etc.) placé à une position fixe dans l'environnement de capture et de marqueurs, qui eux sont placés sur l'acteur étant le sujet de la capture de mouvement. Ces marqueurs servent à définir le mouvement.

Il existe deux catégories principales de marqueurs :
— **les marqueurs passifs** (le marqueur ne peut contribuer de façon active à la détection de sa position.) : un marqueur optique passif réfléchit la lumière et une caméra détecte la position des marqueurs en les filmant.
— **les marqueurs actifs** (le marqueur lui-même a un dispositif interne permettant de faciliter la détection de sa position.) : par exemple, les marqueurs optiques actifs émettent de la lumière infrarouge et la caméra filme dans l'infrarouge, ce qui produit des images où seulement les marqueurs sont visibles.

Enfin, il existe des technologies basées sur d'autres systèmes : marqueurs inertiels (accéléromètres), mécaniques (comme un exosquelette qui permet de connaître directement les paramètres du squelette d'animation) ou magnétiques (la variation du champ magnétique autour des marqueurs est analysé pour avoir leur position dans l'espace).

FIGURE 2.8 – Exemple de capture de mouvements par marqueurs optiques passifs Vicon

2.2 Classification automatique d'actions

D'après [22], une vidéo est définie comme "un ensemble d'images successives qui nous permettent de visualiser une scène dynamique". L'analyse vidéo est utilisée dans différentes applications. Ainsi, on peut classer les analyses vidéos en différentes classes selon leur objectif : la reconnaissance d'actions humaines, la surveillance, qu'elle soit publique (détection d'agressions, analyse de mouvements de foule...) ou privée (protection des personnes âgées, surveillance d'animaux, etc) [22].

L'approche traditionnelle consiste à utiliser des descripteurs de caractéristiques (SIFT, SURF, BRIEF, etc.) pour la détection d'objets. Ces caractéristiques sont les éléments intéressants, c'est-à-dire descriptifs ou informatifs contenus dans les images. Plusieurs algorithmes de vision par ordinateur, tels que la détection des bords, la détection des coins ou la segmentation des seuils, peuvent être impliqués dans cette étape. Le plus grand nombre possible de caractéristiques sont extraites des images et ces caractéristiques forment une définition (connue sous le nom de *bag-of-words*) de chaque classe d'objet. Ces définitions sont recherchées dans d'autres images. Si un nombre important d'éléments d'une liste de mots se trouve dans une autre image, l'image est classée comme contenant cet objet spécifique (par exemple, une chaise, un cheval, etc.). La difficulté de cette approche traditionnelle est qu'il faut choisir les éléments qui sont importants dans chaque image donnée. Le nombre de classes à classer augmentant, l'extraction des caractéristiques devient de plus en plus lourde. En outre, chaque définition de caractéristique nécessite de traiter une grande quantité de paramètres, qui doivent tous être affinés manuellement.

La classification d'actions est étudiée depuis longtemps dans les domaines de la vision par ordinateur. Depuis les premiers travaux il y a trois décennies [164, 325], les chercheurs ont fait d'importants progrès dans ce domaine dans différents domaines d'applications, comme la télésurveillance, la reconnaissance de la langue des signes ou encore la reconnaissance d'actions sportives [66]. Il existe de nombreuses méthodes de classification automatique d'actions basée sur des données de type squelette. Nous allons détailler quelques méthodes pour la classification automatique d'actions basée sur le squelette.

FIGURE 2.9 – Exemple de pipeline pour la reconnaissance d'actions à partir de squelettes [59]

2.2.1 Machine à vecteurs de support

Une machine à vecteurs de support (*Support Vector Machine*, SVM) est un algorithme d'apprentissage supervisé [58] utilisé pour la reconnaissance d'action [254, 27, 308] sur des points d'intérêt 3D. L'objectif d'un SVM est de trouver un hyperplan parmi plusieurs possibilités dans un espace à N dimensions (N étant le nombre de caractéristiques) pour classer distinctement les points de données (voir Fig. 2.10).

Comme expliqué dans [22], pour créer une séparation dans un espace de caractéristiques entre deux ensembles de données, on détermine un **hyperplan** optimal séparant ces deux classes dans l'espace des caractéristiques. L'objectif est alors de classer correctement les données en se trouvant le plus loin possible de l'hyperplan.

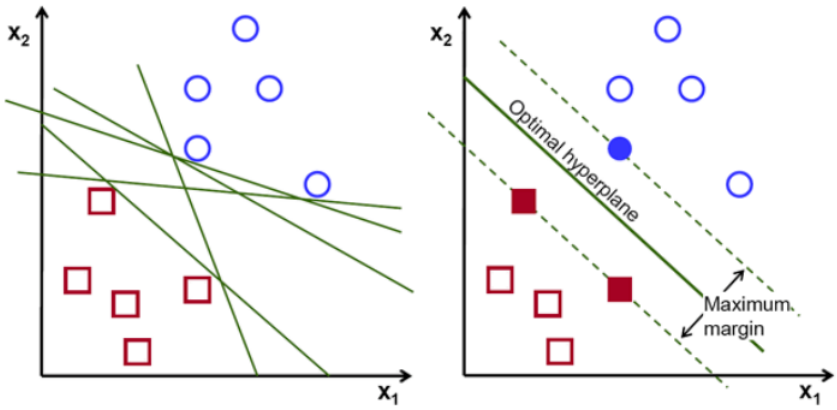


FIGURE 2.10 – Principe d'un SVM

Si les données ne sont pas linéairement séparables, on les projette alors dans un autre espace de dimension supérieure en utilisant une fonction noyau. Cette méthode est appelée *kernel trick* (astuce du noyau en français). En choisissant le bon noyau, les caractéristiques deviennent linéairement séparables.

2.2.2 k-Plus Proches Voisins

La méthode des k-Plus Proches Voisins ou k-NN (*k-Nearest Neighbors, kNN*), est une méthode de classification supervisée [144, 310]. L'algorithme kNN part du principe que les éléments similaires existent à proximité les unes des autres. Dans un espace de caractéristiques, un nouvel échantillon sera labellisé en fonction de la classe la plus représentée parmi celles de ses k plus proches voisins. L'objectif de l'algorithme des k-NN est d'identifier les plus proches voisins d'un point donné, afin de pouvoir attribuer un label de classe à ce point. Pour déterminer les points de données les plus proches d'un point donné, il faut calculer la distance entre ce point et les autres points. Ces mesures de distance aident à former les frontières de décision, qui divisent les points d'interrogation en différentes régions. On représente souvent les frontières de décision par des diagrammes de Voronoï. La mesure de distance la plus couramment utilisée est la **distance euclidienne** : elle mesure une ligne droite entre le point requête et un autre point mesuré. D'autres méthodes peuvent être utilisées : la distance de Manhattan, de Minkowski ou encore de Hamming [20].

FIGURE 2.11 – Principe d'un kNN

La méthode des k-NN est simple à implémenter avec peu de paramètres (le nombre de voisins k et la mesure de distance), et permet facilement d'ajouter de nouveaux exemples dans la base d'apprentissage. Néanmoins, en raison de la « malédiction de la dimensionnalité », le nombre de données nécessaires pour avoir une bonne estimation croît de manière exponentielle avec le nombre de dimensions, c'est-à-dire avec la complexité de la représentation des données. De ce fait, la méthode des k-NN est également plus sujette au surapprentissage. De plus, le choix du nombre de voisins pris en compte k, influe sur les résultats : un k trop petit donnera un système sensible au bruit, tandis qu'un k trop grand aura du mal à discriminer les éléments [185].

2.2.3 Dynamic Time Wraping

Les méthodes précédentes peuvent être combinées à d'autres techniques, par exemple en utilisant la distorsion temporelle dynamique (*Dynamic Time Wraping, DTW*) comme mesure de distance. La DTW est utilisée pour comparer la similarité entre deux séries temporelles comme les mouvements d'une articulation dans le temps [18, 257]. Pour cela, on détermine, pour chaque élément d'une séquence, le meilleur élément correspondant dans l'autre séquence relativement à un certain voisinage et à une certaine métrique.

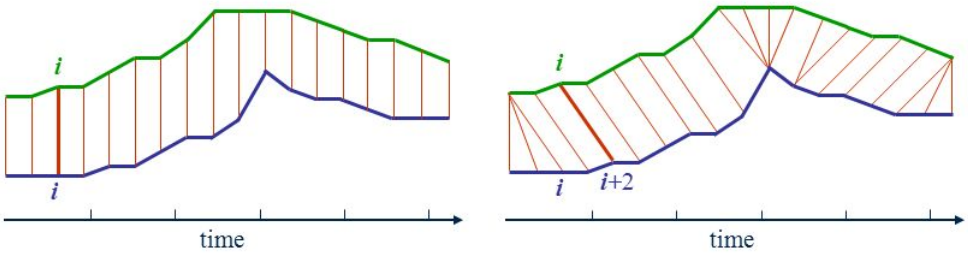


FIGURE 2.12 – Principe d'un DTW

La Fig. 2.12 montre un exemple de DTW. À gauche, on peut voir un alignement entre deux séquences temporelles par calcul d'une distance euclidienne qui permet d'aligner le i-ème point de la série en vert avec le i-ème point de la série en bleue. À droite, on a un alignement non linéaire par DTW entre les motifs correspondants des deux courbes, même s'ils ne sont pas alignés sur l'axe du temps représenté en bas.

2.2.4 Hidden Markov Models

Les modèles de Markov Cachés (textitHidden Markov Models, HMM sont utilisés pour la reconnaissance de gestes, car ils peuvent prendre en compte la dimension temporelle dans l'exécution d'un geste [138].

D'après [92], les HMMs sont des systèmes probabilistes à états finis qui reposent sur le principe de dépendance entre deux observations successives, ce qui permet de modéliser l'évolution dynamique d'un système aléatoire. La propriété fondamentale des chaînes de Markov est que l'évolution future du système ne dépend que de l'état

actuel dans lequel il se trouve. Il s'agit de la propriété dite markovienne. Les chaînes de Markov ont été utilisés pour reconnaître des actions dans une suite d'images. Ainsi Yamato et al. ont proposé une méthode combinant un algorithme de clustering puis des HMMs [325]. Les HMMs ont également été utilisés par Achard et al. pour reconnaître des actions à partir de silhouettes [4]. Zhu et Pun ont utilisé des HMMs pour reconnaître des gestes de la main en temps réel avec des images de profondeur [341]. Xia et al. ont classé des gestes avec des HMMs [323].

2.2.5 Forêts aléatoires

L'algorithme de la forêt aléatoire [28] combine plusieurs arbres de décision aléatoires et agrège leurs prédictions en faisant la moyenne. Une forêt aléatoire est un ensemble de classificateurs composé d'arbres de décision générés en utilisant deux sources différentes de randomisation. Premièrement, chaque arbre de décision individuel est formé sur un échantillon aléatoire des données originales. La deuxième source de randomisation appliquée dans la forêt aléatoire est l'échantillonnage par attributs. Pour cela, à chaque division de nœud, un sous-ensemble de variables d'entrée est sélectionné aléatoirement pour rechercher la meilleure division. Les forêts aléatoires ont montré d'excellentes performances dans des contextes où le nombre de variables est beaucoup plus grand que le nombre d'observations. En outre, cette méthode est suffisamment polyvalente pour être appliquée à des problèmes de régression et de classification, et s'adapte facilement à diverses tâches d'apprentissage [21].

L'algorithme des forets aléatoires utilise une technique appelée *bagging*, une technique qui consiste à assembler un grand nombre d'algorithmes avec de faibles performances individuelles, appelés *weak learners*, pour en créer un beaucoup plus efficace : les *strong learners*. Le principe de cet algorithme est donc que plusieurs algorithmes de faibles performances peuvent être plus efficaces qu'un seul grand algorithme. Les algorithmes de faibles performances peuvent varier de nature et de performances variées. En revanche, ils doivent être indépendants les uns des autres. Appliqué aux forets aléatoires, l'assemblage des *weak learners* (les arbres) en *strong learners* (la forêt) se fait par vote : chaque *weak learners* va donner une réponse (un vote), et la prédiction du *strong learner* sera la moyenne de toutes les prédictions émises.

Les algorithmes de renforcement de gradient (*gradient boosting*) combinent des *weak learners* en un *strong learner* de manière itérative. Les algorithmes de Boosting se basent sur le même principe que ceux de *Bagging*. Cependant, pour le *Boosting*, les algorithmes ne sont plus indépendants : chaque *weak learner* est entraîné pour corriger les erreurs des *weak learners* précédents. Parmi les algorithmes de Boosting les plus célèbres, on peut citer Adaboost, LightGBM et XGBoost [17]. XGBoost est un algorithme qui utilise réside des arbres décisionnels comme *weak learners*. Les arbres qui ne sont pas assez bons sont "élagués", dans le sens où on leur coupe des "branches", jusqu'à ce qu'ils soient suffisamment performants. Dans le cas contraire, ils sont supprimés. Cette méthode est appelée le *pruning* (élagage). De cette façon, XGBoost n'utilise que de bons *weak learners*. De plus, XGBoost est optimisé pour rendre les différents calculs nécessaires à l'application d'un *Gradient Boosting* rapides grâce à des hyperparamètres [43].

Conclusion

Cette partie a démontré que le traitement des mouvements du corps joue un rôle clé dans les domaines de la reconnaissance d'actions et l'informatique affective. Le premier est essentiel pour comprendre comment les personnes agissent dans un environnement, tandis que le second tente d'interpréter les émotions des personnes en se basant sur leurs postures et leurs mouvements. Toutefois, de nombreux défis scientifiques persistent dans l'analyse de mouvements dans le cadre de scènes du monde réel : manque de données, squelettes de haute qualité dans des conditions réelles, segmentation incorrecte et incomplète des personnes, occlusions, etc.

Le second chapitre a montré que l'utilisation de capteurs avec systèmes de détection de points d'intérêts intégrée comme les caméras Kinect ou Vicon permet une captation relativement facile et non intrusive. Bien que les caméras soient sensibles aux occultations, elles permettent d'avoir d'avoir une meilleure connaissance de la scène et donc de l'environnement. L'utilisation de caméras de profondeur permet de s'émanciper des changements de luminosité et d'avoir des informations plus directes sur les mouvements de l'opérateur. Ces données peuvent ensuite être exploitées par un algorithme d'apprentissage automatique profond comme nous l'expliquerons dans la partie suivante.

Deuxième partie

Apprentissage automatique profond pour la reconnaissance d'actions

Chapitre 3

Apprentissage automatique profond

"However, machines of this character can behave in a very complicated manner when the number of units is large."

— Alan Turing

Sommaire

Dès 1950, Alan Turing s'interrogea dans son célèbre article *Computing Machinery and Intelligence* [293] sur la capacité des ordinateurs à être intelligents (dans le sens d'être capable d'imiter un humain), et introduisit le concept de ce qui est maintenant appelé le test de Turing. Le terme d'**Intelligence Artificielle (IA)** apparaît pour la première fois en tant que domaine de recherche à l'occasion d'un colloque scientifique organisé sur le campus de l'université de Dartmouth, New Hampshire (États-Unis) [196]. Organisée par Marvin Minsky et John McCarthy, cette conférence a réuni vingt chercheurs de différentes disciplines, dont Claude Shannon. Pour Minsky, l'IA est la « construction de programmes informatiques capables d'accomplir des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains ». Demis Hassabis, fondateur de DeepMind [1], définit plus simplement l'IA comme étant la « science de rendre les machines intelligentes » [107].

De nos jours, l'intelligence artificielle est définie comme la capacité d'un ordinateur à effectuer des tâches communément associées à l'intelligence. Il existe différentes techniques d'IA que nous allons introduire dans ce chapitre. La plupart des méthodes de détection automatique des émotions, de la douleur et de reconnaissance d'actions reposent actuellement sur des méthodes dites d'Intelligence Artificielle (IA) appelé **apprentissage automatique** (*Machine Learning*) et **apprentissage automatique profond** (*Deep Learning*).

Dans ce chapitre, nous allons dans un premier temps expliquer les bases de l'apprentissage automatique et de l'apprentissage automatique profond, avant de décrire le principe de fonctionnement des principaux type de réseaux de neurones : CNNs, RNNs, Transformer et GNNs.

3.1 Apprentissage Automatique

L' **apprentissage automatique** est un sous-domaine de l'IA (voir Fig. 3.1) qui vise à donner à un programme informatique la capacité d'apprendre, sans que cela soit explicitement formulé. L'**apprentissage automatique profond** est lui-même un sous-ensemble de l'apprentissage automatique, basé sur des réseaux neuronaux profonds (voir Fig. 3.2).

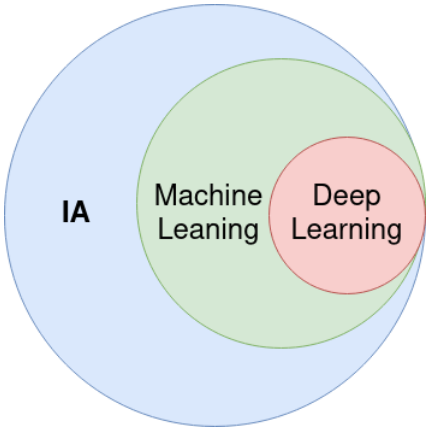


FIGURE 3.1 – Intelligence artificielle

1. https ://www.deepmind.com

De façon générale, l'apprentissage automatique est une méthode où un programme entraîne un modèle prédictif à partir de données d'entrée et utilise ce modèle entraîné pour effectuer des prédictions utiles à partir de nouvelles données (c'est-à-dire jamais vues auparavant), issues de la même distribution que celles utilisées pour entraîner le modèle [2]. L'apprentissage automatique utilise ces données pour détecter des tendances, s'ajuster en conséquence, et donner de nouveaux résultats, comme représenté dans la figure 3.2. L'apprentissage automatique permet à un programme informatique d'apprendre à réaliser des tâches complexes, potentiellement difficiles à résoudre avec des programmes classiques. Le processus d'apprentissage se définit comme la capacité à exécuter cette tâche en s'améliorant avec l'expérience [116]. Le processus d'apprentissage est donc distinct de la tâche en elle-même, l'apprentissage étant le moyen d'atteindre la capacité à accomplir cette tâche. Par exemple, si l'on veut qu'un robot soit capable de marcher, alors la tâche à apprendre est la marche : on peut essayer d'écrire directement un programme qui spécifie exactement la procédure pour marcher (comme c'est le cas en programmation classique), ou bien, on peut programmer le robot pour qu'il apprenne seul à marcher (ce qui est le cas en apprentissage automatique).

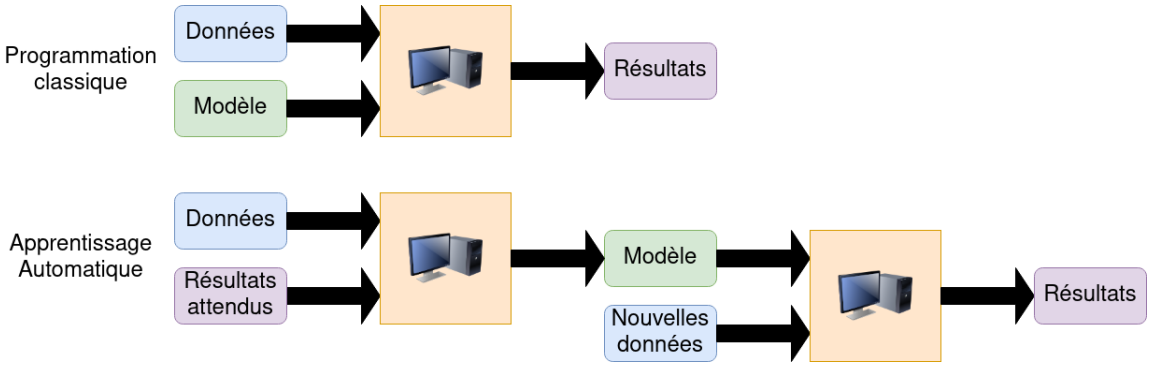


FIGURE 3.2 – Apprentissage automatique

Type d'apprentissage

Les algorithmes d'apprentissage automatique sont souvent classés selon le type d'apprentissage :

- **apprentissage supervisé** : ces algorithmes prennent en entrée une paire donnée d'entrée-label pour chaque échantillon de données, où le label décrit la sortie correcte attendue. Les algorithmes d'apprentissage non supervisé prennent en entrée un ensemble de données contenant de nombreuses caractéristiques (*features*) pour déduire des motifs sous-jacents de cet ensemble de données, par exemple en divisant le jeu de données en groupes d'exemples similaires (*clustering*). L'objectif est alors de résumer et identifier des tendances intéressantes du jeu de données.
- **apprentissage non supervisé** : dans cette méthode d'apprentissage, l'entrée n'est ni classée ni labéllisée. L'objectif ici est donc de trouver des caractéristiques communes entre les échantillons fournis afin de les regrouper de la bonne manière et de fournir le bon résultat.

2. https ://developers.google.com/machine-learning/glossary ?hl=fr

— **apprentissage semi-supervisé** : cette méthode combine les méthodes d'apprentissage supervisé et non supervisé en utilisant une petite quantité de données labellisées et une plus grande quantité de données non labellisées.
— Certains algorithmes d'apprentissage automatique ne se contentent pas d'utiliser un ensemble de données fixé. Ainsi, dans les algorithmes d'**apprentissage par renforcement,** un agent interagit avec un environnement, de sorte qu'il existe une boucle de rétroaction entre le système d'apprentissage de l'agent et ses expériences [282].

Dans le cadre de cette thèse, nous nous intéresserons uniquement à l'apprentissage supervisé. Les algorithmes d'apprentissage supervisé prennent en entrée un ensemble de données. Chaque exemple est également associé à un label que l'on va chercher à prédire. Par exemple, le jeu de données « Iris » [103] est annoté avec trois différentes espèces d'iris [3]. Un algorithme d'apprentissage supervisé utilisant ce jeu de données peut apprendre à classer les iris uniquement selon ces trois espèces différentes. Un label peut indiquer si une image contient un chien ou un chat, quels mots ont été prononcés dans un enregistrement audio, quel est le sentiment général d'un tweet, ou encore quel type d'action est effectué dans une vidéo. Les labels peuvent être obtenus en demandant à des humains (qui peuvent être des experts ou non) d'annoter manuellement les données. La labellisation des données est donc un processus long et coûteux, et une source potentielle de biais et d'erreurs.

L'apprentissage automatique permet de réaliser de nombreuses tâches différentes. Les plus courantes sont la classification et la régression. Pour la **classification**, le programme informatique doit trouver les k catégories auxquelles appartient une entrée. Pour résoudre cette tâche, l'algorithme d'apprentissage automatique doit produire une fonction $f : \mathbb{R}^n \to \{1, \ldots, k\}$. Lorsque $y = f(x)$, le modèle attribue une entrée décrite par le vecteur x à une catégorie identifiée, ou bien f sort une distribution de probabilité sur les classes. Un exemple de tâche de classification est la reconnaissance d'objets, dans laquelle l'entrée est une image (généralement décrite comme un ensemble de valeurs de pixel), et la sortie est un code numérique identifiant l'objet dans l'image. Pour une **régression**, l'objectif est de prédire une valeur numérique à partir d'une certaine entrée. Pour résoudre cette tâche, l'algorithme d'apprentissage automatique doit produire une fonction $f : \mathbb{R}^n \to \mathbb{R}$. Ce type de tâche est similaire à la classification, sauf que le format de la sortie est différent. Un exemple de régression est la prédiction des prix des maisons pour le marché immobilier.

Depuis plusieurs années, les méthodes d'apprentissage automatique permettant de manipuler de grands volumes de données se sont popularisées. Pour les tâches de classification, l'apprentissage supervisé est principalement utilisé puisque plusieurs jeux de données annotées sont disponibles. Les systèmes les plus performants utilisent principalement les **réseaux de neurones profonds**. Le principal défi de l'apprentissage automatique est que l'algorithme doit être performant sur de nouvelles entrées inédites, et pas seulement celles sur lesquelles le modèle a été entraîné. La capacité à obtenir de bonnes performances sur des entrées précédemment non

3. https ://archive.ics.uci.edu/ml/datasets/iris

observées est appelée généralisation. Le principe de cet apprentissage est de calculer une certaine mesure de l'erreur à chaque niveau d'apprentissage sur l'ensemble des données d'apprentissage. L'objectif est donc de réduire cette erreur.

Dans le cas d'un apprentissage supervisé, plusieurs étapes sont nécessaires. Il faut d'abord traiter les données brutes pour les nettoyer, les homogénéiser et les étiqueter avec les labels correspondants. Ensuite, il faut séparer les données en plusieurs jeux de données de façon équitable entre chaque label : un jeu pour l'**entraînement** du modèle, un pour la **validation**, puis un pour le **test** avec de nouvelles données jamais vues par le modèle, la majorité des données étant toujours réservées pour l'entraînement et le reste étant réparties de façon équitable entre la validation et le test. Les jeux d'entraînement et de validation servent à ajuster les paramètres du modèle, tandis que le jeu de test sert à évaluer l'efficacité et la performance du modèle sur des nouvelles données non utilisées pour son élaboration grâce à des indicateurs comme la précision, le rappel et la F-mesure, qui seront présentés dans la section suivante.

Deep Learning

Le **deep learning**, ou apprentissage automatique profond, est une sous-catégorie de l'apprentissage automatique dont le but est d'extraire automatiquement des caractéristiques pour représenter des données grâce des réseaux de neurones artificiels (*artificial neural network* ou ANN). Les ANN reposent sur des neurones artificiels appelés perceptrons, qui sont en fait des unités de calcul de fonctions mathématiques organisées en couches. Le principe des réseaux de neurones est de reprendre une modélisation mathématique simplifiée de la façon dont les neurones biologiques fonctionnent et de leur organisation en couches [3]. L'apprentissage passe alors par différents niveaux de représentation, d'où l'emploi du terme *deep* ou « profond ». On parle de réseau de neurones profonds ou DNN *Deep Neural Network* quand le nombre de couches de neurones cachées est supérieur à 1, soit plus de trois couches neurones en tout (une couche d'entrée, une couche cachée et une couche de sortie). Le récent regain d'intérêt pour les méthodes d'apprentissage profond s'explique par le fait qu'il a été démontré qu'elles surpassent les techniques de pointe antérieures dans plusieurs tâches, ainsi que par l'abondance de données complexes provenant de différentes sources (visuelles, audio, médicales, sociales et capteurs).

Il existe différents types d'architectures profondes, chacune ayant sa spécificité. Par exemple, les algorithmes à base de réseaux de neurones profonds ont fait leurs preuves dans de nombreuses applications d'imagerie et vision par ordinateur, notamment grâce aux réseaux de neurones convolutifs (ou réseau de neurones à convolution) qui ont été popularisés principalement par leur efficacité dans la classification d'images lors de challenges internationaux. Ainsi, actuellement, le classement des défis type Imagnet est dominé par les réseaux neuronaux convolutifs et les techniques d'apprentissage profond détaillées dans la section suivante.

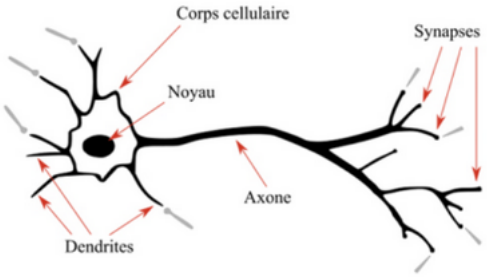
3.2 Réseau de neurones

3.2.1 Historique

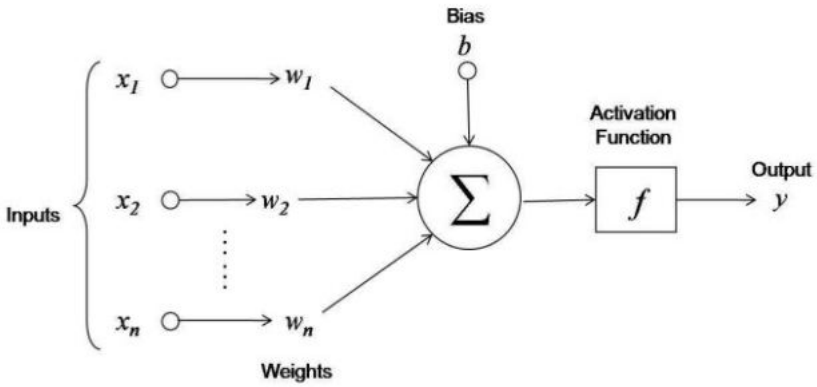
Le perceptron

La **cybernétique**, tel que défini en 1948 par Norbert Wiener dans son ouvrage éponyme, avait l'ambition de créer un système qui simule l'environnement humain et a permis le développement initial des réseaux neuronaux[319]. La cybernétique souhaitait formaliser et modéliser tous les phénomènes qui mettent en jeu des mécanismes de traitement de l'information grâce au concept de **rétroaction** (*feedback*), c'est-à-dire l'idée d'une boucle d'information qui, dans un système, permet de corriger une action pour la rendre plus efficace. Dès 1943, McCulloch et Pitts tentent de comprendre comment le cerveau pourrait produire des modèles très complexes en utilisant des modèles inspirés des neurones. Ils ont alors proposé un modèle d'un neurone appelé modèle MCP (pour *McCulloch Pitts Neuron Model*) [197].

Un neurone biologique se compose principalement de quatre éléments, à savoir les dendrites, le soma, l'axone et les synapses, qui permettent de recevoir, d'élaborer et de transmettre des signaux électriques (voir figure 3.3a). Ces derniers sont reçus par l'intermédiaire des dendrites et du soma, puis ils sont émis le long de l'axone. La synapse est la structure qui permet de mettre en relation un axone et d'autres dendrites de neurones.



(a) Neurone biologique



(b) MCP

FIGURE 3.3 – Analogie entre neurone et MCP

Ce modèle mathématique a été le point de départ au développement des neurones artificiels en réseaux, suivi par une série de contributions majeures menant à l'ère

de l'apprentissage profond que nous connaissons aujourd'hui. En 1957, Franck Rosenblatt développa un algorithme d'apprentissage supervisé de classifieurs binaires (c'est-à-dire séparant deux classes) appelé **perceptron** (voir Fig. 3.3b) [242]. Le perceptron est un neurone artificiel muni d'une règle d'apprentissage qui permet de déterminer automatiquement les poids. Si le problème est linéairement séparable, le perceptron permet de trouver une séparatrice entre les deux classes. Dans les réseaux de neurones artificiels, les perceptrons peuvent donner un message de sortie en fonction de ceux reçus en entrée. Le réseau neuronal est un réseau de neurones artificiels inter-connectés, où la sortie d'un neurone donné peut être l'entrée d'autres neurones. L'apprentissage se produit en activant de façon répétée certaines connexions neuronales par rapport à d'autres, ce qui renforce ces connexions. Cela les rend plus susceptibles de produire un résultat souhaité compte tenu d'un intrant donné. Cet apprentissage implique une rétroaction : lorsque le résultat souhaité se produit, les connexions neuronales qui causent ce résultat sont renforcées. L'architecture des réseaux de neurones artificiels regroupe un ensemble d'unités élémentaires, les perceptrons, de façon interconnecté : ces neurones artificiels sont connectés entre eux pour former un graphe orienté. Par comparaison avec les réseaux de neurones biologiques, les connexions entre les nœuds du graphe représentent les synapses. Ces connexions sont pondérées par des poids ajustés durant la phase d'apprentissage afin de minimiser la différence entre la sortie du réseau (l'hypothèse) et la sortie (la référence attendue) [92].

Chaque perceptron de la couche d'entrée va prendre en entrée des **données** x, les multiplier par un **poids** w, en faire la somme, ajouter un **biais** b et transformer le tout par une **fonction d'activation** f. La « mise en marche » (ou activation) du perceptron est simulée par cette fonction d'activation qui utilise donc la somme pondérée de toutes les entrées de la couche précédente pour générer une valeur de sortie y. Autrement dit, le perceptron est une unité de calcul de type *feed-forward*, ce qui signifie que le flux de données est unidirectionnel de l'entrée vers la sortie. Un j-ème perceptron générique j peut recevoir de multiples entrées qui sont multipliés par des poids et ensuite additionnées (c'est-à-dire la somme pondérée), conformément à l'équation suivante somme pondérée), conformément à l'équation suivante :

$$o_j = \omega_{j0} + \sum_{i=1}^{n} \omega_{ji} x_i \tag{3.1}$$

où xi représente la i-ième entrée, w_{ji} est le poids associé à cette i-ième entrée et w_{j0} est le biais pour cette j-ème unité. Le **biais** est le paramètre qui permet de décaler la fonction d'activation vers la gauche ou la droite dans l'espace vectoriel. La sortie du perceptron o_j est ensuite utilisée en conjonction avec une fonction d'activation, générant la sortie finale y_j.

Il existe plusieurs fonctions d'activation, une des plus communes est la fonction sigmoïde. On peut également citer la tangente hyperbolique (notée *tanh*) ou la fonction ReLU (*Rectified Linear Unit*) [181]. Cette fonction d'activation doit avoir une caractéristique d'activation : une fois que l'entrée est supérieure à une certaine valeur, la sortie doit changer d'état (de 0 à 1, de -1 à 1 ou de 0 à > 0 selon la fonction d'activation).

On peut ainsi regrouper leur paramètre par ligne et former une matrice. Un réseau de neurones est donc une succession de couches de neurones, telles que le nombre de neurones sur chaque couche est égal au nombre d'entrées des neurones de la couche suivante. La première couche contient des neurones qui transmettent les entrées fournies au réseau, il s'agit de la couche d'entrée. La dernière couche est appelée couche de sortie et fournit le résultat du réseau. Les autres couches sont appelées couches cachées.

Perceptron multicouches

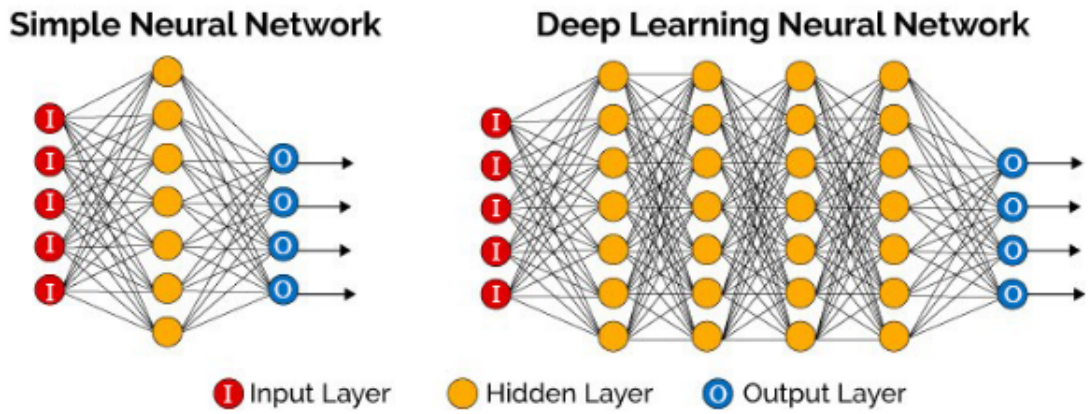


FIGURE 3.4 – Réseaux de neurones [321]

Le **Perceptron multicouches** (*Multilayer Perceptron*, MLP) est un réseau de perceptrons disposés sur plusieurs couches conçues pour traiter des données de façon non linéaire [243]. Ce réseau de neurones standard se compose donc de nombreuses unités de calculs connectées, appelées neurones, chacune produisant une séquence d'activations à valeur réelle. Les unités d'entrée sont activées selon les données d'entrées, puis les autres unités par des connexions avec des poids à valeur réelle provenant d'unités précédemment actives. L'apprentissage consiste donc à trouver les poids qui font que le réseau de neurones présente le comportement souhaité (détection d'objets, classification, régression, etc). Selon le problème et la façon dont les unités sont connectées, un tel comportement peut nécessiter de longues chaînes causales d'étapes de calcul, où chaque étape transforme (souvent de façon non linéaire) l'activation globale du réseau. Dans ce type de réseau de neurones artificiels, les différents neurones d'une couche vont donc transmettre à la couche suivante $N+1$ leur valeur de sortie calculée, et ainsi de suite de couche en couche, jusqu'à la valeur de sortie du réseau (*output*) (voir figure 3.4). Ce genre de progression est dit à propagation avant ou *feedforward* (ou encore *forwardpass*) car l'information ne se déplace que vers l'avant, c'est-à-dire des nœuds d'entrée, puis par les couches cachées vers les nœuds de sortie. Il n'y a pas de cycles ou de boucles dans ce genre de réseau. Les réseaux de neurones profonds se sont développés dans les années 1960. Ivakhnenko et Lapa ont publié le premier algorithme d'apprentissage général et fonctionnel pour les perceptrons multicouches [141], puis en 1971 Ivakhnenko décrivit un réseau profond à 8 couches [142]. Comme les réseaux profonds ultérieurs, les réseaux d'Ivakhnenko ont ainsi appris à créer des représentations internes des données entrantes [168, 253].

Neocognitron

Les années 1980 ont également vu la naissance des **réseaux de neurones à convolutions (CNN)** avec le **Neocognitron** [105] (voir Fig. 3.5). Le neocognitron est un réseau de neurones multicouches utilisé pour la reconnaissance de caractères manuscrits japonais et d'autres taches de reconnaissance de formes. Inspiré par la structure hiérarchique du cortex visuel décrit en 1949 par Hubel et Wiesel [139], Il reprend le principe de structure en couches en ajoutant le principe de cellules reliées entre elles. Chaque couche extrait certaines formes géométriques de l'image d'entrée, puis, au fur et à mesure que l'on s'enfonce dans le réseau, il commence à les combiner entre elles pour reconstituer une forme. Les cellules sont liées de telle sorte qu'elles ne réagissent que lorsqu'elles reçoivent un signal de la couche précédente (voir Fig. 3.5). Le neocognitron a servi d'inspiration pour les réseaux neuronaux convolutifs présentés dans la section suivante.

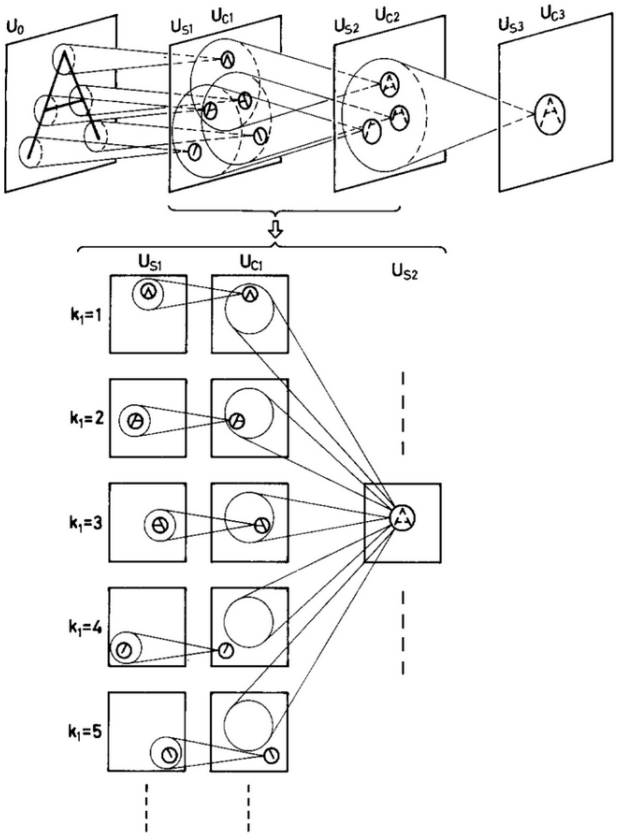


FIGURE 3.5 – Neocognitron [105]

Aujourd'hui, les architectures CNNs sont largement utilisées pour la vision par ordinateur. Le champ réceptif des CNNs (généralement rectangulaire) dotée d'un vecteur de poids donné (un filtre) est déplacé pas à pas sur un réseau bidimensionnel de valeurs d'entrée, comme les pixels d'une image. La matrice résultant des événements d'activation ultérieurs de cette unité peut alors fournir des entrées aux unités de niveau supérieur, et ainsi de suite. En raison de la réplication massive des poids, relativement peu de paramètres peuvent être nécessaires pour décrire le comportement de ces couches convolutionnelles, qui alimentent généralement des couches de sous-échantillonnage composées d'unités dont les connexions à poids fixe

proviennent de voisins physiques dans les couches convolutionnelles inférieures. Les unités de sous-échantillonnage utilisent le *Spatial Averaging* pour devenir actives si au moins une de leurs entrées est active ; leurs réponses sont insensibles à certains petits décalages de l'image. Weng (1993) [315] a ensuite remplacé le *Spatial Averaging* par le *Max-Pooling* (MP), qui est largement utilisé aujourd'hui. Dans ce cas, une couche ou une matrice bidimensionnelle d'activations unitaires est divisé en matrices rectangulaires plus petites. Chacune est remplacée dans une couche de sous-échantillonnage par l'activation de son unité la plus active.

Rétropropagation

La **rétropropagation** a également été développée au début des années 1960 [157, 30, 31]. Dreyfus et Linnainmaa ont publié des méthodes de dérivation de la rétropropagation basée sur la règle de la chaîne, encore utilisée aujourd'hui [72, 180]. Werbos a publié la première application de la rétropropagation aux réseaux de neurones [316]. Durant la période 1980-1990, les ordinateurs sont devenus beaucoup plus rapides et plus accessibles dans les laboratoires universitaires. Des expériences ont alors démontré que la rétropropagation dans les réseaux de neurones peut effectivement produire des représentations internes utiles dans les couches cachées [246]. Le fonctionnement de la rétropropagation est développé dans la section 3.2.2.

Réseaux de neurones à convolutions

En 1986, Geoffrey Hinton et al. ajoutent des couches cachées et développent les réseaux de neurones artificiels multicouches à propagation avant (*feedforward neural network*). En 1989, LeCun et al. [166] ont été les premiers à appliquer la rétropopagation à des **réseaux neuronaux convolutifs** (convolutional neural network, CNN) de type Neocognitron, obtenant de bonnes performances sur la base de données **MNIST** (*Modified ou Mixed National Institute of Standards and Technology*, il s'agit d'une base de données de chiffres écrits à la main) [4]. Des CNNs similaires ont été utilisés commercialement dans les années 1990. Les recherches sur les CNN se sont poursuivies pendant les huit années suivantes et, en 1998, Yann LeCun et al. ont proposé un CNN appelé **LeNet** qui était plus performant que tous les autres modèles [167].

Les succès du deep learning ont continué dans le domaine de la vision par ordinateur, notamment avec les travaux de l'équipe de Geoffrey Hinton à l'université de Toronto et dans le domaine de la reconnaissance vocale avec les travaux de George E. Dahl et al. en 2010. Les techniques d'apprentissage profond ont alors commencé à surpasser celles d'apprentissage automatique classique dans plusieurs domaines, pour plusieurs raisons :

- les très grandes quantités de données (on parle alors de « Big Data ») à traiter favorisent l'apprentissage profond,
- et de grandes bases de données ont été mises à la disposition des chercheurs,
- les ordinateurs sont devenus plus puissants et plus rapides grâce aux processeurs multicœurs : l'augmentation de la puissance de calcul a permis de

4. http ://yann.lecun.com/exdb/mnist/

déployer des réseaux de neurones efficaces,
— de nouveaux modèles et algorithmes ont fait leur apparition,
— de nouvelles bibliothèques telles que Tensorflow ou Pytorch ont fait leur apparition favorisant la diffusion et le partage de ces modèles.

Des CNNs efficaces parallélisés sur des cartes graphiques type GPU [51] ont encore amélioré considérablement le record du MNIST, atteignant pour la première fois le niveau des performances humaines (soit environ 0,2% d'erreurs) [52]. Pour détecter des actions humaines dans des vidéos de surveillance, Yang et al. ont développé un CNN tridimensionnel (3D CNN), combiné à des algorithmes de machines à vecteurs de support (SVM), faisait partie d'un système plus vaste utilisant une approche de type sac de caractéristiques pour extraire des régions d'intérêt. Ce système a obtenu de bons résultats lors du concours TREC Video Retrieval Evaluation (TRECVID) 2009 [328]. En 2011, un ensemble de CNN entraîné sur des cartes graphiques a également été le premier système à réaliser une reconnaissance visuelle de formes dans une compétition, à savoir le concours de reconnaissance de panneaux de signalisation IJCNN 2011 [49]. Le système était deux fois meilleur que les humains et trois fois meilleur que le concurrent le plus proche.

Ainsi en 2012, Krizhevsky et al. ont remporté le concours ImageNet Large Scale Visual Recognition Challenge (ILSVRC) grâce à un réseau de neurones appelé **Alexnet** [163]. D'autres progrès sur ImageNet ont été réalisés grâce à des variantes de ces systèmes [335, 283, 273]. Un CNN a également été le premier réseau de neurones profond à remporter un concours sur la découverte d'objets visuels dans de grandes images [48], à savoir le concours ICPR 2012 sur la détection de mitoses dans les images histologiques du cancer du sein. Dans ce cas, les CNNs profonds sont formés sur des patchs étiquetés de grandes images, puis utilisés comme détecteurs de caractéristiques pour être utilisées dans des scènes visuelles inconnues, en utilisant diverses rotations et facteurs de zoom. Un CNN similaire a été le premier *Deep Learner* à remporter un concours de segmentation d'image [50]. Le CNN a appris à prédire pour chaque pixel s'il appartient à l'arrière-plan. Depuis d'autres concours internationaux ont été remportés à l'aide de CNNs [146].

Réseaux de neurones récurrents

Les architectures de réseaux de neurones précédents ne peuvent pas corréler les informations dans le temps. Par exemple, dans la classification d'événements à différents moments d'une vidéo, un MLP ne peut utiliser les événements précédents dans la séquence vidéo pour prédire les événements suivants. Le **réseau neuronal récurrent** (Recurrent neural network, RNN) a été proposé pour résoudre ce problème, en utilisant des boucles internes qui permettent aux informations de persister comme une sorte de mémoire. Plusieurs travaux [270, 277] ont démontré la puissance et l'efficacité théorique des RNN qui seraient capables d'émuler n'importe quel algorithme, étant Turing-complet. Siegelmann propose ainsi une machine de Turing universelle théorique à partir d'un RNN à 886 neurones cachés [271]. Néanmoins, en réalité, il est difficile d'atteindre cette puissance théorique, car ce genre de réseau est difficile à entraîner en pratique (pour des raisons de temps de calculs notamment).

Alors que les réseaux de neurones simples et les CNN permettent de traiter des

vecteurs de taille fixe, les RNN permettent de traiter des séquences de vecteurs (en entrée et/ou en sortie), et notamment des séquences temporelles [252]. Des variantes de la rétropropagation ont été développées pour les RNN supervisés ([320, 239, 317]). Pendant l'apprentissage par « rétropropagation à travers le temps », le RNN est « déplié » en un réseau qui a essentiellement autant de couches qu'il y a de pas de temps dans la séquence observée des vecteurs d'entrée. Les inconvénients de la rétropropagation sont devenus évidents lorsque le problème du gradient a été identifié et analysé [130, 131] : Avec les fonctions d'activation standard, les signaux d'erreur cumulatifs rétropropagés diminuent de manière exponentielle en fonction du nombre de couches (ou de pas de temps), ou au contraire explosent. Ce problème commun à tous les réseaux de neurones est plus apparent dans les RNNs, à cause de la profondeur liée à leur récursivité.

Une des solutions les plus efficaces permettant de résoudre le problème de calcul du gradient est l'amélioration du concept des RNNs, les neurones de type **Long Short Term Memory** (LSTM). Les LSTMs sont développés depuis les années 1990 [132, 109, 118]. Les LSTM sont conçus de telle sorte que les erreurs rétropropagées ne peuvent ni disparaître ni exploser, mais persistent pendant des milliers de pas, voire plus. Ainsi, les LSTM peuvent apprendre des tâches d'apprentissage complexes qui nécessitent de mémoriser des événements qui se sont produits il y a des milliers de pas de temps, là où les RNNs standards précédents avaient échoué. Il est possible d'élaborer des architectures de LSTM spécifiques à ce problème [15]. Les LSTMs bidirectionnels (BRNN) [255] sont conçus pour des séquences d'entrée dont le début et la fin sont connus à l'avance, comme des phrases parlées à étiqueter par leurs phonèmes. Les directed acyclic graph (DAG) RNNs [13] généralisent les BRNNs à plusieurs dimensions. En 2009, un LSTM est devenu le premier RNN à remporter des concours internationaux de reconnaissance de l'écriture manuscrite [120]. Hannun et al. [127] ont utilisé des RNNs pour gagner un challenge en reconnaissance vocale, sans utiliser de méthodes traditionnelles de traitement de la parole telles que les modèles de Markov cachés (*Hidden Makov Models* ou HMM). Contrairement aux HMMs et aux RNNs, un LSTM peut apprendre à reconnaître une langue en étant sensible au contexte. En 2007, les LSTMs ont commencé à révolutionner la reconnaissance vocale, en surpassant les HMMs traditionnels dans les tâches de repérage de mots-clés [100]. En 2013, les LSTMs ont obtenu les meilleurs résultats dans le célèbre test de reconnaissance de phonèmes TIMIT [119]. Des hybrides de méthodes traditionnelles et de LSTM ont obtenu les meilleures performances connues en reconnaissance vocale [247, 176]. Les LSTMs ont également contribué à améliorer l'état de l'art dans de nombreux autres domaines, notamment la génération de légendes d'images, en étant combiné avec des CNNs [297], la traduction automatique [190], la synthèse vocale [94, 336], l'analyse syntaxique pour le traitement du langage naturel [110], et de nombreuses autres applications [248].

3.2.2 Principes mathématiques

Dans un réseau de neurones, les couches sont divisées en trois catégories selon leur position à l'intérieur du réseau, appelées respectivement couche d'entrée, couche cachée et couche de sortie.

Dans un premier temps, le réseau fonctionne de façon *feed-forward* : les unités

d'une couche sont connectées à celles de la couche suivante, déplaçant les informations dans un chemin unidirectionnel de la couche d'entrée vers la sortie. L'élément clef dans les réseaux de neurones artificiels est le poids des différentes liaisons entre les nœuds du réseau.

Pour déterminer les meilleurs poids, il existe différentes techniques dont la plus courante est la rétro-propagation (*backpropagation*). En comparant la valeur de sortie obtenue à celle que l'on aurait dû avoir, on détermine une erreur que l'on va chercher à diminuer afin d'améliorer la valeur de sortie. Un calcul en chaîne de dérivés permet de déterminer le **gradient** de l'erreur, qui indique le sens (augmentation ou diminution). Ensuite, on modifie en conséquence les poids synaptiques et les biais du réseau pour faire en sorte que cette erreur soit la plus petite possible : les poids contribuant à engendrer une erreur importante se verront modifier de manière plus significative que les poids qui ont engendré une erreur marginale. La rétro-propagation va donc permettre d'optimiser les poids synaptiques. Lors de l'entraînement du réseau, on va répéter ce processus à chaque itération, ce qui constitue l'**apprentissage**.

En résumé, un réseau, composé de couches de neurones liées les unes aux autres, cherche à mettre en correspondance des données d'entrée avec des prédictions de labels. Une fonction de perte (*loss function*) compare ensuite ces prédictions aux véritables labels cibles, afin de vérifier dans quelle mesure ces prédictions du réseau correspondent à ce qui était attendu, ce qui produit une « valeur de perte ». Il existe différentes fonctions de perte :

- Les fonctions de perte de régression sont utilisées dans la modélisation prédictive de régression de régression, qui consiste à prédire une quantité à valeur réelle.
- Les fonctions de perte de classification binaire sont utilisées pour les problèmes de modélisation où les échantillons se voient attribuer l'une des deux étiquettes.
- Les fonctions de perte de la classification multiclasses sont utilisées dans les problèmes de modélisation où les échantillons sont affectés à l'une des différentes classes.

Il existe plusieurs algorithmes différents d' *optimizers*, dont l'un des plus connus est « Adam », qui servent à déterminer comment le réseau sera mis à jour en fonction de la fonction de perte et met en œuvre une variante spécifique de **descente de gradient stochastique** (SGD). L' *optimizer* utilise cette valeur de perte pour mettre à jour les poids du réseau afin d'obtenir de meilleures prédictions.

Pour évaluer les capacités d'un algorithme d'apprentissage automatique, il est nécessaire d'avoir une mesure quantitative de ses performances. Habituellement, cette mesure de performance est spécifique à la tâche exécutée par le système. Pour des tâches telles que la classification, la performance du modèle est souvent évaluée avec la précision du modèle. La précision est simplement la proportion d'exemples pour lesquels le modèle produit la sortie correcte. On peut également mesurer le taux d'erreur, la proportion d'exemples pour lesquels le modèle produit une sortie incorrecte.

Selon certains auteurs, un inconvénient des réseaux de neurones est leur côté « boite noire » [39]. En effet, il est difficile de comprendre leur fonctionnement, car contrairement aux systèmes experts où l'on a accès aux règles qui expliquent et justifient les choix de l'algorithme, dans les réseaux de neurones, il est difficile de comprendre les modifications des poids dans les couches cachées.

3.3 Réseau de neurones convolutifs

Un CNN peut être composé de plusieurs centaines de couches cachées, où chaque couche apprend différentes caractéristiques des données d'entrées. À partir de l'entrée, plusieurs filtres sont appliqués, obtenant une sortie utilisée comme entrée de la couche suivante. Au début, les filtres extraient des caractéristiques de base (par exemple, la luminosité, les bords, etc.), mais plus tard, en variant les itérations de formation, ils commencent à devenir plus complexes, définissant chaque classe considérée d'une manière unique. Les couches principales, utilisées pour l'apprentissage de caractéristiques complexes, sont principalement au nombre de deux, à savoir les couches de convolution et les couches de mise en commun (*pooling*). Enfin, une phase de classification utilise les résultats de l'apprentissage pour faire des prédictions, en utilisant une couche dense, c'est-à-dire entièrement connectée, comme on peut le voir sur la figure 3.6.

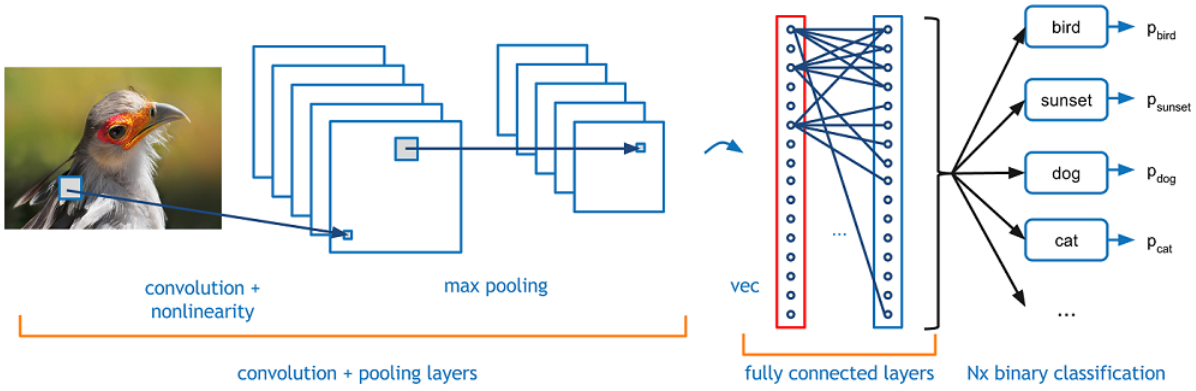


FIGURE 3.6 – Exemple de CNN

L'architecture d'un CNN est donc structurée en une série de différentes couches : les couches convolutives et les couches de regroupement (pooling). La couche de convolution est l'élément principal des CNNs. Son objectif est de détecter la présence de caractéristiques dans les images d'entrée grâce à un filtrage par convolution. Celui-ci consiste à faire glisser une fenêtre sur l'image d'entrée et à calculer le produit de convolution entre la caractéristique et chaque partie de l'image balayée (voir figures 3.7) [92].

Comme expliqué dans []blanc2018description, la couche de regroupement (*pooling*) est située entre les couches convolutives et permet d'appliquer à chacune des cartes de caractéristiques une réduction de leur taille, tout en préservant les caractéristiques les plus importantes (par exemple en ne conservant que les valeurs

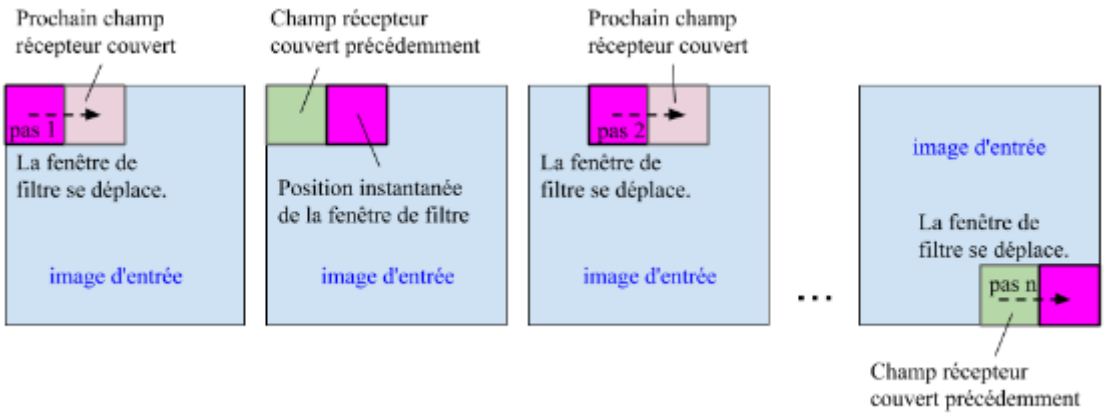


FIGURE 3.7 – Glissement de la fenêtre de filtre sur l'image d'entrée

maximales, voir figure 3.8b), ce qui de réduire le nombre de paramètres du réseau (et donc les calculs nécessaires) [92]. La couche de convolution est caractérisée par trois hyper-paramètres :

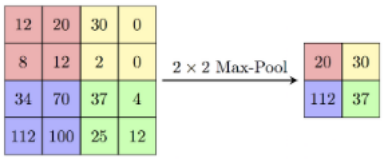
— la profondeur de la couche, c'est-à-dire le nombre de noyaux de convolution (ou nombre de neurones associés à un même champ récepteur),
— le pas (*stride*) : il contrôle le chevauchement des champs récepteurs. Plus le pas est petit, plus les champs récepteurs se chevauchent et plus le volume de sortie sera grand,
— le remplissage (*padding*) : cela consiste à ajouter une couche supplémentaire à la bordure d'une image, en général constituée de zéros (on parle de *zero padding*).

Au sein de la structure convolutive, les premiers noyaux de convolution vont donc détecter des caractéristiques simples (trait, courbe, etc). Les noyaux suivants vont détecter des caractéristiques plus complexes (œil, bouche, oreille, etc) en combinant les caractéristiques simples détectées par les premiers noyaux de convolution, et ainsi de suite. La relation spatiale entre les caractéristiques détectées par les phases successives de convolution est partiellement assurée par un autre type de couches de neurones : les couches de pooling. Ces couches permettent de compresser l'information en réduisant la taille de l'image et donc d'élargir le champ de vision du réseau (voir figure 3.8a).

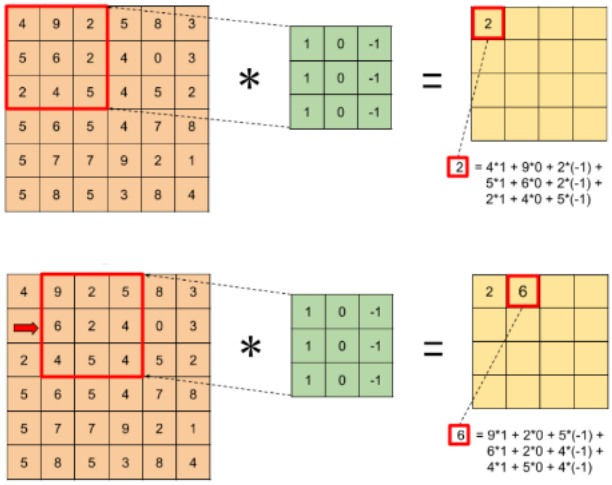
Les couches convolutives réduisent la taille de la sortie. Lorsque l'on veut conserver la taille de l'image d'entrée et conserver les informations présentées dans les coins, on peut utiliser des couches de remplissage qui permettent d'ajouter des lignes et des colonnes supplémentaires sur les bords extérieurs des images. Ainsi, la taille des données d'entrée restera similaire à celle des données de sortie (voir Fig. 3.9).

Enfin, un CNN classique est composée de couches entièrement connectées disposées à la fin du réseau et qui va permettre de classifier l'image.

Des méthodes ont été proposées pour augmenter les bases de données (translations, des réflexions horizontales sur des sous-parties d'images de la base de données, modification des intensités de chacun des canaux de couleur rouge, vert et bleu) [267].



(a) Exemple de Pooling



(b) Exemple de convolution

FIGURE 3.8 – Principe de fonctionnement d'un CNN

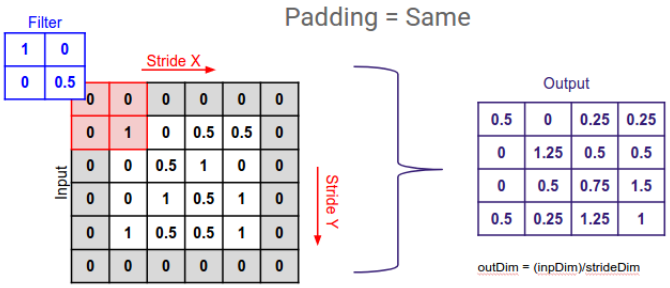


FIGURE 3.9 – Zero padding

En résumé, le fonctionnement des CNNs repose donc sur plusieurs opérations :

— Convolution : cette couche applique à l'image d'entrée une série de filtres convolutifs (ou *kernel*, noyaux) afin d'apprendre des caractéristiques spécifiques. Chaque filtre glisse à travers l'image et il est utilisé pour effectuer une opération de convolution, générant une nouvelle image filtrée appelée carte de caractéristiques. L'opération de convolution est une multiplication matricielle par éléments, où le premier terme est l'image et l'autre le filtre.
— Fonction d'activation type ReLU, sigmoïde ou Tanh : on calcule les valeurs d'activation de chaque entrée des cartes, des caractéristiques obtenues. Cela permet une phase d'apprentissage non linéaire.
— Pooling : permet de réduire le nombre de paramètres du réseau en appliquant un sous-échantillonnage non linéaire sur les cartes de caractéristiques
— Couche dense entièrement connectée : où chaque neurone reçoit une entrée

de tous les neurones de la couche précédente (c'est-à-dire qu'il est densément connecté). Cette couche donne un vecteur de dimension K, où K est le nombre de classes considérées, et qui contient l'analyse des caractéristiques de haut niveau générées par les couches précédentes.

— Softmax : afin d'obtenir la classification finale, une fonction type softmax prend en entrée un vecteur de dimension K et le normalise en une distribution de probabilité. La classe avec la probabilité la plus élevée représente la prédiction pour l'image d'entrée.

Le même principe peut être utilisé pour le traitement des données textuelles, même si d'autres types de réseau sont également employés [331, 158]. Les CNNs sont le type de réseaux de neurones le plus souvent utilisés en vision par ordinateur, mais beaucoup d'autres structures existent, comme les RNNs par exemple. Bien qu'efficace, le deep learning impose la contrainte d'avoir une grande base de données labellisées pour effectuer son apprentissage supervisé.

3.4 Long Short-Term Memory

Les réseaux récurrents utilisent l'information précédente avant de traiter l'élément suivant dans une séquence linéaire. De façon générale, l'architecture d'un RNN présente une unité de réseau qui analyse, à un pas de temps générique t, une entrée x_t, et génère la sortie h_t. En outre, l'unité A transmet l'information d'un pas de temps à l'autre, générant une boucle. Ainsi un RNN peut être représenté comme une chaîne d'unités, où chaque unité transmet des informations à son successeur, comme le montre la figure 3.10 : soient $x_1, \ldots, x_t \in \mathrm{R}_n$ des données d'entrées, le réseau calcule alors des sorties $h_1, \ldots, h_t \in \mathrm{R}_n$ selon l'équation suivante :

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{3.2}$$

où h_t représente l'état à l'instant t pour un neurone caché, σ indique la fonction d'activation sigmoïde, W_{xh} est une matrice de poids utilisée entre les couches d'entrée et cachées, W_{hh} est une matrice de poids au niveau des couches récursives, W_{hy} désigne une matrice de poids entre les couches cachées et de sortie, et enfin le vecteur b_h est le biais.

Pour modéliser des données séquentielles, les RNNs conservent un état caché qui résume leur historique que l'on peut modéliser sous forme de cycle sur un graphe de neurones. Ce type d'architectures permet de traiter des séquences de vecteurs d'entrée, et non pas seulement des données isolées n'ayant pas de signification temporelle. L'avantage des RNNs réside dans leur capacité à prendre en compte le contexte passé lors du traitement de l'information. Cependant, ce type de réseaux a des difficultés à analyser de longues séquences. En effet, l'erreur obtenue avec la rétropropagation du gradient décroît (on parle de dissipation du gradient), ou au contraire augmente d'une manière exponentielle (on parle alors d'explosion du gradient) [24].

Une nouvelle version des RNNs a donc été développée pour éviter ces inconvénients : les LSTMs (*Long Short Term Memory*. La cellule LSTM est caractérisée par un nœud central, contenant l'état (ou mémoire) interne de la cellule, et 3 « portes » (gates). Ces portes permettent de gérer la mémoire de l'information séquentielle

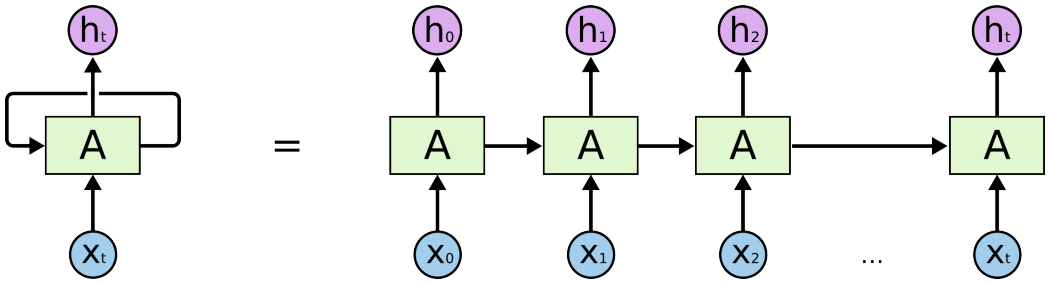


FIGURE 3.10 – Exemple de RNN pliés (à gauche) et dépliés (à droite)

(portes d'entrée et d'oubli) et également l'état interne de chaque sortie (porte de sortie) [24].

L'architecture LSTM, contrairement aux RNNs classique, a donc été conçue pour tirer parti des dépendances à long terme, comme son nom l'indique. Un LSTM capable de retenir des informations (appelées état de la cellule) sur de longues périodes de temps, en utilisant des « portes » dédiées à l'intérieur des unités individuelles. L'innovation majeure du LSTM consiste donc à transmettre l'état de la cellule directement à l'unité suivante via des opérations linéaires, en choisissant les informations à conserver ou à supprimer par le biais de portes dédiées :

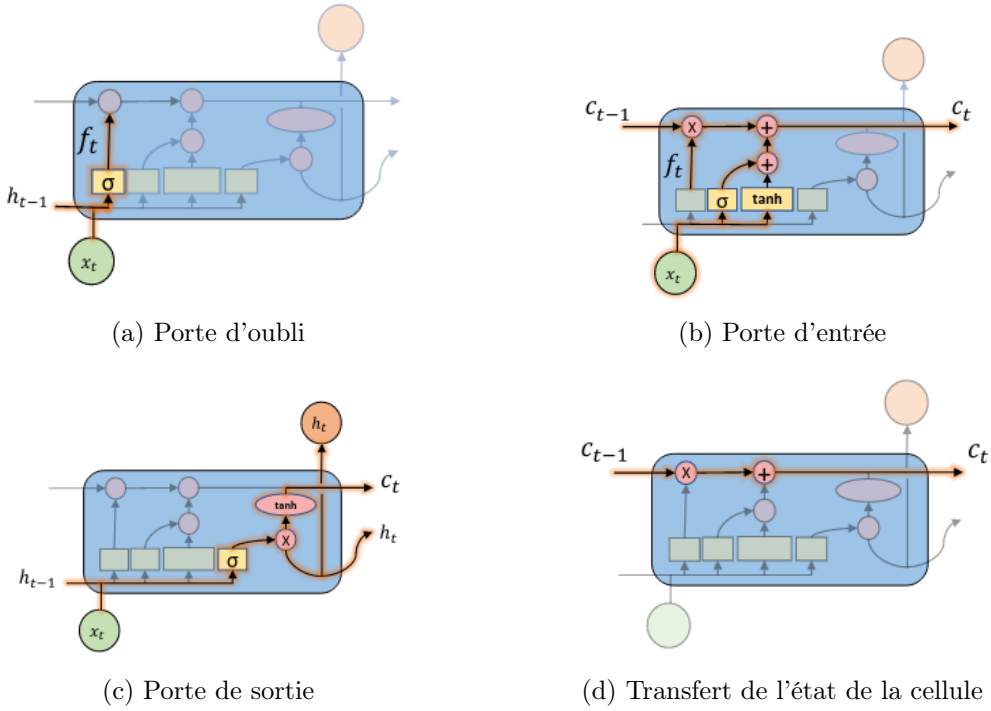
— "forget gate" (porte d'oubli) : elle décide si l'information, provenant de l'entrée x_t et de l'état caché précédent h_{t-1}, doit être conservée dans l'état de la cellule. Une fonction sigmoïde σ est appliquée à x_t et h_t, fournissant soit 0, soit 1 comme sortie. Lorsqu'un 1 est retourné, l'information est conservée dans la cellule d'état, sinon elle est oubliée.

— "input gate" (porte d'entrée) : La porte d'entrée décide des informations qui peuvent être stockées dans l'état de la cellule. Deux chemins différents sont implémentés à cet effet, où le premier applique une fonction sigmoïde pour décider de ce qui doit être mis à jour, tandis que le second utilise une fonction tanh pour créer un vecteur contenant les nouvelles valeurs candidates pour le nouvel état. Les sorties de ces deux chemins sont ensuite multipliées et combinées avec l'état précédent c_{t-1} et le vecteur f_t obtenu par la porte d'oubli.

— « output gate » (porte de sortie) : Cette porte est utilisée pour décider de la sortie de l'unité, et est basée sur une version filtrée de l'état actualisé de la cellule. Une fonction sigmoïde est utilisée pour déterminer quelles composantes de l'état de la cellule doivent être sorties, tandis qu'une fonction tanh est appliquée à l'état cellulaire actualisé pour normaliser les valeurs entre [-1, 1]. Les deux sorties sont ensuite multipliées et seules les parties choisies seront utilisées comme sortie pour une unité donnée.

Les LSTMs ont montré leur efficacité dans de nombreuses applications et sont considérés comme l'approche état-de-l'art dans diverses applications comme le TAL ou encore la traduction automatique.

À partir des RNNs, des modèles de type **seq2seq** (pour *sequence to sequence*) ont été développés dans le domaine du traitement du langage naturel [281]. De manière générale, l'objectif de ces modèles est de transformer une séquence d'entrée en une nouvelle séquence. Parmi les exemples de tâches réalisées par des modèles de type



(a) Porte d'oubli

(b) Porte d'entrée

(c) Porte de sortie

(d) Transfert de l'état de la cellule

FIGURE 3.11 – LSTM

seq2seq, on peut citer la traduction automatique entre plusieurs langues, en texte ou en audio, ou encore la génération de dialogues de type question-réponse. Un modèle seq2seq a généralement une **architecture encodeur-décodeur**, composée de :

- Un encodeur qui traite la séquence d'entrée et compresse l'information dans un vecteur de contexte d'une longueur fixe. Cette représentation est censée être un bon résumé de la signification de l'ensemble de la séquence source.
- Un décodeur qui est initialisé avec le vecteur de contexte pour émettre la sortie transformée. Les premiers travaux n'utilisaient que le dernier état du réseau de l'encodeur comme état initial du décodeur.
- L'encodeur et le décodeur sont tous deux des réseaux neuronaux récurrents.

3.5 Transformer

Un inconvénient majeur de ce type d'architecture seq2seq est que le vecteur de contexte est de longueur fixe, et donc il est dans l'incapacité de se souvenir de longues phrases. Souvent, il en a oublié la première partie une fois qu'il a terminé de traiter l'ensemble de l'entrée. Le **mécanisme d'attention** a été développé pour résoudre ce problème et aider à mémoriser de longues phrases sources dans la traduction automatique neuronale (*Neural Machine Translation* ou NMT). Plutôt que de construire un vecteur de contexte unique à partir du dernier état caché de l'encodeur, le mécanisme d'attention consiste à créer des liens entre le vecteur de contexte et l'entrée source entière. Les poids de ces connexions sont adaptables pour chaque élément de sortie. Comme le vecteur de contexte a accès à la totalité de la séquence d'entrée, il n'y a plus de phénomène d'oubli. L'alignement entre la source et la cible est appris et contrôlé par le vecteur de contexte [12].

Le vecteur de contexte consomme trois éléments d'information :
— les états cachés de l'encodeur,
— les états cachés du décodeur,
— l'alignement entre la source et la cible.

Au lieu de transmettre la séquence de sortie de la couche précédente directement à l'entrée de la couche suivante, le mécanisme d'attention est un intermédiaire déterminant quels éléments de l'entrée sont pertinents pour un élément particulier de la sortie. Grâce à l'attention, les dépendances entre les séquences source et cible ne sont plus limitées par la distance entre les deux. Compte tenu de la grande amélioration de l'attention dans la traduction automatique, elle s'est rapidement étendue au domaine de la vision par ordinateur [124] et les chercheurs ont commencé à explorer diverses autres formes de mécanismes d'attention [189, 29, 209].

Le Transformer est également un modèle de type encoder-decoder (voir Fig. 3.12) [296]. L'encoder est composé de 6 couches identiques, où chaque couche est composée d'un système d'auto-attention à têtes multiples et d'un simple réseau de type feed-forward entièrement connecté en fonction de la position. Chaque couche adopte une connexion résiduelle et une normalisation de couche. Toutes les couches produisent des données de la même dimension.

L'encodeur génère une représentation basée sur l'attention avec la capacité de localiser un élément d'information spécifique dans un contexte large. Il se compose d'une pile de 6 modules, chacun contenant deux sous-modules, d'une couche d'auto-attention à têtes multiples et d'un réseau *feed-forward* entièrement connecté par points. Le Transformer applique la même transformation linéaire (avec les mêmes poids) à chaque élément de la séquence. On peut également considérer cela comme une couche convolutionnelle avec un filtre de taille 1. Chaque sous-module possède une connexion résiduelle et une normalisation de la couche. Tous les sous-modules produisent des données de la même dimension d (dans l'article originel, $d = 512$).

La fonction du décodeur du Transformer est de récupérer les informations de la représentation codée. L'architecture est assez similaire à celle de l'encodeur, sauf que le décodeur contient deux sous-modules d'attention multitêtes au lieu d'un. Le premier sous-module d'attention multitêtes est masqué afin d'empêcher les positions de s'occuper du futur.

L'attention peut donc être interprétée comme un vecteur de poids d'importance qui permet d'estimer à quel point chaque élément d'une séquence est corrélé avec les autres éléments. Les données d'entrée sont appelées **valeurs** V. Un mécanisme (entraînable) attribue une **clé** K à chaque valeur. Ensuite, à chaque sortie, un autre mécanisme attribue une **requête** Q. Ces notions de *query, key, et value* sont des abstractions pour modéliser l'attention. Ces noms viennent d'un type de structure de stockage de données appelé clé-valeur. Le mécanisme d'attention repose sur un produit scalaire : la sortie est une somme pondérée des valeurs, où le poids attribué à chaque valeur est déterminé par le produit scalaire de la requête avec toutes les clés. Le calcul de l'attention consiste donc à calculer un score. Le score détermine le degré d'attention à accorder aux autres parties de la phrase d'entrée lorsque nous encodons un mot à une certaine position. Le score est calculé en prenant le produit scalaire

du vecteur de la requête avec le vecteur clé du mot respectif que nous évaluons.

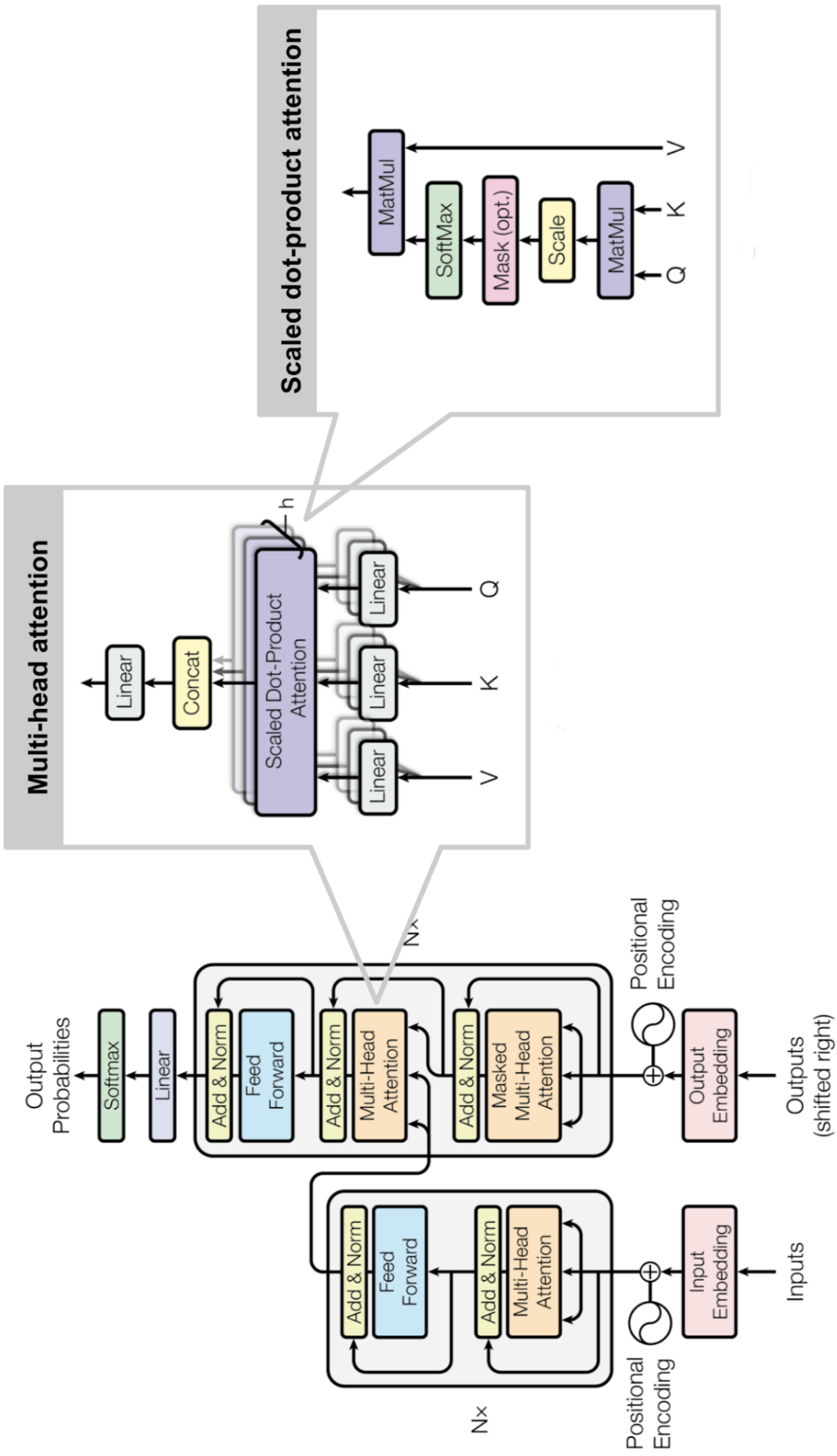


FIGURE 3.12 – Transformer et mécanisme d'attention multitêtes [296]

3.6 Réseaux de neurones à graphes

Les réseaux neuronaux à graphes (GNN) ont été introduits en 2005, et ont commencé à gagner en popularité au cours des cinq dernières années. Les GNN sont capables de modéliser la relation entre les nœuds d'un graphe et d'en produire une représentation numérique. L'importance des GNN est considérable, car il existe un grand nombre de données du monde réel qui peuvent être représentées sous forme de graphe : les réseaux sociaux, les composés chimiques, les cartes, les systèmes de transport, pour n'en citer que quelques-uns. Les graphes sont des structures de données utilisées pour décrire des systèmes complexes, des entités qui ont des interactions entre elles. Un graphe est une collection d'objets (c'est-à-dire des nœuds), ainsi qu'un ensemble d'interactions (c'est-à-dire des arêtes) entre des paires de ces objets. De manière générale, les graphes permettent de décrire et représenter des entités qui ont des interactions entre elles. Par exemple, des graphes peuvent être utilisés pour représenter des réseaux de transports, des réseaux sociaux, des molécules, ou encore des formes 3D. Les graphes peuvent être utilisés pour différentes taches : prédiction sur les nœuds ou les arêtes, sur le graphe en entier, classification, génération de nouveaux graphes, etc

La puissance du formalisme graphique réside à la fois dans sa focalisation sur les relations entre les nœuds (plutôt que les propriétés des nœuds individuels), ainsi que dans sa généralité. Ainsi, le même graphe peut être utilisé pour représenter les réseaux sociaux, les interactions entre les médicaments et les protéines, les interactions entre les atomes d'une molécule, par exemple. Les graphes ne se contentent toutefois pas de fournir un cadre théorique, ils offrent une base mathématique pour analyser et comprendre des systèmes complexes du monde réel [125]. L'apprentissage automatique est un des moyens d'analyser les données de ces graphes.

Pour rappel, comme expliqué dans [2] on définit un graphe G par un couple $G = (V, E)$ tel que :

- V est un ensemble fini de sommets (ou *vertices* en anglais)
- E est un ensemble fini d'arêtes (ou *edges*) allant du nœud $u \in V$ au nœud $v \in V$ tel que $(u, v) \in E$
- A la matrice d'adjacence
- L'ordre d'un graphe est son nombre de sommets
- Une boucle est une arête reliant un sommet à lui-même, un graphe dépourvu de boucle est dit élémentaire
- Un graphe simple ne comporte pas de boucle et a au plus une arête entre deux sommets
- un sommet V_i est dit adjacent à un autre s'il existe une arête entre eux, on parle alors de voisins
- le degré d'un sommet est le nombre d'arêtes incidentes à ce sommet
- un graphe est dit complet s'il comporte une arête (V_i, V_j) pour toute paire de sommets $(V_i, V_j) \in E$
- un graphe est dit orienté si les couples $(V_i, V_j) \in E$ sont ordonnées, V_i étant le sommet initial et V_j le sommet terminal, le couple (V_i, V_j) est alors un arc représenté par $V_i \rightarrow V_j$;
- un graphe est non orienté si les couples (V_i, V_j) et (V_j, V_i) sont équivalents,

V_i — V_j est alors une arête.

Une manière de représenter les graphes est d'utiliser une matrice d'adjacence $A \in R^{|v|\times|v|}$. Pour représenter un graphe avec une matrice d'adjacence, nous ordonnons les nœuds dans le graphe de sorte que chaque nœud indexe une ligne et une colonne particulières dans la matrice d'adjacence. On peut représenter la présence d'arêtes A dans cette matrice telle que $A[u, v] = 1$ si $(u, v) \in E$ et $A[u, v] = 0$ sinon. Certains graphes peuvent également avoir des arêtes pondérées, où les entrées de la matrice d'adjacence sont des valeurs réelles plutôt que 0, 1. Par exemple, une arête pondérée dans un graphe d'interaction protéine-protéine peut indiquer la force de l'association entre deux protéines.

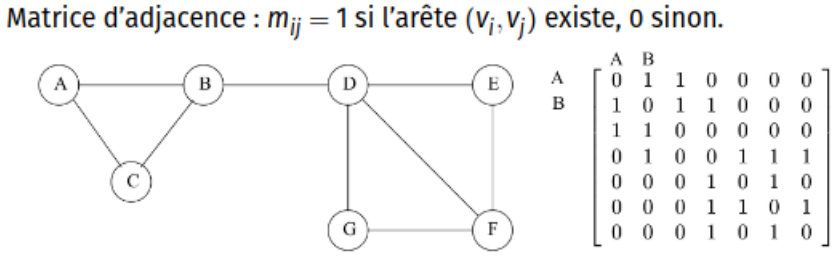


FIGURE 3.13 – Matrice d'adjacence [2]

Les réseaux de neurones classiques type CNN sont conçus pour traiter des données avec des structures simples, de tailles fixes, comme des images (que l'on peut considérer comme des grilles 2D avec une géométrie euclidienne) ou du texte (des chaînes linéaires). Appliquer des réseaux de neurones à des structures plus complexes comme les graphes sont un challenge. En effet, les graphes peuvent être de différentes tailles et ne présentent pas de régularité locale comme les images, il n'y a pas d'ordre pour les nœuds ni points de référence (pas de haut-bas-droite-gauche comme dans une image), et ils peuvent être dynamiques et avoir plusieurs caractéristiques assignées aux nœuds et aux arêtes. De plus, les graphes sont invariants par permutation. On ne peut donc pas directement appliquer les mêmes transformations (comme les convolutions) que sur des images. L'idée derrière les réseaux de neurones à graphes est donc de fusionner/concaténer les matrices d'attributs et d'adjacence afin de les donner en entrée à un réseau de neurones.

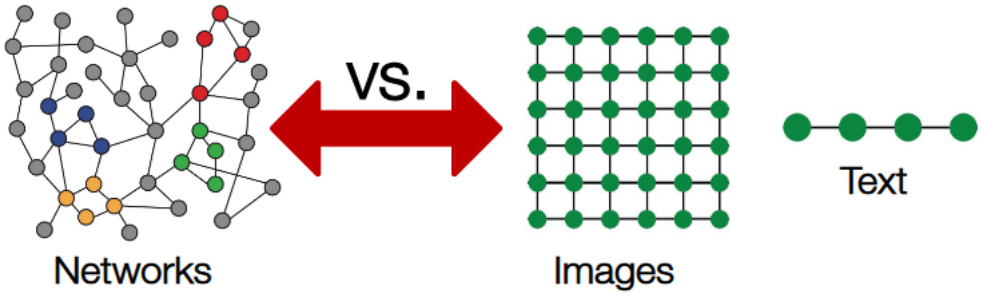


FIGURE 3.14 – Exemple de graphe [1]

Comme dans tout système de machine learning, il s'agit donc d'extraire les caractéristiques pour les nœuds et les arêtes pour entraîner un modèle pour faire de

nouvelles prédictions. Les GNN fonctionnent ensuite grâce au principe d'agrégation et de transmission des messages (« message passing ») : il s'agit de transformer et propager l'information contenue dans le voisinage de nœuds en nœuds. Ainsi pour chaque nœud, on définit un graphe de calcul basé sur son voisinage qui va permettre de propager et transmettre l'information.

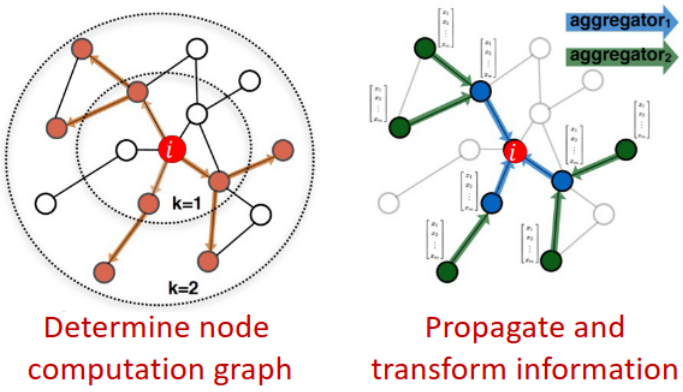


FIGURE 3.15 – Transmission du message dans un GNN [1]

Comme le montre la figure 3.15, le principe du passage de messages du GNN est simple : à chaque itération, chaque nœud agrège les informations provenant de son voisinage local, et au fur et à mesure que ces itérations progressent, chaque nœud intégré contient de plus en plus d'informations provenant d'autres parties du graphe. Plus précisément, après la première itération ($k = 1$), chaque nœud incorporé contient des informations de son voisinage à un saut, c'est-à-dire que chaque intégration de nœud contient des informations sur les caractéristiques de ses voisins immédiats, qui peuvent être atteints par un chemin de longueur 1 dans le graphe. Après la deuxième itération ($k = 2$), chaque nœud contient des informations sur son voisinage à deux sauts ; après k itérations, chaque intégration de nœud contient des informations sur son voisinage à k sauts [1].

En général, les informations se présentent sous deux formes. D'une part, il existe des informations structurelles sur le graphe. Par exemple, après k itérations de *message passing*, l'intégration $h_u^{(k)}$ du nœud u peut coder des informations sur les degrés de tous les nœuds dans le voisinage à k sauts de u. En plus de l'information structurelle, l'autre type d'information clé capturée par l'intégration des nœuds du GNN est l'information basée sur les caractéristiques. Après k itérations de passage de messages, les intégrations de chaque nœud encodent également des informations sur toutes les caractéristiques dans leur voisinage à k sauts. Cette agrégation locale de caractéristiques des GNN est analogue au comportement des noyaux convolutifs des CNN : dans les réseaux neuronaux convolutifs (CNN). Cependant, alors que les CNNs agrègent l'information sur les caractéristiques à partir de tâches définies dans l'espace d'une image, les GNNs agrègent les informations basées sur des voisinages de graphes locaux [126].

Pour l'apprentissage, chaque nœud extrait l'information de tous ses voisins, calcule leur somme et les transmet à l'unité suivante. Cette intégration contient les informations du nœud et celles de tous ses voisins. Au pas de temps suivant, il

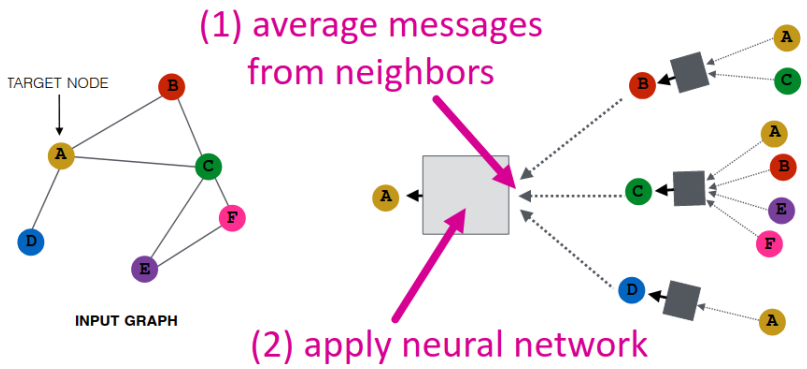


FIGURE 3.16 – Exemples d'agrégation [1]

contiendra également les informations de ses voisins de second ordre. Et ainsi de suite. Le processus se poursuit jusqu'à ce que chaque nœud connaisse tous les autres nœuds du graphe. Chacune des intégrations possède maintenant des informations de tous les autres nœuds. L'étape finale consiste à rassembler tous les enchâssements et à les additionner, ce qui nous donnera une seule intégration pour l'ensemble du graphe [1].

Graph Convolutional Network

Les Graph Convolutional Network ou GCNs généralisent les convolutions pour les données de type graphes. Toutefois, l'opérateur de convolution a dû être repensé. Introduits par Kipf et al., les réseaux convolutifs sur graphes ou GCN utilisent des convolutions similaires à celles des CNN pour les images dans le sens où les paramètres du filtre sont généralement partagés sur tous les emplacements du graphe. Les GCN reposent sur des méthodes dites de passage de messages, dans le sens où les sommets échangent des informations avec leurs voisins et s'envoient donc des « messages » de cette façon [160].

Dans un premier temps, chaque nœud crée un vecteur de caractéristiques qui représente le message qu'il veut envoyer à tous ses voisins. Dans la deuxième étape, les messages sont envoyés aux voisins, de sorte qu'un nœud reçoit un message par nœud adjacent. Comme le nombre de messages varie d'un nœud à l'autre, il faut une opération qui fonctionne pour n'importe quel nombre. Par conséquent, la façon habituelle de procéder est de faire la somme ou de prendre la moyenne. Étant donné les caractéristiques précédentes des nœuds $H^{(l)}$, un GCN est donc définie comme suit :

$$H^{(l+1)} = \sigma\left(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{3.3}$$

où $W^{(l)}$ est le paramètre de poids avec lequel on transforme les caractéristiques d'entrée en messages ($H^{(l)}W^{(l)}$). À la matrice d'adjacence A, on ajoute la matrice d'identité afin que chaque nœud s'envoie également son propre message : =A+I. Enfin, pour prendre la moyenne au lieu de faire la somme, on calcule la matrice qui est une matrice diagonale avec D_{ii} dénotant le nombre de voisins du nœud i. σ représente une fonction d'activation arbitraire, et pas nécessairement la sigmoïde

(on utilise généralement une fonction d'activation basée sur ReLU dans les GNN).

Il existe d'autres types d'architectures de GNN, dont les ST-GCN qui seront développés dans la section 6.4.2.

Conclusion

Dans ce chapitre, plusieurs concepts de deep learning ont été présentés, en commençant par une brève introduction sur le machine learning jusqu'aux architectures de réseaux de neurones profonds les plus courantes. L'utilisation du deep learning pour la reconnaissance d'objets dans les images a fait ses preuves, des recherches ont tenté d'étendre ces résultats à la reconnaissance d'action dans les vidéos. Les caractéristiques sont créées en utilisant les premières couches de cartes de caractéristiques et de sous-échantillonnage, tandis que la classification se fait sur les dernières couches du réseau de neurones. Le deep learning a introduit le concept d'apprentissage dit de bout en bout (« end to end »), où la machine reçoit un ensemble d'images qui sont annotées avec les classes d'objets présentes dans chaque image. Un modèle est ainsi entraîné sur ces données, où les réseaux de neurones analysent automatiquement les caractéristiques les plus descriptives concernant chaque classe d'objet. L'objectif de l'apprentissage automatique est donc de construire une représentation de l'image brute d'entrée de plus en plus haut niveau de couche en couche. On parle alors d'apprentissage profond. Les CNNs sont un type de réseaux de neurones souvent utilisés en vision par ordinateur, mais d'autres structures existent, comme les RNNs par exemple.

Le deep learning est désormais utilisé dans le domaine du traitement numérique des images pour résoudre des problèmes difficiles (par exemple la colorisation, la classification, la segmentation et la détection des images). Les méthodes d'apprentissage profond telles que les CNNs améliorent principalement les performances de prédiction en utilisant des grandes quantités de données.

Toutefois, le deep learning impose la contrainte d'avoir une grande base de données labellisée. Des méthodes ont été proposées pour augmenter les bases de données grâce à différentes méthodes comme des translations, des réflexions horizontales, ou encore des modifications des intensités des couleurs.

Chapitre 4

Analyse de mouvements par Deep Learning

Sommaire

L'objectif des systèmes d'analyse automatique de mouvements est de reconnaître les actions d'un individu à partir des données brutes obtenues par des capteurs ou des caméras. Comme dans de nombreux domaines en vision par ordinateur, l'apprentissage profond s'est aussi récemment appliqué à la reconnaissance d'action, ce qui a permis d'atteindre des résultats qui ont dépassé les méthodes antérieures [272, 307, 98]. L'analyse du comportement humain par deep learning a de nombreux domaines d'application de cette technologie : l'interaction homme-machine et robotique [322], la sécurité et la surveillance vidéo [299], la e-Santé [204], les diagnostics médicaux [251], la reconnaissance de la langue des signes [225] ou les jeux vidéos [193].

Ainsi, on peut trouver dans la littérature des travaux de recherche portant sur différents aspects de l'activité de l'analyse du comportement humain : reconnaissance des gestes et d'actions [272, 98], modélisation des interactions sociales [70, 140], l'analyse des émotions faciales [9] et l'identification des traits de personnalité [149].

4.1 Squelettisation par deep learning

Comme expliqué précédemment, la squelettisation pour l'étude dynamique du corps humain permet de traiter des informations importantes pour la reconnaissance d'action. La composante dynamique du squelette peut être naturellement représentée par une série chronologique de positions articulaires humaines, sous la forme de coordonnées 2D ou 3D. Les actions humaines peuvent alors être reconnues en analysant les schémas de mouvement de celles-ci. Les méthodes classiques sont limitées, car elles n'exploitent pas explicitement les relations spatiales naturelles qui existent entre les articulations qui sont cruciales pour comprendre les actions humaines.

Il existe plusieurs algorithmes de détection de pose en temps réel basé sur des CNNs tels que Openpose [1] [33] ou AlphaPose [2] [95, 324, 173, 174].

4.1.1 DeepPose

Toshev et Szegedy ont publié en 2014 un article fondateur dans l'application de l'apprentissage profond à l'estimation de la pose humaine [289]. Leur modèle *Deep Pose* a dépassé la performance de l'état de l'art de l'époque. Dans cette approche, l'estimation de la pose est formulée comme un problème de régression. Dans cette approche, les auteurs cherchent à détecter la pose d'une manière globale, c'est-à-dire que même si certaines articulations sont cachées, elles peuvent être quand même estimées.

Deep Pose est basée sur le réseau de neurones AlexNet (avec 7 couches). Ce modèle affine les prédictions à l'aide de régresseurs en cascade. La pose grossière initiale est affinée et une meilleure estimation est obtenue. Les images sont recadrées autour de l'articulation prédite et transmises à l'étape suivante, de cette façon les régresseurs de pose suivants voient des images à plus haute définition et apprennent ainsi des caractéristiques pour des échelles plus fines, ce qui conduit finalement à

1. https ://github.com/CMU-Perceptual-Computing-Lab/openpose
2. https ://github.com/MVIG-SJTU/AlphaPose

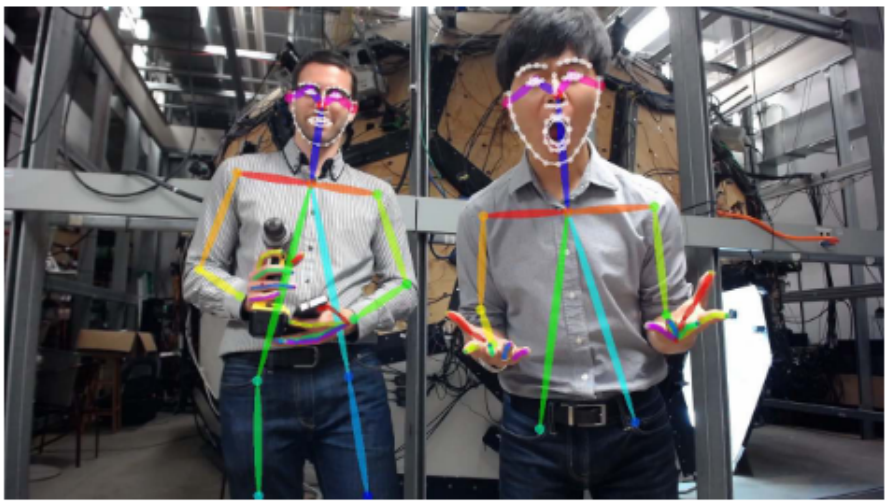


FIGURE 4.1 – Exemple de squelettisation par deep learning par OpenPose[33]

FIGURE 4.2 – Modèle DeepPose [289]

une plus grande précision. Cet article a été un des premiers à utiliser l'apprentissage profond pour l'estimation de la posture humaine et a donné le coup d'envoi à la recherche dans cette direction. D'autres modèles ont ensuite rapidement suivi.

4.1.2 Carte de chaleurs

Les architectures CNNs traditionnelles comprennent des couches de mise en commun (*pooling*) et de sous-échantillonnage (*sub-sampling*) qui réduisent les calculs, augmentent l'invariance et empêchent le surentraînement. Ces avantages de la mise en commun se font au prix d'une réduction de la précision de la localisation des points d'intérêts. Dans *Efficient Object Localization Using Convolutional Networks*[286], les auteurs proposent une nouvelle architecture qui inclut une méthode de « raffinement de la position » : le modèle génère des **cartes thermiques** (encore appelé cartes de chaleurs, ou *heatmaps*) en faisant passer une image dans plusieurs réseaux en parallèle pour capturer simultanément des caractéristiques à différentes échelles. La sortie est une carte thermique discrète au lieu d'une régression continue. Une carte thermique prédit la probabilité que l'articulation soit présente dans un pixel donné [286].

Ce modèle est très efficace et bon nombre des articles qui ont suivi prédisent des cartes thermiques plutôt qu'une régression directe. Ainsi, dans *Convolutional Pose Machines*[313], les auteurs définissent ce qu'ils appellent une **pose machine**, qui se compose d'un module de calcul d'éléments d'image suivi d'un module de prédiction. L'une des principales motivations est d'apprendre les relations spatiales à long terme

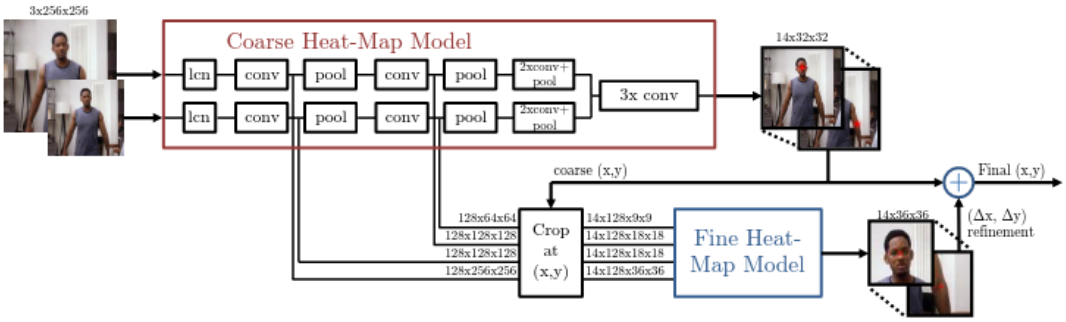


FIGURE 4.3 – Vue générale du modèle [286]

et les auteurs montrent que l'on peut y parvenir en utilisant des champs réceptifs plus larges. Les auteurs ont utilisé une supervision intermédiaire après chaque étape pour éviter le problème de disparition des gradients, qui est un problème courant pour les réseaux profonds. Carreira et al. ont proposé une méthode appelée *Iterative Error Feedback* (IEF) où au lieu de prédire directement les résultats en une seule fois, les auteurs utilisent un modèle autocorrectif qui modifie progressivement une solution initiale en renvoyant les prédictions d'erreurs [36].

FIGURE 4.4 – Séquence de cartes de chaleurs d'une *pose machine* [313]

4.1.3 Réseaux en sabliers

Il existe d'autres architectures pour l'estimation de la pose. *Stacked Hourglass Networks for Human Pose Estimation*[208] est, lui aussi, un article de référence qui a introduit les réseaux de type **hourglasses**. Il s'agit d'un réseau en forme de sabliers empilés puisque le réseau consiste en des étapes de mise en commun et de suréchantillonnage des couches qui ressemblent à un sablier, et qui sont empilées ensemble. La conception du sablier est motivée par la nécessité de saisir l'information à toutes les échelles. Bien que les informations locales soient essentielles pour identifier des caractéristiques comme les mains des visages, une estimation finale de la pose nécessite un contexte global. L'orientation de la personne, la disposition de ses membres et les relations des articulations adjacentes sont parmi les nombreux indices qui sont mieux reconnus à différentes échelles dans l'image (les petites résolutions capturent les caractéristiques d'ordre supérieur et le contexte global).

Le réseau effectue des traitements ascendants (de la haute résolution à la basse résolution) et descendants (de la basse résolution à la haute résolution) répétés avec supervision intermédiaire. Le réseau utilise des connexions de saut pour préserver l'information spatiale à chaque résolution et la transmet à l'échantillonnage

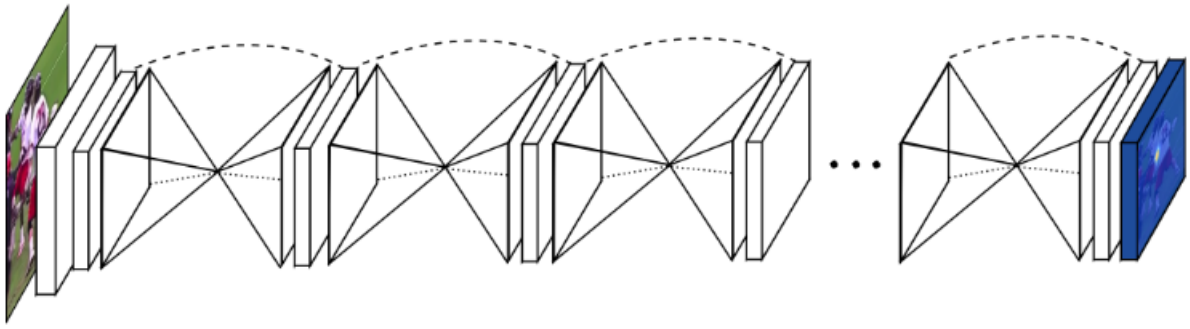


FIGURE 4.5 – Principe des réseaux en sabliers [208]

supérieur, plus bas dans le sablier. La supervision intermédiaire est appliquée aux prédictions de chaque stade de sablier, c'est-à-dire que les prédictions de chaque sablier de la pile sont supervisées, et pas seulement les prédictions finales du sablier.

4.1.4 HRNet

Le modèle HRNet (*High-Resolution Network*) a surpassé toutes les méthodes existantes pour les tâches de détection de points-clés, d'estimation de la position de plusieurs personnes et d'estimation de la position dans l'ensemble de données COCO. HRNet suit une idée simple : alors que la plupart des articles précédents sont passés d'une représentation à haute résolution à une représentation à basse résolution, HRNet maintient une représentation à haute résolution tout au long du processus [154].

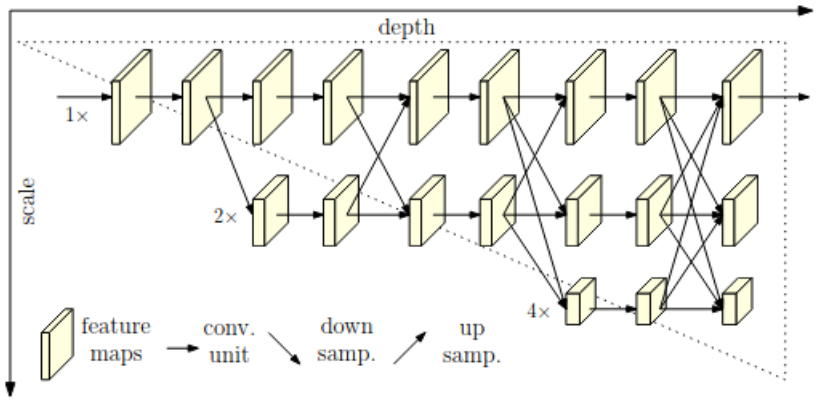


FIGURE 4.6 – Principe de HRNet [154]

L'architecture commence à partir d'un réseau à haute résolution comme première étape, et ajoute progressivement des réseaux à haute et basse résolution un par un pour former plusieurs étages et connecter les réseaux en parallèle. Des fusions multiéchelles répétées sont effectuées par l'échange d'informations à travers des réseaux parallèles multirésolution à maintes reprises tout au long du processus. Un autre avantage est que cette architecture n'utilise pas de supervision intermédiaire de la carte thermique, contrairement aux réseaux en sabliers précédents.

4.2 Reconnaissance d'actions par réseaux convolutionnels

Comme expliqué précédemment, une action ou un geste est un processus dynamique. La dimension temporelle est donc primordiale pour modéliser des actions et des comportements complexes. Dans ce contexte, les auteurs ont proposé plusieurs stratégies, telles que le sous-échantillonnage ou l'agrégation de caractéristiques locales (voir section 3), ou encore la modélisation de séquences temporelles [272, 98]. Pour ce faire, les chercheurs ont notamment utilisé les réseaux neuronaux récurrents. Aujourd'hui, les LSTMs sont une partie importante des modèles de réseaux de neurones profonds pour la modélisation des séquences d'images en reconnaissance d'actions [145, 184]. Toutes ces méthodes ont contribué à améliorer l'état de l'art pour la reconnaissance des actions et des gestes. Asadi-Aghbolaghi et al. ont ainsi proposé une taxonomie de la reconnaissance de gestes et d'action par deep learing [195] en quatre groupes (voir Fig. 6.16) :

— Le premier groupe est constitué des **CNN 2D**, qui sont capables d'exploiter l'information spatiale.
— Dans les méthodes du second groupe, en plus des images RGB, on extrait les caractéristiques de mouvement 2D à partir du **flux optique** (*optical flow*). Ces caractéristiques sont ensuite utilisées dans un deuxième réseau de neurones [272, 312, 113]. Ces réseaux sont donc dits **réseaux à deux flux** (*two stream networks*).
— Le troisième groupe utilise des **CNN 3D** [145, 182] qui ajoutent la dimension temporelle aux CNNs.
— Enfin, le quatrième groupe combine des réseaux convolutionnels (2D ou 3D) avec en plus une modélisation de séquence temporelle grâce aux **RNNs** [172, 340].
— En parallèle de ces méthodes basée sur le traitement des images RGB, les méthodes basées sur les **squelettes** se sont, elles aussi, développées avec les réseaux à graphes.

Ces différentes méthodes seront détaillées dans les sections suivantes.

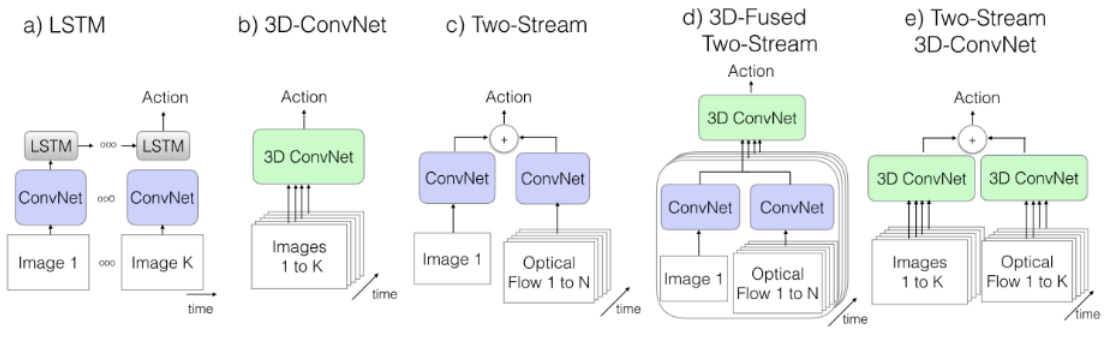


FIGURE 4.7 – Différentes méthodes pour la reconnaissance d'actions [153, 38]

Bases de données

Il existe de nombreuses bases de données internationales disponibles pour la reconnaissance d'actions. Ces bases de données diffèrent par le nombre de sujets humains (de quelques centaines à plusieurs millions), le bruit de fond, les variations d'apparence et de pose, le mouvement de la caméra, le nombre et le type d'actions (sportives, gestes du quotidien, interaction avec des objets), de vues (première ou troisième personne), de modalité (images RGB, RGBD, squelettes), et différents environnements (contrôlés ou non) [162] (voir le tableau 4.14 en annexe).

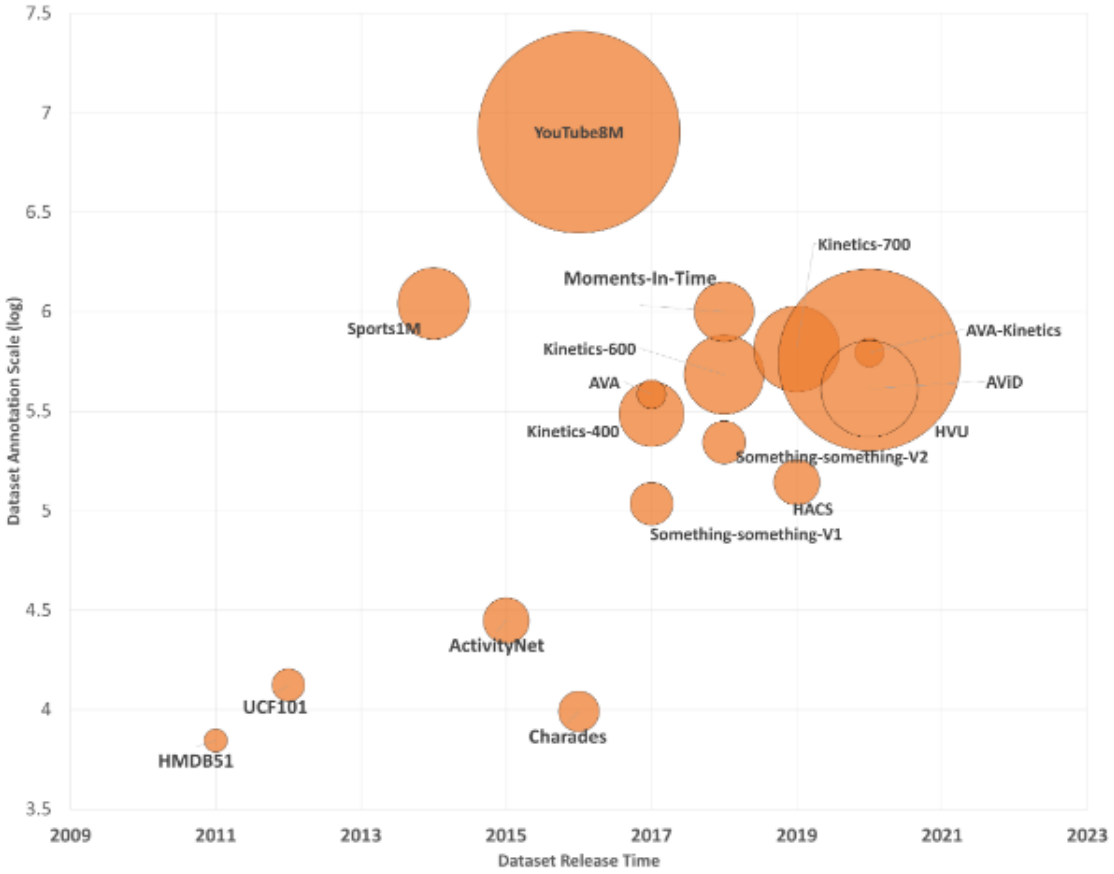


FIGURE 4.8 – Principales bases de données en HAR [162]

Comme on peut le voir sur la figure 4.8, le nombre de vidéos et de classes et de labels augmentent rapidement, par exemple, de 7000 vidéos avec 51 classes dans HMDB51 à 8 millions vidéos avec 3 862 classes dans **YouTube8M**. De même, le rythme auquel de nouveaux ensembles de données sont publiés est en augmentation : 3 jeux de données ont été publiés entre 2011 et 2015, contre 13 de 2016 à 2020. Sur ce graphique, une des plus importantes base de données est **Kinetics**, créé par l'équipe de Google DeepMind en 2017 pour fournir un moyen de former des modèles d'apprentissage machine conçus pour l'analyse et la classification de vidéos et d'actions. Cette base de données contient 400 classes différentes avec au moins 400 vidéos différentes pour chaque classe. Les données se concentrent principalement sur l'action humaine et sont divisés en plusieurs classes :

— La classe Personne-Personne se concentre principalement sur les interactions entre les groupes de personnes,

— la classe Personne se concentre sur les actions effectuées par une seule personne
— et la classe Personne-Objet qui capture l'interaction homme-objet.

Au sein de ces classes, Kinetics est encore divisé en sous classes. Chaque classe "parent" possède un label qui regroupe ses classes "enfants" par similarité. Par exemple, dans la classe Personne-objet, la classe parent « tissus » comprend les classes enfants suivantes : bandage, lessive, pliage des vêtements, pliage des serviettes, repassage, etc.

Comme développée dans [59], lors de la *Computer Vision and Pattern Recognition Conference* (CVPR) 2014, Karpathy et al. ont proposé une méthode basée sur des réseaux de neurones convolutifs pour reconnaître des actions dans des vidéos [152]. Différentes méthodes de fusion de données ont été testés par les auteurs pour prendre en compte plusieurs images afin d'entraîner leurs CNN. Les meilleurs résultats sont obtenus avec une *slow fusion* : les premières couches de convolution extraient des informations de 4 images consécutives, puis ces couches sont fusionnées entre elles. Ainsi, la couche résultante contient des informations provenant de 10 images successives. Néanmoins, les résultats obtenus par cette méthode ne sont pas meilleurs que ceux utilisant des méthodes de traitement d'images classiques [306]. D'autres approches échantillonnent une ou plusieurs images de la vidéo entière, puis appliquent un modèle 2D pré-entraîné sur des jeux de données d'images plus grands, tels que ImageNet [163] pour chacune de ces images, séparément. Enfin, ils classent les actions en faisant la moyenne des résultats [279, 309].

4.2.1 Temporal Segment Networks

Wang et al. ont cherché à améliorer l'architecture de Karpathy et al. précédente [309]. Ainsi, ils proposent d'échantillonner les clips de manière éparse dans la vidéo pour mieux modéliser le signal temporel au lieu de l'échantillonnage aléatoire dans toute la vidéo. Pendant l'entraînement et la prédiction, une vidéo est divisée en K segments de durée égale. Des extraits sont prélevés au hasard dans chacun des K segments. Pour la prédiction finale au niveau de la vidéo, les auteurs ont exploré plusieurs stratégies. La meilleure stratégie était de combiner séparément des dizaines de flux temporels et spatiaux (et d'autres flux si d'autres modalités d'entrée sont impliquées) en établissant une moyenne et de fusionner les scores spatiaux et temporels finaux en utilisant une moyenne pondérée et en appliquant la fonction softmax à toutes les classes (voir figure 4.9).

Une autre partie importante de ce travail consistait à résoudre le problème du surapprentissage (dû à la petite taille des jeux de données) et à démontrer l'utilisation de techniques désormais courantes comme la *batch normalization*, le *dropout* et le *pretraining* pour y remédier.

4.2.2 Reconnaissance d'actions par réseaux de neurones convolutifs 3D

Les convolutions 3D permettent de capturer des caractéristiques discriminantes des dimensions spatiales et temporelles, tout en maintenant la structure temporelle

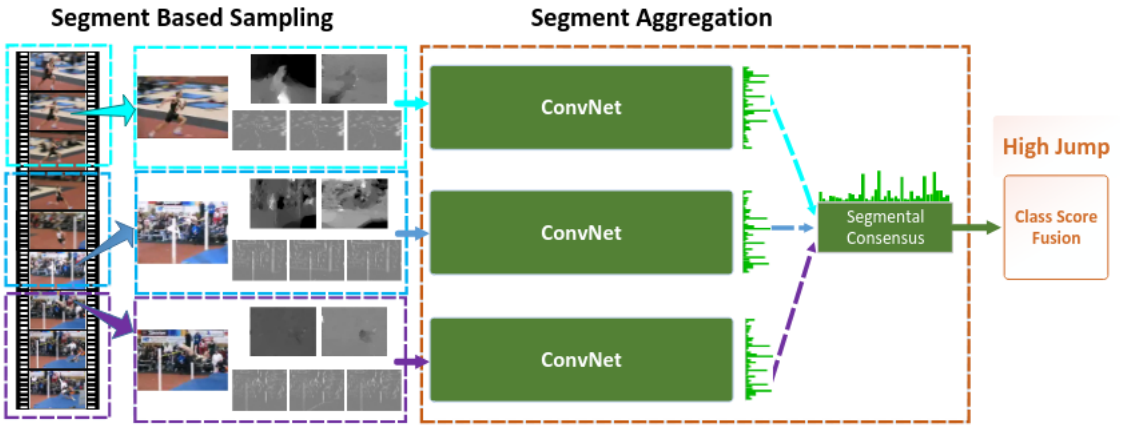


FIGURE 4.9 – Temporal Segment Networks [309]

en contraste à des couches convolutives 2D. Les caractéristiques spatio-temporales extraites par ce type de modèles semblent surpasser les modèles 2D formés sur les mêmes images. Ainsi en 2013, Ji et al. ont proposé une méthode basée sur des réseaux de neurones convolutifs 3D (Convolutional Neural Networks 3D ou 3D-CNN) pour reconnaître des actions filmées par des caméras de vidéos surveillances [146]. Cette méthode a dépassé l'état de l'art basé sur les CNNs 2D.

Les 3D-CNNs [292] sont donc similaires aux réseaux convolutionnels traditionnels en 2D, mais avec des filtres spatio-temporels : 2 dimensions spatiales plus une dimension temporelle). Ils permettent de créer directement des représentations hiérarchiques de données spatio-temporelles. Cependant, il y a donc beaucoup plus de paramètres que pour les CNNs 2D, les réseaux CNNs 3D sont donc plus difficiles à entraîner. De plus, il y a moins de modèles pré-entraînés disponibles avec ce type de réseaux, il faut donc souvent entraîner à partir de zéro.

Une autre approche intéressante est celle de Tran et al. [291] : le principe de cette méthode est de trouver un intermédiaire entre les CNN classiques qui fonctionnent très bien sur les images, et les 3D-CNN qui fonctionnent bien pour les vidéos, mais avec un coût de calcul très important et ne permettent pas d'utiliser le pré-entrainement. Pour cela, les auteurs utilisent des blocs constitués d'une convolution spatiale 2D suivie d'une convolution temporelle 1D (voir figure 4.10).

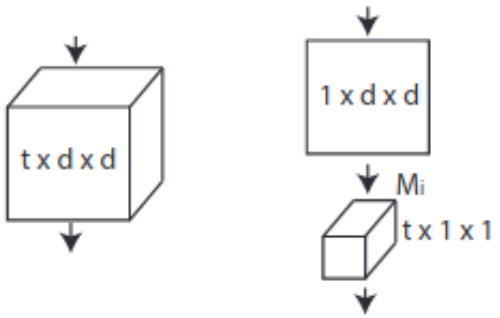


FIGURE 4.10 – convolution 2D+1D et convolution 3d

Sur ce schéma, une convolution 3D est effectuée à l'aide d'un filtre de taille $t \times$

$d \times d$ où t désigne l'étendue temporelle et d la largeur et la hauteur spatiales. En revanche, un bloc convolutionnel (2+1)D est composé d'une convolution 2D spatiale suivie d'une convolution 1D temporelle.

4.2.3 Reconnaissance d'actions par réseaux à deux flux

Simonyan et al. ont proposé une architecture fusionnant deux CNNs : un CNN spatial entraîné sur des images RGB classiques et un CNN temporel entraîné sur des flux optiques [272]. Cette méthode permet d'avoir des informations sur le mouvement ainsi que sur l'apparence. La fusion de données des sorties des deux CNN se faitpar SVM ou bien par moyennage. Les auteurs s'appuient sur les travaux précédents de Karpathy et al. présentés précédemment (voir section 4.2). Étant donné la difficulté des architectures profondes à apprendre les caractéristiques de mouvement, les auteurs ont explicitement modélisé les caractéristiques de mouvement sous la forme de vecteurs de flot optique. Ainsi, au lieu d'un réseau unique pour le contexte spatial, cette architecture comporte deux réseaux distincts : un pour le contexte spatial (préformé), et un pour le contexte de mouvement. L'entrée dans le réseau spatial est une seule image de la vidéo.

Les auteurs ont expérimenté l'entrée dans le réseau temporel et ont constaté que le flux optique bidirectionnel empilé sur 10 images successives était le plus performant. Les deux flux ont été formés séparément et combinés en utilisant un SVM. La prédiction finale est la moyenne des images échantillonnées. Ce modèle [272] utilise de courtes vidéos instantanées (snapshots) de vidéos en faisant la moyenne des prédictions d'une seule image RGB et d'une pile de 10 images de flux optiques calculées auparavant, après les avoir passées dans deux répliques d'un CNN pré-entrainé avec ImageNet.

Bien que cette méthode ait amélioré les performances en capturant explicitement les mouvements locaux, elle présente quelques inconvénients : comme les prévisions au niveau de la vidéo ont été obtenues en calculant la moyenne des prévisions sur des clips échantillonnés, les informations temporelles à long terme manquent encore dans les caractéristiques apprises. De plus, les clips sont échantillonnés de manière uniforme à partir des vidéos, ils souffrent d'un problème de fausse attribution d'étiquette. La vérité de base de chacun de ces clips est supposée être la même que celle de la vidéo, ce qui peut ne pas être le cas si l'action se produit juste pour une petite durée dans toute la vidéo. Enfin, la méthode consiste à précalculer les vecteurs de flot optique et à les stocker séparément. Ils obtiennent de meilleurs résultats que Karpathy et al. et des résultats similaires à ceux obtenus par Wang et al.

Le problème du calcul du flux optique au préalable est un point crucial, Zhu et al. [342] propose donc une nouvelle approche. En effet, l'utilisation du flux optique dans l'architecture à deux flux a rendu obligatoire le précalcul du flux optique pour chaque base échantillonnée. Les auteurs ont donc généré un flux optique avec le plus grand nombre de fps et le moins de paramètres possible sans nuire à la précision. Ainsi, pour la partie temporelle, un réseau de génération de flux optique (MotionNet) a été rajouté.

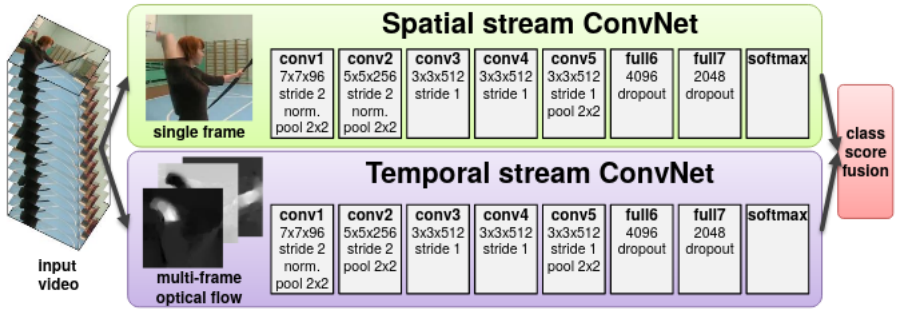


Figure 1: **Two-stream architecture for video classification.**

FIGURE 4.11 – CNN 2 streams de Simonyan et al.[272]

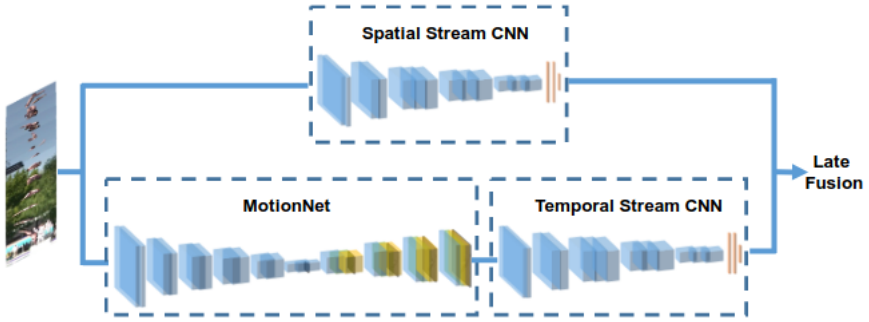


FIGURE 4.12 – Two-stream CNNs de Zhu et al. [342]

À noter qu'il est possible de combiner les approches 2 streams et 3D CNN, comme l'on fait Carreira et al. [38]. Cet article a ainsi introduit une nouvelle architecture pour la classification des vidéos, entrainées sur le dataset Kinetics. Ce modèle appelé **I3D** reprend l'idée des convolutions 3D pour construire un modèle à deux flux. Ainsi, au lieu d'un seul réseau 3D, les auteurs utilisent deux réseaux 3D (un pour les images classiques et un pour le flot optique), pré-entrainés sur des modèles 2D (Imagenet et Kinetics) [3].

4.3 Utilisation de réseaux de neurones récurrents

Sachant que les LSTMs sont conçus pour apprendre les variations dans le temps, Zhang et al. [338] ont énuméré huit types de caractéristiques géométriques contenues dans une pose et sont indépendantes du temps, contrairement à des caractéristiques comme la vitesse et l'accélération (voir figure 5.6).

L'utilisation de caractéristiques géométriques a donc plusieurs intérêts : Zhang et al. ont montré que l'utilisation de ses caractéristiques donnait de meilleurs résultats que l'utilisation des coordonnées des articulations et l'entraînement a besoin de moins données.

CNNs et RNNs peuvent être utilisés ensemble. La pose est utilisée comme un

3. https ://www.deepmind.com/open-source/i3d-model

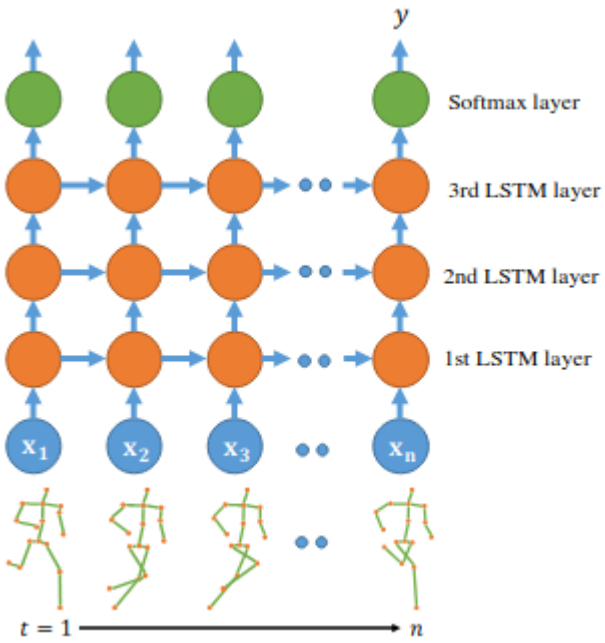


FIGURE 4.13 – Modèle de Zhang et al.[338]

flux d'entrée, fournissant des indices importants pour la discrimination des classes d'activité ; elle sert aussi d'entrée pour le modèle traitant le flux RGB. Les caractéristiques ainsi apprises servent d'entrée au mécanisme d'attention, qui pondère chaque aperçu de sortie en fonction d'une importance estimée par rapport à la tâche générale. Le modèle de flux RGB est un réseau récurrent (de type LSTM), tandis que la représentation de la pose est apprise à l'aide d'un réseau de neurones convolutif prenant comme entrée une sous-séquence de la vidéo (la séquence de squelettes)[14].

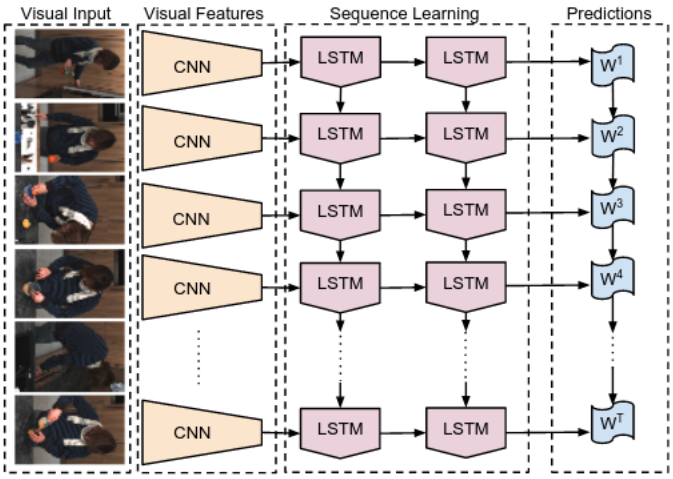


FIGURE 4.14 – Exemple de combinaison entre CNNs et LSTMs [71]

Les avantages de cette méthode sont doubles : une représentation de la pose sur une large plage temporelle permet au mécanisme d'attention d'attribuer une importance estimée à chaque point d'observation et à chaque instant en tenant compte de la connaissance sur toute la fenêtre temporelle. Les informations ainsi

extraites sur les poses sont suffisantes pour apprendre une représentation globale. Cependant, il est nécessaire de trouver une représentation hiérarchique qui respecte les relations spatio-temporelles des articulations.

4.4 Reconnaissance d'actions par réseaux de neurones à graphes

Les méthodes précédentes manquent de modélisation structurelle. L'espace des poses humaines 2D est très structuré en raison des proportions des parties du corps, des symétries gauche-droite, des contraintes d'interpénétration, des limites articulaires (par exemple, les coudes ne se plient pas en arrière) et de la connectivité physique (par exemple, les poignets sont étroitement liés aux coudes). La modélisation de cette structure devrait permettre d'identifier plus facilement les points clés visibles et d'estimer ceux qui sont occlus.

Les données du squelette peuvent représenter le corps humain comme une séquence de vecteurs de coordonnées des principales articulations du corps. Parmi toutes les méthodes de reconnaissance d'actions basées squelette, les GNNs sont l'une des approches les plus populaires [99]. En effet, ces méthodes basées sur les GCN utilisent les données spatiales et temporelles en exploitant les informations contenues dans la structure topologique naturelle du squelette humain. Ces méthodes se sont avérées être plus robustes aux changements de lumière et de fonds que celles basées sur des CNNs et LSTMs.

Plusieurs bases de données ont été développées et sont accessibles au public pour l'HAR basée sur le squelette, comme Kinetics [4] [153, 37] et NTU [5] [183].

Dans ces jeux de données, les modèles les plus performants sont des réseaux à graphes tels que les réseaux convolutifs à graphes adaptatifs à deux flux **ST-GCN** *Spatio-Temporal Graph Convolutional Network* et les **2s-AGCN**(*2 streams Adaptative Graph Convolutional Network* [211].

On définit donc un ordre topologique des articulations dans un corps humain comme un chemin cyclique connecté aux articulations. Le chemin lui-même n'est pas hamiltonien, car chaque nœud peut être visité plusieurs fois : une fois lors d'un passage vers l'avant sur un membre et une fois lors d'un passage vers l'arrière sur le membre jusqu'à l'articulation à laquelle il est attaché. Les doubles entrées dans le chemin sont importantes, car elles garantissent que le chemin préserve les relations de voisinage.

Conclusion

Plusieurs méthodes de squelettisation automatique sont disponibles en open source. Needham et al. ont évalué ces méthodes et ont montré que la précision

4. https ://www.deepmind.com/open-source/kinetics

5. https ://rose1.ntu.edu.sg/dataset/actionRecognition/

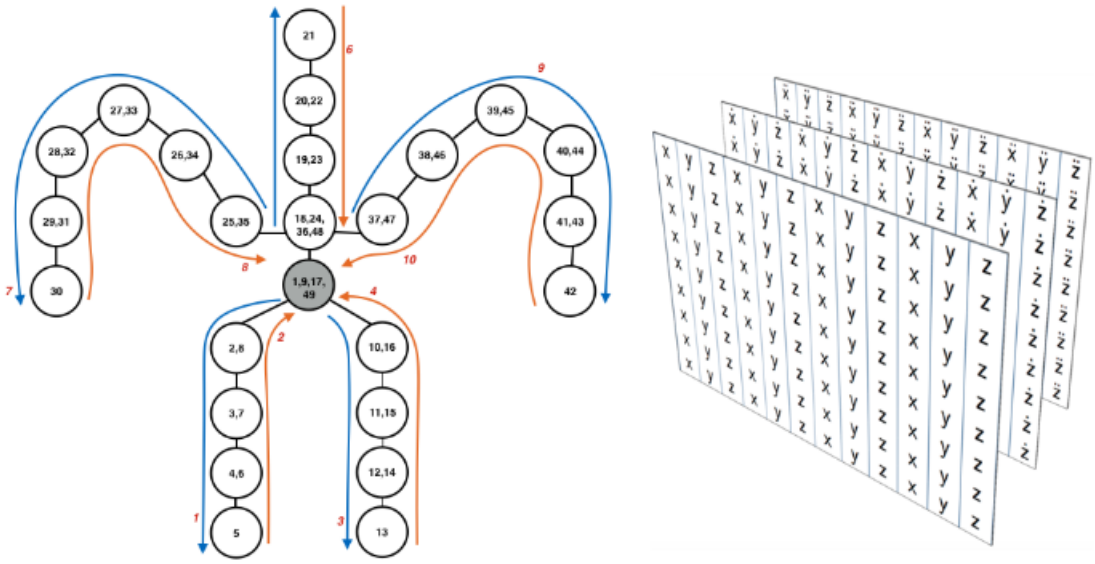


FIGURE 4.15 – Exemple de topologie des points d'intérêts pour la création du squelette [14]

est encore légèrement inférieure aux techniques précédentes [206]. Néanmoins, la détection automatique des points d'intérêts peut faciliter les opportunités pour les chercheurs qui sont prêts à accepter une légère diminution de la performance d'estimation de pose afin de récupérer des données à partir de vidéos plus facilement.

L'analyse de mouvements par deep learning est un problème largement étudié en vision par ordinateur. Les études portant sur la posture se font plutôt dans le cadre des études sur la détection et l'estimation de la posture humaine [339, 207], la reconnaissance d'activités [56] ou encore l'étude de la marche [147]. La plupart des modèles utilisés sont basés sur des images RGB traitées par CNNs. Toutefois, l'analyse de mouvements à partir de données de type squelette présentent de noombreux avantages.

Dans l'analyse de mouvements à partir de squelettes, l'état de l'art est constitué de modèles de type ST-GCN permettant de capturer les dépendances spatiales et temporelles de ces données. Ce type de données n'a en revanche jamais été utilisé pour l'analyse de la douleur.

Troisième partie

Modélisations

Chapitre 5

Analyse de mouvements pour l'affective computing

Sommaire

Cette thèse a pour objectif de proposer un nouveau modèle pour **détecter la douleur à partir de la posture** issue de données vidéos. Ce modèle doit être utilisable par l'entreprise Lucine, il faut donc privilégier les données vidéos et éviter les techniques trop coûteuses ou complexes qui ne pourront pas être déployés auprès de nombreux patients : l'utilisation de capteurs physiques de motion capture est donc à éviter. Le problème étudié ici n'est pas tant de détecter la posture elle-même (c'est-à-dire la position du patient dans l'espace) que de détecter la douleur (qui modifie la façon dont une personne se tient et bouge, comme vu dans la section 6). La solution semble donc de s'inspirer à la fois des travaux de reconnaissance de gestes et d'action, de ceux traitant uniquement de la détection de la posture, mais aussi de la reconnaissance des émotions.

En effet, la compréhension des **émotions exprimées par le corps** est un problème de recherche complexe et très difficile qui exige des chercheurs qu'ils utilisent des connaissances et des approches provenant de différentes disciplines. Les émotions exprimées par le corps sont des phénomènes qui évoluent dans le temps. Ainsi, la vision par ordinateur, l'apprentissage automatique et l'analyse des mouvements peuvent jouer un rôle essentiel dans la résolution de ce problème.

La **reconnaissance automatique des émotions** dans un contexte réel (et non en laboratoire) est encore peu explorée, notamment au niveau de la compréhension du contexte ou encore de la distinction entre les émotions réelles et les émotions simulées. L'objectif principal de l'*affective computing* à l'heure actuelle est de réussir à comprendre les émotions exprimées par le corps dans le cadre de l'activité quotidienne des personnes[311]. Les comportements humains et surtout les émotions exprimées par le corps sont des phénomènes liés au temps. Il faut donc des théories et des modèles qui incorporent cet aspect dynamique.

Dans ce chapitre, nous présentons le travail que nous avons réalisé à partir de la base de données BoLD. Ce travail a fait l'objet d'une présentation lors de la journée *Rencontres des Jeunes Chercheurs en Intelligence Artificielle* (RJCIA) organisée lors de la *Plate-Forme Intelligence Artificielle* (PFIA) 2021.

5.1 Analyse de mouvements pour l'affective computing

L'expression corporelle est définie comme l'affect humain exprimé par des mouvements et/ou des postures du corps. Les recherches antérieures sur l'analyse des mouvements du corps se sont surtout concentrées sur la reconnaissance des activités humaines. Cependant, l'état émotionnel d'une personne est une autre caractéristique importante qui est souvent transmise par les mouvements du corps [75]. En effet, des études en psychologie ont montré que le mouvement et le comportement postural sont des caractéristiques utiles pour identifier les émotions humaines [266, 264, 265, 161, 250, 285]. Le tableau 5.1 montre quelques exemples de comportements caractéristiques de certaines émotions.

La reconnaissance automatique de l'expression corporelle humaine dans des situa-

tions réelles (sans contrainte) est extrêmement complexe étant donné la compréhension incomplète des relations entre les expressions émotionnelles et les mouvements du corps. Pour que des programmes informatiques puissent apprendre à reconnaître le langage corporel des humains, il faut donc développer un travail interdisciplinaire entre l'informatique, les statistiques, la psychologie et l'anthropologie.

Émotions	Gestes et postures
Joie	corps en extension, épaules relevées, bras levés ou écartés du corps
Intérêt	mouvement latéral de la main et du bras, bras tendu dans le plan frontal main vers la tête
Surprise	mains couvrant la bouche, secousse de la tête, recul du corps
Ennui	menton levé, tête vers l'arrière, posture relâchée, tête penchée sur le côté visage couvert avec les mains
Dégoût	épaules en avant, tête vers le bas, haut du corps affaissé bras croisés devant la poitrine, mains près du corps
Colère	épaules soulevées, ouverture et fermeture de la main bras tendus de face, doigt tendu

Tableau 5.1 – Exemple d'expressions corporelles d'émotions

La reconnaissance automatique de l'expression corporelle au travers d'images et de vidéos est très difficile pour plusieurs raisons. Il est extrêmement compliqué de collecter des données et de constituer un jeu de données avec des expressions corporelles avec des annotations de qualités, et ce d'autant plus que les émotions sont sujettes à différentes interprétations selon le contexte ou la culture, comme nous l'avons montré dans la section 1.3. De plus, chaque articulation a plusieurs degrés de liberté, ce qui augmente l'hétérogénéité de comportements et de mouvements. Il y a également toutes les problématiques classiques rencontrées en traitement d'image : les différences de perspective, de fonds, etc. Un autre problème est qu'il n'existe pas de *gold standard* pour l'analyse des expressions corporelles, il n'existe pas d'équivalent des FACS et des AUs [77] comme il existe pour la reconnaissance faciale. Enfin, les expressions corporelles sont des mouvements subtils et composites, que l'on peut classer en trois grandes catégories : les mouvements fonctionnels, les mouvements artistiques et les mouvements communicatifs.

Malgré les progrès significatifs réalisés récemment dans l'estimation des poses 2D/3D (voir précédemment dans le chapitre 2), beaucoup d'études en analyse de mouvements pour l'affective computing se sont limitées aux systèmes de capture de mouvement qui reposent sur le placement de marqueurs optiques actifs ou passifs sur le corps du sujet pour détecter les mouvements, en raison de plusieurs problèmes. Tout d'abord, ces méthodes d'estimation basées sur la vision ont un problème de gigue, c'est-à-dire qu'il existe une variation de latence entre la posture réelle et son estimation. De plus, bien qu'une certaine précision ait été atteinte dans les challenges classiques d'estimation de pose, les critères utilisés dans ces benchmarks ne sont pas conçus pour la reconnaissance des émotions qui exige une précision nettement supérieure. Par conséquent, les erreurs dans les résultats générés par ces méthodes se propagent, car l'estimation de la pose est une première étape dans le pipeline d'analyse de la relation entre le mouvement et l'émotion.

5.2 Reconnaissance des émotions à partir de la posture

La reconnaissance des émotions est une problématique intéressante, car proche de celle de l'étude de la douleur et des bases de données pour la détection des émotions à partir de la posture humaine sont disponibles. Par exemple, Randhavane et al. ont cherché à identifier des émotions perçues à travers la marche [232] : 4 catégories d'émotions de bases (joie, colère, tristesse et neutre), la valence et l'excitation. Les auteurs ont choisi les LSTMs, car il existait déjà de tels modèles pour analyser des séquences vidéos (voir section 4.3) et pour l'étude des mouvements du corps humain comme le cycle de marche [188].

5.2.1 Base de données BoLD

La reconnaissance automatique de l'expression corporelle humaine dans des situations réelles est compliquée étant donné que en général les modèles d'IA ont une compréhension incomplète et limitée de la relation entre les expressions émotionnelles et les mouvements corporels. Pour remédier à cela, Luo et al. ont créé un nouveau jeu de données appelé BoLD (*Body Language Dataset*) comprenant des vidéos (des extraits de films venant de youtube) [187]. Les auteurs ont organisé un concours dans le cadre du *First international workshop on bodily expressed emotion understanding* de la conférence **ECCV 2020**. L'objectif était de développer des méthodes avancées d'apprentissage pour l'analyse et l'interprétation automatique des émotions à partir des mouvements du corps humain en conditions réelles. Pour cela, les auteurs ont construit un jeu de données, entièrement annoté, ainsi qu'une plateforme d'évaluation et de benchmarking de modèles pour comparer les performances des modèles en vue de cet objectif.

D'après les auteurs, la reconnaissance automatique de l'expression corporelle comme problème de recherche est très difficile pour plusieurs raisons :

- il est difficile de recueillir un ensemble de données sur l'expression corporelle avec des annotations de haute qualité. La compréhension et la perception des émotions à partir d'observations concrètes sont souvent soumises au contexte, à l'interprétation, à l'ethnicité et à la culture, comme nous l'avons développé en section 1.3.
- l'expression corporelle est subtile et composite, elle peut être couplée à des mouvements fonctionnels.
- il n'y a souvent pas de gold standard pour l'étude des expressions corporelles.
- l'anatomie humaine offre de nombreux degrés de liberté. Travailler avec des données vidéo « in-the-wild » pose des défis techniques supplémentaires tels que le niveau élevé d'hétérogénéité dans les comportements des personnes, l'arrière-plan de la vidéo très encombré, et les différences souvent substantielles entre l'image, la perspective de la caméra, et la pose de la personne dans le cadre.

Les capacités informatisées de reconnaissance de l'expression des émotions corporelles ont le potentiel de permettre un grand nombre d'applications innovantes, notamment la gestion et la recherche d'informations, la sécurité publique, l'amé-

lioration des soins aux patients et les médias sociaux. Dans le cadre de ce défi, les chercheurs doivent développer un modèle permettant aux ordinateurs de comprendre les émotions humaines à partir de données spatio-temporelles et de données sur les poses humaines provenant du dataset BoLD.

Pour construire ce jeu de données, les auteurs ont choisi les films inclus dans un jeu de données public, *Atomic Visual Actions* (AVA) [1] qui contient une liste d'extraits de films. Ces films bruts ont été extraits de *YouTube* et divisés en plusieurs courtes scènes avant d'utiliser d'autres méthodes pour localiser et suivre chaque personne à travers différentes images de la scène. Pour faciliter le suivi, la même personne dans chaque clip a été marquée avec un numéro d'identification unique. Ainsi, chaque clip a été traité par un estimateur de pose image par image pour positionner des points de repères sur le corps humain (*landmarks*). Enfin, les annotations des émotions de chaque personne dans ces clips ont été réalisées grâce à la plateforme de *crowdsourcing* en ligne *Amazon Mechanical Turk* (AMT). À partir d'une courte séquence vidéo de quelques secondes (6s en moyenne), le modèle doit évaluer 26 catégories d'émotions : Paix, Affection, Estime, Anticipation, Engagement, Confiance, Bonheur, Plaisir, Excitation, Surprise, Sympathie, Doute/confusion, Déconnexion, Fatigue, Embarras, Désir, Désapprobation, Aversion, Ennui, Colère, Sensibilité, Tristesse, Inquiétude, Peur, Douleur, Souffrance, et 3 valeurs émotionnelles continues : *Valence, Arousal, Dominance* (voir Fig. 5.1).

1. https ://research.google.com/ava/

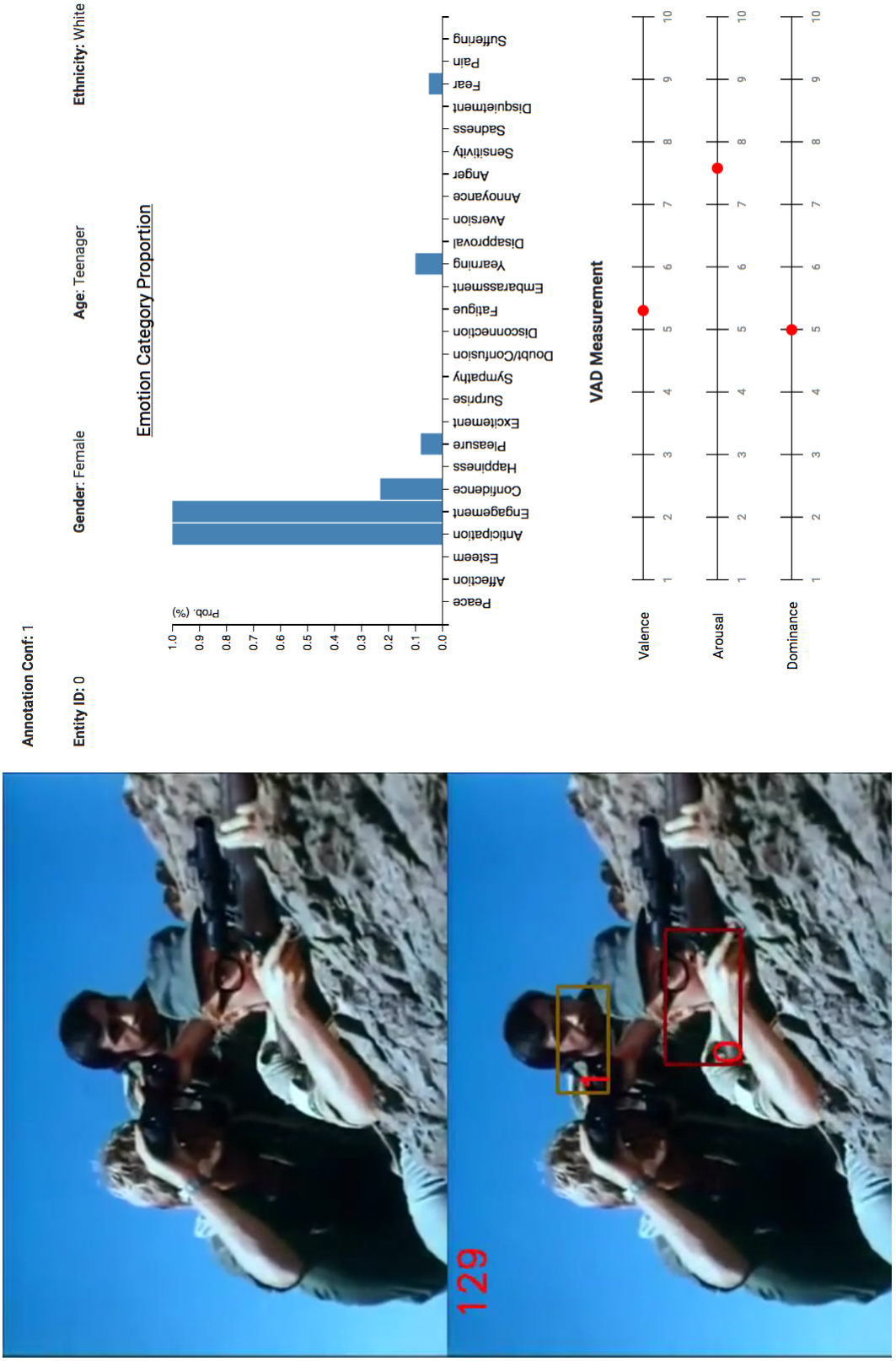


FIGURE 5.1 – Exemple de données issues de BoLD

Luo et al. ont eu pour but de créer un nouveau dataset pour l'étude de la reconnaissance des émotions à partir de la posture dans des conditions réelles [187]. La modélisation automatique de l'expression corporelle nécessite généralement trois étapes : la détection humaine, l'estimation et le suivi de la pose, et l'apprentissage de la représentation. Dans un tel système, le ou les humains sont détectés image par image dans une vidéo et leurs points de repère corporels sont extraits par un estimateur de pose. Par la suite, si plusieurs personnes apparaissent dans la scène, les poses d'une même personne sont associées le long de toutes les images. La pose de chaque personne étant identifiée et associée à travers les images, et une représentation appropriée des caractéristiques de chaque personne est extraite. Les auteurs ont utilisé les caractéristiques de l'analyse des mouvements de Laban (*Laban Movement Analysis*, LMA). Luo et al. ont donc cherché à construire un système de reconnaissance d'émotions en assemblant plusieurs modèles. Tout d'abord deux *Temporal Segment Networks* (TSNs, voir section 4.2.1) : un pour le visage et un pour le corps, puis un modèle utilisant l'**analyse de mouvement Laban** qui utilise quatre composantes pour enregistrer les mouvements du corps humain : le corps, l'effort, la forme et l'espace. La catégorie de corps représente les caractéristiques structurelles et physiques des mouvements du corps humain. Elle décrit les parties du corps qui bougent, celles qui sont connectées, celles qui sont influencées par d'autres, et de façon générale l'organisation du corps. La catégorie Effort décrit l'intention inhérente à un mouvement. La forme décrit les formes statiques du corps, la façon dont le corps interagit avec quelque chose et dont le corps change vers un certain point dans l'espace, et la façon dont le torse change de forme pour soutenir les mouvements du reste du corps.

Les auteurs ont testé différentes méthodes qu'ils ont classées de la façon suivante :
— apprentissage à partir de squelette par LMA puis forets aléatoires (*random forest*)
— apprentissage à partir de squelette par deep learing avec des ST-GCN
— apprentissage à partir des pixels (images brutes) par analyse de la trajectoire classique (par des *histogram of flow* (HOF) et des *motion boundary histograms* (MBH) suivis de SVM pour la classification
— apprentissage à partir des pixels par deep learning avec notamment *2-streams networks* utilisant le flux optique et des TSNs

Modèle	mR2	mAP	mRA	ERS
LMA	0.075	13.59	57.71	0.216
ST-GCN	0.044	12.63	55.96	0.194
Two stream Resnet	0.084	17.04	62.70	0.240
TSN corps	0.095	17.02	62.70	0.247
TSN corps + visage	0.101	17.31	63.46	0.252
TSN corps + LMA	0.101	16.76	62.75	0.249
TSN corps + visage + LMA	0.103	17.14	63.52	0.253
I3D	0.098	15.37	61.24	0.241

Tableau 5.2 – Résultats obtenus par les différentes méthodes de Luo et al. [187] (mR2 : mean of R2, mAP(%) : average precision/area under precision recall curve (PR AUC), mRA(%) : mean of area under ROC curve (ROC AUC), ERS : emotion recognition score)

Comme nous pouvons le voir dans le tableau 5.2, la méthode avec la meilleure performance est celle utilisant un TSN avec un score R2 moyen de 0,095, une précision moyenne de 17,02 %, une AUC (*Area Under the Curve*) moyenne de 62,70% et un ERS (*Emotion Recognition Score*) de 0,247. À partir de ces résultats, les auteurs ont construit leur système de recognition **ARBEE** (*Automated Recognition of Bodily Expression of Emotion*) en assemblant les meilleurs modèles (TSN et des forets aléatoires avec les caractéristiques de LMA). Toutefois, l'ajout des caractéristiques LMA semble avoir une faible valeur ajoutée par rapport aux modèles TSN seuls [187]. Bien que constituant l'état de l'art dans le domaine de la reconnaissance d'actions, les architectures de types 2 streams (RGB et flot optique) sont moins efficaces que les TSN, quelle que le soit le modèle utilisé (I3D, Resnet). Ceci est peut-être dû au fait que pour la classification d'actions, les caractéristiques spatiales sont plus importantes que les informations temporelles pour la classification, alors que dans le cas de l'affective computing (émotions ou douleur), l'information temporelle est primordiale. Comme expliqué en section 4.3 les réseaux de type LSTM ont été spécifiquement conçues pour traiter des séquences, nous avons donc exploré cette piste sur la base de données BoLD.

5.2.2 Analyse des données

Le jeu de données BoLD [2] [187] est constitué de 150 vidéos extraites de YouTube divisées en 18 927 clips au total. Chaque clip dure en moyenne 6s avec 30 fps (*frame per second*), soit environ 180 images à analyser par vidéo. Pour chaque clip vidéo, nous avons un fichier de données stockant une matrice N par 56 où N est le nombre d'images de la vidéo correspondante, puis pour chaque ligne, la première colonne est l'horodatage, la deuxième colonne est l'identifiant de la personne, et le reste sont les résultats de l'estimation de la pose 2D pour chacune des 18 articulations, obtenus grâce à **OpenPose** (voir Fig. 5.2) [33]. Les labels sont stockés sur un fichier dans le dossier des annotations, où la première colonne est le nom de la vidéo, la deuxième est l'identifiant de la personne (une vidéo peut avoir plusieurs identifiants de personnes), les colonnes 3 et 4 sont les images, et le reste sont les émotions (les 26 catégories puis les trois émotions continues).

Comme nous pouvons le voir dans l'histogramme 5.3, il existe une grande disparité de représentations entre les différentes classes d'émotions. L'émotion la plus représentée est l'*Engagement* présente dans 1334 vidéos. On trouve ensuite *Confidence* (1049 vidéos) et *Happiness* (986). On peut ensuite regrouper les émotions *Peace, Affection, Engagement, Doubt/Confusion, Annoyance* qui sont toutes présentes aux environs de 800 vidéos chacune. Puis, on a un groupe vers 600 vidéos pour *Pleasure, Excitement, Disapproval, Disquietment*, et un groupe vers les 400 vidéos avec les émotions *Esteem, Anger, Sadness, Surprise*. *Fear* est présente dans 321 vidéos. Enfin, on a un groupe d'émotions présentes dans environ 200 vidéos avec *Sympathy, Aversion, Sensitivity, Suffering*. Les affects les moins présentes sont *Yearing, Pain* et Embarrassement avec respectivement 74, 88 et 90 vidéos.

La matrice de corrélation montre qu'il y a plusieurs groupes d'émotions corrélées

2. https ://cydar.ist.psu.edu/emotionchallenge/dataset.php

Point de repère	Numéro
Nez	0
Cou	1
Épaule droite	2
Coude droit	3
Poignet droit	4
Épaule gauche	5
Coude gauche	6
Poignet gauche	7
Hanche droite	8
Genou droit	9
Cheville droite	10
Hanche gauche	11
Genou gauche	12
Cheville gauche	13
Œil droit	14
Œil gauche	15
Oreille droite	16
Oreille gauche	17

TABLEAU 5.3 – Liste des landmarks d'OpenPose

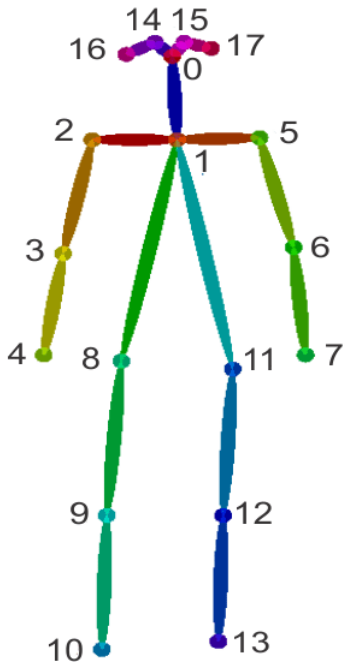


FIGURE 5.2 – Illustration d'un squelette obtenu avec OpenPose [108].

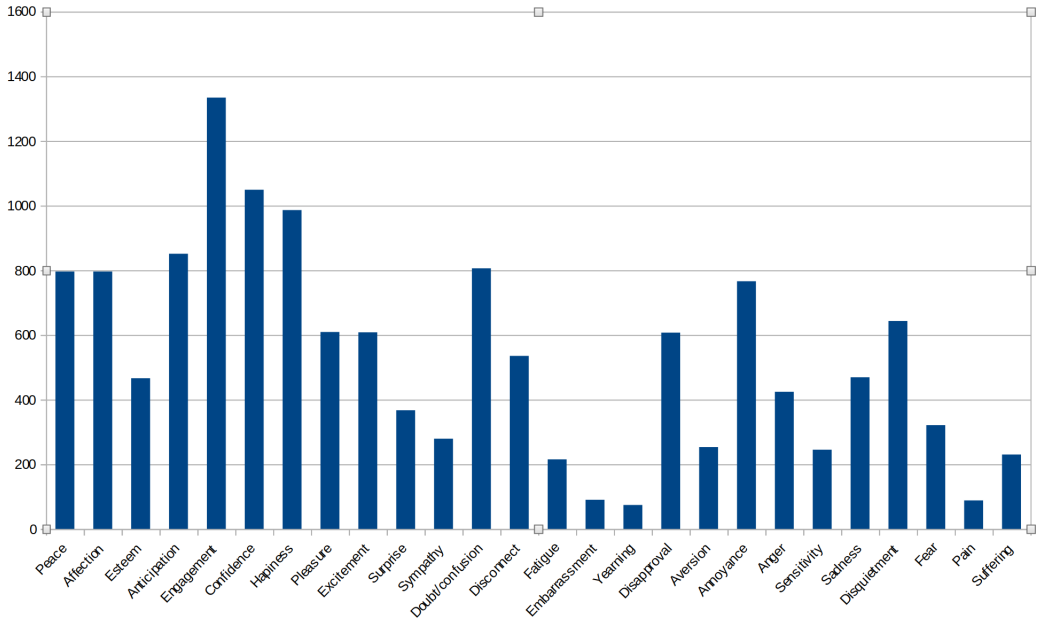


FIGURE 5.3 – Nombre de vidéos pour chacune des 26 classes d'émotions dans BoLD.

entre elles :

- *Happiness, Pleasure* et dans une moindre mesure *Excitement*
- *Disapproval, Aversion, Annoyance* et *Anger*
- *Sensitivity* et *Sadness*
- *Pain* et *Suffering*.

De plus, l'expression d'une mesure de la valence est corrélée avec deux affects : Happiness et Pleasure (voir Fig. 5.4).

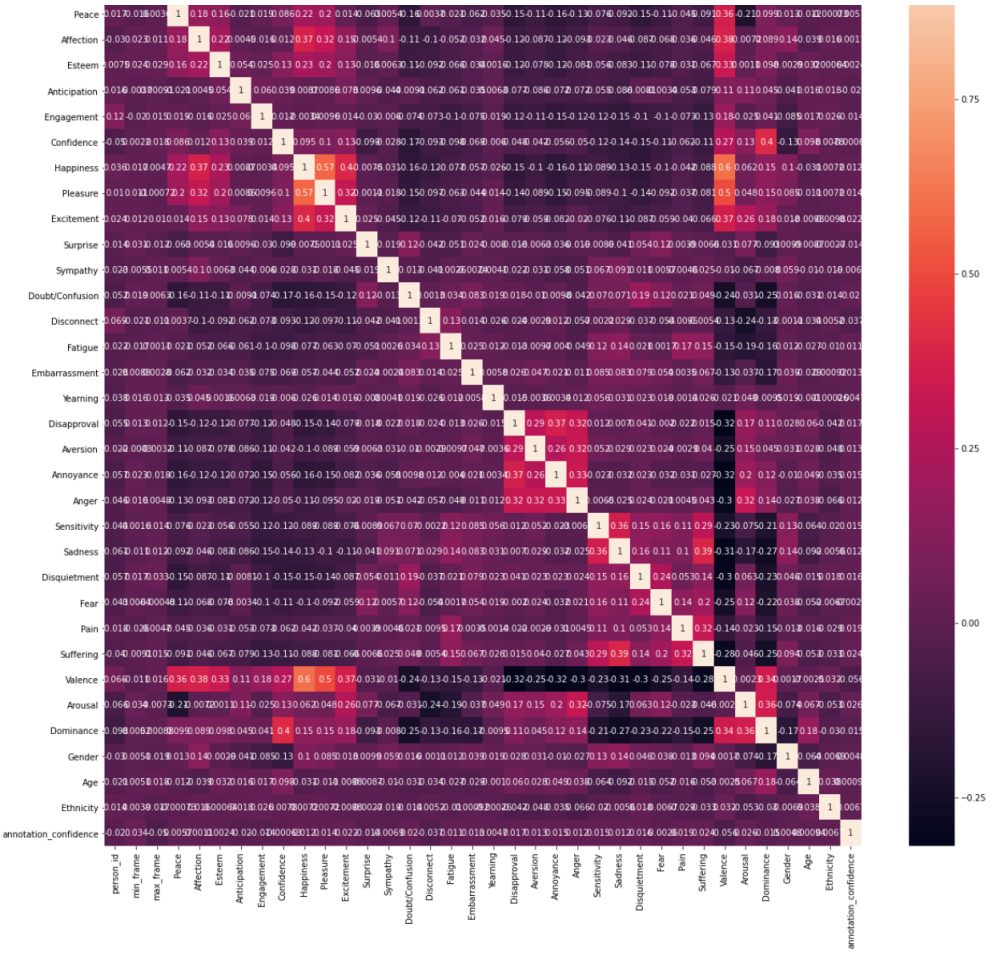


FIGURE 5.4 – Matrice de confusion des émotions de BoLD

Après avoir réalisé cette analyse de données sur la base de données BoLD, nous allons présenter le modèle que nous avons utilisé pour prédire à la fois les catégories d'émotions et les mesures de la *Valence, Arousal, Dominance* comme demandé aux participants du challenge.

5.3 Modèle

Pour élaborer notre modèle capable d'extraire les informations pertinentes pour la classification des affects, nous avons étudié différentes propositions. D'après Yao et al., les caractéristiques issues de la pose ont des performances supérieures aux autres méthodes, car il y a moins de variations au sein d'une même classe comparé aux images RGB-D, et l'information étant déjà de haut niveau, l'apprentissage est plus simple [329]. Toutefois, la construction du squelette de façon automatique à l'aide d'algorithme de type OpenPose [33] a un coût en termes de calculs et peut-être source d'erreurs. D'autre part, sachant que les LSTMs sont conçus pour apprendre les variations dans le temps, Zhang et al. ont énuméré huit types de caractéristiques géométriques contenues dans une pose et indépendantes du temps, contrairement à des caractéristiques comme la vitesse et l'accélération [337]. Zhang et al. ont montré que l'utilisation de ses caractéristiques donnait de meilleurs résultats que l'utilisation des coordonnées des articulations et était plus compréhensible pour les humains, en

particulier les angles entre les segments du corps (voir Fig. 5.6). De plus, l'entraînement a besoin de moins données et permet l'utilisation de modèle simple.

Nous retenons donc la mise en œuvre d'un modèle LSTM ayant comme données d'entrée les angles caractérisant la pose des sujets dans les images.

5.3.1 Construction des caractéristiques géométriques

Dans le jeu de données BoLD, nous avons 18 927 fichiers, chacun contenant les séquences de la localisation x, y dans l'espace des points d'intérêts et le degré de confiance c de la détection, formaté en $x1, y1, c1, x2, y2, c2\ldots$, où c est le degré de confiance entre 0 et 1 . De chaque ligne, nous extrayons les coordonnées x et y des points d'intérêts. Nous obtenons une liste de 11 segments comme indiqué sur le tableau 5.4. On peut ensuite calculer des angles entre 2 de ces segments ainsi constitués (c'est-à-dire entre 3 points de l'espace), comme indiqué dans le tableau 5.5).

Tableau 5.4 – Landmarks pour chacun des segments du squelette

Segment	Landmark 1	Landmark 2
segment 1	nez	cou
segment 2	épaule droite	coude droit
segment 3	coude droit	poignet droit
segment 4	épaule gauche	coude gauche
segment 5	coude gauche	poignet gauche
segment 6	hanche droite	genou droit
segment 7	genou droit	cheville droite
segment 8	hanche gauche	genou gauche
segment 9	genou gauche	cheville gauche
segment 10	œil droit	œil gauche
segment 11	oreille droite	oreille gauche

Les premières expérimentations ont montré que les données concernant la partie inférieure du corps (hanches, genoux et chevilles) dans la base BOLD sont trop entachées de bruit. En effet, nous avons constaté que de nombreuses positions de points de repères anatomiques concernant les membres inférieurs (hanches, genoux et chevilles) données par OpenPose sont erronés ou absentes. Nous avons donc décidé de nous concentrer sur le haut du corps (c'est-à-dire les points de repères situés au-dessus des hanches) afin de garder des valeurs d'angles fiables et non nulles (voir Fig. 5.5). Nous avons donc conservé les angles 1 à 6 (voir Tableau 5.5) [234].

5.3.2 Le modèle proposé

Zhang et al. [337] ont montré que les réseaux LSTM sont efficaces lorsqu'ils sont appliqués à la reconnaissance d'actions basée sur les squelettes. Wang et al. [303, 304] ont montré que les réseaux LSTMs peuvent être utilisés pour la détection des comportements de protection. Inspirés par Sutskever et al. [281], nous avons donc essayé

TABLEAU 5.5 – Points de repères anatomiques d'OpenPose utilisés pour calculer les angles d'intérêt

Angles	Premier point	Deuxième point	Troisième point
Angle 1	nez 0	cou 1	épaule droite 2
Angle 2	nez 0	cou 1	épaule gauche 5
Angle 3	cou 1	épaule droite 2	coude droit 3
Angle 4	cou 1	épaule gauche 5	coude gauche 6
Angle 5	épaule droite 2	coude droit 3	poignet droit 4
Angle 6	épaule gauche 5	coude gauche 6	poignet gauche 7
Angle 7	cou 1	hanche droite 8	genou droit 9
Angle 8	cou 1	hanche gauche 11	genou gauche 12
Angle 9	hanche droite 8	genou droit 9	cheville droite 10
Angle 10	hanche gauche 11	genou gauche 12	cheville gauche 13

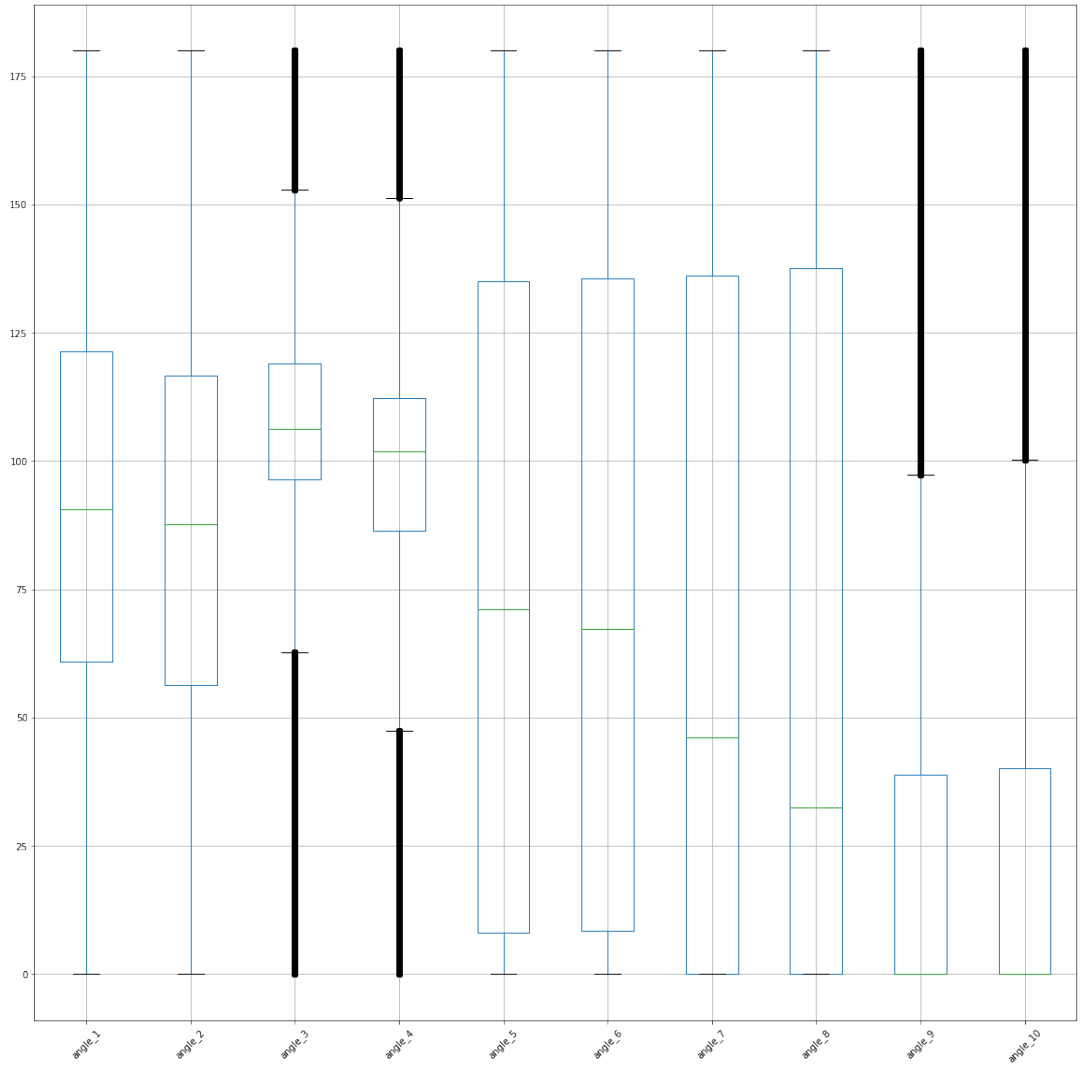


FIGURE 5.5 – Histogrammes des valeurs des angles calculées

une approche basée sur les LSTMs dans une architecture de type *Encoder Decoder*.

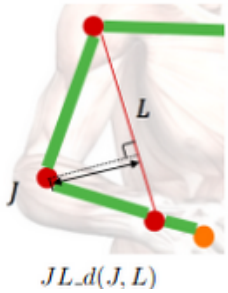


FIGURE 5.6 – Angle entre 2 segments du corps humain [337]

En effet, comme expliqué en section 4.3 ces modèles de type seq2seq sont capables de transformer une séquence d'entrée en une nouvelle séquence avec efficacité.

Comme montré dans le schéma 5.7, dans notre modèle, l'encodeur et le décodeur sont des LSTMs multicouches. Une vidéo est donc représentée par une séquence d'ensembles de coordonnées des angles que nous avons calculés, chaque séquence contenant les informations de pose pour une image donnée. Le réseau reçoit ces informations et génère une sortie par pas de temps, créant ainsi un modèle de séquence à séquence. Pour résumer, le vecteur de caractéristiques d'entrée sur un pas de temps est la liste des angles pour une image donnée, et la sortie est une suite d'émotions identifiées dans la séquence. En ce qui concerne la configuration du réseau, on a défini 3 couches de LSTMs pour l'*encoder* et le *decoder*, avec 16 unités cachées par couches. Les poids sont mis à jour avec l'algorithme d'optimisation Adam. Un *drop out* a été utilisé avec la probabilité d'activation à 0.5. Un taux d'apprentissage de 0.003 et un *batch size* de 16 ont été utilisés. Notre modèle a été entraîné pendant 30 époques [234].

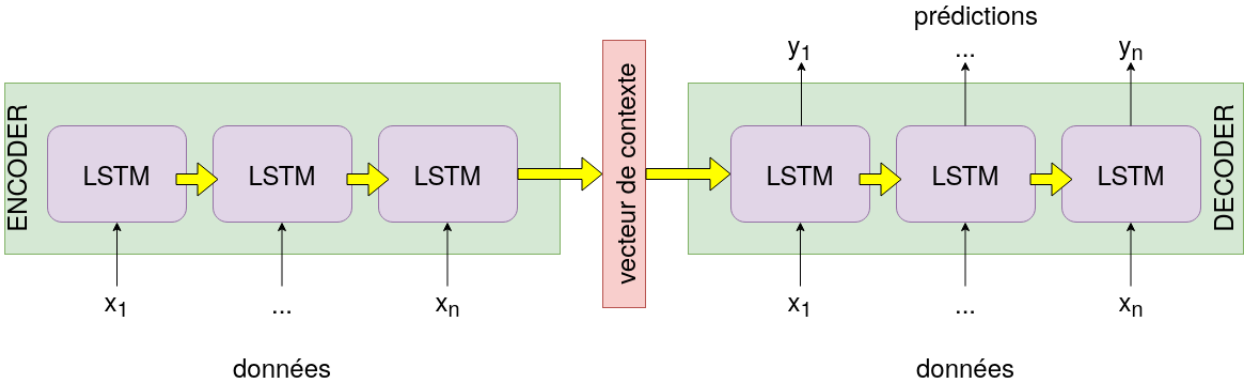


FIGURE 5.7 – Notre modèle Encoder-Decoder

5.4 Résultats

Nous avons utilisé notre modèle pour deux taches différentes. La première est une classification pour les 26 catégories d'émotions, avec comme fonction de perte l'entropie croisée (*Cross Entropy*), et comme métriques la précision moyenne (mAP) et la moyenne de l'air sous la courbe ROC (mRA). La deuxième tâche est de prédire

les valeurs des 3 émotions continues par une régression, avec comme fonction de perte l'erreur quadratique moyenne (*Mean Square Error*), et comme métrique la moyenne du Coefficient de détermination (mR^2). Nous avons également utilisé la métrique globale *Emotion Recognition Score* (ERS) proposé par Luo et al., tel que :

$$ERS = 1/2(mR + 1/2(mAP + mRA)) \tag{5.1}$$

Les performances de notre modèle sont malheureusement restées en deçà de celles du modèle de Liao et al. (voir tableau 5.6). Toutefois, l'avantage de notre modèle qui est relativement simple est qu'il nécessite moins de temps de calcul (90 minutes) et moins de puissance de calcul par rapport aux modèles utilisant les images (comme les TSN) ou le flot optique (tel que I3D). On peut faire l'hypothèse que notre modèle est plus sensible à la qualité des données fournies. Ainsi, les systèmes de détection automatiques de points d'intérêts, bien que prometteurs, sont encore à améliorer, notamment en ce qui concerne les membres inférieurs (hanches, genoux, chevilles et pieds). Les caractéristiques géométriques ne sont pas suffisantes pour pleinement exprimer la topologie du corps humain qui est primordiale pour ce genre de problème [234].

Néanmoins, les données de type squelette sont intéressantes, car elles permettent d'obtenir des résultats proches de ceux utilisant des images tout en étant beaucoup moins volumineux et sont exploitables par des modèles simples basés sur des LSTMs nécessitant moins de temps et de puissance de calcul que ceux basés sur des CNNs utilisant des images.

Tableau 5.6 – Résultats

Modèle	mAP	mRA	mR^2	ERS
Luo et al.	17.14	63.52	0.103	0.253
Encoder-Decoder	16.74	51.18	0.08	0.185

Enfin, on soulignera que l'annotation des émotions par des personnes non expertes est problématique. Le nombre de 26 catégories d'émotions semble trop important, il est difficile de dissocier certaines émotions proches. Une solution pour améliorer les résultats serait donc de n'utiliser que les 6 émotions de bases d'Ekman par exemple (la colère, le dégoût, la peur, le bonheur, la tristesse et la surprise). De plus, en ce qui concerne les vidéos issues d'*AVA*, on notera que les émotions que l'on cherche à détecter ne sont pas des émotions réelles, mais jouées par des acteurs, ce qui peut introduire des biais d'interprétation, les émotions étant jouées plus que ressenties.

C'est pour ces raisons que dans le chapitre suivant, nous nous sommes tournés vers la détection de la douleur en utilisant d'autres bases de données sur lesquelles nous avons testé notre modèle d'Encoder-Decoder, mais aussi que nous avons enrichi notre proposition avec d'autres modèles utilisant des mécanismes d'attention spatiales et temporelles issus du modèle le Transformer. Nous avons également cherché à modéliser la structure topologique du corps humain à l'aide d'un graphe servant de données d'entrée à un modèle basé sur un ST-GCN combiné à des utilisant des mécanismes d'attention.

Chapitre 6

Reconnaissance du comportement douloureux

"Les mouvements de l'homme extérieur vous révèlent les changements survenus dans l'homme intérieur."

— Bernard De Clairvaux

Sommaire

Dans ce chapitre, nous nous concentrons sur la détection de la douleur à partir des mouvements. Pour cela, nous avons utilisé la base de données UI-PRMD et nous proposons un modèle basé sur un ST-GCN auquel nous ajoutons des mécanismes d'attention. Ce travail a été accepté pour le workshop *Learning with few or without annotated face, body and gesture data* de la conférence *Face and Gesture 2023*.

6.1 Les mouvements dans le contexte de la douleur

On a défini la posture comme l'ensemble des diverses positions physiques adoptées par une personne au cours du temps. Elle peut être naturelle ou non, volontaire si la personne peut la choisir, ou contrainte, dans le cas contraire. D'après [59], un geste est défini comme l'exécution d'un mouvement par une personne dans un but précis. Un geste est donc un mouvement du corps ayant du sens. Il peut provenir de mouvements des doigts, des mains, des jambes ou du corps entier, avec l'intention de communiquer des informations et/ou d'interagir avec l'environnement. Cependant, la signification d'un geste peut être différente d'une culture à l'autre, et peut dépendre de l'état individuel d'un individu. Les gestes constituent donc un sous-espace des mouvements du corps humain [59]. Ici, nous nous intéresserons donc aux postures et aux gestes caractéristiques de la douleur (voir tableau 6.1). La plupart des mouvements du corps qui ont été identifiés comme étant liés à la douleur servent à se protéger contre d'autres dommages et à minimiser la douleur. Il s'agit notamment des réflexes de protection, des frottements, des torsions et de la protection.

Signes posturaux	**Mouvements**
Orientation de la tête	Lacet, roulis, tangage (*yaw, pitch, roll*, voir Fig. 6.1)
Visage	FACS, AUs
Mains	Grattements, mouvements des doigts
Orientation du corps	Oscillations, en avant/en arrière
Rythme	mouvements rapides ou ralentis, pause

Tableau 6.1 – Signes posturaux de la douleur

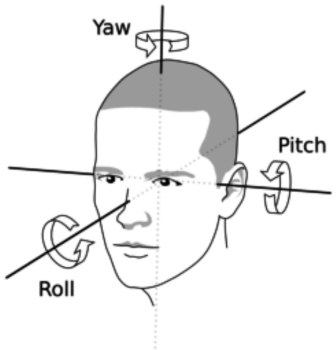


Figure 6.1 – Orientation de la tête [143]

Les comportements concrets de mouvements corporels varient en fonction du type de population médicale. Par exemple, dans le cas d'une lombalgie chronique, les mouvements peuvent être caractérisés par la garde ou la raideur, l'hésitation,

l'appui ou le soutien, une action brusque ou la boiterie, ou ils peuvent servir à frotter ou à stimuler une partie du corps affectée. L'outil d'observation de la douleur en soins intensifs (*Critical-Care Pain Observation Tool*, ou CPOT) comprend plusieurs catégories de mouvements : la protection (mouvement lent, toucher le site de la douleur, rechercher l'attention par des mouvements) et l'agitation (tenter de s'asseoir, se déplacer en boitant). En outre, elle comprend également l'évaluation de la tension musculaire.

Parmi les **signes cliniques** de la posture douloureuse, on peut citer [198, 169, 96, 202] :

— les réflexes de retrait,
— les comportements (voir Tableau 6.2)
— une altération des performances motrices : accélération du mouvement, amplitude et vitesse de pointe, allongement du temps de réaction,
— une modification de la tension musculaire avec une raideur musculaire (due en partie à l'activité réflexe), hypo/hypertonie, spasticité, spasme
— surcharge musculaire chronique
— utilisation inefficace/non nécessaire des muscles
— maux de tête de type tension (TTH), torticolis spasmodique.

Comportement	**Définitions**
Protection	Rigidité, raideurs
Hésitation	Mouvements continus stoppés, perte de la fluidité
Renforcement, Support	Position dans lequel un membre est maintenu, sans ou avec support
Actions abruptes	Mouvements soudains et spontanés
Boiterie	asymétrie de la foulée, répartition inégale de la charge
Stimulations	Massages de la zone douloureuse, frottements

Tableau 6.2 – Comportements évocateurs de la douleur

La détection automatique de la posture et du comportement douloureux a notamment été étudiée par Nadia Bianchi-Bertouze et son équipe de l'*University College London* (UCL) dans le cadre de l'étude des douleurs chroniques aux lombaires [1], notamment avec le challenge EmoPain [201]. D'après cette équipe, la douleur chronique, la peur des blessures et de la douleur, ainsi que l'anxiété conduisent la personne à engager des parties de son corps d'une manière qui n'est pas nécessaire ou efficace sur le plan biomécanique, mais qui peut accroître le sentiment de contrôle et réduire la peur. Ainsi, par exemple, les personnes souffrant de lombalgies chroniques ont tendance à engager différentes parties du corps à différents stades de mouvements plutôt que de manière synergique, malgré le fait que ces stratégies rendent le mouvement plus difficile à exécuter (en raison du sentiment du patient d'un meilleur contrôle sur le mouvement, et/ou de la peur de la douleur ou des blessures, ou encore de capacités réduites de proprioception/coordination).

1. http ://www.emo-pain.ac.uk/publications.html

6.2 Données disponibles

Les ensembles de données de référence présentés en section 1.5.3 visaient à détecter les indices faciaux de la douleur, comme la base de données *UNBC-McMaster Shoulder Pain Expression Archive*, ou la douleur expérimentale, ce qui peut paraître limitant, comme la base de données *Biovid Heat Pain*, alors que le mouvement du corps est une modalité essentielle à prendre en compte pour évaluer l'expérience de la douleur. Cette absence de données sur les mouvements du corps nous a conduits à nous tourner vers d'autres sources de données comme nous le verrons dans la section suivante.

Du côté de la méthodologie, les modèles d'évaluation automatique de la douleur existants poursuivent différents objectifs. Les plus courants sont la détection de la présence ou non de douleur (classification binaire) ou l'évaluation de l'intensité de la douleur (le plus souvent sur une échelle de 0 à 10 semblable à l'EVS présentée en section 1.2.3). La vérité terrain *ground truth* peut provenir de l'auto-déclaration, de l'observation ou de la connaissance de la procédure d'induction de la douleur par des experts. En ce qui concerne la posture, il existe peu de bases de données disponibles : *Biovid* par exemple, contient des données centrées uniquement sur la tête et le haut du corps. Seule *EmoPain* contient des données sur l'ensemble du corps, pour cette raison, nous nous sommes concentrés dans un premier temps sur cette base de données.

6.2.1 Emopain

La base de données **Emopain** a été créé à partir des travaux de Aung et al. [11] et Olugbade et al. [212] sur des personnes atteintes de douleurs lombaires chroniques (*Chronic Lombaire Back Pain, CLBP*) dans le but d'évaluer leur capacité à réaliser des exercices de rééducation sous le contrôle de praticiens experts (kinésithérapeutes et psychologues) en les comparant avec des patients sains. Les patients sont répartis comme indiqués dans le tableau 6.3.

EmoPain contient trois catégories de caractéristiques extraites à partir de différentes parties du corps venant de capteurs de motion capture et d'EMG de surface :

- des informations posturales décrites par 13 angles,
- des informations basées sur la vitesse et l'énergie calculée à partir de la somme des carrés des vitesses angulaires à chacun des 13 angles,
- des niveaux d'activité musculaire utilisant l'enveloppe supérieure du signal rectifié de chacune des 4 sondes sEMG.

Toutes ses informations permettent d'obtenir un vecteur de caractéristiques en 30 dimensions (13+13+4) indépendant des proportions du squelette.

Les kinésithérapeutes ont utilisé des signaux provenant de différentes modalités (corps, visage, tête, yeux) pour évaluer le MRSE (*movement related self-efficacy*, l'auto-efficacité liée au mouvement d'une personne.

Le nouveau modèle que nous proposons est basé sur le travail de Wang et al. [303]. Ils ont étudié le **comportement de protection** au sein des activités, plutôt que l'activité elle-même, afin de comprendre si un tel comportement peut être détecté indépendamment du type d'activité exercée dans un ensemble de cinq activités quotidiennes considérées comme exigeantes par les personnes souffrant de

Groupe	Patients sains	Patients CBLP
Entraînement	6	10
Validation	3	4
Test	3	4

TABLEAU 6.3 – Répartitions des patients dans EmoPain

douleur chronique. Pour cela, ils ont utilisé les données dynamiques et temporelles grâce des techniques de *mocap* utilisant des gyroscopes et des accéléromètres, combinées à de l'électromyographie de surface. Leur modèle est basé sur les LSTMs (voir Fig. 6.2) afin de mieux saisir l'aspect dynamique pour la détection automatique du comportement de protection, tâche non prise en compte par les réseaux basés sur des convolutions. D'après ces auteurs, l'architecture optimale est un réseau à trois couches avec 32 unités cachées dans chacune des cinq catégories de comportement de protection qui sont combinées en une classe spéciale appelée « comportement de protection ». Les résultats de la tâche de détection est donc binaire, permettant de distinguer les comportements protecteurs d'un côté et non protecteurs de l'autre.

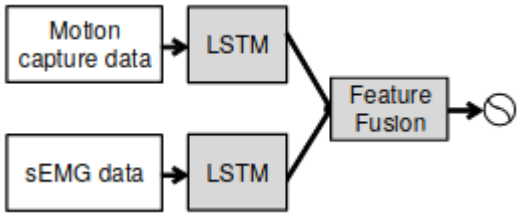


FIGURE 6.2 – Réseau LSTMs de Wang et al. [303]

Wang et al. ont ensuite proposé un second modèle plus complexe basé sur des mécanismes d'attention [304] : une architecture de réseau de neurones « end to end » appelée BANet qui peut, à travers différents types de mouvements, apprendre « quand » (grâce à l'attention temporelle) et « lesquels » (via l'attention spatiale) des sous-ensembles des articulations anatomiques contribuent le plus à la détection du comportement de protection. Les données d'entrées de BANet sont les angles locaux des articulations et leurs énergies (le carré de la vitesse angulaire), chaque angle étant calculé à partir de trois points anatomiques pertinents (voir figure 6.3).

En 2020, dans le cadre de la conférence *Face and Gesture*[2], un challenge a été organisé autour de cette base de données[3]. Ce challenge était le premier défi international portant sur la détection de la douleur et des comportements associés. Il est basé sur le jeu de données EmoPain, qui contient des données sur le visage et les mouvements de participants réels souffrant de douleurs chroniques et pratiquants une activité physique. Le challenge EmoPain 2020 consistait en trois tâches principales axées sur la reconnaissance de la douleur à partir des expressions faciales et des mouvements du corps, ainsi que sur la reconnaissance des mouvements du corps liés à la douleur. Les gagnants de la deuxième tâche (reconnaissance de la

2. https ://fg2020.sunai.uoc.edu/

3. https ://wangchongyang.ai/EmoPainChallenge2020/

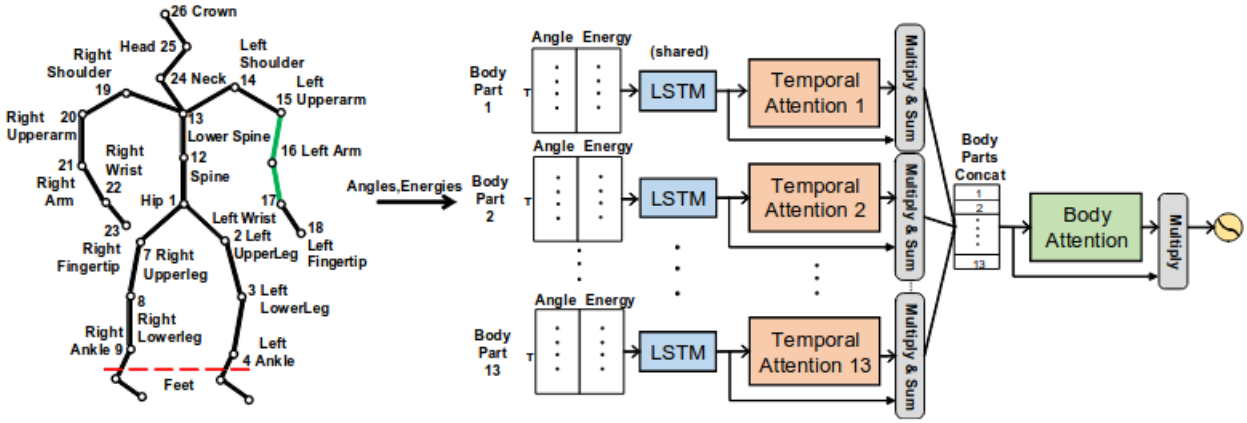


FIGURE 6.3 – BANet [304]

douleur à partir des mouvements) et de la troisième (classification du comportement de mouvement) sont Canavan et Uddin [294] qui ont proposé un modèle de fusion multimodal et multiniveau pour détecter le comportement protecteur et estimer les niveaux d'intensité de la douleur de manière séquentielle. Ils ont utilisé des classificateurs XGBoost et des forêts aléatoires. D'autres modèles ont été proposés, la plupart basés sur des réseaux LSTMs [177, 332]. En comparant différents modèles (LSTM, Transformer et GCN), Radouane et al. ont montré que les GCN était les plus performants [229].

Malheureusement, la base de données EmoPain n'est pas disponible pour les recherches académiques en dehors de la compétition. Les différentes études au cours de la compétition ayant utilisé une partie d'EmoPain ont montré l'intérêt des mécanismes d'attention et des réseaux de neurones de type GCN. Les bases de données sur la posture et le comportement douloureux étant difficile à obtenir et peu nombreuses, nous nous sommes donc tournés vers d'autres jeux de données tels que ceux portant sur la rééducation physique.

6.2.2 Utilisation de la base de données UI-PRMD

La **rééducation physique** est définie par l'OMS comme « un ensemble d'interventions visant à optimiser le fonctionnement et à réduire l'incapacité des personnes atteintes de problèmes de santé en interaction avec leur environnement » [4]. Selon l'OMS, environ 2,4 milliards de personnes souffrent de problèmes de santé qui nécessitent des traitements de réadaptation physique, et ce nombre est en constante augmentation au fil des ans.

La réadaptation réduit l'impact des problèmes de santé, notamment de différents types de maladies et de blessures, et peut compléter les interventions médicales et chirurgicales. L'un des objectifs de la réadaptation physique est d'obtenir les meilleurs résultats possibles et de minimiser ou de ralentir les effets invalidants des affections, telles que les maladies cardio-vasculaires, les cancers et le diabète. La réadaptation est un investissement, avec un rapport coûts/bénéfices intéressant pour

4. https ://www.who.int/news-room/fact-sheets/detail/rehabilitation

les patients et la société. Elle peut permettre d'éviter ou de réduire des hospitalisations coûteuses. La réadaptation est également utilisée à des fins de prévention et permet aux individus de participer à l'éducation et à la responsabilisation en matière d'autogestion, en dotant les personnes de stratégies d'autogestion et des produits d'assistance dont elles ont besoin et en traitant la douleur ou d'autres complications. La réadaptation aide les patients à rester indépendants à domicile et réduit au minimum le besoin d'aide financière ou d'aide aux soins qui en découle.

La réadaptation physique est donc une spécialité médicale qui se concentre sur la restauration des fonctions corporelles de la manière la plus sûre et la plus efficace possible. Au cours d'une séance de rééducation, le comportement des patients reflète leur état de santé et constitue un indicateur important du résultat du traitement. La création d'un système automatique permettant d'évaluer la qualité du mouvement humain de manière objective et fiable peut être utilisée en médecine pour établir un diagnostic différentiel, choisir le traitement adéquat ou surveiller le patient.

En effet, le suivi automatique des activités de réadaptation physique peut aider les travailleurs de la santé en matière de sécurité, d'analyse des tâches quotidiennes, de soutien et de formation de leurs patients. L'objectif de la surveillance automatique des activités de réadaptation physique est de reconnaître l'activité réalisée, l'intensité avec laquelle elle est réalisée ou sa qualité. Il existe peu de jeux de données internationaux liés aux domaines de la réadaptation physique automatique sont disponibles pour la recherche scientifique. En effet, dans le domaine médical, la collecte de grands ensembles de données d'exercices de réhabilitation de patients souffrant d'une déficience ou d'une blessure est plus difficile en raison de problèmes de confidentialité et de sécurité. Par conséquent, seules quelques données publiques pour l'évaluation de la réadaptation sont actuellement disponibles. L'émergence d'ensembles de données plus vastes et plus complets, tels que UI-PRMD, fournit une base pour la recherche.

La quantification du degré d'exactitude de l'exécution des exercices prescrits est importante pour le développement d'outils et de dispositifs de soutien à la réadaptation à domicile. L'évaluation du mouvement est généralement réalisée en comparant l'exécution d'un exercice par un patient à l'exécution souhaitée par des personnes en bonne santé. L'évaluation de la rééducation physique présente de nombreuses applications telles que la surveillance automatisée ou le suivi du comportement à partir de vidéos. L'évaluation de la qualité des mouvements de réadaptation permet aux professionnels de la santé de prendre des mesures appropriées s'il y a lieu de les améliorer. L'évaluation automatique des exercices de réadaptation physique vise à donner un score de qualité à partir d'une séquence de mouvements corporels comme entrée pour une machine ou un algorithme d'apprentissage profond.

Dans le cadre de l'étude UI-PRMD, 10 sujets ont effectué 10 mouvements de rééducation (voir Fig. 6.4) : squat profond (*deep squat*), franchissement de haie (*hurdle step*), fente avant (*inline lunge*), fente de côté (*side lunge*), lever de chaise (*sit to stand*), élévation active de la jambe en position debout (*standing active straight leg raise*), abduction active de l'épaule en position debout (*standing shoulder abduction*), extension active des épaules en position debout (*standing shoulder extension*), rotation interne-externe des épaules en position debout (*standing shoulder internal-external rotation*), et enfin élévation des épaules en position debout (*standing shoulder scaption*). Chaque mouvement est répété 10 fois, à la fois correctement

et incorrectement. Pour les mouvements incorrects, tous les sujets ont exécuté les mouvements de manière sous-optimale [295].

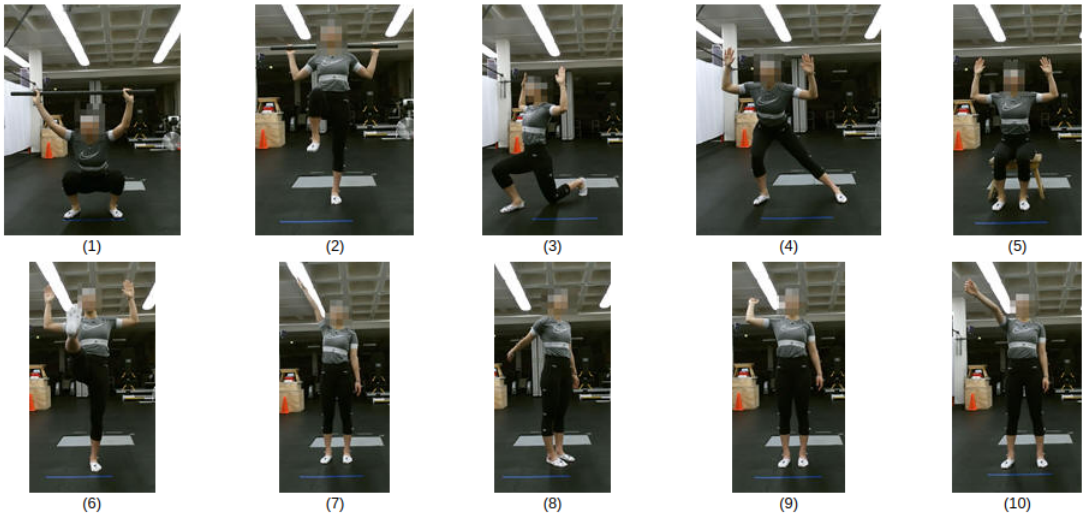


FIGURE 6.4 – Exercices utilisés dans UIPRMD [295]

L'acquisition des données a été faite avec 2 systèmes : une caméra Vicon et une caméra de profondeur Kinect V2. Les données sont présentées sous forme de positions et d'angles des articulations du corps dans les modèles fournis par ces systèmes mocap. Chaque séquence dure environ 20 secondes. La fréquence d'images de la mocap était de 100 Hz avec Vicon et de 30 Hz avec Kinect.

6.3 Construction du graphe

La plupart des architectures de modèles utilisant les bases de données EmoPain et UI-PRMD sont basées sur des CNNs et/ou des RNNs qui calculent les dépendances temporelles entre les images, mais en ignorant la structure topologique du corps humain. Or le squelette humain peut être considéré comme un graphe acyclique dirigé, naturellement structuré, où chaque point d'intérêt (ou articulation) est représentée par un nœud du graphe et est connectée aux autres articulations par des arêtes (voir Fig. 6.5). Chacune de ces articulations possède différentes caractéristiques, telles que les coordonnées tridimensionnelles et/ou les angles d'Euler. Ces valeurs sont données pour chaque image appartenant à une séquence vidéo.

Pour rappel, comme expliqué en section 3.6, les réseaux à graphes sont spécialement conçus pour extraire les informations des nœuds (ici les points d'intérêts du corps humain représentés dans le squelette), des arêtes, et de leurs relations afin de faire des prédictions basées sur les caractéristiques extraites. Les réseaux convolutifs à graphes généralisent la convolution des images aux graphes et ont été adoptés avec succès dans les taches de reconnaissance d'actions. Ici les données brutes issues du squelette d'une image sont toujours une séquence de vecteurs, où chaque vecteur représente les coordonnées 3D des articulations humaines. Une vidéo est donc considérée comme une séquence d'images, avec un squelette par image. Étant donné une séquence de coordonnées, nous définissons V comme le nombre d'articulations

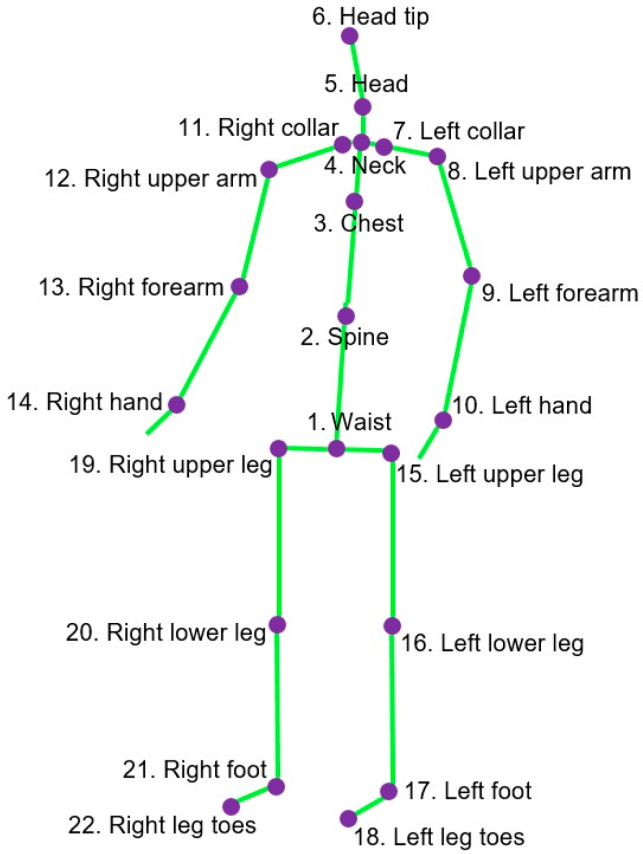


FIGURE 6.5 – Squelette de données issues de UI-PRMD [295]

représentant chaque squelette et T comme le nombre total de squelettes composant la séquence. Pour représenter la séquence, on construit un graphe spatio-temporel (voir Fig. 6.6) G tel que :

$$G = (N, E) \tag{6.1}$$

où

$$N = v_{ti} | t = 1 \ldots T, i = 1 \ldots V \tag{6.2}$$

représente l'ensemble de tous les nœuds v_{ti} du graphe, c'est-à-dire les points d'intérêts/articulations du corps tout au long de la séquence temporelle.

E représente l'ensemble de toutes les connexions entre les nœuds, où

$$E_S = (v_{ti}, v_{tj}) | i, j = 1 \ldots V, t = 1 \ldots T \tag{6.3}$$

est composé par les connexions spatiales du squelette à chaque intervalle de temps t, pour toute paire d'articulations (i, j) reliées dans le squelette.

$$E_T = (v_{ti}, v_{(t+1)i}) | i = 1 \ldots V, t = 1 \ldots T \tag{6.4}$$

représente toutes les connexions temporelles entre les articulations le long d'images consécutives.

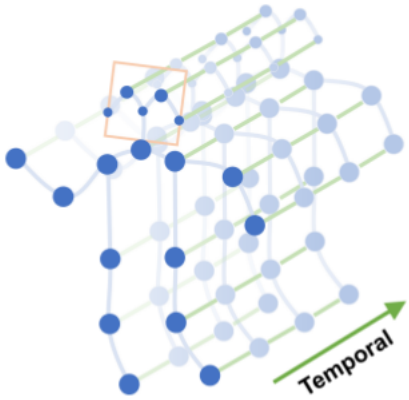


FIGURE 6.6 – Graphe spatio-temporel du squelette [326]

6.4 Notre modèle

À partir des éléments du graphe que nous venons de décrire, nous proposons d'utiliser des mécanismes d'attention pour l'évaluation de la réadaptation physique. Nous avons utilisé un réseau Transformer standard avec un mécanisme d'auto-attention, puis nous avons tiré parti des squelettes de graphes comme entrées pour un réseau de type **ST-GCN** (*Spatial – Temporal Graph Convolutinal Network*) avec des **mécanismes d'attention spatiale et temporelle**.

6.4.1 Mécanisme d'attention

Comme nous l'avons évoqué dans la section 3.5, Vaswani et al. [296] ont proposé de nombreuses améliorations du mécanisme d'attention grâce à un modèle seq2seq sans unités de réseau récurrentes. Ce modèle appelé **Transformer** est entièrement construit sur les mécanismes d'auto-attention sans utiliser d'architecture récurrente alignée sur la séquence. Le Transformer considère la représentation codée de l'entrée comme un ensemble de paires clé-valeur, toutes deux de même dimension que la longueur de la séquence d'entrée. Les clés et les valeurs sont les états cachés du codeur. Dans le décodeur, la sortie précédente est comprimée dans une requête et la sortie suivante est produite en mettant en correspondance cette requête et l'ensemble des clés et des valeurs.

L'**auto-attention** (*self attention*), également appelée intra-attention, est un mécanisme d'attention qui met en relation différentes positions d'une même séquence afin de calculer une représentation de cette même séquence. Ce modèle a été démontré efficace dans la lecture automatique, le résumé de texte ou la génération de descriptions d'images. Il s'agit là de la version la plus classique de l'attention. C'est une attention contextuelle, avec un produit scalaire. L'attention est une moyenne de valeurs associées aux clés correspondant à une requête.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6.5}$$

où Q, K, et V sont des matrices contenant respectivement les vecteurs de requête

(query), de clé (key) et de valeur (value) prédits, et d est la dimension de canal des vecteurs de clés.

Le composant principal du Transformer est le mécanisme d'**auto-attention à têtes multiples** (*multihead attention*). Au lieu de calculer l'attention une seule fois, le mécanisme multitêtes exécute l'attention par produit scalaire plusieurs fois en parallèle. Les sorties indépendantes de l'attention sont simplement concaténées et transformées linéairement dans les dimensions attendues. Selon Vaswani et al., ce mécanisme d'attention multitêtes permet au modèle de traiter conjointement des informations provenant de différents sous-espaces de représentation à différentes positions. Mathématiquement, on note :

$$Multihead(Q, K, V) = [head_1 \ldots head_h]W^O \tag{6.6}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{6.7}$$

où W_i^Q, W_i^K, W_i^V et W^O sont des matrices de paramètres que le modèle va apprendre.

Le modèle Transformer original ne peut pas exploiter le graphe des données du squelette puisqu'il traite les données du squelette comme une séquence linéaire de vecteurs. Cette représentation des données du squelette sous la forme d'une séquence de vecteurs ne peut pas exprimer pleinement la dépendance entre les articulations corrélées du corps humain, ces méthodes ignorent les dépendances biomécaniques entre les articulations et les parties du corps. Afin de proposer une adaptation du Transformer au graphe, nous présentons la conception d'un nouveau modèle basé sur un ST-GCN.

6.4.2 Spatial Temporal Graph Convolutional Networks

L'idée de l'architecture à deux flux provient de la reconnaissance d'action basée sur les images RGB, où le flux optique est extrait d'une vidéo pour modéliser la dépendance temporelle entre les images. Dans les réseaux de graphes à deux flux, les informations de mouvement du squelette sont exploitées et combinées aux informations spatiales pour améliorer les performances.

Le modèle ST-GCN [327] a été le premier à appliquer des GCN pour une tache de reconnaissance d'action à partir de squelettes. Une séquence de squelettes est habituellement représentée par les coordonnées 2D ou 3D de chaque articulation humaine présentes dans les images composant une séquence vidéo. Ces squelettes peuvent être obtenus à partir de dispositifs de capture de mouvement ou d'algorithmes d'estimation de pose. À partir de ces séquences d'articulations du corps (sous forme de coordonnées 2D ou 3D donc), Yan et al. ont construit un graphe spatio-temporel, où les articulations sont les nœuds du graphe, les arêtes sont les connexions naturelles des articulations du corps humain, et où chaque articulation est reliée à elle-même dans l'image suivante de la séquence vidéo (voir figure 6.7).

Les données d'entrée du ST-GCN sont donc les vecteurs de coordonnées des nœuds du graphe. De multiples couches d'opérations de convolution spatio-temporelle sont appliquées aux données d'entrée et génèrent des cartes de caractéristiques de

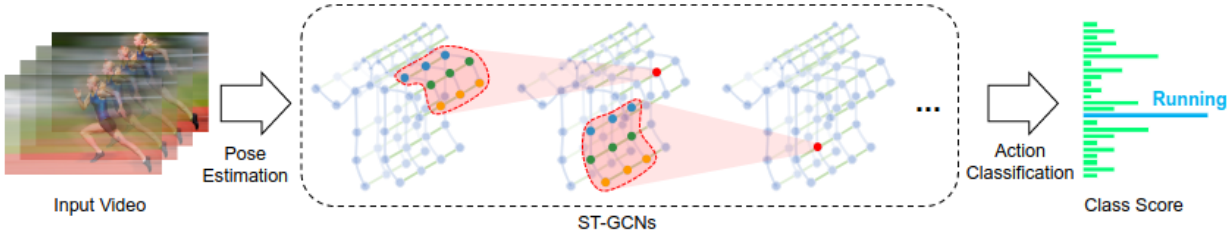


FIGURE 6.7 – ST-GCN

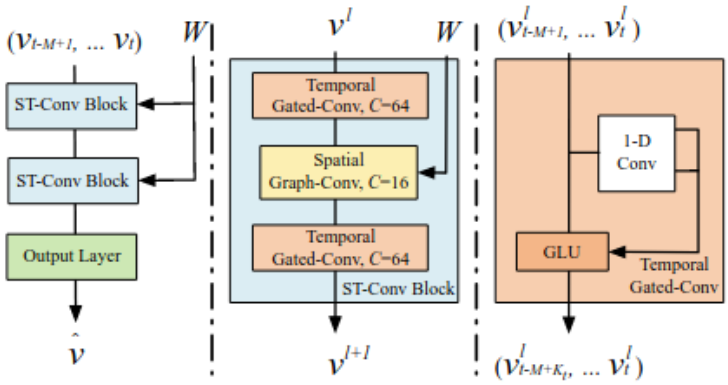


FIGURE 6.8 – Bloc spatio-temporel d'un ST-GCN [327]

haut niveau sur le graphe qui seront ensuite classés par un classifieur standard SoftMax dans la catégorie d'action correspondante. Comme expliqué dans [263], un ST-GCN est structuré comme une hiérarchie de blocs spatio-temporels empilés (voir Fig. 6.8), qui sont composés en interne d'une convolution spatiale (GCN) suivie d'une convolution temporelle (TCN) qui peut être résumée comme suit :

$$f_{out} = \sum_{k}^{K_s} (f_{in} A_k) \tag{6.8}$$

$$A_k = D_k^{-\frac{1}{2}} \left(\tilde{A}_k + I \right) D_k^{-\frac{1}{2}} \tag{6.9}$$

$$D_{ii} = \sum_{k}^{K_s} \left(\tilde{A}_k^{ij} + I_{ij} \right) \tag{6.10}$$

où K_s est la taille du noyau sur la dimension spatiale, $\tilde{A}_k$ est la matrice d'adjacence du graphe non orienté représentant les connexions corporelles. I est la matrice d'identité et W_k est une matrice de poids entraînable. Le réseau de convolution temporelle (TCN) est implémenté sous la forme d'une convolution 2D $1 \times K_t$ opérant sur les dimensions (V, T) du volume d'entrée (C_{in}, V, T), où K_t est le nombre d'images considérées dans le champ récepteur du noyau.

6.4.3 Notre modèle final

L'objectif de ce travail est de créer un modèle permettant de combiner une architecture à deux flux des ST-GCN et le mécanisme d'attention du Transformer. Nous nous sommes donc inspirés du modèle Spatial and Temporal Transformer networks (ST-TR) [227] pour construire un réseau convolutif à deux flux Spatio-Temporal Graph qui prend des entrées de graphes et utilise l'auto-attention du modèle Transformer. L'architecture ST-TR est similaire à celui proposé par Shi [263], à savoir un ST-GCN, mais ici certaines couches comportent des mécanismes d'attention spatiale et temporelle similaires, mécanisme de self-attention du modèle Transformer. Dans le modèle ST-TR, le squelette est considéré comme similaire à un sac de mots. Nous pouvons donc utiliser l'auto-attention du Transformer pour extraire des regroupements de nœuds codant la relation entre les articulations environnantes, comme les mots d'une phrase en TAL. Les modules d'attention spatiale et d'attention temporelle sont utilisés pour extraire les corrélations sur les deux dimensions. L'auto-attention découvre automatiquement les relations articulaires pertinentes pour prédire la qualité d'un mouvement.

De plus, l'analyse des changements dans les articulations au cours d'une séquence permet au modèle d'apprendre des relations à long terme sur différentes images, tout comme les relations entre les phrases en langage naturel. Les avantages de cette méthode sont doubles : une représentation de la pose sur une large plage temporelle permet au mécanisme d'attention d'attribuer une importance estimée à chaque point d'observation et à chaque instant en tenant compte de la connaissance sur toute la fenêtre temporelle. Les informations ainsi extraites sur les poses sont suffisantes pour apprendre une représentation globale. Cependant, il est nécessaire de trouver une représentation hiérarchique qui respecte les relations spatio-temporelles des articulations.

Comme expliqué dans [227], l'auto-attention opère sur chaque paire de nœuds en calculant un poids pour chacun d'eux qui représente la force de leur corrélation. Ces poids sont ensuite utilisés pour évaluer la contribution de chaque articulation du corps v_{ti}, proportionnellement à la pertinence du nœud par rapport à tous les autres.

Comme le montre la figure 6.9, sur chaque flux, les premières couches extraient des caractéristiques de bas niveau par le biais de 3 couches ST-GCN standard. À chaque couche successive, sur le flux spatial (Spatial Transformer, S-TR), l'auto-attention spatiale (Spatial Self Attention, SSA) est utilisée pour extraire les informations spatiales, suivie d'une convolution 2D sur la dimension temporelle (TCN), tandis que sur le flux temporel (Temporal Transformer, T-TR), l'auto-attention temporelle (Temporal Self Attention, TSA) est utilisée pour extraire les informations temporelles, tandis que les caractéristiques spatiales sont extraites par un GCN. Les deux flux diffèrent par la manière dont les mécanismes d'auto-attention proposés sont appliqués : le SSA opère sur le flux S-TR, tandis que le TSA sur le flux T-TR.

La formule pour l'attention spatiale est :

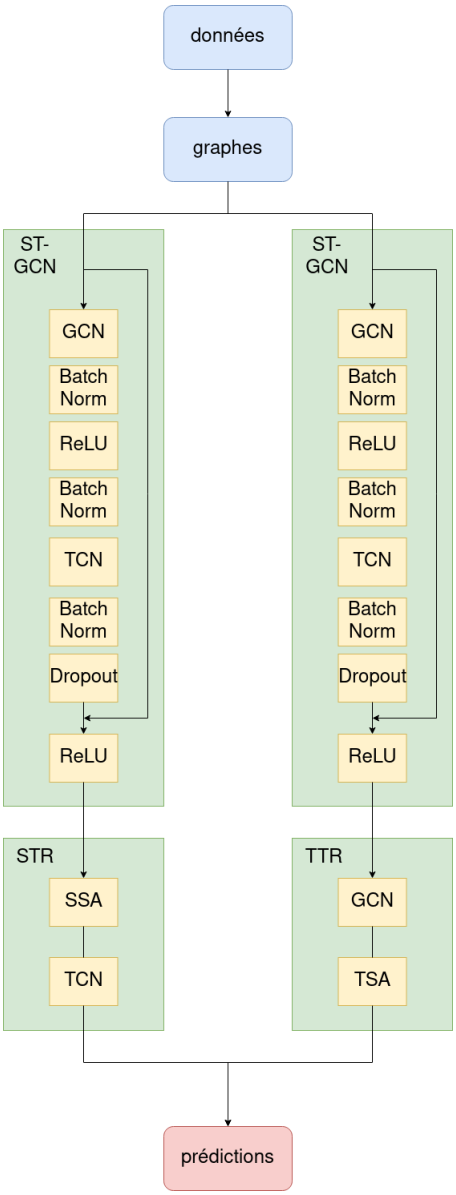


FIGURE 6.9 – Notre modèle

$$Attention = \sum_j softmax_j \left(\frac{q_i \cdot k_j^T}{\sqrt{d_k}} \right) v_j^t \tag{6.11}$$

où v_i et v_j sont deux points différents du squelette.

Pour l'attention temporelle, la formule est la même, v_i et v_j sont le même point d'intérêt v à deux instants différents.

Dans le bloc S-TR, le module SSA se concentre sur les relations spatiales entre les articulations. La sortie du module SSA est transmise à un module convolutif 2D sur la dimension temporelle (TCN) afin d'extraire les caractéristiques temporelles pertinentes :

$$\textbf{S-TR}(x) = \text{Conv}_{2D(1\times K_t)}(\textbf{SSA}(x)) \tag{6.12}$$

Le flux temporel se concentre sur la découverte des relations temporelles inter-images. Comme pour le bloc S-TR, à l'intérieur de chaque couche T-TR, un sous-

module de convolution de graphe standard est suivi par le module d'auto-attention temporelle :

$$\textbf{T-TR}(x) = \textbf{TSA}(GCN(x)). \tag{6.13}$$

Sur les deux flux, les caractéristiques des nœuds sont d'abord extraites par un réseau résiduel à trois couches, où chaque couche traite l'entrée sur la dimension spatiale par GCN, et sur la dimension temporelle par un TCN standard. SSA et TSA sont ensuite appliqués aux blocs S-TR et T-TR dans les couches suivantes en remplacement des modules d'extraction de caractéristiques GCN et TCN respectivement. Les blocs S-TR et T-TR sont entraînés de bout en bout séparément avec leurs couches d'extraction de caractéristiques correspondantes. Les sorties des sous-réseaux sont finalement fusionnées en additionnant leurs scores de sortie softmax pour obtenir la prédiction finale pour la classification. Plizzari et al. ont montré que ce type de modèle donne de meilleurs résultats que les modèles ST-GCN ou 2s-AGCN classiques [227].

6.5 Résultats

Pour évaluer les performances de notre modèle, nous avons utilisé UI-PRMD pour prédire un score de qualité du mouvement avec différents modèles : le modèle hiérarchique de Liao et al., un encodeur-décodeur, un Transformer, XGBoost, et notre modèle. Toutes nos expériences sont menées avec la bibliothèque d'apprentissage profond PyTorch [5].

Le modèle hiérarchique introduit par Liao et al.(2020) dans [179], avec des scores d'écart absolu moyen (Mean Absolute Deviation, MAD) et d'écart quadratique moyen (Root Mean Square Deviation, RMSD). Le modèle de Liao et al. exploite les caractéristiques spatiales des mouvements humains par un traitement hiérarchique des déplacements articulaires des différentes parties du corps via une série de sous-réseaux qui fusionnent progressivement les vecteurs caractéristiques extraits. Le réseau contient à la fois des couches convolutives pour l'apprentissage des dépendances spatiales et des couches récurrentes pour l'encodage des corrélations temporelles dans les données de mouvement. Les couches hiérarchiques initiales du modèle utilisent des filtres convolutifs unidimensionnels en stries pour apprendre les dépendances spatiales dans les mouvements humains. Elles sont suivies par une série de couches récurrentes LSTM pour modéliser les corrélations temporelles dans les représentations apprises.

Pour tester les LSTMs, nous avons repris l'architecture de notre modèle Encoder-Decoder utilisé sur la base de données BoLD. Ce modèle prend une séquence linéaire des coordonnées 3D de chaque point clé p dans une séquence de n images d'une vidéo et sort les valeurs de score estimées de l'exercice donné effectué dans la vidéo. Comme dans notre modèle, l'encodeur et le décodeur sont des LSTMs multicouches. Le LSTM ne prend qu'un seul élément de la séquence à la fois, donc si la séquence d'entrée à une longueur de l, le LSTM a besoin de l pas de temps pour lire la séquence entière. À chaque pas de temps, le décodeur génère un élément de sa séquence

5. https ://pytorch.org/

de sortie en fonction de l'entrée reçue et de son état actuel, tout en mettant à jour son propre état pour le pas de temps suivant.

Dans le cadre du concours EmoPain Challenge 2020 [76], XGBoost a été le meilleur modèle pour la détection des comportements de protection et l'estimation de la douleur à partir des informations sur les mouvements humains [294]. XGBoost (qui signifie *Extreme Gradient Boosting*) est un algorithme qui combine de manière itérative les prédictions de plusieurs apprenants faibles, tels que les arbres de décision, pour produire un modèle beaucoup plus robuste à l'aide du boosting de gradient. Nous avons décidé d'utiliser un modèle XGBoost de scikit-learn [6].

Pour tous ces modèles, les entrées sont les caractéristiques 3D des articulations provenant du système mocap, c'est-à-dire des vecteurs linéaires de coordonnées articulaires pour chaque répétition d'un exercice. Pour notre modèle, nous avons utilisé la descente de gradient stochastique (SGD) avec le momentum de Nesterov (0.9) comme stratégie d'optimisation. La taille du lot (*batch size*) est de 32. Le taux d'apprentissage est fixée à 0,0001. Nous avons entrainé notre modèle pendant 100 époques. Le résultat est un score numérique de la qualité du mouvement (compris entre 0 et 1) pour l'ensemble de l'exercice.

Nous avons repris les mêmes métriques que Liao et al. pour pouvoir comparer nos résultats :

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}| \tag{6.14}$$

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2} \tag{6.15}$$

Nous avons ensuite testé notre modèle sur une classification binaire (mouvement correct/incorrect) afin de se rapprocher de la problématique de détection de la douleur. Nous avons utilisé le F1 score comme métrique tel que :

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{6.16}$$

où pour rappel, dans le cadre d'une classification binaire, on note :

		Classe réelle	
		Positif	Négatif
Classe prédite	Positif	TP, *True Positive* (vrai positif)	FP, *False positif* (faux positif)
	Négatif	FN, (*False Negative*, faux négatif) (faux négatif)	TN, (*True Negative*, vrai négatif) (vrai négatif)

6. https ://scikit-learn.org

	Liao et al.	XGBoost	Encoder Decoder	Transformer	Graph Transformer
MAD	0.0172	0.0168	0.0267	0.0169	**0.0167**
RMSD	0.0273	0.0238	0.0233	0.0186	**0.0143**
F1 score	-	0.80	0.81	0.83	**0.85**

TABLEAU 6.4 – Prédiction de la qualité du mouvement sur UI-PRMD (Mean Absolute Deviation, MAD et Root Mean Square Deviation, RMSD) et de classification (F1 score)

On peut observer dans le tableau 6.4 que notre modèle surpasse les autres modèles. Si nous considérons le MAD, on peut observer que les résultats du Transformer et de XGBoost sont proches. En ce qui concerne la classification, nos résultats montrent que notre modèle surpasse tous les modèles avec un score F1 de 0,85, et que le modèle Transformer surpasse le modèle basé sur les LSTMs.

Ces résultats sont à mettre en perspective avec ceux obtenus pour le même type de classification binaire sur EmoPain où le modèle BANet de Wang et al. ont obtenu un F1 score de 0.84 [304], et lors du challenge EmoPain où le modèle proposé par les gagnants Radouane et al. a obtenu un F1 score de 0.53 [229].

Dans ce chapitre, nous avons donc construit des graphes à partir de données de *mocap* pour conserver les informations spatio-temporelles naturellement contenues dans la structure du squelette humain. Nous avons validé notre modèle sur UI-PRMD, un jeu de données pour l'évaluation de la rééducation physique. Nous avons démontré l'utilité d'une méthode d'apprentissage profond qui combine un réseau ST-GTCN et le mécanisme d'auto-attention du modèle Transformer. L'utilisation d'un réseau à graphe type ST-GCN intégrant des mécanismes d'attention spatial et temporel permet d'améliorer l'évaluation de la qualité du mouvement.

Conclusion

Les travaux réalisés dans le cadre de cette thèse ont porté sur la conception et l'implémentation d'un modèle d'apprentissage profond pour la détection de la douleur à partir de la posture à partir de données vidéos. Cette problématique a fait ressortir plusieurs sous-problèmes :

— quel type de données sont le plus adaptées ?
— comment prendre en compte l'aspect dynamique ?
— quelle architecture de réseaux de neurones choisir ?

Le comportement du patient pendant une séance d'exercices de rééducation reflète son état de santé et constitue donc un indicateur important du résultat du traitement. Le suivi des patients permet aux médecins et aux kinésithérapeutes de suivre les progrès des patients et donc d'adapter le traitement en conséquence.
Pour répondre à tous ces sujets, nous avons implémenté des réseaux de neurones profonds pour analyser la posture à partir de données extraites de vidéos. Ces travaux ont donné lieu à deux publications dans des conférences nationale et internationale que nous résumerons dans la section qui suit.

Contributions et diffusion des travaux de thèse

Nos premiers résultats dans le chapitre 5 ont démontré que les caractéristiques de la posture telles que les coordonnées des points d'intérêts ou les angles entre les segments du corps permettaient de déterminer l'état émotionnel d'une personne. Nous avons montré l'efficacité des architectures de type Encoder-Decoder et Transformer pour l'informatique affective. Ces modèles ont été validés sur le jeu de données BoLD. Les résultats ont été publiés dans *Emotion Recognition through body pose using LSTM neural networks* de la Rencontre des Jeunes Chercheurs en Intelligence Artificielle (RJCIA) organisée dans le cadre de la conférence Plate-Forme Intelligence Artificielle (PFIA) 2021

Dans le chapitre 6 nous proposons d'utiliser le mécanisme d'auto-attention du modèle Transformer en l'intégrant dans une architecture de type ST-GCN. Nous avons construit des graphes à partir de données *mocap* afin de conserver les informations spatio-temporelles naturellement contenues dans la structure du squelette humain. Nous avons validé notre modèle de réseau de graphes à deux flux sur UI-PRMD, un jeu de données pour l'évaluation de la réadaptation physique, pour prédire un score de qualité et des tâches de classification binaire. Ces résultats ont été présentés lors de la *International Conference on Automatic Face and Gesture Recognition 2023* dans le *Workshop on Learning with few or without annotated face, body and gesture data.*

Limites des travaux

Les modèles obtenus par apprentissage automatique profond comme les nôtres sont souvent critiqué pour leur côté boite noire, c'est-à-dire l'absence d'explicabilité, crucial dans le contexte médical. De nombreuses recherches essaient de résoudre ce problème grâce à la visualisation de caractéristiques par exemple. Une autre limite concerne la taille des jeux de données utilisées pour l'entraînement de nos modèles. Les réseaux de neurones nécessitent de très grandes quantités de données. Or, dans le contexte médical, il est extrêmement compliqué de récolter et de stocker des données privées venant de patients.

Perspectives

Pour les travaux futurs, nous prévoyons d'inclure d'autres caractéristiques géométriques, tel que l'énergie cinétique ou encore l'accélération, dans notre modèle afin de l'améliorer et de le tester sur d'autres jeux de données de réadaptation physique.

Une solution pour améliorer la reconnaissance de la douleur serait de combiner la reconnaissance faciale, posturale et vocale de la douleur, ou encore d'ajouter d'autres marqueurs physiologiques de la douleur comme l'ECG ou l'EEG, et d'intégrer au modèle d'autres données sur le patient, comme les scores de questionnaires de personnalité [54] par exemple.

Récolter et annoter des données médicales annotées et conformes au RGPD est extrêmement complexe. Une solution serait de générer de nouvelles données grâce à des techniques telles que les GANs. De plus, chaque type de vérité terrain ayant ses forces et ses faiblesses, la meilleure option serait de créer des bases de données avec différents types de vérité terrain : auto-évaluation par des patients et évaluation par des experts, avec différentes échelles (échelles EVS ou PSPI, etc), et d'évaluer/comparer les systèmes de reconnaissance avec les types de vérité de terrain disponibles.

Enfin, une autre approche sera intéressante à explorer dans le contexte médical : l'apprentissage fédéré (Federated Learning). Il s'agit d'un cadre d'apprentissage automatique dans lequel plusieurs clients (comme des appareils mobiles) forment en collaboration un modèle tout en gardant leurs données décentralisées. L'apprentissage fédéré incarne les principes de collecte et de réduction des données, et peut atténuer bon nombre des risques et des coûts systémiques liés à la vie privée résultant des approches traditionnelles et centralisées de l'apprentissage automatique.

Régulation des systèmes d'IA dans la Santé

Plusieurs questions demeurent en suspens et attendent les réponses des régulateurs dans le domaine de la Santé et de l'informatique :

- les médecins vont-ils prescrire ces médicaments numériques ?
- voudront-ils entrer dans une relation clinique numérique avec leurs patients ?
- quid des effets secondaires non négligeables, notamment une dépendance aux écrans ou aux environnements numériques ?

— quel type de pharmacovigilance utiliser pour assurer un usage sécuritaire de ces médicaments ?

Les enjeux des médicaments numériques ne graviteront donc pas seulement autour de la santé du patient, mais concerneront aussi la protection et la valorisation de ces données. Des projets sont à l'étude pour mettre en place une véritable « santé publique numérique ». Ainsi, depuis 2018, l'Union Européenne (UE) a pour objectifs d'accompagner les pers et les entreprises dans le renforcement et le développement de l'IA dans le but de garantir une sécurité et des droits fondamentaux. Entre mars et décembre 2018, l'UE a ainsi organisé des parties prenantes en matière d'éthique pour une IA digne de confiance. Les résultats ont été publiés dans un livre blanc en février 2020. La même année, un ensemble de résolutions contenant des recommandations tel que la responsabilité civile, les droits de propriété intellectuelle, et les aspects éthiques de l'IA. En 2021, commission pour harmoniser règles en matières d'IA, la considération pénale a également été prise en compte dans le cadre de son utilisation par les autorités judiciaires. Enfin, en septembre 2022, les premières propositions de directives sur la responsabilité en matière d'IA. Des normes et des standards seront amenés à être mis en place dans les prochaines années.

Annexes

Thérapie numérique et douleur

	Digital Health	Digital Medicine	DTx
Definition	Digital health includes technologies, platforms, and systems that engage consumers for lifestyle, wellness, and health-related purposes and capture, store, or transmit health data and/or support life science and clinical operations.	Digital medicine includes evidence-based software and/or hardware products that measure and/or intervene in the service of human health.	DTx products deliver evidence-based therapeutic interventions to prevent, manage, or treat a medical disorder or disease.
Clinical evidence	Typically do not require clinical evidence.	Clinical evidence is required for all digital medicine products.	Clinical evidence and real-world outcomes are required for all DTx products.
Regulatory oversight	These products do not meet the regulatory definition of a medical device and do not require regulatory oversight.	Requirements for regulatory oversight vary. Digital medicine products that are classified as medical devices require clearance or approval. Digital medicine products used as a tool to develop other drugs, devices, or medical products require regulatory acceptance by the appropriate review division.	DTx products must be reviewed and cleared or certified by regulatory bodies as required to support product claims of risk, efficacy, and intended use.
Examples	Lifestyle apps, electronic medical record systems, personal health records, telehealth	Pacemaker, insulin pump, ingestible sensors, digital components integrated with drugs, remote monitoring tools	DTx that deliver a medical intervention to treat/manage/prevent a disease

DTx: digital therapeutics

FIGURE 6.10 – Définitions de la santé numérique, la médecine numérique et thérapie numérique [65]

Table 1. Biomarker Definitions with Present and Future Examples for Pain

Type of Biomarker	Definition	Pain Examples (top rows, present; bottom rows, future)
Diagnostic	To detect or confirm the presence of a disease or condition.	QST, EEG, intra-epidermal nerve fiber density
		Microneurography, neuroimaging, genetics
Monitoring	To assess status of a disease or condition or effect of a medical product by any biomarker that is measured serially.	QST, compound levels in plasma, CSF
		Neuroimaging, EEG, intra-epidermal nerve fiber density
Pharmacodynamic/ Response	To show that a biological response occurs in an individual exposed to a medical product.	QST, neuroimaging, EEG, changes in cytokines
		Specific mechanistic/biochemical pain drivers, intra-epidermal nerve fiber density
Predictive	To identify individuals with more likely than individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product.	Genetics
		Neuroimaging, EEG, intra-epidermal nerve fiber density
Prognostic	To identify likelihood of a clinical event, disease recurrence, or progression in patients with disease of interest.	Genetics
		Neuroimaging, EEG, intra-epidermal nerve fiber density
Safety	Measured before or after an exposure to a medical product to indicate likelihood, presence, or extent of toxicity.	Treatment related, e.g., sedation, tolerance, constipation, respiratory depression
		Neuroimaging, EEG
Susceptibility/Risk	Potential for developing a disease or medical condition	Genetics
		Neuroimaging, EEG

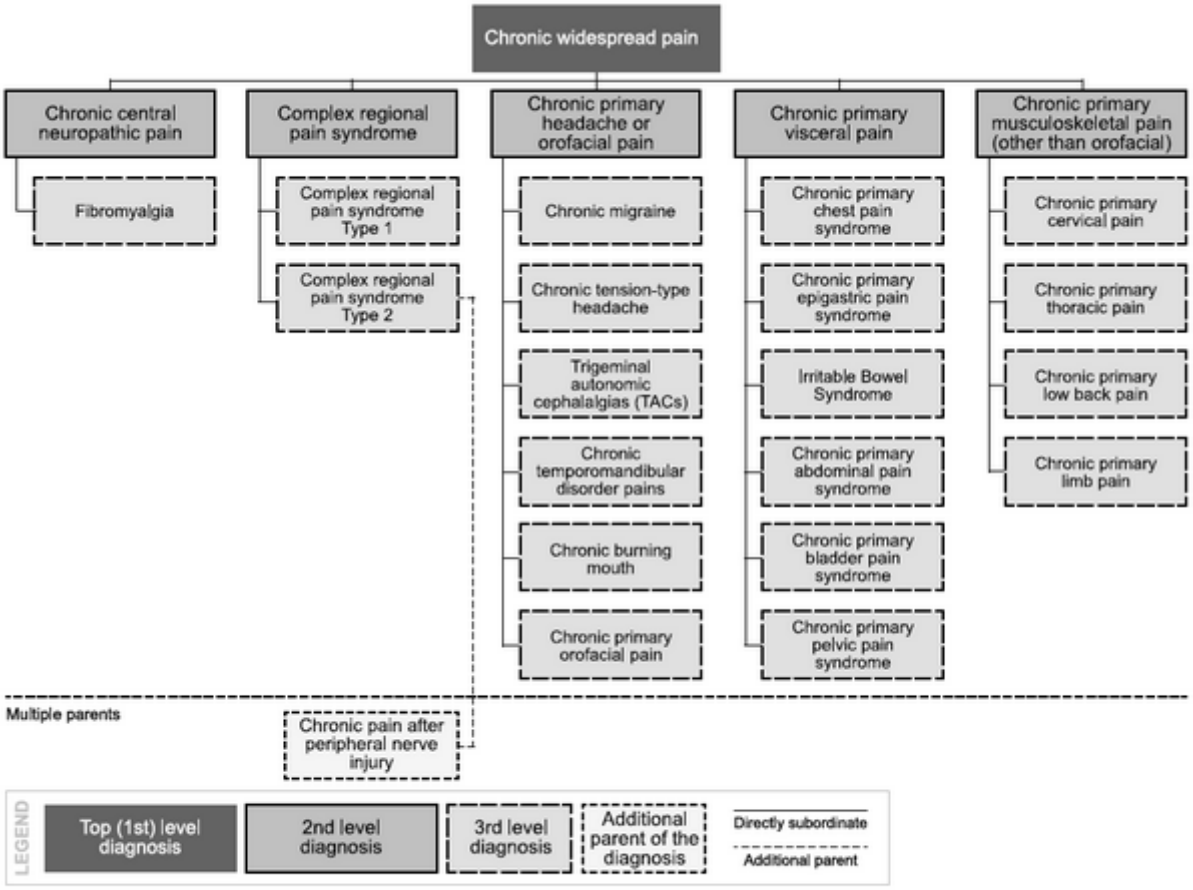
Adapted from "BEST (Biomarkers, Endpoints, and other Tools) Resource," a publication produced by the joint FDA-NIH Biomarker Working Group, December, 2016 (FDA-NIH Biomarker Working Group, 2016).

FIGURE 6.11 – Exemples d'utilisation des Marqueurs de la douleur [290]

	Symp-tom	Signs (= Objectively Observable Features)	Physical Basis	Examples
CP: Canonical Pain				
PCT: Pain with Concordant Tissue Damage	Pain	Manifestation of tissue damage Report of pain concordant with stimulus sufficient to cause this tissue damage Protective response	Activation of nociceptive system through peripheral tissue damage	Primary sunburn Pain from strained muscle Pain from fracture Pulpitis
VP: Variant Pain				
PNT: pain with peripheral trauma but no concordant tissue damage	Pain	Report of pain associated with stimulus intensity insufficient to cause tissue damage	Activation of pain system through cognitive mechanisms regarding threat of tissue damage, the latter often based on peripheral non-nociceptive input to the CNS	Secondary sunburn without tissue damage Myofascial pain disorder Tension-type headache Chronic back pain
NN: neuropathic nociception (pain with no peripheral trauma)	Pain	Report of pain No identifiable pathological peripheral stimulus History of probable causes	Disordered nociceptive system Neuropathic (for example in result of demyelination of nerve fibers)	Trigeminal neuralgia Post-herpetic neuralgia Diabetic neuropathy
PRP: Pain-Related Phenomena Without Pain				
PBWP: pain behavior without pain		Sick role behaviors accompanied by normal clinical examination Report of pain discordant with physical signs Grossly exaggerated pain behaviors Identified external incentives	Description of pain relates to mental states such as anxiety, rather than peripheral tissue locus Misinterpretation of sensory signals by the emotional or cognitive systems Deception by patient	Factitious pain Malingering Anxiety-induced pain report
TWP: tissue-damage without pain		Manifestation of tissue damage normally of the sort to cause pain No reported pain	suppression of pain system by one or other mechanism	Stress associated with sudden emergencies Physiological damping of the pain process caused by adrenalin Placebo induced opioid analgesia Genetic insensitivity to pain

Table 1: Types of Pain and of Pain-Related Phenomena

FIGURE 6.12 – Classification de différents types de douleurs selon les signes cliniques.



Source: Nicholas MK, et al. The IASP classification of chronic pain for ICD-11: chronic primary pain. Pain. 2019 Jan.

FIGURE 6.13 – Classification ICD-11 de la douleur

Allodynia: pain due to a stimulus that does not normally provoke pain. Note: The stimulus leads to an unexpectedly painful response.
Analgesia: absence of pain in response to stimulation which would normally be painful.
Dysesthesia: an unpleasant abnormal sensation, whether spontaneous or evoked. Note: Special cases of dysesthesia include hyperalgesia and allodynia.
Hyperalgesia: increased pain from a stimulus that normally provokes pain.
Hyperesthesia: increased sensitivity to stimulation, excluding the special senses.
Hyperpathia: a painful syndrome characterized by an abnormally painful reaction to a stimulus.
Hypoalgesia: diminished pain in response to a normally painful stimulus.
Hypoesthesia: decreased sensitivity to stimulation, excluding the special senses.
Paresthesia: an abnormal sensation, whether spontaneous or evoked. Note: paresthesia is to be used to describe an abnormal sensation that is not unpleasant.

Table 1 - Pain terms analyzed

Figure 1 - **IASP pain assessment terminology hierarchy**

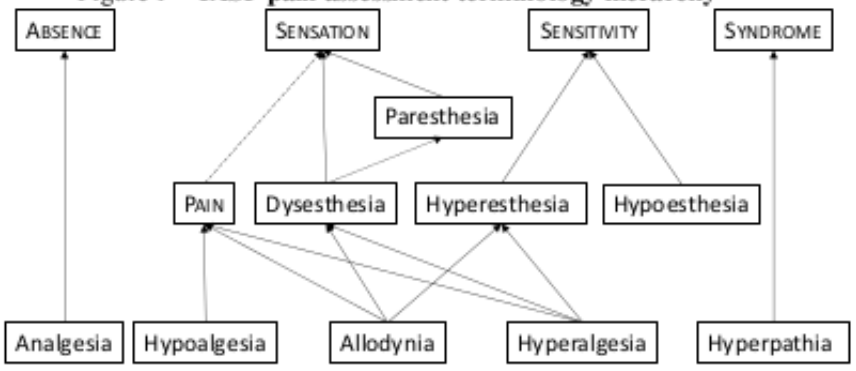


FIGURE 6.14 – Différents types de douleur selon la terminologie de l'IASP

Database	Subjects	Stimuli	Data Modalities (D) / Annotation (A)
UNBC-McMaster Shoulder Pain [165]	25 adult shoulder pain patients	200 range of motion tests with affected and unaffected limbs	**D:** video of face (low resolution, includes social interaction / talking) **A:** self-report (VAS, sensory & affective verbal scales), observer-assessed pain intensity (OPI), affected/unaffected limb, FACS coding
BioVid Heat Pain [56], [168], [169]	90 healthy adults (age 20-65)	14k heat pain (4 intensities × 20 repetitions × 2 parts × 90 participants); emotion elicitation, posed expression	**D:** video of face, EDA, ECG, sEMG (trapezius muscle; corrugator and zygomaticus for part B) **A:** stimulus (calibrated per person)
BP4D-Spontaneous [163]	41 healthy adults (age 18-29)	41 cold pressor task; emotion elicitation	**D:** video of face (color & 3D) **A:** stimulus, FACS coding
BP4D+ [170]	140 healthy adults (age 18-66)	140 cold pressor task; emotion elicitation	**D:** video of face (color, 3D, thermal), heart rate, respiration rate, blood pressure, EDA **A:** stimulus, FACS coding
MIntPAIN [80]	20 healthy adults (age 22-42)	2k electrical pain (40 stimuli in 4 intensities × 2 trials × 20 participants)	**D:** video of face (color, depth, thermal) **A:** stimulus (calibrated per person), self-report (VAS)
COPE [171]	26 neonates (age 18-36 hours)	60 heel lancing for blood collection; non-painful stimuli	**D:** 204 photographs of face **A:** category (pain, rest, cry, air puff, or friction)
YouTube [172]	142 infants (age 0-12 months)	immunizations (injection)	**D:** 142 videos with audio **A:** FLACC observer pain assessment
IIIT-S ICSD [173]	33 infants (age 3-24 months)	immunizations (injection) and other pain causes; non-painful cry causes	**D:** 693 audio cry samples **A:** category annotated by doctors and parents (pain, discomfort, hunger/thirst, and three others)
EmoPain[A] [63]	22 chronic lower back pain patients (age $\mu = 50$) + 28 healthy controls (age $\mu = 37$)	physical exercises (therapy scenarios)	**D:** video, audio, motion capture, sEMG (trapezius, lumbar paraspinal muscles) **A:** self report, pain intensity assessed by naive observers from face, presence of pain behaviors assessed by experts from body movement
SenseEmotion[A] [174]	45 healthy adults (age $\mu = 26$)	8k heat pain (3 intensities × 30 repetitions × 2 stimulus sites × 45 participants); emotion elicitation	**D:** video of face, audio, EDA, ECG, sEMG (trapezius muscle), RSP **A:** pain and emotion stimulus (pain calibrated per person)
X-ITE pain[A] [175]	134 healthy adults (age 18-50)	24k phasic pain, 804 tonic pain (both by heat and electical stimulation, each with 3 intensities)	**D:** video of face (color, thermal), video of body (color, depth), audio, EDA, ECG, sEMG (trapezius, corrugator, zygomaticus) **A:** pain stimulus (calibrated per person)

[A] Announced to be published, but not yet available. Check website in table caption for updates.
ECG: electrocardiogram EDA: electrodermal activity sEMG: surface electromyography FACS: Facial Action Coding System RSP: Respiration

FIGURE 6.15 – Bases de données sur la douleur [318]

Reconnaissance d'actions

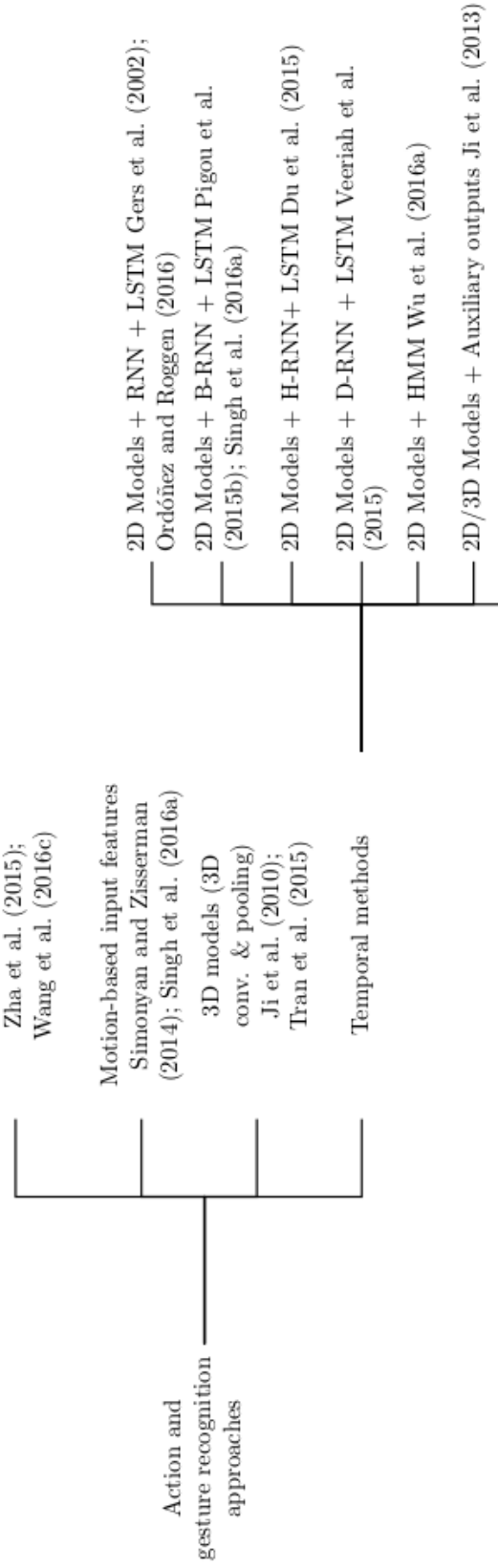


FIGURE 6.16 – Taxonomie des approches par deep learning pour la reconnaissance de gestes et d'actions [195]

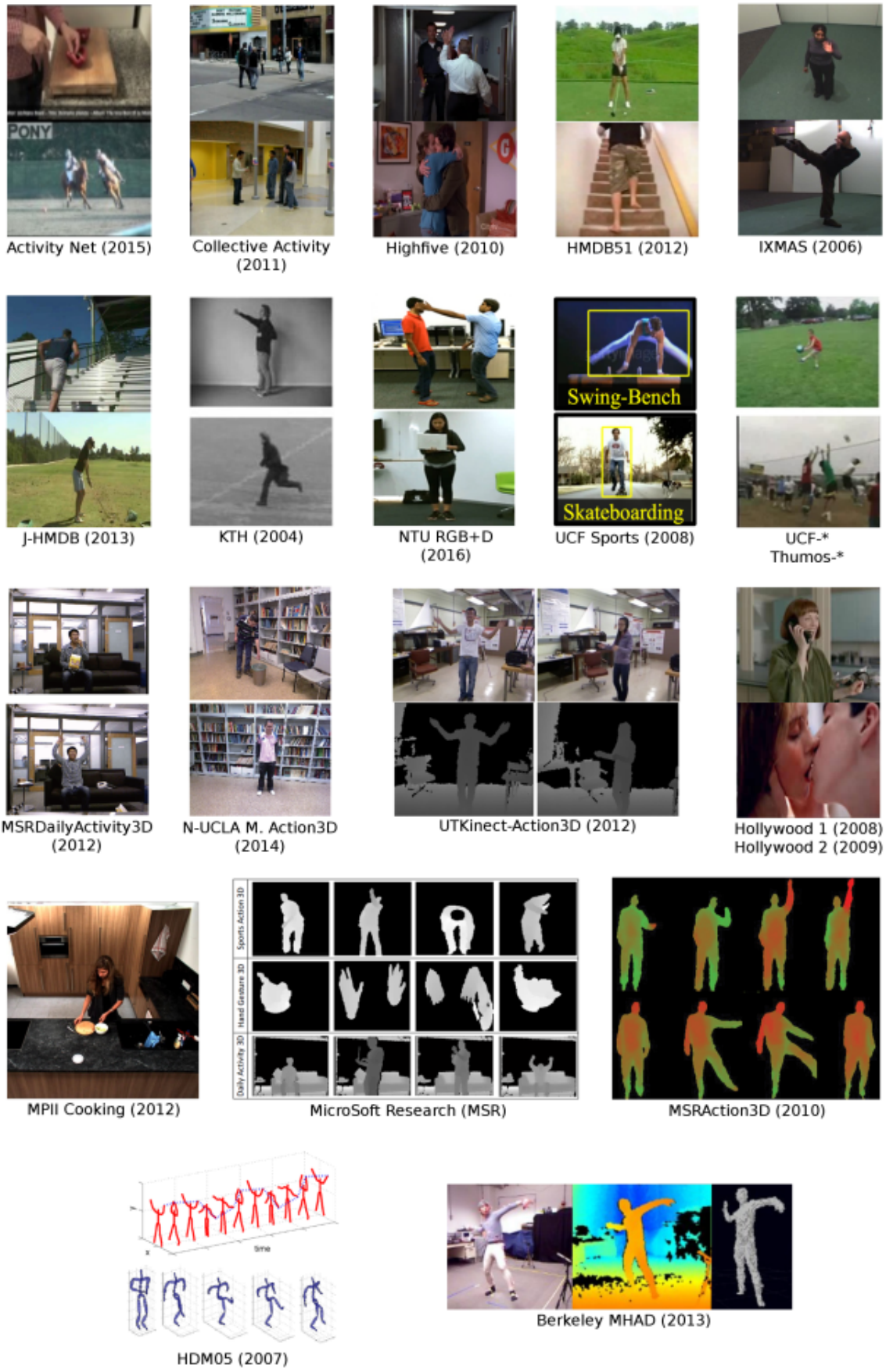


FIGURE 6.17 – Exemples d'images issues de différentes bases de données sur la reconnaissance d'action [195]

Dataset	year	Acquisition device	Seg/Con	Modality	#Class	#Subjects	#Samples	#Views	Metric
CMU Mocap	2001	Mocap	Seg	RGB,S	45	144	2,235	1	Accuracy
HDM05	2007	Mocap	Seg	RGB,S	130	5	2337	1	Accuracy
MSR-Action3D	2010	Kinect v1	Seg	S,D	20	10	567	1	Accuracy
MSRC-12	2012	Kinect v1	Seg	S	12	30	594	1	Accuracy
MSR DailyActivity3D	2012	Kinect v1	Seg	RGB,D,S	16	10	320	1	Accuracy
UTKinect	2012	Kinect v1	Seg	RGB,D,S	10	10	200	1	Accuracy
G3D	2012	Kinect v1	Seg	RGB,D,S	5	5	200	1	Accuracy
SBU Kinect Interaction	2012	Kinect v1	Seg	RGB,D,S	7	8	300	1	Accuracy
Berkeley MHAD	2013	Mocap Kinect v1	Seg	RGB,D,S,Au,Ac	12	12	660	4	Accuracy
Multiview Action3D	2014	Kinect v1	Seg	RGB,D,S	10	10	1475	3	Accuracy
ChaLearn LAP IsoGD	2016	Kinect v1	Seg	RGB,D	249	21	47933	1	Accuracy
NTU RGB+D	2016	Kinect v2	Seg	RGB,D,S,IR	60	40	56880	80	Accuracy
ChaLearn2014	2014	Kinect v1	Con	RGB,D,S,Au	20	27	13858	1	Accuracy JI etc.
ChaLearn LAP ConGD	2016	Kinect v1	Con	RGB,D	249	21	22535	1	JI
PKU-MMD	2017	Kinect v2	Con	RGB,D,S,IR	51	66	1076	3	JI etc.

FIGURE 6.18 – Résumé de bases de données RGB-D[224]

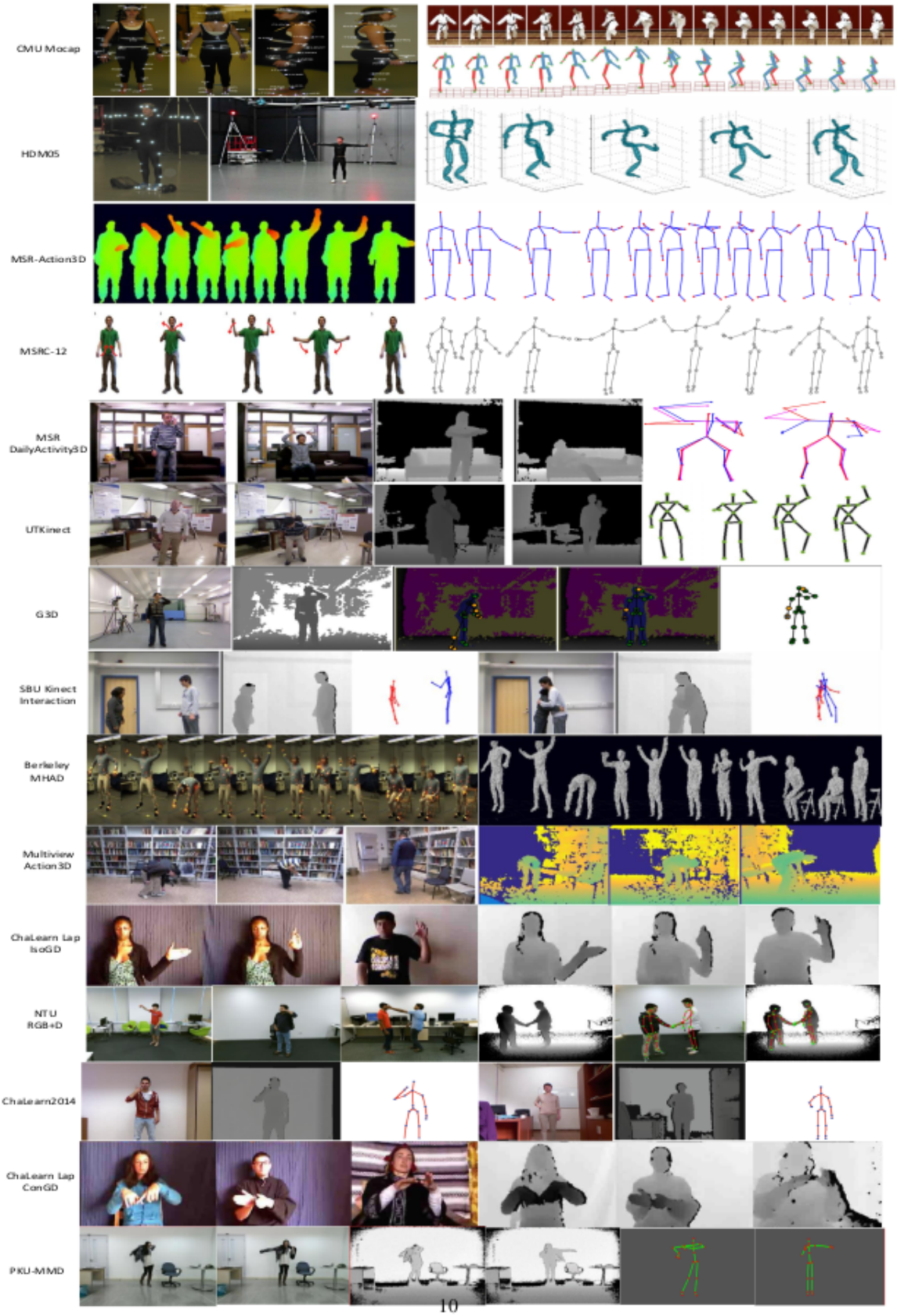


FIGURE 6.19 – Exemples de données RGB-D[224]

Bibliographie

[1] Cs224w : Machine learning with graphs, stanford university. `https://web.stanford.edu/class/cs224w/`.

[2] Raphaël fournier-s'niehotta, apprentissage sur graphes, cnam. `https://cedric.cnam.fr/vertigo/Cours/RCP217/docs/RCP217-GraphML1.pdf`.

[3] L. F Abbott and Peter. Dayan. Computational and mathematical modeling of neural systems. *Theoritical Neuroscience, The MIT press.*, 2005.

[4] Catherine Achard, Xingtai Qu, Arash Mokhber, and Maurice Milgram. Action recognition with semi-global characteristics and hidden markov models. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 274–284. Springer, 2007.

[5] Pascal Bourdon David Helbert. Adrien Raison, Théo Biardeau. Face expressions understanding by geometrical characterization of deep human faces representation. *Image Processing : Algorithms and Systems Conference, IST Electronic Imaging, Jan 2023, San Francisco, United States. ⟨hal-03832842⟩.*

[6] Abby Alpert, William N Evans, Ethan MJ Lieber, and David Powell. Origins of the opioid crisis and its enduring impacts. *The Quarterly Journal of Economics*, 137(2) :1139–1179, 2022.

[7] Anthony J. Alvarado. Cultural diversity : Pain beliefs and treatment among mexican-americans, african-americans, chinese-americans and japanese-americans. 2008.

[8] P. H. et al. Andersen. Can a machine learn to see horse pain ? an interdisciplinary approach towards automated decoding of facial expressions of pain in the horse. 2017.

[9] Rodrigo Araujo and Mohamed S Kamel. A semi-supervised temporal clustering method for facial emotion analysis. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014.

[10] Andreas Aristidou, Panayiotis Charalambous, and Yiorgos Chrysanthou. Emotion analysis and classification : Understanding the performers' emotions using the lma entities. *Computer Graphics Forum*, 34, 04 2015.

[11] Min SH Aung, Sebastian Kaltwang, Bernardino Romera-Paredes, Brais Martinez, Aneesha Singh, Matteo Cella, Michel Valstar, Hongying Meng, Andrew

Kemp, Moshen Shafizadeh, et al. The automatic detection of chronic pain-related expression : requirements, challenges and the multimodal emopain dataset. *IEEE transactions on affective computing*, 7(4) :435–451, 2015.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*, 2014.

[13] Pierre Baldi and Gianluca Pollastri. The principled design of large-scale recursive neural network architectures–dag-rnns and the protein structure prediction problem. *The Journal of Machine Learning Research*, 4 :575–602, 2003.

[14] Fabien Baradel, Christian Wolf, and Julien Mille. Pose-conditioned spatio-temporal attention for human action recognition. *CoRR*, abs/1703.10106, 2017.

[15] Justin Bayer, Daan Wierstra, Julian Togelius, and Jürgen Schmidhuber. Evolving memory cell structures for sequence learning. In *International conference on artificial neural networks*, pages 755–764. Springer, 2009.

[16] Cyrille Beaudry. Analyse et reconnaissance de séquences vidéos d'activités humaines dans l'espace sémantique. vision par ordinateur et reconnaissance de formes. *Université de La Rochelle,2015. Français. NT : 2015LAROS042 el-01661437*, 2015.

[17] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3) :1937–1967, 2021.

[18] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA :, 1994.

[19] Kelley Bevers, Lynette Watts, Nancy D. Kishino, and Robert J. Gatchel. Pain the biopsychosocial model of the assessment , prevention , and treatment of chronic pain. 2016.

[20] Nitin Bhatia et al. Survey of nearest neighbor techniques. *arXiv preprint arXiv :1007.0085*, 2010.

[21] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2) :197–227, 2016.

[22] Katy Blanc. *Description de contenu vidéo : mouvements et élasticité temporelle*. PhD thesis, Université Côte d'Azur, 2018.

[23] Vincent Bonnet. Modélisation des coordinations posturales chez l'humain. *UNIVERSITÉ MONTPELLIER II - SCIENCES ET TECHNIQUES DU LANGUEDOC*, 2009.

[24] Mohamed Bouaziz. *Réseaux de neurones récurrents pour la classification de séquences dans des flux audiovisuels parallèles*. PhD thesis, Université d'Avignon, 2017.

[25] Lubomir Bourdev and Jitendra Malik. Poselets : Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1365–1372, 2009.

[26] Xiaopeng Ning Xu Xu Boyi Hu, Chong Kim. Using a deep learning network to recognize low back pain in static standing. *Ergonomics, DOI :10.1080/00140139.2018.1481230*, 2018.

[27] Matteo Bregonzio, Tao Xiang, and Shaogang Gong. Fusing appearance and distribution information of interest points for action recognition. *Pattern Recognition*, 45(3) :1220–1234, 2012.

[28] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

[29] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv :1703.03906*, 2017.

[30] Arthur E Bryson. A gradient method for optimizing multi-stage allocation processes. In *Proc. Harvard Univ. Symposium on digital computers and their applications*, volume 72, page 22, 1961.

[31] Arthur E Bryson and Yu-Chi Ho. *Applied optimal control : optimization, estimation, and control*. Routledge, 1969.

[32] B.M. Waller E. Zimmermann A.M. Burrows Caeiro, C.C. and M. Davila Ross. Orangfacs : A muscle based facial movementcoding system for orangutans. *International Journal of Primatology. 34, 115-129.*, 2013.

[33] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose : Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[34] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriu, Luca Romeo, and Federica Verdini. The kimore dataset : Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7) :1436–1448, 2019.

[35] Enrico Capobianco. On digital therapeutics. *Frontiers in Digital Humanities*, 2 :6, 2015.

[36] J. Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. 07 2015.

[37] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset, 2019.

[38] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition ? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[39] Davide Castelvecchi. Can we open the black box of ai ? *Nature News*, 538(7623) :20, 2016.

[40] Werner Ceusters. An alternative terminology for pain assessment. *CEUR Workshop Proceedings. 1309.*, 2015.

[41] C Richard Chapman and Jonathan Gavrin. Suffering and its relationship to pain. *Journal of palliative care*, 9(2) :5–13, 1993.

[42] Shofer FS Dean AJ Hollander JE-Baxt WG Robey JL Sease KL Mills AM Chen, EH. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Acad Emerg Med.*, 15(5) :414-8, 2008.

[43] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. Xgboost : extreme gradient boosting. *R package version 0.4-2*, 1(4) :1–4, 2015.

[44] Guilhem Chéron. *Structured modeling and recognition of human actions in video.* Theses, PSL Research University, December 2018.

[45] Stephane Cholet, Helene Paugam-Moisy, Sebastien Regis, and Lionel Prevost. Informatique affective : Classification de l'etat emotionel. In *Extraction et Gestion des Connaissances*, 2017.

[46] Shehzan Haider Chowdhury, Murshed Al Amin, AKM Mahbubur Rahman, M Ashraful Amin, and Amin Ahsan Ali. Assessment of rehabilitation exercises from depth sensor data. In *2021 24th International Conference on Computer and Information Technology (ICCIT)*, pages 1–7. IEEE, 2021.

[47] Ivan Laptev Christian Schuldt and Barbara Caputo. Recognizing human actions : a local svm approach. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., volume 3, pages 32–36 Vol.3. IEEE, 2004.*, 2004.

[48] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25, 2012.

[49] Dan Ciresan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural networks*, 32 :333–338, 2012.

[50] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer, 2013.

[51] Dan C. Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. High-performance neural networks for visual object classification. *CoRR*, abs/1102.0183, 2011.

[52] Dan C. Ciresan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. *CoRR*, abs/1202.2745, 2012.

[53] Giusti A. Cireşan DC. Mitosis detection in breast cancer histology images with deep neural networks. *Medical Image Computing and Computer-Assisted Intervention – MICCAI, page 411–418*, 2013.

[54] Torkil Clemmensen and Lene Nielsen. Proceedings of the 5th danish human-computer interaction research symposium. *Copenhagen Business School, Department of Informatics, Working Papers*, 01 2005.

[55] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[56] Paul Compagnon. Sequence metric learning : Application to human activity recognition. *Artificial Intelligence [cs.AI]. Université de Lyon, 2021. NNT : 2021LYSEI033 tel-03407186*, 2021.

[57] Schilling G. Bausch C. Nadstawek J.-Wartenberg H. C. Wegener I. Geiser F. Imbierowicz K. Liedtke R. Conrad, R. Temperament and character personality profiles and personality disorders in chronic pain patients. *Pain, 133(1-3), 197–209. https ://doi.org/10.1016/j.pain.2007.07.024*, 2007.

[58] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

[59] Eva Coupeté. *Reconnaissance de gestes et actions pour la collaboration homme-robot sur chaîne de montage*. PhD thesis, Paris Sciences et Lettres (ComUE), 2016.

[60] Alan Cowen, Gautam Prasad, Misato Tanaka, Yukiyasu Kamitani, Vladimir Kirilyuk, Krishna Somandepalli, Brendan Jou, Florian Schroff, Hartwig Adam, Jeffrey Brooks, and Dacher Keltner. Title : How emotion is experienced and expressed in multiple cultures : A large-scale experiment. 10 2021.

[61] Kenneth D Craig. The facial expression of pain better than a thousand words ? *APS Journal*, 1(3) :153–162, 1992.

[62] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion, 2017.

[63] Langford et al. Dale. Coding of facial expressions of pain in the laboratory mouse. *Nature methods, 7. 447-9.10.1038/nmeth.1455.*, 2010.

[64] Lebelt D.Stucke D-Canali E Dalla Costa E, Minero M and Leach MC. Development of the horse grimace scale (hgs) as a pain assessment tool in horses undergoing routine castration. *PLoS One ; 9 : e92281*, 2014.

[65] Amit Dang, Dimple Arora, and Pawan Rane. Role of digital therapeutics and the changing future of healthcare. *Journal of Family Medicine and Primary Care*, 9(5) :2207, 2020.

[66] Edmond Boyer. Daniel Weinland, Rémi Ronfard. A survey of vision-based methods for action repre-sentation, segmentation and recognition. *Computer Vision and Image Understanding, Elsevier, 2011,115 (2), pp.224-241.*, 2011.

[67] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3048, 2013.

[68] Beatrice de Gelder. Why bodies ? twelve reasons for including bodily expressions in affective neuroscience. *Phil. Trans. R. Soc. B (2009) 364, 3475–3484 doi :10.1098/rstb.2009.0190*, 2009.

[69] Beatrice de Gelder and Ruud Hortensius. The many faces of the emotional body. *Research and Perspectives in Neurosciences*, 21 :153–164, 11 2014.

[70] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines : Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4772–4781, 2016.

[71] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[72] Stuart Dreyfus. The numerical solution of variational problems. *Journal of Mathematical Analysis and Applications*, 5(1) :30–45, 1962.

[73] DTA. Digital therapeutics : combining technology and evidence-based medicine to transform personalized patient care. 2018.

[74] Carlos Arango Duque, Olivier Alata, Rémi Emonet, Hubert Konik, and Anne-Claire Legrand. Mean oriented riesz features for micro expression classification. *Pattern Recognition Letters*, 135 :382–389, 2020.

[75] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human emotion recognition : Review of sensors and methods. *Sensors*, 20(3) :592, 2020.

[76] Joy O Egede, Siyang Song, Temitayo A Olugbade, Chongyang Wang, C De C Amanda, Hongying Meng, Min Aung, Nicholas D Lane, Michel Valstar, and Nadia Bianchi-Berthouze. Emopain challenge 2020 : Multimodal pain evaluation from facial and bodily expressions. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 849–856. IEEE, 2020.

[77] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[78] Gernot Ernst. The myth of the'mediterranean syndrome' : do immigrants feel different pain ? *Ethnicity & health*, 5(2) :121–126, 2000.

[79] Bahadori MT et al. Multi-layer representation learning for medical concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016.

[80] Franck Dernoncourt et al. De-identification of patient notes with recurrent neural networks. 2016.

[81] Gulshan V et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 2016.

[82] Hillel Aviezer et al. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science 338 , 1225 (2012); DOI : 10.1126/science.1224313*, 2015.

[83] Matthew Ung et al. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac Symp Biocomput., 20 :132–143*, 2015.

[84] Stephens ZD et al. Big data :astronomical or genomical ? *PLoS Biol, 13(7) :e1002195*, 2015.

[85] Tuttle et al. deep neural network to assess spontaneous pain from mouse facialexpressions. *Molecular pain vol. 14*, 2018.

[86] Zitnik et al. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics,34 :457–466*, 2018.

[87] Zuccon G et al. Medical semantic similarity with a neural language model. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM'14,*, 2014.

[88] Plis S et al. Aliper A. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular Pharmaceutics, 13 :2524–2530*, 2016.

[89] Li L et al. Miotto R. Deep patient : An unsupervised representation to predict the future of patients from the electronic health records. *Nature Scientific Reports, 6 :26094*, 2016.

[90] Dzamba M et al. Wallach I. Atomnet : A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. 2015.

[91] Chen H et al. Yu L. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging, 36 :994–1004*, 2017.

[92] Caroline Etienne. *Apprentissage profond appliqué à la reconnaissance des émotions dans la voix.* PhD thesis, Université Paris Saclay (COmUE), 2019.

[93] Caroline Etienne. Apprentissage profond appliqué à la reconnaissance des émotions dans la voix. *Intelligence artificielle [cs.AI]. Université Paris Saclay (COmUE)*, 2019.

[94] Yuchen Fan, Yao Qian, Frank K Soong, and Lei He. Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4475–4479. IEEE, 2015.

[95] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE : Regional multi-person pose estimation. In *ICCV*, 2017.

[96] Simona Farina, Michele Tinazzi, Le Pera Domenica, and Massimiliano Valeriani. Pain-related modulation of the human motor cortex. *Neurological research*, 25 :130–42, 04 2003.

[97] Adnan Farooq and Chee Sun Won. A survey of human action recognition approaches that use an rgb-d sensor. *IEIE Transactions on Smart Processing and Computing*, 4(4) :281–290, 2015.

[98] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[99] Miao Feng and Jean Meunier. Skeleton graph-neural-network-based human action recognition : A survey. *Sensors*, 22(6) :2091, 2022.

[100] Santiago Fernández, Alex Graves, and Jürgen Schmidhuber. An application of recurrent neural networks to discriminative keyword spotting. In *International Conference on Artificial Neural Networks*, pages 220–229. Springer, 2007.

[101] Bruce A Ferrell. Pain management. *Clinics in geriatric medicine*, 16(4) :853–873, 2000.

[102] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Petros Koutras, Gerasimos Potamianos, and Petros Maragos. Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction. *CoRR*, abs/1901.01805, 2019.

[103] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2) :179–188, 1936.

[104] Hironobu Fujiyoshi, Alan J Lipton, and Takeo Kanade. Real-time human motion analysis by image skeletonization. *IEICE TRANSACTIONS on Information and Systems*, 87(1) :113–120, 2004.

[105] K. Fukushima. Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernetics 36, 193–202*, 1980.

[106] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real-time human pose tracking from range data. In *European conference on computer vision*, pages 738–751. Springer, 2012.

[107] Benoît Georges. Intelligence artificielle : de quoi parle-t-on ? *Constructif, 2019/3 (N° 54), p. 5-10. DOI : 10.3917/const.054.0005.*, 2019.

[108] Santiago Gerling Konrad, Mao Shan, Favio Masson, Stewart Worrall, and Eduardo Nebot. Pedestrian dynamic and kinematic information obtained from vision sensors. 06 2018.

[109] Felix A Gers and E Schmidhuber. Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE transactions on neural networks*, 12(6) :1333–1340, 2001.

[110] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv :1602.06291*, 2016.

[111] Silvio Giancola, Matteo Valenti, and Remo Sala. *A survey on 3D cameras : Metrological comparison of time-of-flight, structured-light and active stereoscopy technologies*. Springer, 2018.

[112] Gadi Gilam, James J. Gross, Tor D. Wager, Francis J. Keefe, and Sean C. Mackey. What is the relationship between pain and emotion ? bridging constructs and communities. *Neuron*, 107(1) :17–21, 2020.

[113] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 759–768, 2015.

[114] K. B. Gleerup and C. Lindegaard. Recognition and quantification of pain in horses : A tutorial review. *Equine Vet Educ, 28 : 47-57. doi :10.1111/eve.12383*, 2016.

[115] Laetitia Gond. Système multi-caméras pour l'analyse de la posture humaine. *Université Blaise Pascal - Clermont-Ferrand II, 2009. Français. NNT : 2009CLF21922.*, 2009.

[116] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[117] Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv :1312.6082*, 2013.

[118] Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.

[119] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE, 2013.

[120] Emmanuele Grosicki and Haikal El Abed. Icdar 2009 handwriting recognition competition. pages 1398–1402, 01 2009.

[121] FDA-NIH Biomarker Working Group. Best (biomarkers, endpoints, and other tools) resource. *Silver Spring (MD) : Food and Drug Administration (US), Co-published by National Institutes of Health (US), Bethesda (MD).*, pages 1–1, 10 2016.

[122] Leach M.Minot EO Stewart M Guesgen MJ, Beausoleil NJ and KJ. Stafford. Coding and quantification of a facial expression for pain in lambs. *Behav Processes 2016 ; 132 : 49–56.*, 2016.

[123] Hatice Gunes, Caifeng Shan, Shizhi Chen, and Yingli Tian. 14 bodily expression for automatic affect recognition. 2014.

[124] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision : A survey. *Computational Visual Media*, pages 1–38, 2022.

[125] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3) :1–159, 2020.

[126] William L Hamilton. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, 14(3) :1–159, 2020.

[127] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech : Scaling up end-to-end speech recognition. *arXiv preprint arXiv :1412.5567*, 2014.

[128] Chris Harris and Mike Stephens. A combined corner and edge detection. *TheFourth Alvey Vision Conference (1988), pages 147 – 151.*, 1988.

[129] Renqiao Zhang Heng Yang and Peter Robinson. Human and sheep facial landmarks localisation by triplet interpolated features. *DOI : 10.13140/RG.2.1.2939.1841*, 2015.

[130] Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma thesis, TU Munich*, 1991.

[131] Sepp Hochreiter. Recurrent neural net learning and vanishing gradient. *International Journal Of Uncertainity, Fuzziness and Knowledge-Based Systems*, 6(2) :107–116, 1998.

[132] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8) :1735–1780, 1997.

[133] Kelly M et al. Hoffman. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences of the United States of America : 4296-301*, vol. 113, 2016.

[134] Collins M.Bell A Reid J Scott EM Holden E, Calvo G and Nolan AM. Evaluation of facial expression in acute pain in cats. *J Small Anim Pract 2014 ; 55 : 615–621*, 2017.

[135] Brian Holt, Eng-Jon Ong, Helen Cooper, and Richard Bowden. Putting the pieces together : Connected poselets for human pose estimation. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 1196–1201. IEEE, 2011.

[136] Ji Sun Hong, Chris Wasden, and Doug Hyun Han. Introduction of digital therapeutics. *Computer Methods and Programs in Biomedicine*, 209 :106319, 2021.

[137] Kristina Host and Marina Ivašić-Kos. An overview of human action recognition in sports based on computer vision. *Heliyon*, page e09633, 2022.

[138] Jiankun Hu, Xinghuo Yu, Dong Qiu, and Hsiao-Hwa Chen. A simple and efficient hidden markov model scheme for host-based anomaly intrusion detection. *IEEE network*, 23(1) :42–47, 2009.

[139] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3) :574, 1959.

[140] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016.

[141] AGe Ivakhnenko and VG Lapa. Cybernetic predicting devices. ccm information corporation. *First working Deep Learners with many layers, learning internal representations*, 1965.

[142] Alexey Grigorevich Ivakhnenko. Polynomial theory of complex systems. *IEEE transactions on Systems, Man, and Cybernetics*, (4) :364–378, 1971.

[143] Tommi Jantunen, Johanna Mesch, Anna Puupponen, and Jorma Laaksonen. On the rhythm of head movements in finnish and swedish sign language sentences. pages 850–853, 05 2016.

[144] Neziha Jaouedi, Noureddine Boujnah, Oumayma Htiwich, and Med Salim Bouhlel. Human action recognition to human behavior analysis. In *2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 263–266. IEEE, 2016.

[145] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1) :221–231, 2012.

[146] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1) :221–231, 2013.

[147] Juan C. Moreno Joana Figueiredo, Cristina P. Santos. Automatic recognition of gait patterns in human motor disorders using machine learning : A review. *Medical Engineering and Physics 0 0 0 (2018) 1–12*, 2018.

[148] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception Psychophysics 14, 201–211*, 1973.

[149] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion : Inferring communicative intents of images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–223, 2014.

[150] Achuta Kadambi, Ayush Bhandari, and Ramesh Raskar. 3d depth cameras in vision : Benefits and limitations of the hardware. In *Computer vision and machine learning with RGB-D sensors*, pages 3–26. Springer, 2014.

[151] Byeongkeun Kang, Subarna Tripathi, and Truong Q Nguyen. Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 136–140. IEEE, 2015.

[152] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[153] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv :1705.06950*, 2017.

[154] Sun Ke, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. 02 2019.

[155] Flecknell PA.and Leach MC. Keating SCJ, Thomas AA. Evaluation of emla cream for preventing pain during tattooing of rabbits : changes in physiological, behavioural and facial expression responses. *PLoS One ; 7 : e44437*, 2012.

[156] Francis J Keefe, Mark Lumley, Timothy Anderson, Thomas Lynch, and Kimi L Carson. Pain and emotion : new research directions. *Journal of clinical psychology*, 57(4) :587–607, 2001.

[157] Henry J Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10) :947–954, 1960.

[158] Y. Kim. Convolutional neural networks for sentence classification. 2014.

[159] Yeon-Wook Kim, Kyung-Lim Joa, Han-Young Jeong, and Sangmin Lee. Wearable imu-based human activity recognition algorithm for clinical balance assessment using 1d-cnn and gru ensemble model. *Sensors*, 21(22) :7628, 2021.

[160] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv :1609.02907*, 2016.

[161] Andrea Kleinsmith and Nadia. Bianchi-Berthouze. Affective body expression perception and recognition : A survey. *Affective Computing, IEEE Transactions on. 4. 15-33. 10.1109/T-AFFC.2012.16.*, 2013.

[162] Yu Kong and Yun Fu. Human action recognition and prediction : A survey. *International Journal of Computer Vision*, 130(5) :1366–1401, 2022.

[163] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. 25, 2012.

[164] Yasuo Kuniyoshi, Hirochika Inoue, and Masayuki Inaba. Design and implementation of a system that generates assembly programs from visual recognition of human action sequences. In *EEE International workshop on intelligent robots and systems, towards a new frontier of applications*, pages 567–574. IEEE, 1990.

[165] Ivan Laptev and Tony Lindeberg. Space-time interest points. *Proceedings NinthIEEE International Conference on Computer Vision, 1 :432–439.*, 2003.

[166] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4) :541–551, 1989.

[167] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[168] Yann LeCun, D Touresky, G Hinton, and T Sejnowski. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28, 1988.

[169] J.P Lefaucheur. The complex relationship between pain and motor cortex. *Brain Stimulation*, 10(2) :382, 2017.

[170] Romain Lejeune. *Prise en charge de la douleur et des symptômes terminaux chez les patients adultes en fin de vie dans le cadre d'une hospitalisation à domicile.* PhD thesis, 2017.

[171] Linda. Leresche. Defining gender disparities in pain management. *Clinical orthopaedics and related research*, vol. 469,7, 2011.

[172] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using lstm and cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 585–590. IEEE, 2017.

[173] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose : Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[174] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik : A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021.

[175] Larry Li et al. Time-of-flight camera–an introduction. *Technical white paper*, (SLOA190B), 2014.

[176] Xiangang Li and Xihong Wu. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 4520–4524. IEEE, 2015.

[177] Joshi J. Li Y, Ghosh S. Plaan : Pain level assessment with anomaly-detection based network [published online ahead of print. *Journal on Multimodal User Interfaces. 2021 ;1-14.*, 2021.

[178] Kelly Liao, Mallori Henceroth, Qian Lu, and Angie LeRoy. Cultural differences in pain experience among four ethnic groups : A qualitative pilot study. *Journal of Behavioral Health*, 5 :1, 01 2016.

[179] Yalin Liao, Aleksandar Vakanski, and Min Xian. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2) :468–477, 2020.

[180] S Linnainmaa. : The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors.(in finnish) master's thesis, department of computer science, university of helsinki, 1970. 1970.

[181] Zachary Chase Lipton. A critical review of recurrent neural networks for sequence learning. *CoRR*, abs/1506.00019, 2015.

[182] Hong Liu, Juanhui Tu, and Mengyuan Liu. Two-stream 3d convolutional neural network for skeleton-based action recognition. *arXiv preprint arXiv :1705.08106*, 2017.

[183] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120 : A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10) :2684–2701, 2020.

[184] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016.

[185] Wenfei Liu, Jingcheng Wei, and Qingmin Meng. Comparisions on knn, svm, bp and the cnn for handwritten digit recognition. In *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pages 587–590. IEEE, 2020.

[186] Yan Liu. Egcn : An ensemble-based learning framework for exploring effective skeleton-based rehabilitation exercise assessment. In *EGCN : An Ensemble-based Learning Framework for Exploring Effective Skeleton-based Rehabilitation Exercise Assessment*, pages 3681–3687. 2022.

[187] Yu Luo, Jianbo Ye, Reginald B. Adams Jr., Jia Li, Michelle G. Newman, and James Z. Wang. ARBEE : towards automated recognition of bodily expression of emotion in the wild. *CoRR*, abs/1808.09568, 2018.

[188] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. 12 2017.

[189] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv :1508.04025*, 2015.

[190] Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv :1410.8206*, 2014.

[191] Darya Yelshyna Jaime Ferreira Hélder David Silva Luís Rocha Nuno Sousa Luís Costa, Miguel F. Gago and Estela Bicho. Application of machine learning in postural control kinematics for the diagnosis of alzheimer's disease. 2016.

[192] Lu Y Mahmoud M and Robinson P. Estimating sheep pain level using facial action unit detection. *IEEE 2017, pp.394–399*, 2016.

[193] Stefan Marks, JA Windsor, and B Wünsche. Evaluation of the effectiveness of head tracking for view and avatar control in virtual environments. In *2010 25th International Conference of Image and Vision Computing New Zealand*, pages 1–8. IEEE, 2010.

[194] Keela Herr Michelle M. Hilgeman Princess Nash Nayak Polissar Mary Ersek, Moni B. Neradilek and Francis X. Nelson. Psychometric evaluation of a pain intensity measure for persons with dementia. *Pain Medicine, doi : 10.1093/pm/pny166*, 2018.

[195] Marco Bellantonio Hugo Escalante Víctor Ponce-López et al Maryam Asadi-Aghbolaghi, Albert Clapes. Deep learning for action and gesture recognition in image sequences : a survey. *Gesture Recognition, Springer Verlag, pp.539-578, 2017. hal-01678006*, 2017.

[196] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. A proposal for the dartmouth summer research project on artifical intelligence, 1955.

[197] Pitts W. McCulloch, W.S. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5, 115–133*, 1943.

[198] Catherine Mercier and Guillaume Léonard. Interactions between pain and the motor cortex : Insights from research on phantom limb pain and complex regional pain syndrome. *Physiotherapy Canada*, 63(3) :305–14, 2011.

[199] Cyrille Migniot and Fakhreddine Ababsa. Hybrid 3d–2d human tracking in a top view. *Journal of Real-Time Image Processing*, 11(4) :769–784, 2016.

[200] Herve Jegou Mihir Jain and Patrick Bouthemy. Better exploiting motion for bet-ter action recognition. *In 2013 IEEE Conference on Computer Vision and Pattern Recognition, pages 2555–2562. IEEE, jun 2013.*, 2013.

[201] Bernardino Romera-Paredes Brais Martinez-Aneesha Singh Matteo Cella Michel Valstar Hongying Meng Andrew Kemp Moshen Shafizadeh Aaron C. Elkins Natalie Kanakam Amschel de Rothschild Nick Tyler Paul J. Watson Amanda C. de C. Williams Maja Pantic Min S. H. Aung, Sebastian Kaltwang and Nadia Bianchi-Berthouze. The automatic detection of chronic pain-related expression : Requirements, challenges and the multimodal emopain dataset. *IEEE Transactions on Affective Computing (Volume : 7 , Issue : 4 , Oct.-Dec. 1 2016*, 2016.

[202] Gaurav Misra, Edward Ofori, Jae Woo Chung, and Steve Coombes. Pain-related suppression of beta oscillations facilitates voluntary movement. *Cerebral Cortex*, 27 :bhw061, 03 2016.

[203] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016.

[204] Hossein Mousavi Hondori and Maryam Khademi. A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. *Journal of medical engineering*, 2014, 2014.

[205] Carlo Dal Mutto, Pietro Zanuttigh, and Guido M Cortelazzo. *Time-of-flight cameras and microsoft kinect (TM)*. Springer Publishing Company, Incorporated, 2012.

[206] Laurie Needham, Murray Evans, Darren P Cosker, Logan Wade, Polly M McGuigan, James L Bilzon, and Steffi L Colyer. The accuracy of several pose estimation methods for 3d joint centre localisation. *Scientific reports*, 11(1) :1–11, 2021.

[207] Natalia Neverova. *Deep learning for human motion analysis*. PhD thesis, Université de Lyon, April 2016.

[208] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. volume 9912, pages 483–499, 10 2016.

[209] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452 :48–62, 2021.

[210] Bill Noble, David Clark, Marcia Meldrum, Henk ten Have, Jane Seymour, Michelle Winslow, and Silvia Paz. The measurement of pain, 1945–2000. *Journal of Pain and Symptom Management*, 29(1) :14–21, 2005.

[211] Yuya Obinata and Takuma Yamamoto. Temporal extension module for skeleton-based action recognition. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 534–540, 2021.

[212] Temitayo A Olugbade, Nadia Bianchi-Berthouze, Nicolai Marquardt, and Amanda C de C Williams. Human observer and automatic assessment of movement related self-efficacy in chronic pain : from exercise to functional activity. *IEEE Transactions on Affective Computing*, 11(2) :214–229, 2018.

[213] Karen Otte, Bastian Kayser, Sebastian Mansow-Model, Julius Verrel, Friedemann Paul, Alexander U Brandt, and Tanja Schmitz-Hübsch. Accuracy and reliability of the kinect version 2 for clinical measurement of motor function. *PloS one*, 11(11) :e0166532, 2016.

[214] Thierry Paillard. *Posture et équilibration humaines*. De Boeck Superieur, 2016.

[215] Chris Pasero. Margo mccaffery : Resolute and visionary. *Pain Management Nursing*, Vol 19, No 2 (April), 2018 : pp 89-91, 2018.

[216] Mohammadreza Amirian Peter Bellmann-Georg Layher Yan Zhang Maria Velana Sascha Gruss Steffen Walter Harald C. Traue Daniel Schork Jonghwa Kim Elisabeth André Heiko Neumann Friedhelm Schwenker Patrick Thiam, Viktor Kessler. Multi-modal pain intensity recognition based on the senseemotion database. *IEEE Transactions on Affective Computing, DOI 10.1109/TAFFC.2019.2892090,*, 2019.

[217] Sue Peacock and Shilpa Patel. "cultural influences on pain.". *Reviews in pain*, 1,2 :6–9, 2008.

[218] Pr Serge Perrot. La douleur : définitions et concepts. *Livre blanc de la douleur 2017*, page 31, 2017.

[219] Ayoub Al-Hamadi Philipp Werner and Robert Niese. Head movements and postures as pain behavior. *International Journal of Pattern Recognition and Artificial IntelligenceVol. 28, No. 05, 1451008 (2014) Machine Learning, https ://doi.org/10.1142/S0218001414510082*, 2014.

[220] Ayoub Al-Hamadi et al. Philipp Werner. Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing (Volume : 8 , Issue : 3 , July-Sept. 1 2017) , https ://doi.org/10.1371/journal.pone.0192767*, 2017.

[221] Kerstin Limbrecht-Ecklundt-Steffen Walter Harald C. Traue Philipp Werner, Ayoub Al-Hamadi. Head movements and postures as pain behavior. 2018.

[222] R. W. Picard. Affective computing. *M.I.T Media Laboratory Perceptual Computing Section Technical Report*, No. 321, 1995.

[223] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

[224] Philip Ogunbona Jun Wan-Sergio Escalera Pichao Wang, Wanqing Li. Rgb-d-based human motion recognition with deep learning : A survey. 2018.

[225] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign language recognition using convolutional neural networks. In *European conference on computer vision*, pages 572–578. Springer, 2014.

[226] Christian Plagemann, Varun Ganapathi, Daphne Koller, and Sebastian Thrun. Real-time identification and localization of body parts from depth images. In *2010 IEEE International Conference on Robotics and Automation*, pages 3108–3113. IEEE, 2010.

[227] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *International Conference on Pattern Recognition*, pages 694–701. Springer, 2021.

[228] Kenneth Prkachin and Patricia Solomon. The structure, reliability and validity of pain expression : Evidence from patients with shoulder pain. *Pain*, 139 :267–74, 06 2008.

[229] Karim Radouane, Andon Tchechmedjiev, Binbin Xu, and Sebastien Harispe. Comparison of deep learning approaches for protective behaviour detection under class imbalance from mocap and emg data. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 01–08. IEEE, 2021.

[230] Srinivasa N Raja, Daniel B Carr, Milton Cohen, Nanna B Finnerup, Herta Flor, Stephen Gibson, Francis Keefe, Jeffrey S Mogil, Matthias Ringkamp, Kathleen A Sluka, et al. The revised iasp definition of pain : concepts, challenges, and compromises. *Pain*, 161(9) :1976, 2020.

[231] Cohen M et al. Raja SN, Carr DB. The revised international association for the study of pain definition of pain : concepts, challenges, and compromises. *Pain*, 161(9) :1976-1982, 2020.

[232] Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Identifying emotions from walking using affective and deep features. *arXiv preprint arXiv :1906.11884*, 2019.

[233] Maheen Rashid. Interspecies knowledge transfer for facial keypoint detection. *CVPR*, 2017.

[234] K Réby and M Beurton-Aimar. Reconnaissance d'émotions à partir de la posture par lstm. In *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA'21) Plate-Forme Intelligence Artificielle (PFIA'21)*, 2021.

[235] Giuseppe Recchia, Daniela Maria Capuano, Neeraj Mistri, and Roberto Verna. Digital therapeutics-what they are, what they will be. *Acta Sci Med Sci*, 4 :1–9, 2020.

[236] Ilaria Renna, Catherine Achard, and Ryad Chellali. Combination of annealing particle filter and belief propagation for 3d upper body tracking. *Applied Bionics and Biomechanics*, 9, 08 2011.

[237] Jamie L Rhudy and Mary W Meagher. Fear and anxiety : divergent effects on human pain thresholds. *Pain*, 84(1) :65–75, 2000.

[238] de Courval M-L.Mulon P-Y Harvey D Bichot S Gauvin D Livingston A Beaudry F Hélie P Frank D del Castillo JRE Rialland P, Otis C and Troncy E. Assessing experimental visceral pain in dairy cattle : a pilot, prospective, blinded, randomized, and controlled study focusing on spinal pain proteomics. *J Dairy Sci ; 97 : 2118–2134.*, 2014.

[239] AJ Robinson and Frank Fallside. *The utility driven dynamic error propagation network*. University of Cambridge Department of Engineering Cambridge, 1987.

[240] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca. Deep pain : Exploiting long short-term memory networks for facial expression classification. *IEEE Transactions on Cybernetics*, pages 1–11, 2017.

[241] Antonia Barke Qasim Aziz-Michael I. Bennett Rafael Benoliel Milton Cohen Stefan Evers Nanna B. Finnerup Michael B. First Maria Adele Giamberardino Stein Kaasa Eva Kosek Patricia Lavand'homme Michael Nicholas Serge Perrot Joachim Scholz Stephan Schug Blair H. Smith Peter Svensson Johan W.S. Vlaeyen Shuu-Jiun Wang Rolf-Detlef Treede, Winfried Rief. A classification of chronic pain for icd-11. *Pain 156, 1003-1007*, 2015.

[242] Frank Rosenblatt. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6) :386, 1958.

[243] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[244] Mathieu Roy. How is the pain experience "constructed" by the brain ? effects of emotional context on pain perception. *Douleur et Analgésie*, 26 :2–10, 03 2013.

[245] Terri Ashmeade Ruicong Zhi, Ghada Zamzmi Dmitry Goldgof and Yu Sun. Automatic infants' pain assessment by dynamic facial representation : Effects of profile view, gestational age, gender, and race. *J. Clin. Med. 2018, 7, 173 ; doi :10.3390/jcm7070173*, 2018.

[246] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088) :533–536, 1986.

[247] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv :1402.1128*, 2014.

[248] Haşim Sak, Andrew Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Françoise Beaufays, and Johan Schalkwyk. Learning acoustic frame labeling for speech recognition with recurrent neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4280–4284. IEEE, 2015.

[249] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing : Structured-light versus time-of-flight kinect. *Computer vision and image understanding*, 139 :1–20, 2015.

[250] Nikolaos Savva and Nadia Bianchi-Berthouze. Automatic recognition of affective body movement in a video game scenario. In Antonio Camurri and Cristina Costa, editors, *Intelligent Technologies for Interactive Entertainment*, pages 149–159, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[251] Jacob Scharcanski and M Emre Celebi. *Computer vision techniques for the diagnosis of skin cancer*. Springer, 2013.

[252] J. Schmidhuber. Deep Learning. *Scholarpedia*, 10(11) :32832, 2015. revision #184887.

[253] Jürgen Schmidhuber. Who invented backpropagation. *More in [DL2]*, 2014.

[254] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions : a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[255] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11) :2673–2681, 1997.

[256] Loren Arthur Schwarz, Artashes Mkhitaryan, Diana Mateus, and Nassir Navab. Human skeleton tracking from depth data using geodesic distances and optical flow. *Image and Vision Computing*, 30(3) :217–226, 2012.

[257] Samsu Sempena, Nur Ulfa Maulidevi, and Peb Ruswono Aryan. Human action recognition using dynamic time warping. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pages 1–5. IEEE, 2011.

[258] S Cameron Sepah, Luohua Jiang, and Anne L Peters. Long-term outcomes of a web-based diabetes prevention program : 2-year results of a single-arm longitudinal study. *Journal of medical Internet research*, 17(4) :e4052, 2015.

[259] SFMG. Sociologie et antropologie, quels apports pour la médecine générale. *Société Française de Médecine Générale, Documents de Recherche en Médecine Générale n°64*, pages 41–42, 2007.

[260] Min S.H. Aung, Sebastian Kaltwang, Bernardino Romera-Paredes, Brais Martinez, Aneesha Singh, Matteo Cella, Michel Valstar, Hongying Meng, Andrew Kemp, Aaron Elkins, Nick Tyler, Paul Watson, Amanda Williams, Maja Pantic, and Nadia. Bianchi-Berthouze. The automatic detection of chronic pain-related expression : requirements, challenges and a multimodal dataset. *IEEE Transactions on Affective Computing. 99. 1-1. 10.1109/TAFFC.2015.2462830.*, 2015.

[261] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and G. Wang. Ntu rgb+d : A large scale dataset for 3d human activity analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016.

[262] Claude E Shannon, Warren Weaver, and Norbert Wiener. The mathematical theory of communication. *Physics Today*, 3(9) :31, 1950.

[263] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29 :9532–9545, 2020.

[264] Maggie Shiffrar. People watching : visual, motor, and social processes in the perception of human movement. *Wiley Interdisciplinary Reviews : Cognitive Science*, 2(1) :68–78, 2011.

[265] Maggie Shiffrar and Tom Heinen. Athletic ability changes action perception : embodiment in the visual perception of human movement. *Z. Sportpsychol*, 17 :1–13, 2011.

[266] Maggie Shiffrar, Martha D Kaiser, and Areti Chouchourelou. Seeing human movement as inherently social. *The science of social vision*, pages 248–264, 2011.

[267] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1) :1–48, 2019.

[268] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011.

[269] Matheen Siddiqui and Gérard Medioni. Human pose estimation from a single view point, real-time range sensor. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 1–8. IEEE, 2010.

[270] Hava T. Siegelmann. Computation beyond the turing limit. *Science, 268 :545–548*, 1995.

[271] H.T. Siegelmann and E.D. Sontag. On the computational power of neural nets. *Journal of Computer and System Sciences*, 50(1) :132–150, 1995.

[272] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[273] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2015.

[274] Aneesha Singh, Tali Swann-Sternberg, Nadia Bianchi-Berthouze, Amanda Williams, M Pantic, and Paul. Watson. Emotion and pain : interactive technology to motivate physical activity in people with chronic pain. *conference paper, CHI 2012, May 5–10, 2012, Austin, TX, USA.*, 2012.

[275] Ravinder Singh. The pain : How does anthropology look at it ? suffering of body and mind. *Ethnologia Actualis*, 17 :123–139, 12 2017.

[276] Werner Ceusters Louis J. Goldberg Richard K. Ohrbach Smith, Barry. Towards an ontology of pain and of pain-related phenomena. 2011.

[277] Siegelmann Hava T. Sondag, Eduardo D. On the computationnal power of neural nets. *Journal of computer and systems science, 50 :132–150.*, 1995.

[278] Susana G et al Sotocinal. The rat grimace scale : a partially automated method for quantifying pain in the laboratory rat via facial expressions. *Molecular pain vol.7, 55, doi :10.1186/1744-8069-7-55*, 2011.

[279] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605, 2015.

[280] Byunghan Lee. Sungroh Yoo Seonwoo Min. Deep learning in bioinformatics. *Briefings in Bioinformatics,18(5) :851–869*, 2017.

[281] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[282] R. S. Sutton and A. G. Barto. Reinforcement learning : An introduction. *The MIT Press*, 2018.

[283] Christian Szegedy, Scott Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe. Scalable, high-quality object detection. *arXiv preprint arXiv :1412.1441*, 2014.

[284] Mohammad Tavakolian. *Efficient spatiotemporal representation learning for pain intensity estimation from facial expressions.* PhD thesis, University of Oulu, Finland, 2021.

[285] Nicolai Marquardt Temitayo A. Olugbade, Nadia Bianchi-Berthouze and Amanda C. de C. Williams. Human observer and automatic assessment of movement related self-efficacy in chronic pain : from exercise to functional activity. *IEEE Transactions on Affective Computing (Early Access) DOI : 10.1109/TAFFC.2018.2798576*, 2018.

[286] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann Lecun, and Christoph Bregler. Efficient object localization using convolutional networks. pages 648–656, 06 2015.

[287] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, 2015.

[288] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *CoRR*, abs/1406.2984, 2014.

[289] Alexander Toshev and Christian Szegedy. Deeppose : Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013.

[290] Irene Tracey, Clifford J. Woolf, and Nick A. Andrews. Composite pain biomarker signatures for objective assessment and effective treatment. *Neuron*, 101(5) :783–800, 2019.

[291] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[292] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3D : generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

[293] A. M. Turing. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236) :433–460, 10 1950.

[294] M. T. Uddin and S. Canavan. Multimodal multilevel fusion for sequential protective behavior detection and pain estimation. *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 844-848*, 2020.

[295] Aleksandar Vakanski, Hyung-pil Jun, David Paul, and Russell Baker. A data set of human body movements for physical rehabilitation exercises. *Data*, 3(1) :2, 2018.

[296] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[297] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell : A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[298] Lawlis P.Leach M Viscardi AV, Hunniford M and Turner PV. Development of a piglet grimace scale to evaluate piglet pain using facial expressions following castration and tail docking : a pilot study. *Front Vet Sci ; 4 : 230*, 2017.

[299] Sarvesh Vishwakarma and Anupam Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10) :983–1009, 2013.

[300] Joseph Walsh, Niall O' Mahony, Sean Campbell, Anderson Carvalho, Lenka Krpalkova, Gustavo Velasco-Hernandez, Suman Harapanahalli, and Daniel Riordan. Deep learning vs. traditional computer vision. 04 2019.

[301] Keogh E. Walsh J, Eccleston C. Pain communication through body posture : the development and validation of a stimulus set. 2014.

[302] Eric A Wan et al. Time series prediction by using a connectionist network with internal delay lines. In *SANTA FE INSTITUTE STUDIES IN THE SCIENCES OF COMPLEXITY-PROCEEDINGS VOLUME-*, volume 15, pages 195–195. Addison-Wesley publishing co, 1993.

[303] Chongyang Wang, Temitayo Olugbade, Akhil Mathur, Amanda Williams, Nicholas Lane, and Nadia Bianchi-Berthouze. Recurrent network based automatic detection of chronic pain protective behavior using mocap and semg data. 07 2019.

[304] Chongyang Wang, Min Peng, Temitayo Olugbade, Nicholas Lane, Amanda Williams, and Nadia Bianchi-Berthouze. Learning temporal and bodily attention in protective movement behavior detection. 07 2019.

[305] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *2013 IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[306] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition and detection by combining motion and appearance features. 2014.

[307] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015.

[308] Limin Wang, Yu Qiao, Xiaoou Tang, et al. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, 1(2) :2, 2014.

[309] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks : Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[310] Pengbo Wang, Yongqiang Zhang, and Wenting Jiang. Application of k-nearest neighbor (knn) algorithm for human action recognition. In *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, volume 4, pages 492–496. IEEE, 2021.

[311] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing : Emotion models, databases, and recent advances. *Information Fusion*, 2022.

[312] Yifan Wang, Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Two-stream sr-cnns for action recognition in videos. In *BMVC*. York, UK, 2016.

[313] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. 01 2016.

[314] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *CoRR*, abs/1602.00134, 2016.

[315] John J Weng, Narendra Ahuja, and Thomas S Huang. Learning recognition and segmentation of 3-d objects from 2-d images. in 1993 (4th) international conference on computer vision (pp. 121-128), 1993.

[316] Paul J Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4) :339–356, 1988.

[317] Paul J Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4) :339–356, 1988.

[318] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. Automatic recognition methods supporting pain assessment : A survey. *IEEE Transactions on Affective Computing*, pages 1–1, 10 2019.

[319] Norbert Wiener. *Cybernetics or control and communication in the animal and the machine*. Technology Press, 1948.

[320] Ronald J Williams and David Zipser. Gradient-based learning algorithms for recurrent. *Backpropagation : Theory, architectures, and applications*, 433 :17, 1995.

[321] Vinay Williams, Vasileios Argyriou, Peter Shaw, Christoph Montag, Georg Herdrich, Aaron Knoll, and Maximilian Moertl. Development of pptnet a neural network for the rapid prototyping of pulsed plasma thrusters. 09 2019.

[322] Paul A Wilson and Barbara Lewandowska-Tomaszczyk. Affective robotics : modelling and testing cultural prototypes. *Cognitive computation*, 6(4) :814–840, 2014.

[323] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 20–27. IEEE, 2012.

[324] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow : Efficient online pose tracking. In *BMVC*, 2018.

[325] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *CVPR*, volume 92, pages 379–385, 1992.

[326] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[327] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. 01 2018.

[328] Ming Yang, Shuiwang Ji, Wei Xu, Jinjun Wang, Fengjun Lv, Kai Yu, Yihong Gong, Mert Dikmen, Dennis J Lin, and Thomas S Huang. Detecting human actions in surveillance videos. In *TRECVID*, 2009.

[329] Angela Yao, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Does human action recognition benefit from pose estimation ? pages 67.1–67.11, 01 2011.

[330] Hongnian Yu, Shuang Cang, and Yan Wang. A review of sensor selection, sensor devices and sensor deployment for wearable sensor-based human activity recognition systems. In *2016 10th international conference on software, knowledge, information management & applications (skima)*, pages 250–257. IEEE, 2016.

[331] Chin Lin. Yu Sheng Lou. Artifical intelligence learning semantics via external ressources for classifying diagnosis codes in discharge notes. *J Med Interne Res, 19(11)*, 2017.

[332] Xinhui Yuan and Marwa Mahmoud. Alanet : Autoencoder-lstm for pain and protective behaviour detection. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 824–828. IEEE, 2020.

[333] Pietro Zanuttigh, Giulio Marin, Carlo Dal Mutto, Fabio Dominio, Ludovico Minto, and Guido Maria Cortelazzo. Time-of-flight and structured light depth cameras. *Technology and Applications*, pages 978–3, 2016.

[334] Dimsdale JE. Zatzick, DF. Cultural variations in response to painful stimuli. *Psychosom Med.*, 52(5) :544–57, 1990.

[335] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[336] Heiga Zen and Haşim Sak. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474. IEEE, 2015.

[337] S. Zhang, X. Liu, and J. Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157, 2017.

[338] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157. IEEE, 2017.

[339] Aichun Zhu. Articulated human pose estimation in images and video, détection et suivi de la posture humaine dans les images fixes et les vidéos. 2016.

[340] Aichun Zhu, Qianyu Wu, Ran Cui, Tian Wang, Wenlong Hang, Gang Hua, and Hichem Snoussi. Exploring a rich spatial–temporal dependent relational model for skeleton-based action recognition by bidirectional lstm-cnn. *Neurocomputing*, 414 :90–100, 2020.

[341] Hong-Min Zhu and Chi-Man Pun. Real-time hand gesture recognition from depth image sequences. In *2012 Ninth international conference on computer graphics, imaging and visualization*, pages 49–52. IEEE, 2012.

[342] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. Hidden two-stream convolutional networks for action recognition. In *Asian conference on computer vision*, pages 363–378. Springer, 2018.