Arts et Métiers Sciences et Technologies

ÉCOLE DOCTORALE SCIENCES ET MÉTIERS DE L'INGÉNIEUR

Laboratoire d'Innovation Numérique pour les Entreprises et les Apprentissages au service de la Compétitivité des Territoires (CESI LINEACT)
Campus de LYON

THÈSE

Présentée par : **Toufik BENMESSABIH**
Soutenue le : **5 février 2026**

Pour obtenir le **Doctorat** *délivré par* :

L'École Nationale Supérieure d'Arts et Métiers

Spécialité : **Informatique**

Reconnaissance d'actions humaines basée apprentissage profond et représentation squelettique : Propositions d'approches hors ligne et en ligne dans un contexte d'Industrie 5.0

Thèse dirigée par :
David BAUDRY

et co-encadrée par :
Rim SLAMA et Vincent HAVARD

Jury

M. Hassen Drira	Professeur des universités, Université de Strasbourg, ICube	Président
Mme. Catherine Achard	Professeur des universités, Sorbonne Université, ISIR	Rapporteure
M. Halim Benhabiles	Enseignant-chercheur HDR, IMT Nord Europe, CERI Systèmes Numériques	Rapporteur
M. David Baudry	Directeur de Recherche, CESI LINEACT de Rouen	Directeur
Mme. Rim Slama	Maîtresse de conférences, Université Gustave Eiffel, ENTPE, EMob-Lab	Co-encadrante
M. Vincent Havard	Enseignant-chercheur HDR, CESI LINEACT de Rouen	Co-encadrant
M. Hedi Tabia	Professeur des universités, Université Paris-Saclay, IBISC	Examinateur

THÈSE

Remerciements

Ces travaux de thèse ont été réalisés au cours de trois années riches sur les plans scientifique et humain. Ils n'auraient pas pu aboutir sans le soutien et la bienveillance de nombreuses personnes à qui je souhaite adresser mes sincères remerciements.

Je tiens tout d'abord à remercier les membres du jury pour l'intérêt qu'ils ont porté à ce travail. Je remercie particulièrement les rapporteurs pour le temps consacré à la lecture et à l'évaluation de mon manuscrit, ainsi que l'ensemble des membres du jury pour leur participation à ma soutenance de thèse.

Je souhaite également exprimer ma profonde gratitude à mon directeur de thèse, David Baudry, ainsi qu'à mes co-encadrants, Rim Salmi et Vincent Havard. Je les remercie pour leur accompagnement, leurs conseils précieux, leur disponibilité et la qualité du travail réalisé ensemble durant ces trois années de thèse. Je souhaite également remercier l'ensemble de mes enseignants qui ont contribué à ma formation tout au long de mon parcours académique.

Je remercie également l'ensemble des collègues du CESI pour leur accueil chaleureux au sein du laboratoire et pour les opportunités d'évolution qui m'ont été offertes, notamment à travers les activités d'enseignement et les formations réalisées au CESI. Je souhaite remercier mes amis : Amin, Souhila, Issam, Sami, Salah, Candice, Fidma, Sofian, Akram, Idir, Ilyas, Sarah, Clément, Baha, Ismail et Ghiless. J'ai passé de très bons moments à vos côtés, que ce soit au laboratoire au quotidien ou lors de randonnées, de footings ou de semi-marathons.

Je tiens également à remercier ma famille, qui m'a toujours soutenu et encouragé tout au long de ce parcours. Mes remerciements vont particulièrement à mes parents, à mon frère et à ma sœur, dont les gâteaux, bien connus de mes collègues du laboratoire, ont souvent apporté une touche de douceur et de convivialité. Je souhaite remercier tout particulièrement ma femme, Oumaima. Sa présence à mes côtés est une source constante de motivation et me pousse chaque jour à donner le meilleur de moi-même. J'adresse également une pensée particulière à mon grand-père Ali, à qui je rends hommage.

Merci à toutes les personnes qui m'ont soutenu de près ou de loin. Je tiens enfin à rendre grâce à Allah.

Résumé

Dans le contexte de l'industrie 5.0, qui place l'humain au cœur du processus de production, notamment pour favoriser la collaboration humain-robot, améliorer l'ergonomie des postes de travail et mettre à jour les jumeaux numériques, la reconnaissance d'actions humaines (RAH) devient une composante importante. Cependant, malgré les avancées récentes de l'apprentissage profond, la RAH demeure un problème ouvert dans ce contexte industriel. Les approches existantes restent encore peu efficaces pour détecter avec précision et en temps réel des actions hétérogènes, présentant une dynamique temporelle importante et réalisées dans des environnements complexes. Les verrous scientifiques majeurs concernent donc la modélisation conjointe des dynamiques spatiales et temporelles des séquences d'actions à partir de jeux de données déséquilibrés, l'interprétabilité des modèles et leur adaptation aux contraintes du temps réel.

Plusieurs contributions ont permis de relever ces défis. Tout d'abord, la modalité squelettique a été exploitée, car elle est à la fois robuste et précise pour représenter le mouvement humain, et la première proposition a permis de l'intégrer dans une stratégie à multi-résolutions temporelles (coarse-to-fine). La méthodologie proposée améliore la distinction entre les actions de courte durée et dépasse les approches existantes en termes de performance. Ensuite, cette méthodologie a été enrichie d'une représentation en graphe, qui conserve la topologie du squelette tout au long de l'architecture du modèle, et intègre un mécanisme d'attention, permettant une représentation des mouvements plus expressive et une interprétabilité des décisions du modèle de segmentation d'actions. Enfin, comme troisième contribution, nous avons proposé une approche en ligne de détection d'actions, basée sur la distillation progressive des connaissances sous forme de leçons (temporelle, d'occultation et d'objet) transférées d'un modèle hors ligne enseignant vers un modèle en ligne étudiant, et adaptée aux contraintes industrielles. Cette stratégie renforce la stabilité temporelle, la robustesse face aux pertes de données dues aux occultations et la distinction des actions liées aux interactions avec les outils

manipulés (objets) lors des tâches d'assemblage. Nos résultats montrent que notre méthode surpasse l'état de l'art, tant quantitativement que qualitativement, sur le jeu de données InHARD.

Ce travail ouvre des perspectives sur des applications industrielles concrètes. Il pourra notamment contribuer à l'évaluation ergonomique des postes de travail, à la collaboration humain-robot par l'assistance en temps réel aux opérateurs, ainsi qu'à la mise à jour de jumeaux numériques. Ces apports s'inscrivent dans une démarche plus large d'intégration de systèmes de perception intelligents, en cohérence avec la transition vers l'industrie 5.0.

Mots-clés : Industrie 5.0, Reconnaissance d'actions, Apprentissage profond, Représentation basée graphe, Distillation des connaissances, Vision par ordinateur.

Abstract

In the context of Industry 5.0, which places humans at the center of production processes, particularly to foster human–robot collaboration, improve workplace ergonomics, and update digital twins, human action recognition (HAR) is becoming a key component. However, despite recent advances in deep learning, HAR remains an open problem in industrial environments. Existing approaches are still not sufficiently effective at accurately detecting heterogeneous actions in real time, especially when performed in constrained environments. The major scientific challenges lie in jointly modeling the spatial and temporal dynamics of action sequences from imbalanced datasets, ensuring model interpretability, and adapting models to real-time constraints.

Several contributions in this thesis addressed these challenges. First, skeletal data were exploited for their robustness and precision in representing human movement, and the initial contribution introduced a multi-resolution temporal strategy (coarse-to-fine) to process this modality. The proposed methodology improves the discrimination of short-duration actions and surpasses existing approaches in performance. Second, this methodology was extended using a graph-based representation that preserves the skeletal topology throughout the model architecture and incorporates an attention mechanism, enabling more expressive motion representations and improving the interpretability of action segmentation decisions. Finally, as a third contribution, we proposed an online action detection approach based on progressive knowledge distillation. This strategy transfers several types of "lessons" (temporal, occlusion-based, and object-related) from an offline teacher model to an online student model, making it suitable for industrial constraints. This approach enhances temporal stability, robustness to occlusions and missing data, and improves the distinction of actions involving tool manipulation during assembly. Our results show that the proposed method outperforms the state of the art, both quantitatively and qualitatively, on the InHARD dataset.

This work opens several perspectives toward concrete industrial applications. It can support ergonomic workstation assessment, human–robot collaboration through real-time operator assistance, and continuous updating of digital twins. These contributions align with a broader vision of integrating intelligent perception systems, in line with the transition toward Industry 5.0.

Keywords : Industry 5.0, Action Recognition, Deep learning, Graph-based representation, Knowledge distillation, Computer vision.

Table des matières

Remerciements **2**

Résumé **4**

Abstract **7**

Liste des tableaux **18**

Liste des figures **24**

Introduction **25**

1 Analyse du Mouvement Humain en Industrie 5.0 : Contexte, Technologies et Jeux de Données **31**

1.1 Introduction 32

1.2 Industrie 5.0 et l'Analyse du mouvement humain 33

1.2.1 De l'industrie 4.0 à l'industrie 5.0 33

1.2.2 Le rôle de l'humain dans l'industrie 5.0 34

1.2.3 Analyse du mouvement humain : Définitions et périmètre 35

1.2.4 Applications de l'analyse du mouvement humain en industrie 35

1.2.4.1 Méthodologie de l'étude systématique de la littérature 36

1.2.4.1.1 Définition des mots-clés et requête de recherche. . . 36

1.2.4.1.2 Critères de sélection. 36

1.2.4.1.3 Processus de filtrage. 37

1.2.4.1.4 Classification des travaux. 37

1.2.4.2 Sécurité 38

1.2.4.3 Ergonomie 39

1.2.4.4 Productivité et efficacité 41

1.2.5 Synthèse 43
1.3 Technologies de capture de mouvement 44
1.3.1 Technologies de capture de mouvement (MoCap) 44
1.3.1.1 Systèmes optiques 44
1.3.1.2 Systèmes non optiques (Unités de mesure inertielle - IMUs) . 46
1.3.2 Capteurs basés sur les signaux musculaires 46
1.3.2.1 Electromyographie (sEMG) 47
1.3.2.2 Myographie de force (FMG) 48
1.3.2.3 Combinaison EMG et FMG 48
1.3.3 Technologies hybrides 48
1.3.3.1 Combinaisons IMU + capteurs optiques 48
1.3.3.2 Combinaisons MoCap + EMG/FMG 49
1.3.4 Synthèse 49
1.4 Jeux de données pour l'analyse du mouvement humain en industrie 50
1.4.1 Aperçu des jeux de données industriels 50
1.4.1.1 Sécurité et collaboration 51
1.4.1.2 Érgonomie et évaluation des risques 51
1.4.1.3 Suivi des tâches et assistance à l'opérateur 52
1.4.1.4 Navigation et flux opérateurs 52
1.4.1.5 Productivité et optimisation de processus 52
1.4.2 Caractéristiques des jeux de données 52
1.4.2.1 Données segmentées (trimmed) vs non segmentées 53
1.4.2.2 Distribution déséquilibrée des données 53
1.4.2.3 Variabilité des exécutions humaines 54
1.4.2.4 Parties du corps, types d'interactions et variations de points de vue 55
1.4.2.5 Compréhension de l'intention humaine 56
1.4.2.6 Technologies et modalités 57
1.4.2.7 Annotation automatique et jumeaux numériques 58
1.4.3 Synthèse 58
1.5 Discussion et principaux défis 59

1.5.1 Évolution de l'analyse du mouvement humain de l'industrie 4.0 à l'industrie 5.0 . . . 59
1.5.2 Choix des capteurs appropriés . . . 60
1.5.3 Recommandations pour l'acquisition de données . . . 62
1.6 Conclusion . . . 64

2 État de l'art sur les approches d'analyse du mouvement humain 66
2.1 Introduction . . . 67
2.2 Taxonomie des approches . . . 67
2.3 Approches hors ligne (Offline) . . . 69
2.3.1 Reconnaissance d'actions sur séquences *trimmed* . . . 70
2.3.2 Segmentation temporelle sur séquences *untrimmed* . . . 70
2.3.2.1 Définition et objectif . . . 70
2.3.2.2 Approches Top-down . . . 71
2.3.2.3 Approches Bottom-up . . . 72
2.3.2.4 Applications industrielles . . . 73
2.3.3 Architectures utilisées en *offline* . . . 74
2.3.3.1 Réseaux récurrents (RNN, LSTM, GRU) . . . 76
2.3.3.2 Réseaux convolutifs 2D et 3D . . . 76
2.3.3.3 Réseaux de convolutions temporelles (TCN) . . . 77
2.3.3.4 Graph Convolutional Networks (GCN) . . . 77
2.3.3.5 Transformers spatio-temporels . . . 78
2.3.3.6 Méthodes hybrides . . . 79
2.3.4 Synthèse . . . 80
2.4 Approches en ligne (Online) . . . 81
2.4.1 Schéma général des approches en ligne . . . 83
2.4.1.1 Extraction de caractéristiques . . . 83
2.4.1.2 Agrégation temporelle . . . 84
2.4.1.3 Prédiction . . . 85
2.4.2 Détection d'actions image par image avec historique (Historical Per-frame Action Detection) . . . 85
2.4.3 Discrimination de l'information pertinente (Information Discrimination) 86
2.4.4 Détection du début d'action (Detection of Action Start) . . . 87

2.4.5 Anticipation pour la classification (Anticipation for Classification) . . 89
2.4.6 Synthèse 90
2.5 Métriques d'évaluation 91
2.5.1 Métriques dans l'état de l'art 91
2.5.2 Métriques retenues 94
2.5.2.1 Mean-over-Frames (MoF) 94
2.5.2.2 Edit-score 94
2.5.2.3 F1@IoU 95
2.5.3 Synthèse 96
2.6 Discussion et tendances 96
2.6.1 Segmentation d'actions sur données non découpées 96
2.6.2 Approches 97
2.7 Conclusion 98

3 Segmentation d'actions dans des vidéos non découpées à partir de données squelettiques 100
3.1 Introduction 101
3.2 Le choix des données squelettiques 102
3.3 Approche méthodologique proposée 103
3.3.1 Le Backbone 104
3.3.1.1 Extracteur multi-caractéristiques 104
3.3.1.1.1 Positions articulaires (p) : 105
3.3.1.1.2 Angles des membres (b) : 106
3.3.1.1.3 Vitesses articulaires (v) : 107
3.3.2 Architecture encodeur-décodeur 108
3.3.2.1 Encodeur 109
3.3.2.1.1 Bloc DoubleConv. 110
3.3.2.1.2 Hiérarchie temporelle : MaxPool + DoubleConv. . . 110
3.3.2.2 Goulot d'étranglement (Bottleneck) 111
3.3.2.3 Décodeur 113
3.3.3 Bloc de Reconnaissance 114
3.3.4 Synthèse 116
3.4 Protocole expérimental 116

3.4.1 Jeux de données . . . 116
3.4.1.1 Le jeu de données InHARD . . . 117
3.4.1.2 Jeu de données IKEA-ASM . . . 118
3.4.2 Protocole d'entraînement . . . 120
3.4.2.1 Prétraitement . . . 120
3.4.2.2 Détails d'implémentation . . . 121
3.5 Comparaison avec l'état de l'art . . . 123
3.6 Études d'ablation et analyse du modèle . . . 127
3.6.1 Ablation sur la pondération des classes . . . 128
3.6.2 Impact de la modalité squelette . . . 129
3.6.3 Ablation sur les caractéristiques du squelette . . . 131
3.7 Conclusion . . . 132

4 Segmentation d'actions basées sur des graphes avec un mécanisme d'attention 134
4.1 Introduction . . . 135
4.2 Modélisation Graphique du Mouvement Humain . . . 136
4.2.1 Réseaux de Convolution Graphique (GCN) . . . 136
4.2.2 Topologie du graphe squelettique . . . 138
4.3 Approche méthodologique proposée . . . 140
4.3.1 Formulation du problème . . . 140
4.3.2 Architecture proposée . . . 141
4.3.3 Vue d'ensemble de l'approche . . . 142
4.3.4 Réseau de convolution spatio-temporel sur graphe . . . 144
4.3.4.1 Convolution spatiale sur graphe. . . . 144
4.3.4.2 Convolution temporelle. . . . 146
4.3.4.3 Intégration dans le G-Encodeur . . . 146
4.3.5 G-Encodeur . . . 146
4.3.6 G-Bottleneck . . . 148
4.3.7 G-Attention . . . 149
4.3.8 G-Decoder . . . 151
4.3.9 Agrégation différentielle . . . 153
4.3.10 Bloc de reconnaissance et de segmentation . . . 154
4.4 Protocole Expérimental . . . 155

4.4.1 Jeux de données . . . 155
4.4.2 Protocole d'entraînement . . . 157
4.5 Comparaison avec l'état de l'art . . . 159
4.5.1 Résultats sur le jeu de données InHARD . . . 159
4.5.2 Résultats sur le jeu de données HA4M . . . 160
4.6 Études d'ablation . . . 161
4.6.1 Influence de la représentation graphique et du mécanisme d'attention 161
4.6.2 Impact de la topologie du graphe squelettique . . . 167
4.6.3 Différentes méthodes d'agrégation . . . 168
4.7 Visualisation et discussion . . . 169
4.7.1 Répartition des actions . . . 169
4.7.2 Variabilité de la durée des actions . . . 171
4.7.3 Variabilité de l'ordre des actions et des gestes . . . 173
4.8 Conclusion . . . 174

5 Détection d'actions en ligne 176
5.1 Introduction . . . 177
5.2 Enjeux de la détection d'actions en ligne . . . 178
5.2.1 Contrainte de temps réel . . . 178
5.2.2 Gestion de l'incertitude et des ambiguïtés . . . 179
5.2.3 Perspectives applicatives . . . 179
5.3 Méthodes de distillation de connaissances (enseignant-étudiant) . . . 180
5.3.1 Distillation au niveau des logits . . . 181
5.3.2 Distillation au niveau des caractéristiques . . . 182
5.3.3 Distillation des relations . . . 184
5.3.4 Analyse des méthodes de distillation . . . 185
5.4 Méthode de distillation proposée . . . 185
5.5 Leçons proposées transmises par le modèle enseignant . . . 187
5.5.1 Leçon temporelle . . . 187
5.5.2 Leçon liée aux objets . . . 188
5.5.3 Leçon liée à l'occultation . . . 191
5.5.4 Leçons cumulées . . . 192
5.5.4.1 Leçons combinées . . . 192

5.5.4.2 Leçons progressives 193
5.6 Résultats expérimentaux et discussion 195
5.6.1 Choix des modèles enseignant–étudiant 195
5.6.2 Protocole expérimental 196
5.6.3 Analyse de la leçon temporelle 198
5.6.4 Analyse de la leçon liée aux objets 199
5.6.5 Analyse de la leçon liée aux occultations 201
5.6.6 Analyse des leçons cumulées 202
5.6.7 Discussion et visualisation 204
5.7 Conclusion 209

Conclusion **211**
5.8 Résumé des contributions 212
5.9 Défis et perspectives futures 213
5.9.1 perspectives scientifiques 213
5.9.2 perspectives applicatives 214

Bibliographie **216**

Liste des tableaux

1.1 Comparaison des systèmes de capture de mouvement pour l'industrie [13] . . 46
1.2 Comparaison des systèmes de détection du signal musculaire pour la capture du mouvement humain en industrie. 47
1.3 Comparaison des technologies hybrides pour l'analyse du mouvement en industrie. 50

1.4 Jeux de données sélectionnés pour l'analyse du mouvement humain en contexte industriel. U : Séquences non segmentées (Untrimmed), T : Séquences segmentées (Trimmed), HO : Interaction Humain-Objet, HR : Interaction Humain-Robot, NI : Pas d'interaction, P : Représentation partielle (main, bras), Up : Haut du corps, F : Corps entier, E : Vue égocentrique, A : Vue allogénique (3e personne), NA : Non applicable, MNP : Modèle non précisé. 51
1.5 Moyenne et écart-type du nombre d'échantillons par classe pour quelques jeux de données. 54

2.1 Approches proposées dans l'état de l'art pour l'analyse du mouvement humain en contexte industriel. P : Privé, RA : Reconnaissance d'Actions, VMM : Modélisation de Markov à longueur variable. 75
2.2 Métriques utilisées pour l'évaluation des approches d'analyse du mouvement humain. TP : Vrai Positif, TN : Vrai Négatif, FP : Faux Positif, FN : Faux Négatif, P : étiquettes prédites, GT : Vérité Terrain, IoU : Intersection over Union, $D(.,.)$: distance de Levenshtein [159] 93

3.1 Synthèse des dimensions à travers les blocs de GSK-C2F (voir Fig. 3.1). La dimension temporelle est notée T et la dimension des canaux C. 116
3.2 Comparaison des scores MoF (Mean-over-Frames) avec les travaux récents sur le jeu de données InHARD. * : méthodes réimplémentées. Les caractéristiques utilisées dans l'architecture de base sont indiquées entre parenthèses : p (positions articulaires), b (angles osseux), v (vitesses articulaires). 124
3.3 Comparaison des scores MoF (Mean-over-Frames) avec les travaux récents sur le jeu de données IKEA-ASM. Les caractéristiques utilisées dans l'architecture de base sont indiquées entre parenthèses : p (positions articulaires), b (angles osseux), v (vitesses articulaires). 127
3.4 Comparaison du score *Mean-over-Frames* (MoF), du nombre de paramètres (en millions) et de la complexité computationnelle (en GMACs) pour différentes approches sur le jeu de données InHARD-13. * : méthodes réimplémentées. . 128
3.5 Étude d'ablation de la stratégie de pondération (WS) dans GSK-C2F(p) sur InHARD-13. 129
3.6 Étude comparative des modalités sur InHARD-13, InHARD-4 et InHARD-3. 130
3.7 Étude d'ablation des caractéristiques du squelette sur InHARD-13. 131

4.1 Comparaison des scores MoF (%) avec les meilleurs résultats rapportés par chaque méthode de l'état de l'art sur le jeu de données InHARD (toutes topologies confondues). * : méthodes ré-implémentées. NA : information non disponible. 160
4.2 Comparaison avec l'état de l'art sur HA4M en utilisant G-EAD. Haut : partie supérieure du corps, RGB-A : alignement RGB-Depth. * : résultats non reproductibles, car la distribution d'entraînement/test et le code non publiés. 161
4.3 Étude d'ablation des différentes architectures basées sur les graphes sur le jeu de données InHARD-13. B_G-C2F : ST-GCN utilisé comme backbone. C2F-E_G-D : ST-GCN intégré à l'encodeur. C2F-E_G-D_G : ST-GCN intégré à l'encodeur, au goulot d'étranglement et au décodeur. G-EAD : architecture encodeur-décodeur avec ST-GCN et mécanisme d'attention. 163
4.4 Comparaison du score moyen par image (MoF), du nombre de paramètres (en millions) et de la complexité de calcul (GMACs) entre des approches à base de graphes évaluées sur InHARD-13 à 21 jointures. * : méthodes ré-implémentées. 166
4.5 Impact des différentes topologies de graphe sur les performances de segmentation (MoF en %) sur le jeu de données InHARD-13 à 21 jointures. 167
4.6 Impact de la stratégie de pooling (PS) sur les performances de segmentation du modèle G-EAD sur le jeu de données InHARD-13 à 21 jointures. 168

5.1 Performances des modèles de détection d'action en ligne : Comparaison entre GSK-C2F et GEAD sur la base InHARD-13 à 21 articulations (complexité, coûts de calcul et performances). 195
5.2 Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 entre le modèle de base (sans distillation) et le modèle étudiant (avec distillation temporelle). 199
5.3 Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 entre le modèle de base (sans distillation) et le modèle étudiant (avec distillation liée aux objets). 200
5.4 Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 entre le modèle de base (sans distillation) et le modèle étudiant (avec distillation liée aux occultations : fb (full-body) → up (upper-body)). 201

5.5 Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 à 53 articulations entre le modèle de base (sans distillation) et les modèles étudiants (avec distillation cumulée : Leçons combinées vs progressives). . . . 202
5.6 Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 à 53 articulations entre le modèle de base (sans distillation) et les modèles étudiants (avec distillation cumulée : Leçons combinées vs progressives). . . . 203
5.7 Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 à 53 articulations entre le modèle de base (sans distillation) et les modèles étudiants (avec distillation cumulée : Leçons combinées vs progressives). . . . 205

Table des figures

1.1 Applications de l'analyse du mouvement humain en industrie. (a) Principales applications considérant la collaboration homme-robot. (b) Sous-domaines liés à la productivité et à l'efficacité. 38
1.2 Technologies MoCap : (a) combinaison optique [63], (b) marqueurs actifs [64], (c) Kinect V2 [4], (d) Kinect V1 [65], (e) Perception Neuron 32 [3]. 45
1.3 Capteurs de suivi musculaire. Exemples de données acquises via : (a) bracelets Myo [18], (b) bracelets FMG [72]. 47
1.4 Statistiques des jeux de données dans le tableau 1.4. (a) Fréquence des parties du corps représentées. (b) Répartition des types d'interactions. (c) Nombre de jeux de données segmentés/non segmentés par an. 53
1.5 Exemples illustrant la similarité entre différentes classes d'actions dans le jeu de données InHARD. 55
1.6 Exemples illustrant la variabilité intra-classe dans le jeu de données InHARD. 56
1.7 Illustrations des jeux de données, regroupées par parties du corps suivies, par types d'interaction et par angles de vue. (a) [15], (b) [48], (c) [69], (d) [4], (e) [3]. 57

2.1 Taxonomie des approches d'analyse du mouvement humain selon le type de séquence, le régime de traitement, l'architecture utilisée et le type d'application. 68

2.2 Taxonomie des approches hors ligne (offline) pour l'analyse du mouvement humain. 69
2.3 Répartition statistique des approches de reconnaissance d'actions selon l'architecture utilisée, d'après la littérature récente (2018-2023). [11] 74
2.4 Exemple d'architecture CNN. 77
2.5 Exemple d'architecture GCN spatio-temporel inspirée de [139]. 77
2.6 Exemple d'architecture Transformer extraite de [142] 78
2.7 Exemple d'architecture hybride. 80
2.8 Taxonomie des approches en ligne (online) pour l'analyse du mouvement humain. 82
2.9 Illustration du calcul du MoF. f_n : frame ou image d'indice n. 94
2.10 Illustration de l'Edit-score : comparaison entre la séquence temporelle prédite et la vérité du terrain, mesurée par la distance de Levenshtein normalisée, afin de quantifier les erreurs d'ordre et de segmentation. 95
2.11 Illustration du F1 score avec un IoU de 50% (la moitier de la taille des segments du GT) sur La séquence "Pred". GT : la sequence segmentée de verité terrain. Pred : la sequence segmentée predite par le model. 96

3.1 Architecture globale de GSK-C2F. La séquence d'entrée illustrative provient du jeu de données InHARD [3]. p : positions articulaires, b : angles des membres, v : vitesses articulaires, Y_l : sortie de la couche l du décodeur. 103
3.2 Représentations squelettiques : (p) positions 3D des articulations en bleu (référencées par rapport au bassin en rouge), (b) angles des membres par rapport à l'axe z, et (v) vitesses articulaires. 105
3.3 Détails des modules : (a) encodeur, (b) goulot d'étranglement et (c) décodeur. Xe : entrée de l'encodeur, X_{ef_out} : sortie finale de l'encodeur, $X^l_{e_out}$: connexions résiduelles de la couche l, X_{b_out} : sortie du bottleneck (utilisée comme entrée du décodeur), X_d : entrée du décodeur, $X^l_{d_out}$: sortie du décodeur de la couche l, Y : prédiction. 109
3.4 Bloc de segmentation et de reconnaissance : pour chaque couche de décodeur l, la sortiea du décodeur $Y(l,t)$ (voir figure 3.3.c), est rééchantillonnée à la longueur temporelle initiale T afin de générer de nouvelles représentations $Y'(l,T)$. Ces caractéristiques sont ensuite fusionnées entre les couches, puis passées par une fonction softmax afin de produire des prédictions d'action, image par image. 114

3.5 Graphique croisé dynamique du nombre d'actions et leurs durées dans le jeu de données InHARD-13 [3]. 118
3.6 Distribution de la durée des actions dans le jeu de données InHARD-13 [3]. Les classes "No action" et "Assemble system" comptent le plus grand nombre d'échantillons et la plus longue durée totale. 119
3.7 Labélisation des actions dans le jeu de données InHARD-13 [3] avec l'outil ANVIL. En vert le label d'action (bas niveau), en jaune le label d'action (haut niveau). 119
3.8 Exemples d'outils de détection de la pose humaine utilisés pour extraire des squelettes à partir de séquences RGB. 121
3.9 Matrices de confusion C2F-TCN (RGB) sur le jeu de données InHARD-13 . . 125
3.10 Matrices de confusion GSK-C2F(pbv) sur le jeu de données InHARD-13 . . . 126
3.11 Segmentation temporelle pour la vidéo "P13-R02" d'InHARD-13 avec GSK-C2F(p) : GT (vérité du terrain), LCW (pondération faible) et No_W (sans pondération). 129
3.12 Segmentation temporelle de la vidéo "P13-R02" d'InHARD-13 : vérité de terrain (GT) et résultats prédits pour chaque modalité : 3D-SK, E-SK et RGB. 131

4.1 Schéma global de l'approche proposée : du signal d'entrée, constitué d'une séquence 3D de squelettes (en haut), jusqu'à la sortie segmentée des actions (en bas). Chaque segment correspond à une classe d'action (par exemple, "Prendre à gauche (Pick left)", "Tourner la feuille (Turn sheet)"). 142
4.2 Vue d'ensemble de l'architecture proposée, nommée G-EAD : d'abord, les séquences vidéo de squelettes sont traitées par un encodeur de graphes (1), suivi d'un goulot d'étranglement (2) pour capturer des caractéristiques multi-résolutions. Ensuite, un mécanisme d'attention (3) est appliqué afin de se concentrer sur les articulations pertinentes. Les séquences sont ensuite traitées par un décodeur de graphes (4), dont la sortie de chaque couche est transmise à un module de regroupement différentiel (5) afin de réduire hiérarchiquement le graphe. Enfin, la segmentation des actions est réalisée par un bloc d'agrégation à prédiction ensembliste (6). 143

4.3 Structure d'une couche d'encodeur (`ST-Down`) : les représentations squelettiques passent d'abord par une convolution graphique (GCN), suivie d'un max-pooling temporel, puis d'une convolution temporelle (TCN). Cette combinaison permet de capter des motifs d'action tout en réduisant la durée de la séquence. A : matrice d'adjacence, $\otimes$: agrégation spatiale par multiplication matricielle. 147

4.4 Illustration du module G-Bottleneck : des opérations de max-pooling parallèles sont appliquées sur l'axe temporel à l'aide de différents noyaux. Chaque branche subit une compression de canaux par convolution, puis un rééchantillonnage à la longueur initiale. L'ensemble est concaténé à la représentation d'origine. 149

4.5 Illustration de G-Attention : d'abord, les requêtes et les clés sont extraites des caractéristiques d'entrée H pour produire une matrice de similarité spatiale. Puis, les scores d'attention sont calculés via une opération softmax et modulés par la matrice d'adjacence A par multiplication élément par élément. Enfin, plusieurs têtes d'attention sont concaténées, suivies d'une connexion résiduelle et d'une activation ReLU. $\odot$: multiplication élément par élément. 150

4.6 Illustration d'une couche du décodeur (ST-Up) : d'abord, l'entrée $\mathbf{X}_{gd}^{l}$ est suréchantillonnée, puis combinée avec la sortie de l'encodeur ($\mathbf{X}^{(l+1)}$ en figure 4.3) via les connexions résiduelles $\mathbf{X}_{ge_out}^{l}$, et enfin traitée par le bloc TCN. La sortie $\mathbf{X}_{gd}^{(l+1)}$ est utilisée comme entrée pour la couche suivante et pour le pooling global lors de la prédiction $\mathbf{Y}^{l}$. 152

4.7 Illustration d'une couche de pooling différentiel (DFpool) : à chaque couche, les embeddings de caractéristiques et les matrices d'adjacence sont générés, réduisant progressivement le nombre d'articulations du squelette (par exemple, de 21 à 13 puis à 1). La sortie X_{out} ainsi que la matrice d'adjacence mise à jour A_{out} sont transmises à la couche suivante. 153

4.8 Bloc de segmentation et reconnaissance : les sorties du décodeur $X(l,t)$ sont rééchantillonnées à T, agrégées via un pooling différentiel, puis fusionnées entre couches avant la fonction softmax finale. 154

4.9 Modalités présentes dans le jeu de données HA4M [65]. 156

4.10 Impact de la taille d'embedding d'augmentation dans le goulot d'étranglement (d^x) et du nombre de têtes d'attention sur la performance en MoF de l'approche G-EAD évaluée sur le jeu de données InHARD. 157
4.11 Évolution de l'architecture. Étape 1 (B_G-C2F) : Intégration du graphe en tant que backbone pour l'extraction des caractéristiques. Étape 2 (C2F-E_G-D) : Ajout de la structure graphique dans l'encodeur afin d'affiner l'extraction des caractéristiques avant le décodage. Étape 3 (C2F-E_G-D_G) : Intégration du graphe également au niveau du décodeur pour enrichir la représentation. Étape 4 (G-EAD) : Introduction d'un mécanisme d'attention spatiale pour capturer plus efficacement les dépendances entre les articulations les plus pertinentes dans les séquences d'actions squelettiques. 162
4.12 Représentation des cartes d'attention pour chaque seconde dans l'intervalle de temps (secondes 6 à 8) de l'action « Prendre à gauche ». Images extraites de la vidéo « P13-R02 » du jeu de données InHARD-13. 165
4.13 Matrice de confusion pour les classes d'actions sous-représentées sur InHARD-13, avec (a) GSK-C2F et (b) le modèle proposé G-EAD utilisant 21 articulations. Id d'action, Classe (Nombre d'échantillons) : 0 : "Aucune action" (6606) ; 1 : "Consulter fiche" (132) ; 3 : "Prendre tournevis" (420) ; 7 : "Prendre réglette" (76) ; 8 : "Poser réglette" (74) ; 10 : "Poser composant" (385) ; 12 : "Prendre sous-ensemble" (39). 170
4.14 Chronologies de segmentation d'actions sur le jeu de données InHARD pour les modèles G-EAD et C2F-TCN (approche RGB) [60] (vidéo P13R02). Chaque barre colorée correspond à une action parmi les 14 classes d'action. GT : Vérité terrain (Ground Truth). 172
4.15 Chronologie de la segmentation dans le jeu de données InHARD (vidéo P13R02 – 100 premières secondes) obtenue par le modèle G-EAD. Chaque barre colorée représente l'une des 14 actions. GT : Vérité Terrain. 173
4.16 Chronologie de la segmentation dans le jeu de données HA4M (vidéo ID035U001) obtenue par le modèle G-EAD. Chaque barre colorée représente l'une des 12 actions. GT : Vérité Terrain. 173

5.1 Vue d'ensemble de l'architecture Enseignant–Étudiant proposée avec distillation par logits. ws : taille de la fenêtre glissante. $\mathcal{L}_{\text{cls}}$: la fonction de perte de classification (voir l'équation 3.23). $\mathcal{L}_{\text{KD-logit}}$: la fonction de perte de distillation de logits (voir l'équation 5.2). 181

5.2 Vue d'ensemble de l'architecture Enseignant–Étudiant proposée avec distillation par caractéristiques. ws : taille de la fenêtre glissante. $\mathcal{L}_{\text{cls}}$: la fonction de perte de classification (voir l'équation 3.24). $\mathcal{L}_{\text{KD-feature}}$: la fonction de perte de distillation des caractéristiques internes (voir l'équation 5.4). 183

5.3 Schéma du processus de distillation de la leçon temporelle : Le modèle enseignant, disposant de l'ensemble de la séquence, fournit un signal temporel cohérent, qui est transmis au modèle étudiant, disposant uniquement d'une fenêtre temporelle limitée, afin d'améliorer la détection d'action en ligne. . . . 189

5.4 Schéma du processus de distillation de la leçon objet : Le modèle enseignant exploite à la fois le contexte temporel global et l'information de l'outil manipulé, ici "Le manuel", qui oriente directement vers la bonne action. Ce signal enrichi est ensuite transmis au modèle étudiant afin d'améliorer la détection d'action en ligne. 189

5.5 Processus de fusion résiduelle (Early Fusion Motion–Tool) entre les embeddings d'objets et les caractéristiques de mouvement. Les tenseurs suivent la forme $(B, T, \cdot)$, où B désigne la taille du batch, T la longueur de la séquence et $\cdot \in \{nbr_{outil}, dim, D\}$ la taille des embeddings. x : caractéristiques squelettiques, e embedding de l'outil manipulé, W_o : matrice projetant l'espace des outils vers celui du mouvement, λ : coefficient de l'outil, x' : caractéristiques de sorite. $\oplus$: addition. $\otimes$: multiplication. 190

5.6 Schéma du processus de distillation de la leçon d'occultation. Le modèle enseignant utilise le contexte temporel complet et un squelette non occulté pour guider un modèle étudiant doté d'un squelette partiellement occulté, améliorant ainsi la détection d'actions en ligne. 192

5.7 Exemple d'un processus de distillation progressive. Cet exemple correspond au modèle (Étudiant P2) présent dans le tableau 5.7. 194

5.8 Schéma du processus de détection d'action en ligne par fenêtre glissante. Chaque couleur dans la séquence de sortie représente une classe d'action. w : la taille de la fenêtre glissante. 197
5.9 Prédictions des modèles hors ligne sur l'intervalle [8 sec - 70 sec] de la séquence (P13-R02) du jeu de données InHARD. (a) Confiance des modèles au fil du temps. (b) Segments temporels prédits (chaque couleur correspond à une action). Enseignant : modèle hors ligne (contexte offline), Enseignant Objet : modèle hors ligne (contexte offline + objet), GT : vérité du terrain. 201
5.10 Prédictions des approches en ligne utilisant différentes configurations de distillation à partir du modèle Enseignant sur l'intervalle [8 sec - 70 sec] de la séquence (P13-R02) du jeu de données InHARD. (a) Confiance des modèles au fil du temps. (b) Segments temporels prédits (chaque couleur correspond à une action). Base : modèle de base (sans distillation), S1 : étudiant avec distillation temporelle, C1 : étudiant avec distillation par occultation, C2 : étudiant avec distillation d'objet, P1 : étudiant avec distillation objet-temporelle progressive, Enseignant : modèle hors ligne, GT : vérité du terrain. 206
5.11 Prédictions des approches de détection d'action en ligne sur l'intervalle [8 sec - 70 sec] de la séquence (P13-R02) du jeu de données InHARD. (a) Confiance des modèles au fil du temps. (b) Segments temporels prédits (chaque couleur correspond à une action). Base : modèle en ligne de base (sans distillation). P1 : modèle étudiant en ligne avec distillation, objet-temporelle progressive. Enseignant : modèle hors ligne. GT : vérité du terrain. 208

Introduction

L'avènement de l'Industrie 5.0 marque un tournant décisif dans l'évolution du secteur manufacturier moderne, en réaffirmant le rôle central de l'opérateur humain au sein des systèmes industriels [1]. Contrairement aux révolutions industrielles précédentes, principalement orientées vers l'automatisation et la maximisation de la production, l'Industrie 5.0 vise à créer un environnement de production plus personnalisé, durable et résilient, où l'opérateur collabore avec la machine dans le processus de production. Cette transition ne vise pas seulement à répondre aux enjeux environnementaux et aux nouvelles dynamiques mondiales, mais également à améliorer le bien-être, la sécurité et l'implication des travailleurs. Dans ce cadre, la compréhension et l'analyse des mouvements humains constituent des leviers fondamentaux pour assurer une coopération humain-machine à la fois sûre, performante et intelligente.

L'analyse du mouvement humain englobe l'étude et la reconnaissance des gestes manuels, des actions et des activités humaines, ainsi que des interactions entre l'opérateur et son environnement, incluant les outils, les robots ou les autres collaborateurs présents dans l'environnement de travail. Elle présente un large éventail d'applications, notamment dans le domaine de l'ergonomie, où elle permet de générer des rapports visant à préserver la santé et la sécurité des opérateurs en réduisant les risques de troubles musculo-squelettiques. Cette analyse joue également un rôle déterminant dans l'assistance aux opérateurs lors de tâches d'assemblage dans des environnements collaboratifs en présence de robots, ce qui contribue à accroître à la fois la productivité et l'efficacité des processus industriels.

Dans le cadre plus général de l'analyse du mouvement humain, plusieurs tâches fondamentales ont émergé pour mieux comprendre et interpréter les activités humaines à partir de données visuelles.

La reconnaissance d'actions humaines (*Human Action Recognition*, HAR) traditionnelle se concentre sur la classification d'actions à partir de séquences vidéo pré-segmentées, dans lesquelles une unique action est déjà isolée et délimitée temporellement [2]. Bien que cette approche ait permis des avancées notables, notamment grâce à l'essor de l'apprentissage profond et à la disponibilité croissante de jeux de données de référence à grande échelle, elle reste éloignée des conditions rencontrées dans les scénarios réels, où les actions s'enchaînent de manière continue, sans frontières temporelles clairement définies.

Pour dépasser cette limitation, le champ d'étude s'est progressivement élargi vers la segmentation et la détection d'actions appliquées à des séquences vidéo non segmentées. **La segmentation d'actions** (Temporal Action Segmentation, TAS) vise à attribuer une étiquette

à chaque instant d'un flux vidéo continu, produisant une partition temporelle complète en segments successifs. Elle est généralement étudiée dans un cadre hors ligne, où l'ensemble de la séquence est disponible lors de l'inférence. À l'inverse, les approches en ligne relèvent principalement du paradigme de **la détection d'actions** (Online Action Detection, OAD), dont l'objectif est de reconnaître l'action en cours à l'instant présent sans accès aux images futures. Ainsi, dans la littérature, les travaux hors ligne sont majoritairement associés à la reconnaissance et à la segmentation temporelle, tandis que les travaux en ligne s'inscrivent principalement dans une logique de détection sous contrainte causale.

Malgré les avancées significatives en matière de reconnaissance et de détection d'actions humaines, de nombreux défis persistent, notamment lorsqu'il s'agit d'analyser des actions complexes dans des environnements industriels réels. Ces situations se caractérisent souvent par des interactions main-objet sophistiquées et par de longues dépendances temporelles, que les modèles standards peinent à modéliser efficacement.

Dans ce contexte, ce mémoire s'intéresse à la problématique de la reconnaissance et de la détection d'actions humaines en environnements industriels, afin de favoriser une collaboration efficace entre humains et robots, tout en apportant des éléments d'analyse pour l'étude ergonomique des postes de travail. L'objectif est ainsi de contribuer à l'amélioration des conditions de travail, de la sécurité et de l'efficacité des opérateurs. Une première phase de l'étude est consacrée à l'exploration des approches existantes de reconnaissance d'actions humaines en ligne à partir de séquences vidéo non segmentées. Cette phase examine les technologies de capture du mouvement, les jeux de données de référence disponibles, ainsi que les métriques couramment utilisées pour l'évaluation de ces modèles. Dans un second temps, afin de répondre à la problématique de la segmentation d'actions dans des vidéos non segmentées, nous explorons des approches d'apprentissage profond fondées sur des structures de graphes, exploitant des données squelettiques afin d'améliorer la reconnaissance d'actions en contexte industriel. Enfin, en nous appuyant sur les connaissances et résultats obtenus, nous étendons notre étude à la détection d'actions en ligne, dans le but de répondre aux exigences d'assistance en temps réel des opérateurs lors de tâches industrielles collaboratives. La mise en œuvre de tels systèmes dans des contextes industriels réels soulève toutefois un ensemble d'exigences applicatives qu'il convient de prendre en compte. Dans les environnements industriels visés par ce travail, et en particulier dans le cadre de tâches d'assemblage collaboratif, l'analyse des actions humaines repose sur des contraintes applicatives fortes. Dans un premier

temps, les approches hors ligne doivent être capables de traiter des séquences vidéo longues et non segmentées, dans lesquelles des actions fines et répétitives s'enchaînent sans frontières temporelles clairement définies. Elles doivent permettre une modélisation précise des dépendances spatiales et temporelles, tout en offrant des performances élevées et une robustesse suffisante face aux variations inter-opérateurs, aux occlusions partielles et aux interactions complexes entre l'opérateur et les objets manipulés. Toutefois, dans un contexte d'assistance à l'opérateur, ces approches hors ligne ne répondent pas à elles seules aux exigences applicatives, puisqu'elles supposent un accès complet à la séquence vidéo. Les approches en ligne introduisent alors des contraintes supplémentaires, en imposant un traitement en temps réel des données, sans accès au futur et avec des ressources de calcul limitées. Elles doivent être capables de fournir des prédictions fiables à chaque instant, tout en préservant un compromis entre performances, frugalité computationnelle et stabilité des décisions, afin de garantir une assistance efficace et sûre pour la collaboration humains-robots dans des environnements industriels.

Afin de répondre à ces enjeux, ce travail, financé par une allocation de l'établissement CESI, apporte trois contributions scientifiques principales. Les deux premières contributions portent sur la reconnaissance et la segmentation d'actions humaines hors ligne dans des environnements industriels, en tenant compte de séquences longues et non segmentées. La première approche repose sur l'extraction de caractéristiques pertinentes à partir des données squelettiques afin de représenter efficacement les mouvements humains. La seconde approche étend ce cadre en introduisant une modélisation du squelette sous forme de graphe, associée à un mécanisme d'attention permettant d'améliorer l'interprétabilité du modèle. Ces approches visent à modéliser efficacement les dépendances spatiales et temporelles des mouvements humains, tout en répondant aux contraintes spécifiques des scénarios industriels. Dans la continuité de ces travaux, une troisième contribution porte sur la reconnaissance d'actions humaines en ligne. Elle s'appuie sur le modèle hors ligne précédemment proposé, dont les connaissances sont transférées vers un modèle en ligne plus léger , plus particulièrement par une fenêtre temporelle plus courte, au moyen d'un mécanisme de distillation. Ce mécanisme intègre des informations de contexte spatio-temporel global ainsi que des informations relatives à l'outil manipulé, afin de répondre aux contraintes du milieu industriel et d'assurer une assistance en temps réel dans des scénarios de collaboration humain-robot.

Plan du mémoire

La suite de ce manuscrit est structurée en quatre chapitres, brièvement décrits ci-dessous :

Chapitre 1 : Ce chapitre introduit le contexte général de l'analyse du mouvement humain dans la transition de l'industrie 4.0 vers l'industrie 5.0. On y présente le rôle accru de l'opérateur au sein des systèmes collaboratifs humain-machine, ainsi que les principaux domaines d'application de cette discipline en environnement industriel. Une revue des dispositifs de capture du mouvement est ensuite proposée, accompagnée d'une analyse comparative des jeux de données existants. Enfin, une discussion met en évidence les défis actuels et formule des recommandations pour la conception et l'exploitation de données adaptées aux besoins de l'industrie 5.0.

Chapitre 2 : Après avoir établi les fondements généraux de l'analyse du mouvement humain en contexte industriel, ce chapitre se concentre plus spécifiquement sur la reconnaissance et la segmentation temporelle des actions. Il met en évidence les enjeux propres à l'interprétation des actions dans des séquences vidéo continues issues de milieux manufacturiers. Une taxonomie des approches existantes est proposée, en distinguant clairement les méthodes *en ligne* et *hors ligne*, et en s'appuyant sur des analyses statistiques permettant d'identifier les tendances actuelles de la recherche ainsi que les lacunes encore présentes. Les travaux récents les plus significatifs sont ensuite examinés, avec une attention particulière portée à leurs performances et à leurs limites face aux contraintes industrielles, telles que le traitement en temps réel, la robustesse aux variations contextuelles ou encore la frugalité. Ce chapitre permet ainsi de resserrer le champ d'étude en identifiant précisément les problématiques ciblées dans le cadre de cette thèse.

Chapitre 3 : Ce chapitre présente notre première contribution, consacrée à la reconnaissance et à la segmentation d'actions humaines dans des vidéos non segmentées, à l'aide d'une approche d'apprentissage profond. En s'appuyant sur les données squelettiques, la méthode proposée se distingue par sa robustesse face aux contraintes propres aux environnements industriels complexes. L'approche est évaluée sur deux jeux de données industriels de référence : *InHARD* [3] et *IKEA ASM* [4]. Nous décrivons, dans un premier temps, le processus d'ingénierie des descripteurs squelettiques, avant de détailler l'architecture du modèle développé afin d'améliorer la segmentation temporelle des actions.

Chapitre 4 : Ce chapitre prolonge la contribution précédente en approfondissant l'intégration de la structure de graphe au sein de l'architecture, cette fois de manière holistique.

Un mécanisme d'attention y est introduit afin de pondérer dynamiquement les informations spatiales et temporelles les plus pertinentes. Cette extension permet non seulement d'améliorer les performances du modèle, mais aussi de renforcer son interprétabilité. Cet aspect s'avère particulièrement important dans le contexte industriel, où la transparence, la fiabilité et l'explicabilité des systèmes sont essentielles.

Chapitre 5 : Dans la continuité du travail présenté au chapitre précédent, ce chapitre aborde la problématique de la détection d'actions humaines en ligne, où la reconnaissance doit s'effectuer en temps réel à partir de flux de données continus. Nous y présentons notre seconde contribution : un modèle d'apprentissage profond basé sur la distillation de connaissances. Plusieurs stratégies de transfert sont explorées, introduisant le concept de leçons distillées issues d'un modèle hors ligne vers un modèle en ligne, et évaluées dans le cadre de scénarios industriels dynamiques.

Enfin, ce manuscrit se conclut par une synthèse des contributions présentées et propose plusieurs perspectives de recherche qui nous semblent prometteuses à explorer.

Chapitre 1

Analyse du Mouvement Humain en Industrie 5.0 : Contexte, Technologies et Jeux de Données

1.1 Introduction

Ce premier chapitre a pour vocation de situer le contexte dans lequel s'inscrit notre travail de recherche. Il propose une vue d'ensemble de l'évolution récente des systèmes de production industriels et met en lumière le rôle croissant de l'analyse du mouvement humain dans cette transformation. L'industrie 4.0, fondée sur la numérisation, l'automatisation et l'exploitation massive des données, a permis d'importants gains en termes de productivité et de flexibilité. Toutefois, ses limites en matière de durabilité, de résilience et de prise en compte de l'opérateur humain ont conduit à l'émergence de l'industrie 5.0 [1]. Celle-ci replace l'humain au centre des processus, en privilégiant la collaboration humain-machine, la personnalisation et l'intégration de considérations sociales et environnementales. L'analyse du mouvement humain s'impose alors comme une composante essentielle pour comprendre, modéliser et anticiper les actions et intentions des opérateurs, afin de concevoir des environnements de travail plus sûrs, ergonomiques et performants. Dans ce cadre, notre étude se concentre plus particulièrement sur les tâches d'assemblage industriel, avec pour objectif la reconnaissance continue des actions réalisées par les opérateurs au cours de l'exécution de leurs activités.

Pour structurer cette réflexion, ce chapitre se décompose en plusieurs parties complémentaires. La section 1.2 retrace l'évolution du paradigme de l'industrie 4.0 vers l'industrie 5.0, définit le cadre industriel dans lequel se situe notre recherche et met en évidence le rôle de l'humain dans ce nouveau contexte, tout en présentant les principales applications de l'analyse du mouvement humain dans un environnement industriel.

Dans un second temps, la section 1.3 se concentre sur les technologies permettant de capter et de modéliser le mouvement humain. Nous y présentons les approches basées sur les caméras, les capteurs inertiels, les signaux physiologiques ou encore les systèmes hybrides, en mettant en évidence leurs avantages, leurs limites et leurs conditions d'intégration en milieu industriel. La section 1.4 s'intéresse ensuite aux jeux de données disponibles dans ce domaine. Nous analysons leurs caractéristiques, les scénarios représentés et les modalités d'annotation, tout en soulignant leurs limites, leurs contraintes et les perspectives d'amélioration. Enfin, la section 1.5 propose une discussion approfondie visant à dégager les technologies les mieux adaptées aux environnements industriels et à identifier les jeux de données les plus pertinents pour évaluer les méthodes de reconnaissance et de segmentation d'actions.

Ainsi conçu, ce chapitre fournit le socle contextuel et méthodologique indispensable pour

appréhender les contributions de cette thèse. Il offre une vue d'ensemble de l'environnement industriel, des applications visées, des outils disponibles et des ressources de données mobilisables, préparant ainsi le terrain pour l'étude des approches d'apprentissage et des architectures de reconnaissance détaillées dans les chapitres suivants.

1.2 Industrie 5.0 et l'Analyse du mouvement humain

1.2.1 De l'industrie 4.0 à l'industrie 5.0

L'industrie 4.0 a marqué un tournant décisif dans l'automatisation des processus de production, grâce à l'intégration de technologies telles que l'Internet des objets (IoT), les systèmes cyber-physiques (CPS), l'intelligence artificielle (IA) et les mégadonnées. Cette transformation a permis de renforcer l'efficacité, la traçabilité, la flexibilité et l'interconnexion des chaînes de valeur, ouvrant la voie à des usines dites "intelligentes" [5].

Toutefois, cette orientation, centrée presque exclusivement sur la productivité et l'autonomie des systèmes, a révélé plusieurs limites. D'une part, la dépendance accrue aux technologies numériques a accentué les enjeux de cybersécurité et de fiabilité des infrastructures [6]. D'autre part, les impacts environnementaux (consommation énergétique, rareté des ressources) et sociaux (risques de déqualification des opérateurs, déshumanisation des environnements de travail) sont devenus de plus en plus préoccupants. Enfin, la crise sanitaire liée au COVID-19 a mis en évidence la vulnérabilité des chaînes mondiales de production et la nécessité de systèmes plus résilients et adaptatifs.

C'est dans ce contexte que s'inscrit la transition vers l'industrie 5.0. Contrairement à l'industrie 4.0, elle ne se limite pas à l'automatisation, mais adopte une vision durable et centrée sur l'humain de la production industrielle. L'humain redevient un acteur central, non pas comme simple surveillant de machines, mais comme collaborateur capable de tirer parti des technologies pour accroître sa créativité et sa capacité de décision. L'industrie 5.0 met ainsi l'accent sur la collaboration humain–machine (par exemple via la robotique collaborative et l'IA explicable), la personnalisation des produits et services, et l'intégration de dimensions éthiques, sociales et environnementales dans les systèmes productifs [6].

En somme, l'industrie 5.0 ne constitue pas une rupture technologique radicale, mais plutôt une évolution des principes de l'industrie 4.0 vers une approche plus équilibrée, qui articule la performance économique, la durabilité écologique et le bien-être humain. Elle ouvre ainsi de

nouvelles perspectives de recherche sur l'interaction humain–machine, la conception de systèmes résilients et responsables, ainsi que sur l'intégration de modèles de production orientés vers la société et l'environnement.

1.2.2 Le rôle de l'humain dans l'industrie 5.0

L'industrie 5.0 place l'humain au centre du système de production en le considérant comme un collaborateur actif des machines intelligentes. Contrairement à l'approche précédente, largement axée sur l'autonomie des systèmes et la maximisation de la productivité, ce nouveau paradigme valorise la complémentarité entre les capacités humaines telles que la créativité, l'adaptabilité, l'expertise tacite et la puissance des technologies numériques [7]. L'objectif n'est donc pas de remplacer l'opérateur, mais de mettre à sa disposition des technologies qui renforcent ses compétences, valorisent son rôle dans le processus industriel et assurent son bien-être ainsi que sa sécurité.

Cette vision s'accompagne d'un besoin croissant en interfaces intuitives, multimodales et adaptatives. La réalité augmentée, le contrôle gestuel ou vocal, les exosquelettes, les dispositifs portables et les assistants cognitifs représentent autant d'outils qui facilitent l'interaction humain–machine, tout en réduisant la charge physique et cognitive [8]-[10]. Parallèlement, des approches telles que l'IA explicable et les jumeaux numériques visent à rendre les décisions automatisées plus transparentes et compréhensibles, renforçant ainsi la confiance des opérateurs.

Un enjeu central réside dans la capacité à comprendre et modéliser les comportements humains afin d'assurer une collaboration fluide et sécurisée. L'analyse du mouvement humain, incluant la capture, la modélisation et l'interprétation des gestes et postures, devient alors essentielle pour adapter les systèmes en temps réel aux besoins des opérateurs, prévenir les risques liés aux postures contraignantes et anticiper leurs intentions [11]. Ainsi, l'industrie 5.0 redéfinit le rôle de l'opérateur : d'un simple surveillant de processus automatisés, il devient un partenaire clé dont l'expertise est augmentée par la technologie. Ce repositionnement ouvre de nouvelles perspectives mais soulève aussi des défis majeurs : concevoir des environnements de travail véritablement centrés sur l'humain, respectueux de ses capacités cognitives et physiques, tout en intégrant des considérations sociales et éthiques. Dans cette perspective, l'analyse du mouvement humain s'impose comme un levier essentiel pour comprendre, modéliser et améliorer l'interaction entre l'opérateur et les systèmes industriels.

1.2.3 Analyse du mouvement humain : Définitions et périmètre

Dans l'industrie 5.0, où l'humain occupe une place centrale, la capacité à comprendre, modéliser et interpréter ses mouvements est cruciale pour garantir des interactions sûres, efficaces et personnalisées avec les systèmes automatisés. L'analyse du mouvement humain devient ainsi un outil fondamental pour soutenir à la fois la performance industrielle et le bien-être des opérateurs.

L'analyse du mouvement humain (*Human Motion Analysis*, HMA) regroupe un ensemble de techniques visant à détecter, suivre, reconnaître et comprendre les gestes et actions d'un individu à partir de données multimodales telles que des vidéos, des capteurs inertiels, des signaux musculaires ou des représentations squelettiques [12]. Elle couvre un large spectre de tâches allant de la reconnaissance de gestes simples (mouvements de la main, posture) à l'identification d'actions complexes impliquant des interactions entre l'humain, des objets et l'environnement. Dans les contextes industriels, cette discipline permet de modéliser les comportements de l'opérateur en vue de les exploiter à des fins d'assistance, d'analyse ergonomique ou de supervision des processus [11]. Les différentes applications de l'analyse du mouvement humain seront détaillées dans la section 1.2.4, afin d'illustrer concrètement son rôle et son potentiel dans des contextes variés de l'industrie.

1.2.4 Applications de l'analyse du mouvement humain en industrie

L'industrie 4.0 et, plus récemment, l'industrie 5.0 soulignent la nécessité pour les systèmes industriels d'intégrer des notions de résilience, de durabilité et d'approche centrée sur l'humain, en accordant une priorité à la sécurité et au bien-être des individus [1]. Pour atteindre ces objectifs, l'analyse du mouvement humain revêt une importance capitale et occupe une place centrale dans le domaine industriel [13], [14].

Elle permet, par exemple, de surveiller en continu la conformité des procédures d'assemblage, en s'assurant que chaque étape est correctement réalisée et conforme aux normes de qualité, ce qui réduit le risque d'erreurs ou de défauts [15]. Un autre apport majeur concerne l'assistance dynamique fournie par les systèmes intelligents et les robots collaboratifs. En interprétant en temps réel les actions de l'opérateur, ces dispositifs peuvent ajuster leur comportement afin d'apporter un soutien ciblé : par exemple en adaptant la vitesse, la trajectoire ou la force appliquée lors d'une tâche conjointe [11], [16]. Enfin, l'analyse du mouvement peut

fournir un retour ergonomique immédiat, en alertant l'opérateur lorsqu'une posture contraignante est détectée et en proposant des ajustements pour prévenir la fatigue et améliorer le confort [17]. Ces travaux trouvent des applications concrètes dans divers secteurs, notamment la robotique et l'automatisation, la formation, la production, la logistique ou encore l'aéronautique [18]-[20].

Afin d'identifier, caractériser et structurer ces différentes applications, nous avons mené une étude systématique de la littérature scientifique récente, présentée dans notre article intitulé "Online Human Motion Analysis in Industrial Context : A Review [11]".

1.2.4.1 Méthodologie de l'étude systématique de la littérature

Cette étude repose sur une méthodologie de revue systématique visant à recenser et analyser les travaux consacrés à l'analyse du mouvement humain en contexte industriel.

1.2.4.1.1 Définition des mots-clés et requête de recherche. La première étape a consisté à définir un ensemble de mots-clés représentatifs des axes principaux de la recherche. Ces mots-clés ont été regroupés en quatre catégories :

- **Industrie** : *industr**, *manufactur**, *robot* ;
- **Humain** : *human*, *operator* ;
- **Analyse du mouvement** : *"action recognition"*, *"gesture recognition"*, *"motion analysis"*, *"movement recognition"*;
- **Modalités de données** : *skeleton*, *video*, *RGB*, *depth*, *camera*, *IMU*, *MoCap.*

Ces termes ont été combinés dans la requête suivante, appliquée aux bases de données **Scopus** et **Scinapse** :

```
(industr* OR manufactur* OR robot) AND (human OR operator) AND ("ac-
tion recognition" OR "gesture recognition" OR "motion analysis" OR "mo-
vement recognition") AND (skeleton OR video OR RGB OR depth OR camera
OR MoCap OR IMU)
```

1.2.4.1.2 Critères de sélection. Les publications ont été retenues selon les critères suivants :

- étude portant sur l'**analyse du mouvement humain**, incluant la reconnaissance d'actions et de gestes ainsi que les interactions opérateur–environnement ;

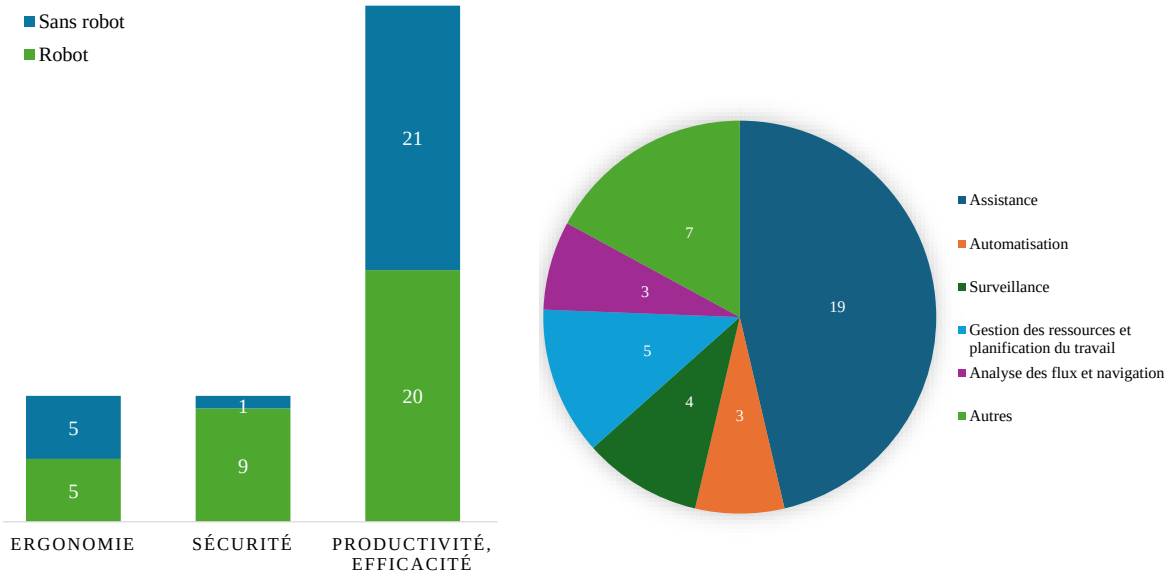
— application explicite dans un **contexte industriel** (exclusion des domaines du sport, de la santé ou de la surveillance) ;
— mention explicite des **modalités de données** utilisées (RGB, profondeur, squelette, MoCap, IMU, etc.) ;
— utilisation de **méthodes d'apprentissage profond** ;
— publications rédigées en anglais, parues entre **2018 et mai 2023**, dans des revues ou conférences scientifiques relevant de l'ingénierie ou de l'informatique.

1.2.4.1.3 Processus de filtrage. Le processus de sélection s'est déroulé en cinq phases successives :

1. **Recherche initiale** : 706 articles identifiés à partir de la requête.
2. **Élimination des doublons** : 658 articles uniques conservés.
3. **Filtrage préliminaire** : examen des titres, des mots-clés et des résumés selon les critères d'inclusion (270 articles retenus).
4. **Analyse approfondie** : lecture détaillée du contenu afin de vérifier la pertinence scientifique et thématique (114 articles conservés).
5. **Extension par citations** : ajout d'articles pertinents identifiés à partir des références et citations, portant le total final à **140 publications**.

1.2.4.1.4 Classification des travaux. Sur les 140 publications sélectionnées après l'ensemble du processus de filtrage, seules celles proposant un cas d'usage concret ou une implémentation appliquée ont été retenues pour l'analyse statistique présentée dans la figure 1.1. Les travaux de type revue, article théorique, cadres méthodologiques généraux ou ne présentant aucun scénario expérimental n'ont pas été intégrés dans cette figure afin de garantir une classification strictement fondée sur des applications effectives. Les articles retenus ont ensuite été classés selon trois grandes catégories d'applications : Érgonomie, sécurité, productivité et efficacité.

Dans les sous-sections suivantes, nous présentons plus en détails les principales catégories d'applications identifiées dans la littérature, à savoir la sécurité, l'ergonomie et la productivité/efficacité.



(a) (b)

FIGURE 1.1 – Applications de l'analyse du mouvement humain en industrie. (a) Principales applications considérant la collaboration homme-robot. (b) Sous-domaines liés à la productivité et à l'efficacité.

1.2.4.2 Sécurité

Garantir la sécurité des opérateurs dans les environnements industriels est un enjeu crucial, car les accidents surviennent fréquemment lorsque les machines opèrent à proximité de parties sensibles du corps comme la tête ou les bras. Pour limiter ces risques, plusieurs approches ont été appliquées aux robots, telles que la détection de collisions et la planification réactive des mouvements, communément appelées *speed and separation monitoring (SSM)* [21] et *power and force limitation (PFL)* [22]. La majorité des travaux (9 sur 10) portent sur des cas impliquant l'utilisation de robots, tandis qu'une seule étude [23] analyse les mouvements d'opérateurs dans une station de travail sans robots. La plupart des recherches dans ce domaine sont référencées dans le contexte de l'industrie 4.0.

Ces méthodes peuvent être regroupées en deux catégories. La première consiste à placer des capteurs directement sur les robots, appelés capteurs de présence (non visuels), complétés par des caméras pour surveiller l'activité humaine. Par exemple, ABU AL-HAIJA et AL-SARAIREH [24] ont proposé un modèle autonome de détection de contact homme-robot afin de prévenir les collisions dangereuses lors de tâches collaboratives. De même, KWON, JIN et LEE [25] ont présenté une méthode permettant de déterminer la localisation et la survenue d'une

collision afin d'éviter des impacts abrupts et imprévisibles. Birjandi, Kühn et Haddadin [26] ont introduit un système de détection basé sur la fusion de capteurs, améliorant la précision tout en réduisant l'incertitude de détection. Enfin, Zhang, Liu, Chang et al. [27] ont proposé un modèle de prévention des collisions capable de prédire la trajectoire future de l'opérateur lors d'un assemblage de moteur, permettant ainsi une planification et une exécution robotique en temps réel.

La seconde catégorie repose sur la pose de capteurs sur l'opérateur, combinés à des caméras supplémentaires, afin de collecter les données nécessaires aux algorithmes de détection de collisions. Par exemple, Grushko, Vysockỳ, Heczko et al. [28], dans le but de prévenir les collisions entre le robot et l'opérateur, ont choisi de prédire la trajectoire future du robot. Pour cela, ils ont utilisé un système de notification vibrante qui s'activait lorsque la main de l'opérateur s'approchait de cette trajectoire anticipée, avertissant ainsi l'opérateur du risque de collision. Pastor, Lin-Yang, Gómez-de-Gabriel et al. [29] ont proposé une base de données haptique permettant d'estimer précisément la localisation de la prise lors de la saisie d'un membre humain, avec des applications possibles en robotique d'assistance et de sauvetage. Enfin, Delamare, Duval et Boutteau [23] ont introduit un nouveau jeu de données pour le suivi des opérateurs, offrant des informations plus fines sur leurs mouvements afin d'améliorer les conditions de sécurité en industrie.

1.2.4.3 Ergonomie

Dans les environnements où les opérateurs jouent un rôle central, le bien-être des travailleurs est une priorité. Pour y répondre, l'analyse du mouvement humain peut être utilisée afin d'estimer des indices ergonomiques permettant de suggérer des postures optimales et de corriger les postures nocives pouvant conduire à des troubles musculosquelettiques (TMS) [30]. L'HMA contribue également à l'automatisation intelligente, en réduisant les tâches répétitives, longues et physiquement contraignantes, afin de préserver la santé mentale et physique des opérateurs [6].

De nombreuses études s'appuient sur des indices ergonomiques bien établis, tels que le *Rapid Upper Limb Assessment (RULA)*, l'indice le plus couramment utilisé [31], qui évalue posture, force et mouvements liés aux tâches sédentaires [32]. Le *Rapid Entire Body Assessment (REBA)* est similaire dans son principe, mais s'applique au corps entier (nuque, mains, épaules, jambes, torsion des membres) [31]. L'*Ergonomic Assessment WorkSheet (EAWS)* est

une méthode plus complète prenant en compte la force de travail, la fréquence et la répétition des gestes [31]. Le *Ovako Working posture Analyzing System (OWAS)* identifie et évalue les postures contraignantes et détermine la nécessité de mesures correctives [33]. Enfin, l'indice du *National Institute for Occupational Safety and Health (NIOSH)* vise à déterminer la charge maximale pouvant être manipulée manuellement lors d'une activité [31]. D'autres scores ergonomiques existent également, tels que l'index *OCRA (Occupational Repetitive Actions)*, le *Quick Exposure Check (QEC)* ou encore le *Strain Index (SI)*, qui visent respectivement l'évaluation des gestes répétitifs, l'exposition globale du corps et les contraintes sur les mains et les poignets. Toutefois, comme le souligne la littérature, ces approches restent moins représentées dans les publications de l'état de l'art que les méthodes classiques telles que RULA ou REBA [34].

Ces scores ergonomiques ont été utilisés dans plusieurs travaux. Certains incluent la présence de robots [35], [36], tandis que d'autres portent uniquement sur l'ergonomie humaine [17], [37], [38]. La majorité de ces études se concentrent sur le contexte de l'industrie 4.0, seules quelques-unes étendant leur analyse à l'industrie 5.0. Sedlar, Stepanova, Skoviera et al. [38] ont, par exemple, introduit un jeu de données pour estimer la pose d'outils portables lors de tâches industrielles. Cette base a permis de développer des méthodes robustes à l'occlusion et a démontré une amélioration notable de la précision de pose. De leur côté, Havard, Jeanne, Lacomblez et al. [35] ont étudié l'ergonomie d'un poste manuel en fonction de la position d'un bras robotisé utilisé comme "troisième bras" pour maintenir un produit en assemblage. Leur dispositif a permis d'identifier les zones d'assemblage générant des postures fortement sollicitantes et de proposer une configuration du robot réduisant sensiblement la fatigue de l'opérateur. Kim, Peternel, Lorenzini et al. [36] ont analysé l'usage d'un bras robotisé pour soutenir un objet à polir, ce qui a permis d'estimer des paramètres ergonomiques tels que le couple articulaire, la fatigue musculaire et la manipulabilité, afin d'identifier les risques et de proposer une meilleure configuration posturale. Manghisi, Evangelista et Uva [37], quant à eux, ont développé *ErgoVR*, un outil de visualisation 3D pour évaluer en temps réel la posture des opérateurs et détecter rapidement les gestes à risque. Enfin, Bortolini, Faccio, Gamberi et al. [17] ont conçu un système d'analyse du mouvement (MAS) permettant de mesurer les mouvements squelettiques et d'en déduire les indices ergonomiques mentionnés, ce qui a permis d'identifier des zones à risque sur des chaînes d'assemblage réelles.

1.2.4.4 Productivité et efficacité

L'optimisation de la productivité et de l'efficacité demeure un objectif majeur dans les secteurs industriels. Pour y parvenir, l'analyse du mouvement humain constitue un outil clé. Identifier et comprendre les mouvements dans un environnement industriel est toutefois complexe, en particulier lorsque les tâches sont réparties entre plusieurs acteurs, comme dans les scénarios de collaboration homme–robot. Dans ce contexte, le robot agit comme un facilitateur, en assistant l'opérateur dans des tâches telles que le transfert d'outils ou la manipulation de charges lourdes [39]. Pour que cette collaboration soit efficace, le robot doit être capable d'anticiper les intentions humaines et d'adapter son comportement en conséquence [40]. Cette problématique, déjà présente dans l'industrie 4.0, est encore plus centrale dans l'industrie 5.0, où ces capacités deviennent un objectif prioritaire [20].

Les applications de l'HMA en matière de productivité et d'efficacité se déclinent en plusieurs sous-domaines, comme indiqué ci-dessous :

— **Assistance** : améliore la productivité en facilitant le transfert de pièces, d'objets et d'outils lors des processus d'assemblage. Elle permet également d'aider les opérateurs à atteindre des résultats plus rapides et de meilleure qualité. De nombreux travaux se concentrent sur l'assistance aux opérateurs et la transmission d'objets. Par exemple, Castro, Silva et Santos [41] et Matheson, Minto, Zampieri et al. [42] ont mené des analyses approfondies sur la remise d'objets et l'assistance robotique. Des robots collaboratifs ont ainsi été utilisés pour manipuler des outils (tournevis, pièces d'assemblage, etc.), livrer des composants ou encore ajuster l'orientation d'objets tenus par l'opérateur [43]-[45]. D'autres études se sont concentrées sur l'assistance physique (maintien ou manipulation d'objets) lors de tâches industrielles telles que l'assemblage d'un moteur, la soudure, ou l'aide à des opérations sur le poste de travail [3], [20], [46]-[48]. Enfin, certains travaux [49], [50] ont porté sur la prédiction des mouvements humains pour améliorer l'assistance, en intégrant la reconnaissance d'intention et d'objets [51].

— **Surveillance (Monitoring)** : permet d'accroître la productivité en détectant les erreurs lors des tâches d'assemblage et en guidant les opérateurs vers une exécution plus rapide et plus fiable. Rana, Rahman et Smith [14] ont proposé un nouveau jeu de données dédié à la surveillance d'activités présentant des écarts par rapport aux standards

établis dans les environnements industriels, afin d'optimiser le flux de production lors de l'assemblage de tricycles. Zamora, Mauricio-Andrés, Castro-Vargas et al. [52] ont quant à eux étudié l'assemblage manuel de skateboards pour identifier et corriger les erreurs des opérateurs. Enfin, Chen, Wang, Li et al. [53] ont proposé une méthode permettant d'identifier les actions répétitives et d'estimer leur durée, afin de limiter les problèmes de qualité liés aux étapes manquantes ou aux irrégularités dans l'exécution.

— **Gestion des ressources et planification du travail** : concerne l'allocation efficace des ressources disponibles (temps, espace, main-d'œuvre) afin d'optimiser la productivité. Bortolini, Faccio, Gamberi et al. [17] ont proposé un système d'analyse du mouvement (MAS) capable de distinguer les activités à valeur ajoutée et celles sans valeur ajoutée. Hernandez, Valarezo, Cobos et al. [54] ont introduit une méthode hiérarchique de reconnaissance d'actions pour évaluer automatiquement les indicateurs de productivité (disponibilité, performance, efficacité du travail) dans des tâches de conditionnement et de logistique. Enfin, Wang, Qin, Yan et al. [55] ont proposé une approche basée sur des capteurs de vision pour réduire le temps d'exécution et améliorer la qualité.

— **Analyse des flux et navigation** : vise à étudier et simuler les activités récurrentes dans un environnement industriel, dans l'objectif d'optimiser le déplacement des robots. Schreiter, Almeida, Zhu et al. [56] ont simulé des activités typiques en environnement industriel, dans la continuité des travaux de Rudenko, Kucner, Swaminathan et al. [57], où des véhicules à guidage automatique (AGV) étaient utilisés comme obstacles mobiles. De plus, Li, Fan, Zheng et al. [58] ont conçu un bras robotisé capable de saisir, livrer et maintenir des objets (par exemple une boîte à outils) tout en se déplaçant aux côtés de l'opérateur.

— **Automatisation des plans d'assemblage** : consiste à générer automatiquement des séquences d'assemblage à partir de la reconnaissance d'actions. Un exemple marquant est l'IKEA ASM dataset, publié par Upadhyay, Barua, Dubey et al. [59], permettant la génération automatique de plans pour l'assemblage de meubles.

Les études restantes se sont principalement concentrées sur l'optimisation de solutions dédiées à la segmentation, à la reconnaissance et à l'anticipation d'actions complexes. Ces approches optimisées présentent un fort potentiel d'application dans divers domaines in-

dustriels et permettent d'améliorer de manière significative la productivité et l'efficacité des processus. Elles ouvrent également la voie à des systèmes plus robustes et adaptatifs, capables de s'intégrer à des environnements de production réels et dynamiques [60]-[62].

1.2.5 Synthèse

En résumé, la transition de l'industrie 4.0 vers l'industrie 5.0 marque un changement de paradigme majeur : alors que la première était centrée sur l'automatisation et la productivité, la seconde met l'accent sur une approche résiliente, durable et profondément centrée sur l'humain. Dans ce contexte, l'analyse du mouvement humain (HMA) apparaît comme un levier essentiel pour comprendre, modéliser et anticiper les actions des opérateurs, favorisant ainsi des environnements de travail plus sûrs, ergonomiques et efficaces.

L'examen de la littérature révèle trois applications clés de l'HMA en contexte industriel : la **sécurité**, où la détection du contact physique et la prévention des collisions sont cruciales pour protéger les opérateurs dans des environnements potentiellement dangereux ; l'**ergonomie**, qui vise à évaluer et corriger les postures de travail afin d'améliorer le confort et de réduire les risques de troubles musculosquelettiques ; et enfin la **productivité et l'efficacité**, obtenues grâce à la surveillance continue, à l'évaluation des performances et à l'anticipation des mouvements, favorisant une collaboration homme–robot plus fluide et une optimisation des processus industriels.

La majorité des études analysées s'inscrivent dans le cadre de l'industrie 4.0, seules quelques-unes étendent leur portée à l'industrie 5.0. Parmi elles, une nette préférence se dessine pour les applications liées à la productivité et à l'efficacité : pas moins de 41 travaux identifiés portent sur ces aspects. Toutefois, l'accent croissant mis sur la sécurité et l'ergonomie reflète une évolution significative des pratiques industrielles. Si l'industrie 4.0 reconnaissait déjà ces dimensions, l'industrie 5.0 en ferait désormais des piliers fondamentaux, en affirmant une vision équilibrée entre performance, bien-être et durabilité.

Après avoir présenté les différents domaines d'application de l'analyse du mouvement humain en milieu industriel, il convient désormais de s'intéresser aux moyens technologiques permettant sa mise en œuvre. La section suivante est consacrée aux technologies de capture de mouvement, essentielles à l'acquisition et à la modélisation du geste humain.

1.3 Technologies de capture de mouvement

Au cours des dernières années, une multitude de technologies ont été développées pour l'acquisition et l'analyse des mouvements des opérateurs dans un environnement industriel [13]. Parmi celles-ci, les technologies de **capture de mouvement** (MoCap) sont largement utilisées pour l'extraction du squelette humain. En parallèle, certaines études s'appuient sur l'activité musculaire, en particulier au niveau des membres supérieurs, grâce à des capteurs de signaux physiologiques tels que l'**électromyographie** (EMG) [18], [36]. Nous présentons ci-dessous un panorama complet des technologies utilisées pour capturer les mouvements des opérateurs industriels. Enfin, nous détaillons les limites propres à chaque modalité, ainsi que les combinaisons hybrides qui permettent de pallier les faiblesses individuelles dans un environnement industriel.

1.3.1 Technologies de capture de mouvement (MoCap)

Les systèmes de *Motion Capture* (MoCap) constituent la principale source de données pour la modélisation du squelette humain et l'analyse des gestes en milieu industriel. On distingue généralement deux grandes catégories de systèmes MoCap : les systèmes **optiques** et les systèmes **non optiques**. La figure 1.2 présente la taxonomie détaillée des systèmes de MoCap, tandis que le tableau 1.1 en propose une étude comparative détaillant leurs caractéristiques, avantages et limites.

Ces ressources visuelles permettent de mieux appréhender les avantages et les limites de chaque solution technologique et facilitent la prise de décision en choisissant la technologie la mieux adaptée aux contraintes d'un cas d'usage donné.

1.3.1.1 Systèmes optiques

Les capteurs optiques utilisent des caméras pour mesurer indirectement les mouvements corporels à partir d'images. On distingue deux types principaux :

— **Capteurs avec marqueurs :** nécessitent des marqueurs visibles placés sur le corps. Par exemple, Niemann, Reining, Moya Rueda et al. [63] ont utilisé 39 marqueurs sur une combinaison optique MoCap pour l'analyse de tâches logistiques. Ces systèmes offrent une très haute précision, mais sont longs à mettre en place et peu pratiques en industrie, notamment en raison des réflexions qui surviennent sur les objets métalliques

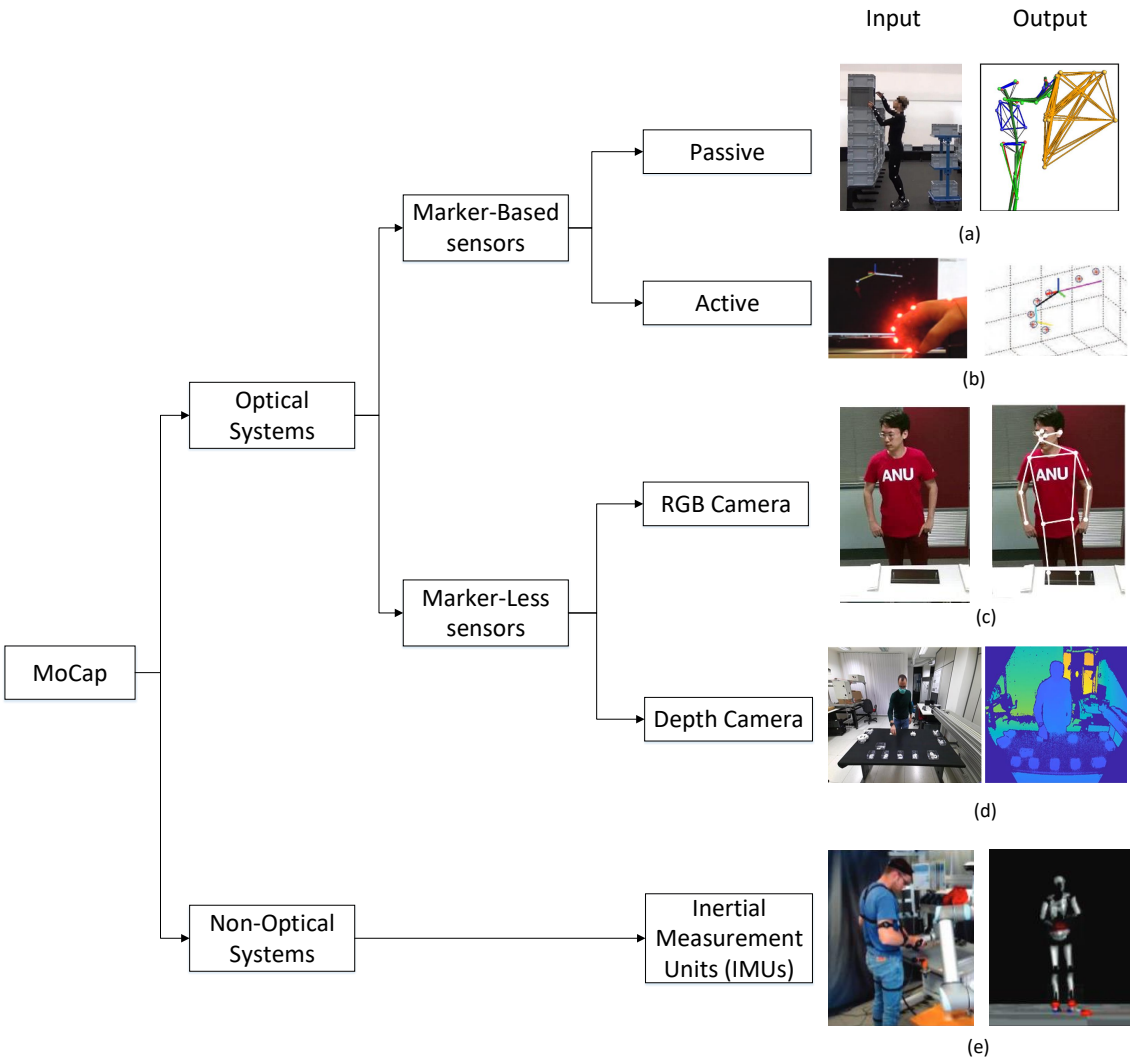


FIGURE 1.2 – Technologies MoCap : (a) combinaison optique [63], (b) marqueurs actifs [64], (c) Kinect V2 [4], (d) Kinect V1 [65], (e) Perception Neuron 32 [3].

souvent présents dans l'environnement industriel [66].

— **Capteurs sans marqueurs :** utilisent des caméras RGB-D comme le Kinect V2 [4] ou Realsense D435i [67] capables de fournir à la fois des images en couleur (RGB) et des informations de profondeur (Depth). Ces capteurs permettent ainsi d'estimer la posture humaine à partir de données 3D, notamment sous forme de nuages de points ou de reconstructions squelettiques (voir figure 1.2(c)). Ils sont faciles à déployer en environnement industriel, mais restent sensibles aux conditions d'éclairage, aux occlusions et à la qualité de la scène. [66].

1.3.1.2 Systèmes non optiques (Unités de mesure inertielle - IMUs)

Les IMUs, légers et portables, permettent de mesurer les mouvements corporels grâce à des accéléromètres et des gyroscopes. Ils sont utilisés notamment dans Al-Amin, Tao, Doell et al. [18], Dallel, Havard, Baudry et al. [3] et Maurice, Malaisé, Amiot et al. [30] ; ils permettent de reconstruire un modèle 3D du squelette humain. Leur principal inconvénient est la dérive des mesures sur de longues durées [68].

Tableau 1.1 – Comparaison des systèmes de capture de mouvement pour l'industrie [13]

Système	Type	Précision	Installation	Coût	Portabilité	Limites	Cas d'usage
Optique	Avec marqueurs	Très élevée (0.1 mm, 0.5°), dépend du nombre et de la position des caméras	Longue et coûteuse	$\geq$ \$5000	Non	Calibrations fréquentes, obstructions	Reconnaissance d'activités [63], [66], [69]
Optique	Sans marqueurs	Faible (statique, 0.0348 m), dépend de la distance à la caméra	Rapide	$\approx$ \$200	Oui	Sensible à l'environnement, variations lumineuses	Segmentation d'actions [4], [67], [70]
Non optique	IMUs	Élevée (0.75°–1.5°), dépend du nombre de capteurs IMU utilisés	Simple	$\approx$ \$50–\$12000	Oui	Dérive, bruit, désalignement	Reconnaissance d'actions [3], [18], [71]

1.3.2 Capteurs basés sur les signaux musculaires

Les capteurs basés sur les signaux musculaires sont couramment utilisés dans les domaines de la santé et du sport, et, plus récemment, dans le secteur industriel. Nous présentons ici deux technologies principales : l'**électromyographie de surface (sEMG)** et la **myographie de force (FMG)**, illustrées en figure 1.3, tandis que le tableau 1.2 présente une étude comparative détaillée de leurs caractéristiques, avantages et limites respectifs.

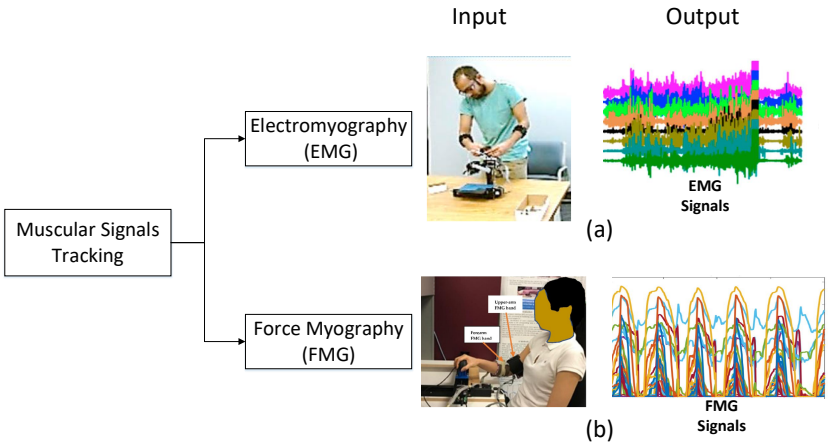


FIGURE 1.3 – Capteurs de suivi musculaire. Exemples de données acquises via : (a) bracelets Myo [18], (b) bracelets FMG [72].

TABLEAU 1.2 – Comparaison des systèmes de détection du signal musculaire pour la capture du mouvement humain en industrie.

Système	Installation	Coût	Portabilité	Limites	Cas d'usage
EMG	Alignement précis des électrodes	$1900–$2400	Oui	Bruit, artefacts	Reconnaissance d'actions [18], estimation de la fatigue musculaire [36]
FMG	Port de bande calibrée	$8–$144	Oui	Glissement du capteur	Reconnaissance de postures [73], reconnaissance de gestes de la main [74]

1.3.2.1 Electromyographie (sEMG)

Cette technique permet l'enregistrement multi-points de l'activité à la surface des muscles en temps réel. Elle offre une analyse fine de l'activité musculaire, très utile en ergonomie, en analyse biomécanique et en reconnaissance de gestes [75]. Des travaux antérieurs ont exploité des capteurs sEMG : AL-AMIN, TAO, DOELL et al. [18] ont positionné ces capteurs sur les bras des opérateurs afin de suivre leur activité musculaire et de reconnaître les actions fines réalisées lors des tâches d'assemblage, tandis que KIM, PETERNEL, LORENZINI et al. [36] s'en sont servis pour estimer la fatigue musculaire. L'EMG mesure directement l'activation musculaire précédant le mouvement [76], mais ces signaux sont sensibles au bruit, aux artefacts, à la sueur et aux mouvements parasites [77].

1.3.2.2 Myographie de force (FMG)

Cette technologie non invasive mesure les contractions musculaires à l'aide de capteurs de pression intégrés à des bracelets portables [72], [78]. Elle permet d'estimer la force musculaire appliquée lors d'une interaction homme-robot [76]. La FMG est plus accessible que l'EMG : moins coûteuse, plus portable et généralement plus simple à déployer [79]. Zakia et Menon [72] ont utilisé cette technologie pour estimer la force interactive lors de la manipulation d'un robot linéaire.

1.3.2.3 Combinaison EMG et FMG

Des solutions hybrides ont été proposées pour tirer parti des points forts des deux modalités. Ke, Huang, Chen et al. [80] ont conçu un capteur EMG-FMG modulaire pour améliorer la reconnaissance de gestes fins. Sun, Ke, Rahmani et al. [81] recommandent la fusion de capteurs pour améliorer la précision de l'acquisition du mouvement humain.

1.3.3 Technologies hybrides

Certaines approches récentes ont proposé de combiner plusieurs modalités pour tirer parti des avantages respectifs des capteurs optiques, inertiels et physiologiques. Ces systèmes dits **hybrides** permettent une meilleure couverture spatiale, une robustesse accrue face aux artefacts de données, ainsi qu'une interprétation plus fine de l'activité humaine.

1.3.3.1 Combinaisons IMU + capteurs optiques

Dallel, Havard, Baudry et al. [3] ont utilisé un système à base d'IMUs (Perception Neuron 32) pour capturer la donnée squelettique, tout en filmant simultanément l'activité avec trois caméras RGB placées selon des vues différentes (côté gauche, côté droit, dessus). Cette combinaison permet une analyse robuste des postures, tout en facilitant les annotations manuelles.

Maurice, Malaisé, Amiot et al. [30] ont combiné les capteurs MoCap Qualisys avec des IMUs ainsi qu'un gant électronique (e-glove) pour surveiller la flexion des doigts. Des caméras RGB complétaient le dispositif pour une capture visuelle de la scène de travail.

Yoshimura, Morales, Maekawa et al. [82] ont utilisé des unités IMU pour capter les accélérations, gyroscopes et quaternions, placées sur les poignets et les bras. En complé-

ment, ils ont installé des caméras RGB (vue supérieure), un capteur Kinect et un capteur LiDAR (vue frontale), permettant ainsi une capture précise du squelette et de la position de l'opérateur, même à distance du poste.

1.3.3.2 Combinaisons MoCap + EMG/FMG

Comme indiqué précédemment, les capteurs IMU et optiques permettent une reconstruction du squelette et une analyse de la posture. Toutefois, ils ne permettent pas de déterminer la charge soulevée ou l'effort musculaire, ce qui limite leur capacité à discriminer certaines actions.

C'est pourquoi plusieurs travaux intègrent des capteurs physiologiques, comme l'EMG ou le FMG, en complément de la capture de mouvement. Par exemple, Kim, Peternel, Lorenzini et al. [36] ont combiné caméras RGB, capteurs EMG et IMUs pour estimer la fatigue musculaire et améliorer la reconnaissance d'action. De leur côté, Al-Amin, Tao, Doell et al. [18] ont utilisé des brassards Myo, qui combinent EMG et IMU dans un même dispositif, pour reconnaître des gestes fins des doigts.

1.3.4 Synthèse

Selon les objectifs d'analyse, le degré de précision recherché et les contraintes opérationnelles (coût, déploiement, perturbation de l'activité), différentes combinaisons peuvent être envisagées. Le recours à des systèmes hybrides, comme présenté dans le tableau 1.3, devient une option de plus en plus pertinente en contexte industriel, notamment pour les applications liées à l'évaluation ergonomique ou à l'interaction humain-robot.

Les technologies de capture présentées précédemment permettent d'acquérir des données essentielles à l'analyse du mouvement humain. Ces données constituent la base de la création de jeux de données spécifiques, dont le contenu et la structure varient selon le cas d'usage industriel visé. Au regard des différentes technologies d'acquisition de mouvement discutées précédemment et des compromis qu'elles impliquent en termes de précision et de coût et de déploiement, nous privilégions dans cette thèse la modalité squelettique comme représentation principale du mouvement. Celle-ci peut être obtenue soit par extraction à partir de cameras RGB, soit directement via des dispositifs dédiés de capture de mouvement (IMU). Ce choix permet de conserver une description fine de la dynamique corporelle tout en maintenant une relative indépendance vis-à-vis des configurations matérielles. La section suivante présente les

principaux jeux de données existants utilisés pour l'analyse du mouvement humain dans le contexte industriel.

TABLEAU 1.3 – Comparaison des technologies hybrides pour l'analyse du mouvement en industrie.

Combinaison de modalités	**Avantages**	**Limites**	**Applications**	**Références**
IMU + cameras RGB	Robustesse, annotations faciles	Calibration, désynchronisation	Assemblage, suivi posture	[3]
IMU + MoCap + e-glove	Haute précision multi-segment	Coût, encombrement	Ergonomie, santé au travail	[30]
IMU + LiDAR + RGB-D	Suivi précis à distance	Intégration complexe	Tâches logistiques	[82]
Myo armband (EMG + IMU)	Fusion capteurs compacts	Glissement capteur, bruit	Reconnaissance gestuelle fine	[18]
EMG + IMU + RGB	Suivi fatigue + posture	Intrusif, bruit EMG	Estimation de charge, fatigue	[36]

1.4 Jeux de données pour l'analyse du mouvement humain en industrie

1.4.1 Aperçu des jeux de données industriels

Les études récentes ont mis en évidence un manque historique de jeux de données publics consacrés à l'analyse du mouvement humain (HMA) dans les contextes industriels. Cependant, cette tendance commence à évoluer avec l'apparition de nouveaux jeux de données permettant d'entraîner et d'évaluer des méthodes de reconnaissance, de détection et de prédiction dans des environnements industriels réalistes.

Ces jeux de données diffèrent selon leurs objectifs, les modalités de données utilisées, les configurations de capteurs, les types d'interactions (par exemple, humain-objet ou humain-robot) et les styles d'annotation. Le tableau 1.4 présente un résumé comparatif des jeux de données publics les plus pertinents dans ce domaine, tandis que la figure 1.7 illustre quelques exemples de trames ainsi que les variations liées au suivi du mouvement humain, aux types d'interactions et aux points de vue de caméra.

Les jeux de données recensés ont été conçus pour répondre à divers objectifs opérationnels. Les sous-sections suivantes en présentent les principales catégories.

TABLEAU 1.4 – Jeux de données sélectionnés pour l'analyse du mouvement humain en contexte industriel. U : Séquences non segmentées (Untrimmed), T : Séquences segmentées (Trimmed), HO : Interaction Humain-Objet, HR : Interaction Humain-Robot, NI : Pas d'interaction, P : Représentation partielle (main, bras), Up : Haut du corps, F : Corps entier, E : Vue égocentrique, A : Vue allogénique (3e personne), NA : Non applicable, MNP : Modèle non précisé.

Dataset	Année	U	T	Sujets	Classes	Échantillons	Vues	Angles	Corps	Technologies	HO	HR	NI	Modalités	Public
IHAD_V1 [83]	2023	✓		-	12	-	1	A	P	UI-326xCP-C	✓			RGB + Skeleton 3D	
HRI30 [67]	2022		✓	11	30	2940	1	A	F	Realsense D435i	✓	✓	✓	RGB	✓
CoAx [48]	2022	✓		6	9	180	1	A	Up	Realsense D435	✓	✓		RGB, Depth, Pointcloud	
OpenPack [82]	2022	✓		16	10	20129	1	A	Up	Kinect, LiDAR	✓			Accéléro, Gyro, BVP, EDA, Profondeur	
InHARD-DT [84]	2022	✓		12	18	4799	3	A	F	HTC Vive Pro Eye, PN32	✓	✓		RGB + Skeleton 3D	✓
HA4M [65]	2022	✓		41	12	4124	1	A	F	Kinect V1	✓			RGB, Depth, IR	✓
Assembly101 [15]	2022	✓		53	1380	1013523	12	E, A	P	8 RGB + 4 mono cam	✓			RGB, Depth, 3D hand	✓
COVERED [85]	2022	✓		-	6	218	4	A	F	Ouster OS0-128 LiDAR			✓	Pointcloud	✓
CAARL [19]	2021	✓		2	7	70	-	A	F	MoCap optique	✓		✓	IR + Skeleton 3D	✓
WGD [69]	2021		✓	8	3	-	8	A	Up	MoCap optoélectronique	✓			RGB + Skeleton 3D	
Contact Detection [21]	2021	✓		-	5	2125	2	A	F	Kinect V2 + capteurs robot		✓	✓	RGB, Depth, Skeleton 3D	✓
LARa [63]	2020	✓		14	8	420	NA	A	F	MoCap optique, IMU	✓		✓	IR + Skeleton 3D	✓
InHARD [3]	2020	✓		16	14	4804	3	A	F	C920 + PN32	✓	✓		RGB + Skeleton 3D	✓
MECCANO [86]	2020		✓	20	61	8858	1	E	P	RealSense SR300	✓			RGB, Depth + Gaze	✓
IKEA ASM [4]	2020	✓		48	33	17577	3	A	F	Kinect V2	✓			RGB, Depth, Skeleton 3D	✓
AndyData-lab [30]	2019	✓		13	6	195	2	A	F	Xsens MVN, Qualisys, e-glove	✓			RGB + Skeleton 3D	✓

1.4.1.1 Sécurité et collaboration

Le jeu de données HRI30 [67] a été conçu pour améliorer la sécurité dans les scénarios d'interaction humain-robot (HRI). Il contient 3000 vidéos annotées couvrant des actions d'interaction humain-objet, des actions sans interaction et des actions collaboratives. Le jeu de données COVERED [85] exploite un système multi-LiDAR pour capturer des nuages de points 3D à haute résolution, ce qui permet la localisation d'humains, de robots et de véhicules autonomes (AGV). Le jeu Physical Human-Robot Contact Detection [21] vise à détecter les contacts physiques afin d'anticiper les collisions.

1.4.1.2 Érgonomie et évaluation des risques

Plusieurs jeux de données ont été développés pour estimer des indices ergonomiques. TAMANTINI, CORDELLA, LAURETTI et al. [69] ont proposé le Working Gesture Dataset (WGD), visant à prévenir les troubles musculosquelettiques (TMS) par l'analyse des gestes professionnels. De plus, MAURICE, MALAISÉ, AMIOT et al. [30] ont présenté le jeu de données AnyData-lab-onePerson, centré sur des activités industrielles où les participants adoptent différentes postures afin de réduire la charge sur l'opérateur et d'améliorer le confort au travail.

1.4.1.3 Suivi des tâches et assistance à l'opérateur

Plusieurs jeux de données visent à reconnaître des actions dans des scénarios industriels. InHARD [3] et sa version équivalente obtenue au sein du jumeau numérique InHARD-DT [84] sont construits autour d'un cas d'usage réel avec captation multi-vues et suivi squelettique. OpenPack [82] vise la reconnaissance de tâches de conditionnement dans des environnements logistiques intelligents. CoAx [48] propose des enregistrements d'interactions humain-robot pour anticiper les actions.

Pour les tâches d'assemblage fin, les jeux de données MECCANO [86] et Assembly101 [15] offrent des données multimodales de haute résolution, en vue frontale ou embarquée, permettant la segmentation d'actions, la détection d'erreurs ou l'évaluation des compétences.

1.4.1.4 Navigation et flux opérateurs

Pour prédire les trajectoires des opérateurs dans les zones de travail partagées, les jeux de données THOR [57], Magni [56] et Delamare et al. [23] utilisent des systèmes MoCap, LiDAR ou ultra-wideband.

1.4.1.5 Productivité et optimisation de processus

IKEA ASM [4] documente des séquences d'assemblage de meubles à l'aide de capteurs Kinect synchronisés. Le jeu de données HA4M [65] soutient la reconnaissance d'actions par assemblage manuel à partir de données multimodales (RGB, profondeur, IR, squelette). LARa [63] et CAARL [19] proposent un suivi contextuel des activités à l'aide de MoCap et d'IMUs.

Au total, plus de 16 jeux de données sont désormais disponibles pour l'analyse du mouvement humain dans un contexte industriel, chacun apportant des perspectives et modalités différentes, et favorisant le développement de solutions robustes et adaptées aux réalités du contexte industriel.

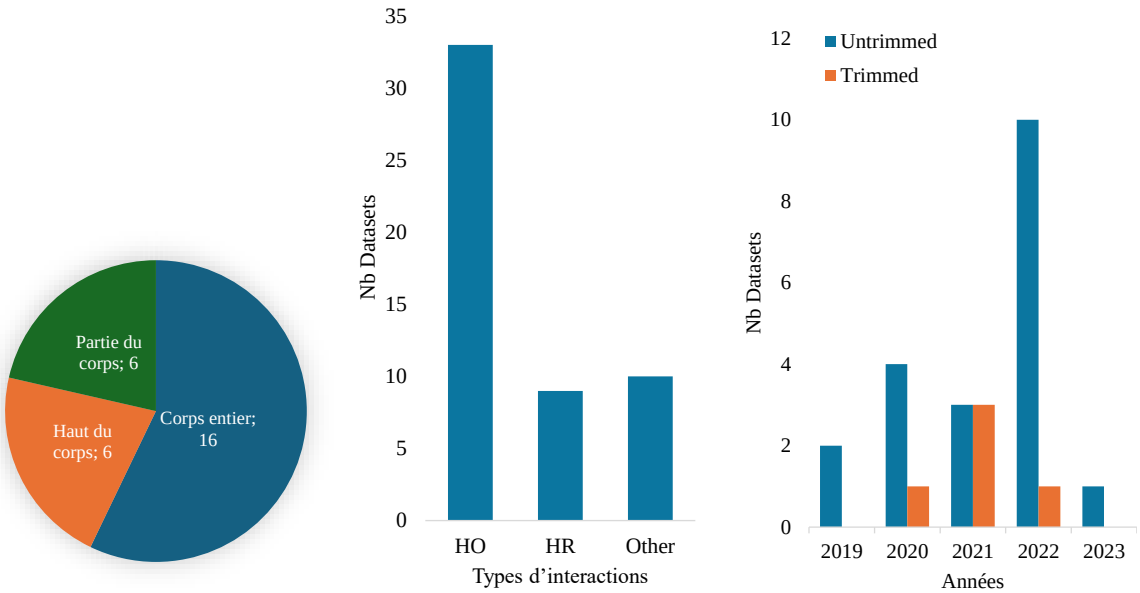
1.4.2 Caractéristiques des jeux de données

L'analyse approfondie des jeux de données présentés dans le tableau 1.4 permet de dégager plusieurs caractéristiques majeures, ainsi que des défis notables à résoudre, afin de concevoir des systèmes robustes d'analyse du mouvement humain en contexte industriel. La suite de

cette section présente les principaux défis identifiés dans l'état de l'art concernant l'analyse du mouvement humain dans un contexte industriel.

1.4.2.1 Données segmentées (trimmed) vs non segmentées

Les modèles d'apprentissage profond actuels, tels que les CNN et les RNN spatio-temporels [67], [87], ont montré des performances remarquables sur des jeux de données *segmentés* (trimmed), tels qu'ActivityNet [88] ou THUMOS [89]. Cependant, comme illustré à la figure 1.4(c), la majorité des jeux de données récents proposent désormais des séquences *non segmentées* (untrimmed). Cette évolution reflète la nécessité croissante de développer des solutions en ligne capables de traiter des flux continus, plus représentatifs des scénarios industriels réels, mais aussi plus complexes à modéliser.



(a) (b) (c)

FIGURE 1.4 – Statistiques des jeux de données dans le tableau 1.4. (a) Fréquence des parties du corps représentées. (b) Répartition des types d'interactions. (c) Nombre de jeux de données segmentés/non segmentés par an.

1.4.2.2 Distribution déséquilibrée des données

La majorité des jeux de données industriels en analyse du mouvement humain présentent un déséquilibre marqué, tant entre les classes (inter-classes) qu'à l'intérieur d'une même classe (intra-classe). Le tableau 1.5 illustre cette variabilité à travers les moyennes et écarts-types

du nombre d'échantillons par action, en excluant la classe « aucune action ». Un écart-type élevé indique une forte disparité entre les classes, ce qui se traduit par une surreprésentation de certaines actions.

Par exemple, dans *Assembly101* [15], les classes « pickup » et « put down » comptent à elles seules autant d'échantillons que l'ensemble des 18 autres classes combinées. De même, dans *CAARL* [19], environ 70 % des actions annotées appartiennent à la catégorie « Handling », tandis que dans *MECCANO* [86], les actions « take » et « check » représentent la moitié des annotations. Bien que *InHARD* [3] présente une distribution plus équilibrée, certaines actions, telles que « assemble system » ou « picking left », demeurent majoritaires. À l'inverse, la version virtuelle *InHARD-DT* [84], issue d'un jumeau numérique, présente une meilleure homogénéité entre les classes.

Enfin, au-delà du déséquilibre en nombre d'échantillons, un déséquilibre temporel intra-classe subsiste : les durées d'exécution des actions varient fortement, allant de quelques secondes à plusieurs dizaines, comme dans l'action « walking with a polisher » [90]. Cette hétérogénéité rend difficile la reconnaissance fiable et en temps réel des gestes dans des environnements industriels complexes.

TABLEAU 1.5 – Moyenne et écart-type du nombre d'échantillons par classe pour quelques jeux de données.

Jeu de données	**Moyenne ± Écart-type**
HA4M [65]	307,42 ± 165
InHARD-DT [84]	369,15 ± 112
InHARD [3]	369,46 ± 347
MECCANO [86]	737,75 ± 793
CAARL [19]	955,50 ± 1245
Assembly101 [15]	42479,17 ± 48666

1.4.2.3 Variabilité des exécutions humaines

Les activités industrielles présentent une forte variabilité liée à la diversité des opérateurs, à leurs niveaux d'expertise et à leurs styles d'exécution. Cette variabilité engendre à la fois des différences inter-classes, où plusieurs actions distinctes présentent des postures ou des mouvements similaires, et des variations intra-classes, où une même action peut être réalisée différemment selon l'opérateur, la cadence ou la précision du geste. La Figure 1.5 met en évidence des cas de similarité inter-classes observés dans le jeu de données *InHARD* [3], où

certaines actions, telles que « Consulter fiche » et « Tourner fiche », présentent des postures très proches. À l'inverse, la Figure 1.6 illustre la variabilité intra-classe pour une même action, telle que « Prendre sous-ensemble », exécutée différemment selon les opérateurs ou les conditions d'enregistrement. Pour remédier à cela, des jeux de données tels que *Assembly101* [15] intègrent des annotations précises du niveau de compétence des opérateurs (experts ou novices), ainsi que la vitesse d'exécution des tâches et les erreurs commises. Ces informations supplémentaires exigent un important travail d'annotation, ce qui accroît considérablement les efforts de création de tels jeux de données et explique en grande partie leur rareté.



(a) " Consulter fiche " et " Aucune action " (vue gauche)

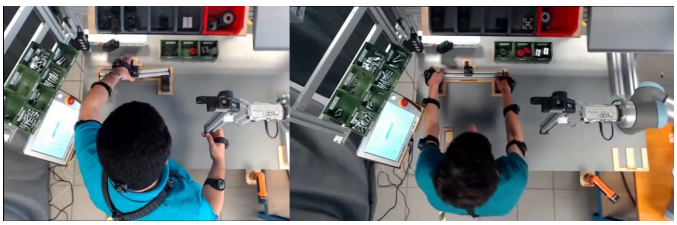


(b) " Tourner fiche " et " Prendre devant soi " (vue gauche)

FIGURE 1.5 – Exemples illustrant la similarité entre différentes classes d'actions dans le jeu de données InHARD.

1.4.2.4 Parties du corps, types d'interactions et variations de points de vue

Différentes configurations de l'analyse du mouvement humain (HMA) peuvent être mises en œuvre selon le cas d'étude considéré, comme illustré à la figure 1.7. Dans les tâches d'assemblage statiques, seules les parties supérieures du corps "bras ou tronc" sont généralement suivies à l'aide de caméras ou de capteurs de signaux, comme dans les jeux de données *MECCANO* [86], *WGD* [69], *Assembly101* [15] et *HAR-V1* [83]. À l'inverse, lorsque l'opérateur se déplace dans son environnement de travail, l'ensemble du corps est suivi à l'aide de capteurs inertiels (IMUs) ou de caméras, comme dans *HRI30* [67], *Magni* [56], *THOR* [57],



(a) " Prendre sous-ensemble " (vue supérieure)



(b) " Poser tige de mesure " (vue gauche)

FIGURE 1.6 – Exemples illustrant la variabilité intra-classe dans le jeu de données InHARD.

COVERED [85], *LARa* [63] et *CAARL* [19]. La reconnaissance des gestes humains en environnement industriel demeure toutefois complexe, notamment en raison des occlusions, des variations d'éclairage, des entités mobiles et des contraintes mécaniques. Pour limiter ces effets, certaines bases de données ont adopté des vues multiples : *WGD* emploie huit caméras RGB disposées à l'avant et à l'arrière de l'opérateur, tandis que *Assembly101* combine des caméras statiques et égocentriques. D'autres, comme *CAARL* et *LARa*, suivent l'opérateur en mouvement dans son espace de travail, capturant ainsi l'ensemble des interactions entre l'humain, les outils et l'environnement. Les figures 1.4(a) et (b) illustrent respectivement la fréquence de représentation des parties du corps et la répartition du nombre de jeux de données selon les types d'interactions observés dans les expérimentations précédemment présentées.

1.4.2.5 Compréhension de l'intention humaine

Certains jeux intègrent une annotation contextuelle pour anticiper les actions futures. Par exemple, dans l'assemblage automobile, connaître le nombre de roues permet d'anticiper les actions « poser roue ». Cependant, cette approche est limitée lorsque l'ordre des actions n'est pas déterminé à l'avance [15]. Les jeux *LARa*, *CAARL* et *IKEA ASM* [4] enrichissent l'analyse en suivant également les objets et outils utilisés.

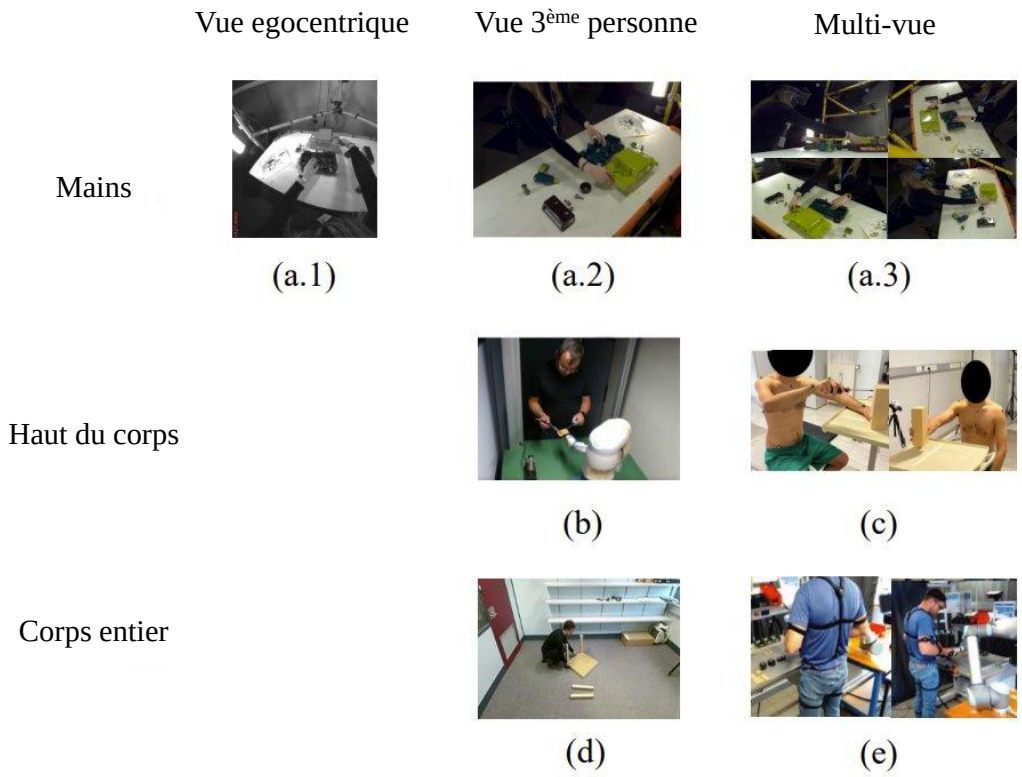


FIGURE 1.7 – Illustrations des jeux de données, regroupées par parties du corps suivies, par types d'interaction et par angles de vue. (a) [15], (b) [48], (c) [69], (d) [4], (e) [3].

1.4.2.6 Technologies et modalités

Les jeux de données étudiés s'appuient sur une large variété de capteurs et de modalités d'acquisition du mouvement humain en contexte industriel. La majorité utilisent des caméras RGB-D pour suivre les opérateurs dans leur environnement (*HRI30* [67], *HA4M* [65], *IHADv1* [83]), tandis que d'autres combinent caméras et capteurs inertiels (IMUs) pour une estimation plus fine des trajectoires et postures (*InHARD* [3], *yDataLab* [71], *CAARL* [19], *LARa* [63]). Certaines approches, comme celle de MOHAMMADI, REZAYATI, VENN et al. [21], exploitent les couples articulaires des robots pour prévenir les collisions homme–robot.

Les modalités exploitées varient également selon les cas d'usage. Certains jeux de données reposent uniquement sur la modalité RGB, comme *CAARL* [19] et *CoAx* [48], tandis que d'autres associent les informations RGB et squelettiques, à l'instar de *WGD* [69], *LARa* [63] et *IKEA ASM* [4]. Enfin, plusieurs jeux de données adoptent une approche multimodale combinant accéléromètres, profondeur, infrarouge et squelette, comme *OpenPack* [82]

et *HA4M* [65].

Les données squelettiques sont souvent privilégiées pour leur robustesse face aux variations de l'environnement et de la morphologie des opérateurs. Toutefois, l'intégration de modalités multiples, bien qu'elle améliore la précision de la reconnaissance, accroît la complexité du traitement et les besoins en ressources computationnelles.

1.4.2.7 Annotation automatique et jumeaux numériques

L'annotation des données constitue une étape coûteuse en temps et en ressources lorsqu'elle est effectuée manuellement. Pour surmonter cette difficulté, Dallel, Havard, Dupuis et al. [84] ont proposé une approche basée sur le jumeau numérique, qui consiste à recréer un environnement industriel virtuel et à associer des déclencheurs à chaque action afin de permettre l'annotation automatique des données. Cette méthode contribue également à équilibrer les jeux de données en reproduisant artificiellement des actions rares. Dans la même perspective, Vysocky, Grushko, Spurny et al. [91] ont généré des données synthétiques destinées à l'apprentissage de la détection des mains sur des images de profondeur et ont proposé un jeu de données dédié aux applications industrielles de suivi et de localisation des mains.

Ces observations sont essentielles pour orienter les futures collectes de données industrielles vers une répartition plus équitable des classes de données, ainsi qu'une précision accrue des annotations.

1.4.3 Synthèse

L'analyse du mouvement humain dans les environnements industriels repose sur l'étude des déplacements des opérateurs, de leurs interactions avec l'environnement, ainsi que de leurs signaux physiologiques. Pour cela, diverses technologies sont mobilisées, notamment les systèmes de capture de mouvement (MoCap) et les capteurs physiologiques basés sur les signaux musculaires.

Ces technologies permettent la génération de différentes modalités de données, telles que le squelette, les images RGB, les images de profondeur, ou encore les signaux EMG et FMG. Parmi celles-ci, les MoCap restent les plus utilisées, en particulier les caméras Kinect et les IMUs. À l'inverse, les capteurs physiologiques sont encore relativement peu répandus.

Un total de 16 jeux de données a été recensé, majoritairement orientés vers des applications industrielles de type assemblage. La plupart proposent des données RGB et squelette, généralement capturées à l'aide de capteurs MoCap. Il est également notable que 14 de ces jeux de données intègrent plusieurs capteurs afin de générer des modalités hybrides, combinant ainsi les avantages de chaque technologie.

Dans le cadre de cette thèse, seuls certains jeux de données répondent pleinement aux contraintes méthodologiques de notre étude, notamment la disponibilité de données squelettiques, d'annotations temporelles fines et de séquences représentatives de tâches d'assemblage. En particulier, les bases InHARD [3] et HA4M [65] fournissent directement des données squelettiques exploitables, bien que HA4M se limite au haut du corps, ce qui permet d'évaluer la robustesse des modèles dans des conditions de visibilité partielle. Par ailleurs, le jeu de données IKEA-ASM [4], bien que reposant principalement sur des données RGB, permet l'extraction a posteriori du squelette, offrant ainsi un cadre pertinent pour étudier la capacité de généralisation de nos modèles à des modalités et des configurations de capture différentes. L'ensemble de ces jeux de données constitue ainsi un socle expérimental cohérent pour l'analyse et la reconnaissance d'actions d'assemblage en environnement industriel.

1.5 Discussion et principaux défis

L'étude des travaux de recherche sur l'analyse du mouvement humain dans divers contextes industriels a permis de dégager des observations précieuses ainsi que des pistes d'amélioration. Nous avons identifié les constats présentés dans les sections suivantes. :

1.5.1 Évolution de l'analyse du mouvement humain de l'industrie 4.0 à l'industrie 5.0

Dans notre revue de la littérature, nous avons relevé 36 articles traitant de l'industrie 4.0 et 9 articles mentionnant l'industrie 5.0, dont 5 sont des articles de synthèse. Ce déséquilibre peut s'expliquer par le fait que la majorité de ces travaux se concentre sur l'amélioration de la productivité et de l'efficacité, en adéquation avec les objectifs de l'industrie 4.0. L'industrie 4.0 se caractérise essentiellement par la numérisation et l'automatisation, visant à optimiser la production et l'efficacité [92], tout en intégrant également des considérations de sécurité et d'ergonomie [17].

Dans notre contexte, l'industrie 5.0 accorde une importance accrue à ces deux aspects en adoptant une orientation nettement plus centrée sur l'humain [93]. Dans cette perspective, les systèmes d'analyse du mouvement humain (Human Motion Analysis - HMA) sont principalement utilisés dans des applications visant le bien-être des opérateurs, notamment par la création d'environnements de travail ergonomiques afin de réduire la charge physique [1]. Ces systèmes permettent d'analyser la posture des travailleurs pour proposer de meilleures postures de travail et d'évaluer également les niveaux de sécurité et de fatigue [94].

Le suivi en temps réel des mouvements des opérateurs est utilisé pour prévenir les accidents et collisions dans des environnements partagés, renforçant ainsi la sécurité dans des contextes collaboratifs dynamiques et facilitant l'interaction homain-robot (HRI) [21]. De plus, il rend possible la personnalisation en adaptant l'environnement de travail aux préférences et besoins individuels des opérateurs [1]. Cette personnalisation favorise un espace de travail plus confortable et inclusif, améliorant in fine le bien-être et la productivité des travailleurs en milieu industriel.

En conclusion, cette transition de l'industrie 4.0 vers l'industrie 5.0 traduit principalement, dans notre contexte, un passage d'une logique centrée sur l'optimisation des processus à une approche plus globale qui prend en compte les besoins de la main-d'œuvre et l'intégration de systèmes autonomes [20].

1.5.2 Choix des capteurs appropriés

Les capteurs basés sur caméra sont sensibles aux occlusions et peuvent être affectés par les conditions de l'environnement de travail telles que l'éclairage, les réflexions ou encore les variations de température [13]. Pour pallier les problèmes liés aux occlusions, l'utilisation de caméras multi-vues constitue une solution viable. En les positionnant de manière stratégique, il est possible de fusionner les données issues de différentes capacités de détection présentes dans l'usine (par exemple caméras embarquées, robots, casques de réalité augmentée, etc.). Cette approche offre une couverture plus large de l'espace de travail et permet de suivre les mouvements les plus importants ainsi que les interactions homain-objet.

Une autre solution pour résoudre les problèmes liés à l'éclairage consiste à recourir aux jumeaux numériques afin de recréer virtuellement l'environnement industriel. Cette technique permet de générer des données avec des paramètres quasiment entièrement contrôlés, capables de simuler différentes conditions d'éclairage ou points de vue de caméras.

De plus, la combinaison de plusieurs capteurs, tels que les IMUs et des systèmes de vision bas coût, peut fournir une solution plus complète permettant de compenser les limites de chaque capteur pris isolément. Bien que capables de fournir des données précises sur le mouvement corporel, les IMUs présentent certains inconvénients tels que des procédures d'installation longues et fastidieuses, ainsi qu'un inconfort pour l'opérateur lorsqu'ils sont portés en situation de travail. Une piste intéressante consiste à utiliser les IMUs plus spécifiquement lors de la phase d'entraînement, tout en libérant les opérateurs de cette contrainte durant la phase de test.

Par ailleurs, les images capturées par les caméras RGB peuvent révéler l'identité des opérateurs, soulevant ainsi des considérations éthiques liées à l'anonymisation des données, dans la mesure où des informations personnelles sont enregistrées. Une solution consiste à masquer ou flouter certaines parties sensibles des images, comme le visage, au cours du traitement des données. De plus, la sécurisation des données peut être assurée par des techniques de chiffrement, puis traitées via des architectures de calcul en périphérie (edge computing) ou dans le cloud [95]. Enfin, une alternative pertinente consiste à n'enregistrer que des informations dérivées telles que les caractéristiques issues de l'estimation de pose et les données relatives aux objets, garantissant ainsi la protection des données sensibles tout en permettant une analyse efficace et l'extraction d'informations utiles.

Dans le cadre de cette thèse, le choix des capteurs est directement guidé par l'application visée, à savoir l'analyse fine des gestes d'assemblage dans un contexte d'Industrie 5.0, où l'interaction homain-machine et l'assistance à l'opérateur occupent une place centrale. Les actions étudiées correspondent à des manipulations précises, souvent caractérisées par de faibles amplitudes de mouvement mais par des configurations articulaires spécifiques du membre supérieur. Dans ce contexte, l'information la plus pertinente n'est pas l'image brute, mais la représentation structurée de la posture humaine. C'est pourquoi les travaux présentés s'appuient sur des données squelettiques, extraites à partir de caméras RGB/Depth ou issues d'IMUs, permettant une description explicite, compacte et anonymisée du mouvement humain. Ce choix est également cohérent avec les bases de données utilisées dans cette thèse, qui reposent sur ces modalités de capture, et qui offrent un cadre réaliste pour l'étude des gestes industriels d'assemblage.

1.5.3 Recommandations pour l'acquisition de données

La disponibilité de jeux de données de référence (*benchmark datasets*) constitue un facteur essentiel pour l'avancement de l'analyse du mouvement humain en contexte industriel. En effet, ces jeux de données offrent une base standardisée permettant de comparer les méthodes et de valider leur précision. De plus, l'accessibilité de tels benchmarks accélère la recherche en épargnant aux chercheurs la tâche chronophage et coûteuse de la collecte de données, leur permettant ainsi de se concentrer sur le développement de solutions innovantes et sur les défis spécifiques aux environnements industriels. Ces jeux de données favorisent également la collaboration interdisciplinaire et garantissent que les recherches répondent aux exigences du secteur. Enfin, il convient de noter que ces benchmarks incluent souvent des métriques d'évaluation prédéfinies, telles que le Mean over Frame (MoF), le score F1 et d'autres métriques détaillées en section 2.5, ce qui facilite la mesure objective des performances et contribue au développement de solutions plus efficaces pour l'industrie.

Cependant, il est constaté que la majorité des jeux de données existants sont collectés dans des environnements de laboratoire contrôlés, ce qui ne reflète pas fidèlement la complexité des environnements industriels réels. Pour pallier cette limite, l'adoption de jeux de données plus diversifiés et complets est recommandée. Ceux-ci devraient intégrer, en plus des données de mouvement humain, des informations contextuelles telles que les outils et objets utilisés [19], différents environnements de postes de travail comme dans [4], ou encore des environnements dynamiques comportant des entités en mouvement (opérateurs, objets, robots, etc.) à l'image de [57]. L'intégration de ces éléments permettrait de développer des solutions plus robustes, propices à un déploiement efficace en conditions réelles.

Un autre problème majeur réside dans le manque de jeux de données industriels couvrant des aspects variés, tels que l'acquisition simultanée de différentes parties du corps (corps entier, partie supérieure, ou encore segments spécifiques comme les mains ou les bras), ou encore des perspectives variées (égocentriques, allocentriques et multivues). L'inclusion de ces aspects est fortement conseillée pour la conception de jeux de données industriels. Les protocoles d'acquisition et de collecte peuvent s'inspirer de jeux existants ayant déjà intégré ces dimensions, comme le jeu NTU-RGB+D120 [96], reconnu pour son protocole de collecte massive de données, Assembly101 [15] offrant de multiples points de vue, ou encore NTU-X [97] avec une représentation dense du visage et du corps entier.

Beaucoup de chercheurs proposent d'utiliser des jeux de données personnalisés et privés pour évaluer leurs approches. Bien que ces jeux soient précieux pour la communauté, leur inaccessibilité limite leur réutilisation et empêche la comparaison directe des résultats. De plus, ils pourraient enrichir les jeux existants en apportant de nouveaux cas d'usage. Il est donc fortement recommandé d'encourager les auteurs à partager leurs données, en appliquant si nécessaire des techniques d'anonymisation pour protéger l'identité des participants.

Dans les jeux de données industriels, certaines actions ont tendance à être répétées plus fréquemment que d'autres, ce qui conduit à des jeux déséquilibrés avec des classes d'actions sur-représentées et d'autres sous-représentées. Pour répondre à cette problématique, l'utilisation de techniques de génération de données synthétiques est proposée afin d'assurer une représentation plus équilibrée des classes. Les générateurs de données synthétiques, tels que les jumeaux numériques [84] ou les réseaux antagonistes génératifs (GANs) [98], constituent une solution prometteuse pour enrichir la diversité des activités industrielles représentées et faciliter l'entraînement de modèles HAR plus robustes et précis.

Enfin, les approches multimodales ont démontré de solides performances lorsqu'elles exploitent des données issues de capteurs variés [81], [99]. Ces données permettent d'intégrer des informations complémentaires telles que différentes vues des postures humaines ou les objets manipulés par les opérateurs. Les jeux de données devraient donc inclure plusieurs modalités comme le RGB, la profondeur, les nuages de points, l'infrarouge, les squelettes et même des signaux physiologiques, pour une représentation riche des mouvements et des interactions. L'exemple du jeu HA4M [65] illustre parfaitement cette approche et constitue une piste à suivre pour l'acquisition future de données multimodales.

Dans le cadre de nos travaux, nous faisons toutefois le choix de privilégier principalement la modalité squelettique. Cette représentation offre un compromis pertinent entre richesse descriptive, compacité des données et robustesse face aux variations d'environnement. Néanmoins, les architectures proposées dans cette thèse sont conçues de manière à pouvoir intégrer ultérieurement des informations complémentaires, telles que les objets manipulés ou d'autres signaux contextuels, ouvrant ainsi la voie à des extensions multimodales futures.

1.6 Conclusion

Ce premier chapitre a présenté un panorama général de l'analyse du mouvement humain dans le contexte de l'industrie 5.0. Après avoir rappelé l'évolution de l'industrie 4.0 vers l'industrie 5.0, nous avons souligné le rôle central de l'humain dans cette nouvelle ère industrielle, où la collaboration humain–machine et la personnalisation des processus de production occupent une place prépondérante.

Nous avons ensuite défini le périmètre de l'analyse du mouvement humain et mis en évidence son importance dans les environnements industriels, en particulier pour la surveillance des tâches, l'assistance aux opérateurs et l'amélioration de la sécurité, en identifiant les défis spécifiques liés aux environnements industriels, tels que la variabilité des gestes, les occlusions ou encore la complexité des chaînes de production.

Un tour d'horizon des technologies de capture de mouvement a permis de comparer les solutions basées sur la vision, les capteurs inertiels, les signaux musculaires ainsi que les approches hybrides, tout en soulignant les compromis entre précision, coût, robustesse et facilité d'intégration dans un environnement industriel. Enfin, l'examen des jeux de données existants a montré que, malgré des avancées significatives, il subsiste un manque de bases de données adaptées aux contextes industriels complexes, dotées d'annotations précises et de conditions réalistes.

Par ailleurs, il convient de préciser le positionnement spécifique de ce travail au regard des cas d'usage industriels visés. Notre étude se concentre sur les tâches d'assemblage industriel, dont l'objectif est de reconnaître en continu les actions réalisées par les opérateurs. Cette analyse du mouvement humain vise plusieurs finalités : évaluer la charge ergonomique, améliorer le bien-être et la sécurité des opérateurs, et assister l'opérateur dans un espace de travail partagé, notamment dans des scénarios de collaboration humain–robot. Dans cette perspective, notre travail s'appuie principalement sur des données squelettiques, qu'elles soient extraites à partir de vidéos RGB ou directement acquises via des systèmes de capture de mouvement. Cette représentation permet de modéliser de manière précise la dynamique corporelle des opérateurs tout en restant indépendante des variations d'apparence visuelle. Les méthodes développées dans les chapitres suivants seront donc analysées au regard de ces modalités et adaptées aux contraintes propres aux environnements d'assemblage réels.

Ces constats ouvrent la voie à l'étude approfondie des méthodes de reconnaissance et de

segmentation d'actions humaines, qui fera l'objet du chapitre suivant. Celui-ci détaillera les approches et architectures utilisées dans la littérature, ainsi que les métriques d'évaluation permettant de mesurer leur performance, en vue d'identifier les solutions les mieux adaptées aux contraintes et exigences de l'industrie 5.0.

Chapitre 2

État de l'art sur les approches d'analyse du mouvement humain

2.1 Introduction

L'analyse du mouvement humain (HMA – *Human Motion Analysis*) constitue un domaine clé en vision par ordinateur et en interaction humain-machine, avec des applications variées allant de la vidéosurveillance à l'assistance collaborative en contexte industriel [1]. Dans le cadre de l'industrie 5.0, les approches doivent répondre à des contraintes spécifiques telles que la robustesse aux conditions complexes (éclairage variable, occlusions, bruit), la faible latence pour les applications en temps réel et la capacité à traiter des séquences longues et hétérogènes [11].

Ce chapitre présente un panorama structuré des approches d'analyse du mouvement humain (HMA), couvrant les principales méthodes de reconnaissance, de segmentation et de détection des actions. Après avoir introduit, dans la section 2.2, une taxonomie générale des approches selon le type de séquence de données, l'architecture du modèle et l'application visée, les sections 2.3 et 2.4 examineront respectivement les méthodes hors ligne et en ligne, en détaillant leurs principes, leurs avantages et leurs limites. Enfin, la section 2.5 présentera les métriques d'évaluation couramment utilisées pour mesurer leurs performances, suivie d'une discussion synthétique en section 2.6 mettant en perspective les tendances actuelles et les défis à relever en contexte industriel.

2.2 Taxonomie des approches

La taxonomie présentée à la figure 2.1 organise les approches d'analyse du mouvement humain selon quatre axes hiérarchiques principaux :

1. **Type de séquence** :
   - — Trimmed : vidéos préalablement découpées pour ne contenir qu'une seule action.
   - — Untrimmed : flux continu comportant plusieurs actions successives ou entrecoupées de périodes sans action.
2. **Approche de traitement** :
   - — Offline : l'ensemble de la séquence est disponible avant l'inférence, permettant une exploitation complète du contexte temporel.
   - — Online : le traitement s'effectue au fur et à mesure de l'arrivée des données, avec un accès limité au passé immédiat et sans accès aux données sur le futur.

3. **Architectures** :
   — Modèles séquentiels tels que les réseaux récurrents (RNN, LSTM, GRU).
   — Réseaux convolutifs (CNN) 2D ou 3D pour la capture des caractéristiques spatio-temporelles.
   — Architectures à base de Transformers exploitant l'attention spatio-temporelle.
   — Graph Convolutional Networks (GCN) pour les données structurées en graphe, notamment les données squelettiques.
   — Architectures hybrides combinant plusieurs paradigmes.
4. **Applications** :
   — Recognition : attribution d'une étiquette unique à une séquence donnée.
   — Temporal Segmentation : étiquetage frame par frame d'un flux continu.
   — Detection : localisation des instants de début et de fin des actions dans un flux vidéo.
   — Anticipation : prédiction de l'action avant qu'elle ne soit complétée ou avant son déclenchement.

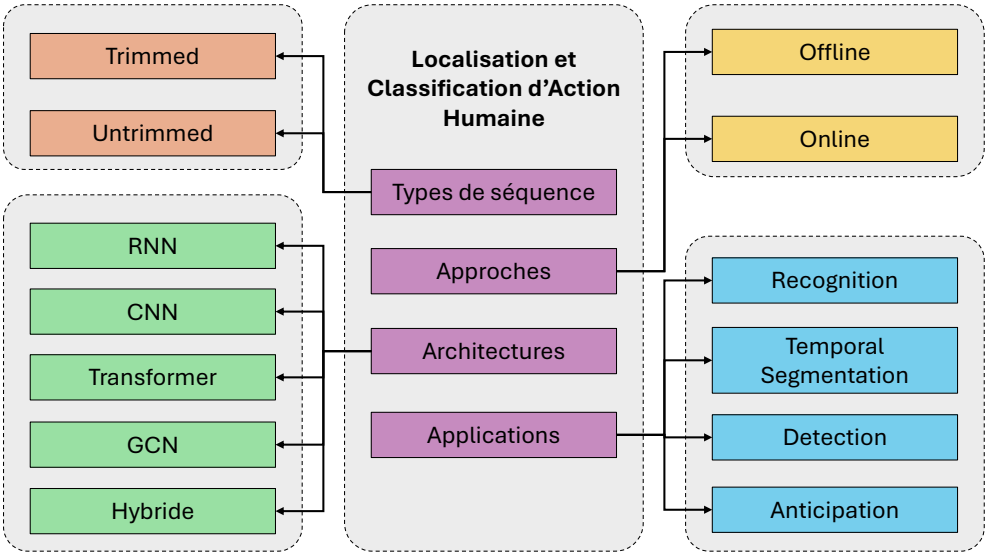


FIGURE 2.1 – Taxonomie des approches d'analyse du mouvement humain selon le type de séquence, le régime de traitement, l'architecture utilisée et le type d'application.

Cette structuration illustre la diversité des stratégies existantes, allant de la classification globale sur des séquences *trimmed* à la segmentation et à la détection dans des flux *untrimmed*, en intégrant des scénarios en ligne ou hors ligne, et en mobilisant des architectures variées. Les sections suivantes examineront chacune de ces branches en détail, en exposant les principes, les

méthodes représentatives, leurs avantages et leurs limites, ainsi que les métriques d'évaluation pertinentes.

2.3 Approches hors ligne (Offline)

L'analyse du mouvement humain peut être menée selon différents régimes temporels, en fonction de la disponibilité et de l'exploitation du contexte des données d'entrée. Dans les approches dites *offline*, l'intégralité de la séquence est accessible avant l'inférence, ce qui permet d'exploiter l'ensemble du contexte temporel de manière non causale. Cette configuration permet à l'algorithme de mobiliser à la fois l'historique et le futur de l'action observée, ce qui maximise la précision des prédictions. La taxonomie présentée à la figure 2.2 illustre la structuration des méthodes hors ligne autour de deux tâches principales : (i) la **reconnaissance d'actions sur séquences segmentées (trimmed)**, et (ii) la **segmentation temporelle d'actions (TAS) sur des séquences non segmentées (untrimmed)**.

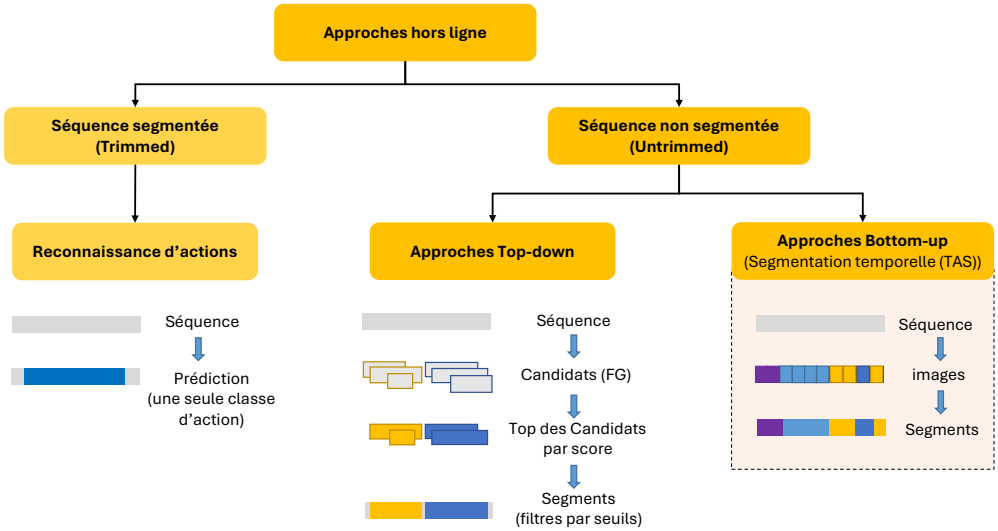


FIGURE 2.2 – Taxonomie des approches hors ligne (offline) pour l'analyse du mouvement humain.

Ces deux catégories se distinguent fondamentalement par la nature de la séquence vidéo : Dans le cas des données *trimmed*, chaque séquence contient une seule action, déjà isolée temporellement, tandis que, dans les données *untrimmed*, la vidéo est continue et comporte plusieurs actions successives, séparées par des transitions et des phases de *non-action*. Nous détaillons dans les sections suivantes les spécificités, les avantages, les limites et des exemples

représentatifs de chacune de ces familles.

2.3.1 Reconnaissance d'actions sur séquences *trimmed*

La reconnaissance d'actions sur des séquences *trimmed* est historiquement l'une des tâches les plus étudiées dans la littérature. Elle consiste à attribuer une étiquette unique à une séquence vidéo préalablement découpée, afin qu'elle ne contienne qu'une seule instance d'action. Dans ce cadre, l'algorithme peut exploiter l'intégralité des images, de la première à la dernière frame, pour extraire des caractéristiques spatiales et temporelles pertinentes. Cette configuration présente plusieurs avantages :

— **Exploitation complète du contexte temporel**, incluant l'historique et la fin de l'action, ce qui améliore la précision de la classification.
— **Entraînement plus stable** que sur des séquences longues et hétérogènes, car l'action est déjà isolée.

En contrepartie, ce type de méthode repose sur la disponibilité de séquences déjà segmentées, ce qui n'est pas toujours réaliste dans des environnements réels, en particulier industriels, où les enregistrements sont continus et rarement annotés de manière découpée. De plus, elles ne fournissent pas d'informations sur le moment précis où l'action commence ou se termine, ce qui limite leur applicabilité à la surveillance en temps réel ou au suivi d'opérations en flux continu.

2.3.2 Segmentation temporelle sur séquences *untrimmed*

Si la reconnaissance *trimmed* a constitué une étape essentielle dans le développement des algorithmes d'analyse du mouvement, elle demeure insuffisante pour traiter des enregistrements continus où plusieurs actions se succèdent. Dans un contexte industriel, par exemple sur une ligne d'assemblage, les opérateurs enchaînent une série de gestes et d'actions au sein d'un même flux vidéo, sans que ceux-ci soient préalablement isolés. C'est précisément dans ce type de scénario que la **segmentation temporelle d'actions** (*Temporal Action Segmentation*, TAS) s'avère indispensable.

2.3.2.1 Définition et objectif

La TAS sur des séquences *untrimmed* vise à attribuer une étiquette d'action à chaque frame, puis à agréger les frames contigues de la même classe pour former des segments.

Contrairement aux approches en ligne, les méthodes *offline* peuvent exploiter des informations futures pour améliorer la précision des frontières d'action, réduire la sur-segmentation et mieux gérer les actions longues ou à transition progressive. La **sur-segmentation** désigne un phénomène où une même action est, à tort, fragmentée en plusieurs segments successifs, souvent en raison d'une variabilité intra-classe ou d'un manque de contexte temporel suffisant. Elle conduit à une multiplication artificielle des prédictions d'actions et dégrade la cohérence temporelle des résultats.

La TAS permet :

— de reconstituer un historique précis des opérations réalisées [59] ;

— d'évaluer la durée et la fréquence des gestes techniques pour l'ergonomie [30] ;

— de détecter les écarts de procédure ou anomalies dans le flux de travail [52] ;

— d'automatiser le comptage et la traçabilité des tâches répétitives [53].

Les premiers modèles de segmentation temporelle reposaient principalement sur des réseaux de neurones récurrents (RNN) tels que les *Long Short-Term Memory* (LSTM) et les *Gated Recurrent Units* (GRU) [11], [100]. Toutefois, ces modèles se sont révélés moins efficaces et plus lents sur le plan computationnel, notamment pour les séquences longues, en raison de leur mémoire fixe [101]. Pour pallier ces limites, des réseaux de convolutions temporelles (TCN), tels que MS-TCN [101], ont été introduits. Ces modèles surpassent les RNN en précision tout en offrant une inférence plus rapide. S'appuyant sur le succès de MS-TCN, de nombreuses études ont affiné et étendu son architecture pour améliorer encore les performances de segmentation.

Actuellement, deux principales catégories de méthodes entièrement supervisées pour la reconnaissance et la segmentation d'actions dans des vidéos non segmentées (*untrimmed*) existent : les approches **top-down** et **bottom-up** [102].

2.3.2.2 Approches Top-down

Les approches top-down commencent par générer des segments candidats (*proposals*) correspondant à des actions potentielles, puis affinent leurs bornes temporelles de début et de fin afin d'obtenir une séquence d'actions segmentées [103]. Par exemple, [104] ont introduit le *Trident-head* afin de modéliser les frontières d'action et de réduire le problème de perte de rang (*rank loss*) dans l'auto-attention, en utilisant une couche de perception à granularité variable (*Scalable-Granularity Perception* (SGP)) qui agrège l'information à travers différentes

granularités temporelles. De même, [105] ont proposé le *Scale FPN* (SFPN), un réseau en forme de U qui capture des caractéristiques vidéo RGB à différentes échelles et améliore le flux d'information entre les couches pour la segmentation d'actions. [106] ont adopté une stratégie *predict-and-refine* : leur méthode génère une prédiction initiale, puis la raffine à l'aide d'un mécanisme d'attention, ce qui permet de débruiter les représentations des segments et de capturer leurs corrélations temporelles.

Ces approches sont généralement implémentées en deux étapes, l'une pour la reconnaissance et l'autre pour la localisation, ce qui rend le modèle légèrement plus performant mais aussi plus complexe que les méthodes en une seule étape. Bien que les méthodes à une étape puissent être entraînées de bout en bout, elles sont fortement sensibles aux réglages d'hyperparamètres et difficiles à appliquer à des scènes longues et denses, le traitement de milliers de propositions nécessitant d'importantes ressources de mémoire [107].

2.3.2.3 Approches Bottom-up

À l'inverse, les approches bottom-up prennent une vidéo non découpée en entrée et produisent directement une séquence d'actions avec une étiquette de classe pour chaque frame. Après l'application d'une fonction d'agrégation, des segments sont formés à partir de ces frames, ce qui aboutit à une séquence finale indiquant le début et la fin de chaque action [108]. [60] ont proposé une architecture encodeur-décodeur nommée C2F-TCN, qui met en œuvre un ensemble *coarse-to-fine* de sorties du décodeur pour atténuer le problème de sur-segmentation dans des vidéos RGB. Dans un autre registre, [109] ont introduit ASPnet, un modèle flexible capable de fusionner plusieurs modalités d'entrée (vidéo RGB et accéléromètre) tout en capturant simultanément la dynamique temporelle à long terme dans des données séquentielles. Les approches bottom-up offrent généralement une meilleure détection des actions de courte durée et de transitions temporelles rapides, là où les méthodes top-down tendent à privilégier une modélisation plus globale de la séquence. Cependant, elles ont tendance à souffrir d'une sur-segmentation à mesure que leur confiance augmente lors de l'entraînement [60]. Ainsi, après avoir présenté les principes et les limites des approches de reconnaissance et de segmentation des actions, la section suivante examine leur potentiel et leurs applications dans le contexte industriel.

2.3.2.4 Applications industrielles

Bien que ces approches aient amélioré la précision de la segmentation d'actions, elles ont principalement été appliquées à des vidéos courtes et à un environnement contrôlé, comme celles des jeux de données THUMOS-14 [89] et ActivityNet [88], ou à des vidéos où seule une partie du corps humain est visible (par exemple les mains dans Epic-KITCHEN [110]). Malgré cette focalisation sur des contextes sportifs ou de la vie quotidienne, la segmentation temporelle des actions présente un fort potentiel pour les applications industrielles, où une compréhension précise des actions humaines complexes est cruciale.

Des travaux récents ont démontré cette pertinence. [111] ont utilisé la segmentation d'actions pour évaluer la posture et les mouvements des ouvriers, afin de soutenir des évaluations ergonomiques et d'améliorer la sécurité des opérateurs. Dans une perspective complémentaire, [112] ont appliqué ces techniques pour compter automatiquement des tâches répétitives sur des lignes d'assemblage, afin de réduire le risque de troubles musculosquelettiques grâce à des évaluations ergonomiques plus précises. [113] ont proposé un cadre de détection de la fatigue lors d'activités d'assemblage manuel, afin de surveiller le bien-être des travailleurs et de prévenir les blessures. Au-delà de la santé des travailleurs, [114] ont exploité la segmentation temporelle pour optimiser l'organisation des flux de travail sur les postes industriels, améliorant ainsi l'efficacité et la productivité des processus.

Cependant, les travaux dans ce domaine restent rares, et le potentiel de la TAS dans un contexte industriel demeure largement à explorer. Contrairement aux environnements simples des activités quotidiennes, les contextes industriels impliquent des séquences d'actions d'assemblage longues et complexes, ce qui constitue un défi plus difficile. Les méthodes existantes peinent à produire des résultats de segmentation précis dans ces environnements, comme l'illustre l'utilisation de C2F-TCN sur le jeu de données Assembly101 [15]. De plus, l'utilisation de la modalité RGB peut affecter la précision de la reconnaissance en raison du bruit induit par les objets et les entités en mouvement dans les postes industriels [11].

Face à ces défis, il est apparu qu'une modalité plus précise est nécessaire pour réduire le bruit environnemental et se concentrer sur les gestes du corps humain. La modalité squelette a été choisie pour sa robustesse face aux changements environnementaux (éclairage, occlusions, etc.) et sa capacité à représenter avec précision le mouvement de l'opérateur [115]. Par conséquent, les approches à base de graphes ont été considérées comme particulièrement

appropriées pour intégrer les données squelettiques et fournir une représentation plus riche des mouvements.

2.3.3 Architectures utilisées en *offline*

Les méthodes *offline*, qui ciblent la reconnaissance d'actions sur des séquences *trimmed* ou la segmentation temporelle sur des séquences *untrimmed*, mobilisent un large éventail d'architectures, comme synthétisé dans le tableau 2.1 (les métriques associées sont détaillées dans la section 2.5 et dans le tableau 2.2 de cette même section).

La répartition statistique des approches de reconnaissance d'actions selon l'architecture utilisée, observée dans la littérature récente, est illustrée à la figure 2.3. Cette visualisation met en évidence la prépondérance de certaines familles de modèles, comme les CNNs, tout en montrant l'émergence de nouvelles approches hybrides, ainsi que des approches basées sur les Transformers et les GCNs.

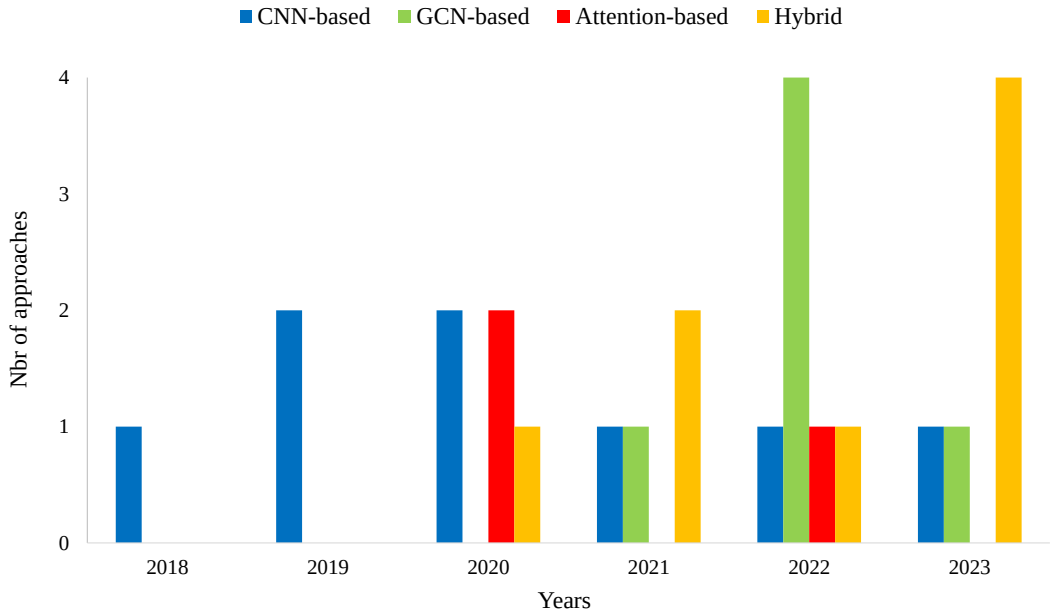


FIGURE 2.3 – Répartition statistique des approches de reconnaissance d'actions selon l'architecture utilisée, d'après la littérature récente (2018-2023). [11]

Dans ce qui suit, nous présentons en détails les principales familles de méthodes utilisées pour la reconnaissance d'actions hors ligne, en mettant en lumière leurs avantages, leurs limites et les architectures représentatives qui les composent.

TABLEAU 2.1 – Approches proposées dans l'état de l'art pour l'analyse du mouvement humain en contexte industriel.
P : Privé, RA : Reconnaissance d'Actions, VMM : Modélisation de Markov à longueur variable.

	Année	Approche	Objectif	Untrimmed	Trimmed	Jeu de données	Métriques					
							Accuracy	F1-score	Latence	IoU	mAP	Edit-score
Basé sur CNN	2023	CNN [14]	reconnaître l'activité humaine à partir de données de trajectoire intérieure bruitées en utilisant un apprentissage semi-supervisé dans un atelier d'assemblage de tricycles	✓		P	-	0.81-0.95	-	-	-	-
		TSM préentraîné sur Kinetics-400 [116]	évaluer différents jeux de données par apprentissage supervisé et auto-supervisé pour la reconnaissance d'actions et de gestes		✓	InHARD [3]	0.881	-	-	-	-	-
		TSN préentraîné sur Kinetics-400 [116]			✓	MECCANO [86]	0.411	-	-	-	-	-
	2022	ResNet-50 [45]	reconnaître les actions et intentions humaines pour optimiser le comportement réactif du robot et l'ergonomie humaine		✓	HRI30 [67]	0.866	-	-	-	-	-
	2021	CollisionNet [25]	détection et localisation de collisions	✓		P	-	0.965	-	-	-	-
		CNN (ResNet) [52]	évaluer l'activité de l'opérateur (identification d'erreurs) lors d'un assemblage manuel	✓		P	-	-	-	-	-	-
	2020	3D-CNN [21]	reconnaître l'action humaine et détecter le contact entre robot et humain pour améliorer la sécurité	✓		P	0.997	-	-	-	-	-
		1D-CNN [21]				PHR Contact Detection Dataset [21]	0.96	-	-	-	-	-
	2019	Regroupement hiérarchique basé sur CNN [55]	reconnaître les actions humaines pour la planification du travail et la productivité	✓		P	0.56	-	-	-	-	-
		CNN à fusion pondérée [18]	développer un système basé sur la fusion de capteurs pour reconnaître les actions humaines lors de tâches d'assemblage		✓	P	0.846	-	-	-	-	-
	2018	DCNN [117]	Reconnaître les mouvements humains et identifier le contexte de l'action associée dans l'assemblage d'un moteur automobile	✓		P (3 actions)	0.96	-	-	-	-	-
Basé sur RNN	2023	LSTM [118]	reconnaître l'activité spatio-temporelle humaine à partir du squelette humain	✓		InHARD (4/14 actions incluant le fond) [3]	0.38 - 0.67	0.68	-	-	-	-
Basé sur GCN	2023	Mask-GCN [61]	reconnaître des actions humaines complexes avec de nouveaux schémas de mouvement		✓	P	0.73	-	-	-	-	-
	2022	STGCN-SWMV [119]	reconnaître les actions humaines	✓		OAD [120]	0.954	0.953	-	0.977	-	-
						UOW [121]	0.934	0.936	0.047	0.958	-	-
		ST-GCN [84]	reconnaître les actions humaines à partir de données physiques et auto-générées		✓	75% InHARD [3] 25% InHARD-DT [84]	0.956	0.955	-	-	-	-
		ST-GCN [40]	reconnaître les actions d'assemblage de l'opérateur lors d'une collaboration homme-robot		✓	P	0.40	-	-	-	-	-
		MTM-STGCN-SW [122]	estimer l'avancement de l'assemblage en étendant les méthodes de RA à des processus multivariants	✓		InHARD (3/14 actions sans fond) [3]	0.593	-	-	-	-	-
	2021	TL-STGCN [58]	reconnaître les actions humaines pour le contrôle réactif et la prise de décision du robot en utilisant l'apprentissage par transfert		✓	P	-	-	-	-	0.956	-
Basé sur l'attention	2022	P3DAttenNet [59]	reconnaître les actions humaines et générer des plans d'assemblage	✓		IKEA ASM [4]	0.692	0.81-0.95	-	-	0.431	-
	2020	HAMLET [123]	reconnaître les activités humaines pour les systèmes robotiques collaboratifs	✓		UTD-MHAD [124]	0.951	-	-	-	-	-
						UT-Kinect [125]	0.974	-	-	-	-	-
						UCSD-MIT [126]	0.815	-	-	-	-	-
		SAM [127]	reconnaître les actions humaines à partir de représentations de caractéristiques issues de capteurs portés sur le corps	✓		PAMAP2 [128]	-	0.95-0.96	-	-	-	-
						Opportunity [129]	-	0.61-0.67	-	-	-	-
						USC-HAD [130]	-	0.50-0.55	-	-	-	-
						Skoda [131]	-	0.93-0.97	-	-	-	-
Hybride	2023	ResNest-34 + LSTM [20]	reconnaître les actions humaines et extraire le contexte de haut niveau des opérations d'assemblage industriel afin de l'intégrer dans des plans d'assemblage collaboratifs	✓		P	0.966	-	-	0.941	-	-
		Attention + CNN + Bi-LSTM [44]	estimer la trajectoire du mouvement humain pour la collaboration en assemblage	✓		P	-	-	-	-	-	-
		CNN-LSTM [132]	reconnaître les actions humaines		✓	InHARD (9/14 actions incluant le fond) [3]	0.85-0.88	-	-	-	-	-
		C2F-TCN [60]	segmenter temporellement les actions et gestes humains en utilisant un apprentissage entièrement supervisé et semi-supervisé	✓		Assembly101 [15]	-	0.212	-	0.5	-	0.324
	2022	ConvTransformer [133]	reconnaître les actions humaines	✓		Opportunity [129]	-	0.443-0.861	-	-	-	-
						PAMAP2 [128]	-	0.858-0.915	-	-	-	-
						DSADS [134]	-	0.846	-	-	-	-
	2021	CNN-VMM [62]	reconnaître et anticiper les actions humaines sur la base de VMM dans les tâches d'assemblage de moteurs	✓		P	0.947	-	-	-	-	-
		GRU + CT-HMM [54]	reconnaître les tâches humaines à deux niveaux, primitives et activités, pour mesurer la performance et le travail manuel		✓	P	0.784-0.925	-	-	-	-	-
	2020	TSN + GCN + Attention [135]	reconnaître les actions humaines dans les flux de travail industriels		✓	P	-	0.816	-	-	-	-

2.3.3.1 Réseaux récurrents (RNN, LSTM, GRU)

Les réseaux neuronaux récurrents (RNN) et leurs variantes, notamment les *Long Short-Term Memory* (LSTM) et les *Gated Recurrent Units* (GRU), ont été parmi les premières architectures profondes à être appliquées à la reconnaissance d'actions vidéo [100]. Ils traitent les séquences image par image et propagent une mémoire interne à travers le temps, ce qui permet de modéliser les dépendances temporelles à long terme.

Cependant, ces architectures présentent plusieurs limitations :

— **Coût computationnel élevé** : l'inférence se fait séquentiellement, ce qui rend difficile la parallélisation.

— **Problèmes de gradient** : les RNN standards souffrent d'explosion ou de disparition du gradient, ce qui affecte l'apprentissage sur de longues séquences.

— **Mémoire fixe** : même avec un LSTM ou un GRU, la capacité de mémorisation reste limitée par la taille de l'état caché.

Malgré ces limites, des variantes optimisées, telles que [101], ont tenté d'améliorer la robustesse des LSTM pour des vidéos plus longues. Toutefois, leur adoption dans des environnements industriels demeure marginale en raison des contraintes de temps réel et de la complexité des gestes à traiter.

2.3.3.2 Réseaux convolutifs 2D et 3D

Les réseaux de neurones convolutifs (CNN) comportent trois types de couches principales : convolutionnelles, de pooling et entièrement connectées. Chacune joue un rôle spécifique dans l'extraction hiérarchique des caractéristiques visuelles [2]. En reconnaissance d'actions, ils sont souvent utilisés pour extraire des caractéristiques spatiales à chaque image avant une agrégation temporelle, captant ainsi efficacement la structure spatiale sans modéliser explicitement la dimension temporelle [136].

Pour pallier cette limite, les CNN 3D [81] ont été introduits, utilisant des filtres spatio-temporels capables de capturer simultanément l'information spatiale et temporelle. Des modèles comme I3D [137] exploitent des réseaux pré-entraînés sur ImageNet, puis remplacent les convolutions 2D par des convolutions 3D, atteignant des performances de pointe sur Kinetics et UCF101.

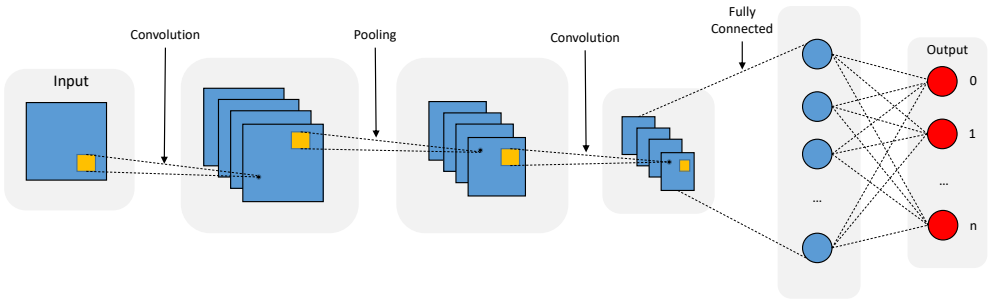


FIGURE 2.4 – Exemple d'architecture CNN.

2.3.3.3 Réseaux de convolutions temporelles (TCN)

Les *Temporal Convolutional Networks* (TCN) [138] remplacent la récurrence par des convolutions causales et dilatées, ce qui permet de capturer des dépendances temporelles à long terme tout en bénéficiant d'une parallélisation efficace sur GPU. Elles ont d'abord été utilisées pour des tâches de segmentation temporelle, mais se sont également révélées performantes pour la reconnaissance sur des séquences non découpées.

2.3.3.4 Graph Convolutional Networks (GCN)

Les *Graph Convolutional Networks* (GCN) ont suscité un intérêt particulier pour les données squelettiques, où les articulations sont représentées comme des nœuds connectés selon la topologie du corps humain. ST-GCN [139] a été l'un des premiers modèles à apprendre directement les relations spatio-temporelles dans un graphe articulé. Depuis, de nombreuses variantes ont été proposées pour mieux modéliser la dynamique articulatoire, introduisant, par exemple, des mécanismes d'attention [140] ou des topologies adaptatives.

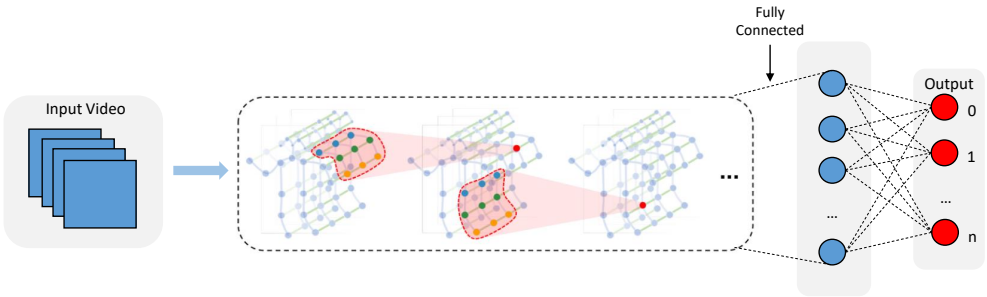


FIGURE 2.5 – Exemple d'architecture GCN spatio-temporel inspirée de [139].

En contexte industriel, l'utilisation des GCN sur des données squelettiques présente deux avantages majeurs : robustesse face aux perturbations visuelles (éclairage, occlusions par les outils) et focalisation sur la dynamique corporelle, ce qui est crucial pour analyser des gestes techniques précis.

2.3.3.5 Transformers spatio-temporels

Les Transformers, initialement développés pour le traitement du langage naturel (NLP), ont rapidement démontré leur efficacité dans divers domaines, notamment la vision par ordinateur et la reconnaissance d'actions humaines [81]. Leur architecture, fondée sur un couple encodeur–décodeur intégrant des mécanismes d'auto-attention et d'attention croisée, permet de modéliser efficacement les dépendances à long terme, de fusionner plusieurs modalités et de traiter des tâches multiples [2]. Appliqués aux séquences squelettiques, ils capturent simultanément les relations spatiales et temporelles entre les articulations. Des travaux récents, tels que P3DAttnNet [59], StarTransformer [141] et HAMLET [123], ont confirmé leur pertinence pour la modélisation spatio-temporelle et multimodale dans des contextes industriels. Cependant, leur coût de calcul élevé demeure un frein à leur utilisation en temps réel sur des séquences longues et complexes [11].

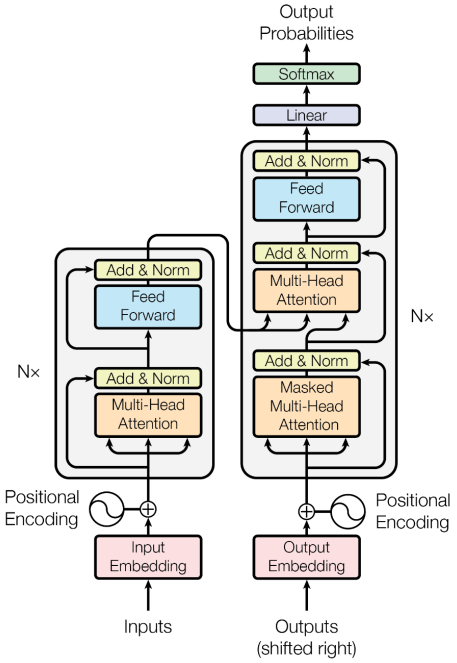


FIGURE 2.6 – Exemple d'architecture Transformer extraite de [142]

.

2.3.3.6 Méthodes hybrides

Les approches hybrides combinent différents types de modèles pour exploiter leurs forces respectives et compenser leurs faiblesses [2]. Par exemple, l'intégration de réseaux convolutifs (CNN) pour l'extraction de caractéristiques spatiales, associée à des modèles séquentiels tels que les LSTM ou les GRU pour la modélisation temporelle, a été largement explorée [132]. Ces architectures CNN-LSTM exploitent la capacité des CNN à extraire des représentations visuelles robustes image par image, tandis que les LSTM capturent la dynamique à long terme à partir de ces représentations. Bien que ces architectures soient efficaces sur des séquences relativement courtes, elles peuvent rencontrer des difficultés lorsque la longueur des séquences augmente, en raison de la complexité computationnelle des RNN et de leurs limites de mémoire.

Une autre variante consiste à remplacer les modules récurrents par des réseaux de convolution temporelle (TCN), ce qui permet un traitement parallèle des séquences et une meilleure exploitation des dépendances temporelles à long terme grâce aux convolutions dilatées [115]. Ce type d'architecture, souvent désigné sous le nom de CNN-TCN, combine la puissance des CNN pour la modélisation spatiale et l'efficacité des TCN pour la modélisation temporelle.

Les modèles hybrides peuvent également inclure des Transformers, soit en remplaçant complètement la partie temporelle par un module d'attention, soit en intégrant l'attention au sein des couches convolutionnelles [44]. Par exemple, certaines approches appliquent un CNN pour extraire des descripteurs locaux, suivis d'un Transformer pour modéliser des relations temporelles à l'échelle globale. Cette combinaison permet de bénéficier de la parallélisation des CNN et de la flexibilité des mécanismes d'attention.

Dans le cas des données squelettiques, les approches hybrides peuvent combiner des réseaux de graphes convolutifs (GCN), qui modélisent explicitement la structure articulatoire, avec des TCN ou des Transformers pour modéliser la dynamique temporelle [135]. Ce type de combinaison est particulièrement intéressant pour l'analyse d'actions complexes dans des environnements industriels, car il exploite à la fois la topologie du corps humain et les relations temporelles à différentes échelles.

Enfin, les stratégies de fusion utilisées dans les architectures hybrides jouent un rôle clé dans leurs performances. Trois schémas de fusion sont couramment employés :

— **Fusion précoce (early fusion)** : les caractéristiques issues de différentes modalités ou

de réseaux sont concaténées dès les premières couches.

— **Fusion intermédiaire (mid-level fusion)** : l'interaction entre les flux se fait à un niveau intermédiaire du réseau, souvent via des mécanismes d'attention croisée.

— **Fusion tardive (late fusion)** : les prédictions de chaque flux sont combinées à la fin, par exemple en calculant une moyenne pondérée ou un vote majoritaire.

Ces approches hybrides offrent une grande flexibilité et permettent d'adapter l'architecture aux contraintes spécifiques de la tâche, qu'il s'agisse de précision, de rapidité ou de robustesse face aux variations de l'environnement.

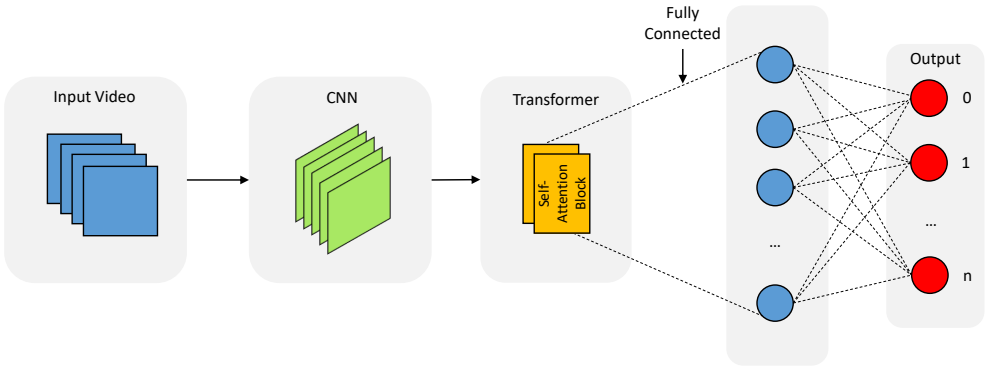


FIGURE 2.7 – Exemple d'architecture hybride.

2.3.4 Synthèse

Les approches hors ligne (*offline*) exploitent pleinement la séquence vidéo avant l'inférence, permettant d'extraire un contexte temporel complet et d'atteindre des performances généralement supérieures à celles des méthodes en ligne. L'analyse menée dans cette section met en évidence la diversité des paradigmes explorés pour la reconnaissance et la segmentation d'actions, allant des modèles séquentiels récurrents (RNN, LSTM, GRU) aux réseaux de convolution spatio-temporels (2D, 3D, TCN), en passant par les *Graph Convolutional Networks* (GCN) et les Transformers, ainsi que leurs combinaisons hybrides. Chaque famille d'architectures présente des atouts distincts : capacité à modéliser des dépendances temporelles à long terme, à extraire des caractéristiques spatio-temporelles riches, ou encore à exploiter explicitement la structure articulatoire du corps humain. Néanmoins, ces méthodes peuvent être coûteuses en calcul et en mémoire, et leur efficacité dépend souvent de la modalité choisie.

En segmentation temporelle, deux grandes approches coexistent : les méthodes **top-down**,

qui procèdent par génération puis par affinage de propositions d'action, et les méthodes **bottom-up**, qui attribuent une étiquette à chaque frame avant l'agrégation en segments. Les premières offrent une localisation plus précise des frontières d'action, tandis que les secondes excellent dans la détection d'actions courtes et à granularité fine, mais sont sujettes à la sur-segmentation. L'introduction d'architectures hybrides, comme C2F-TCN ou ASPnet, a permis d'améliorer la robustesse face aux séquences complexes, mais la majorité des travaux reste concentrée sur des environnements contrôlés ou des vidéos courtes.

Dans les contextes industriels, ces approches hors ligne ont montré leur intérêt pour des tâches telles que l'évaluation ergonomique, la détection de la fatigue ou l'optimisation des flux de travail. Cependant, elles se heurtent à plusieurs défis : la longueur et la complexité des séquences, les gestes fins et la variabilité des conditions de captation. La modalité squelettique, associée aux GCN et à des stratégies de fusion adaptées, constitue une alternative pertinente aux approches strictement basées sur l'image. Sous réserve d'une extraction fiable, elle offre une représentation structurée du corps humain, moins dépendante des variations d'apparence, d'éclairage ou d'arrière-plan. Les progrès récents des algorithmes d'estimation de pose 2D et 3D ont par ailleurs considérablement amélioré la qualité des squelettes extraits à partir de séquences RGB, renforçant ainsi leur exploitabilité en contexte industriel. Cette représentation permet en outre une modélisation explicite des relations articulaires et de la dynamique spatio-temporelle du mouvement.

Cette analyse prépare le terrain à l'étude des approches *en ligne* (*online*), qui, bien qu'elles disposent d'un contexte temporel limité au moment de la prédiction, sont essentielles pour les scénarios en temps réel et les environnements où la réactivité prime sur l'accès complet à la séquence vidéo.

2.4 Approches en ligne (Online)

Dans la section précédente, nous avons étudié la *Temporal Action Segmentation* (TAS) en contexte hors ligne, où l'intégralité de la séquence vidéo est disponible avant l'inférence. Dans ce cadre, les modèles peuvent exploiter l'ensemble du contexte temporel, passé, présent et futur, pour délimiter précisément les segments d'action et attribuer un label à chaque frame.

En contexte *online*, la situation est radicalement différente de celle rencontrée en reconnaissance d'actions hors ligne. À l'instant t, seules les observations passées et la frame courante

sont accessibles, et la prédiction doit être réalisée **de manière strictement causale**, c'est-à-dire sans aucune connaissance du futur. Cette contrainte transforme fondamentalement la tâche de *Temporal Action Segmentation* en un problème de **détection d'actions en ligne** (*Online Action Detection*), où les décisions doivent être prises au fil du temps, et non après la fin de la séquence. L'absence d'accès au futur rend la détection des frontières de début et de fin particulièrement difficile, tandis que la nécessité de produire une prédiction en temps réel impose de fortes exigences en matière de latence. De plus, la grande variabilité intra-classe observée au début d'une action, où les indices discriminants sont encore faibles ou ambigus, complique considérablement la reconnaissance précoce. Ainsi, la reconnaissance d'actions en ligne constitue un défi scientifique majeur : elle requiert des modèles capables d'analyser un flux continu de données tout en conciliant réactivité, robustesse temporelle et fiabilité de la prise de décision. Dans le contexte industriel, ces contraintes prennent une dimension opérationnelle essentielle, car un système d'assistance ou de surveillance doit être en mesure de comprendre et d'interpréter les gestes humains au moment même où ils se produisent, afin d'adapter instantanément sa réponse.

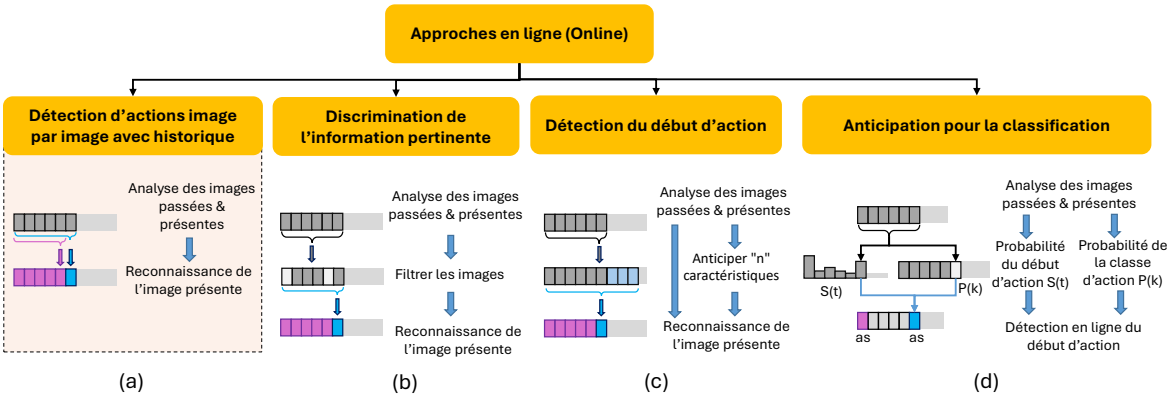


FIGURE 2.8 – Taxonomie des approches en ligne (online) pour l'analyse du mouvement humain.

Dans cette section, nous nous intéressons exclusivement au traitement de séquences vidéo *untrimmed*, c'est-à-dire contenant plusieurs actions enchaînées, comme c'est le cas dans des environnements industriels. Les approches en ligne se sont progressivement diversifiées en fonction de la nature du traitement temporel et du type de sortie qu'elles génèrent. La littérature récente distingue principalement quatre grandes familles de méthodes, illustrées sur la figure 2.8 et détaillées dans les sous-sections suivantes. Il convient de noter qu'il existe

d'autres variantes ou combinaisons méthodologiques, mais les catégories présentées ici constituent les approches les plus fréquemment rencontrées dans les travaux contemporains sur la détection d'actions en ligne.

— **Détection d'actions image par image avec historique (Historical Per-frame Action Detection)** Présentée en sous-section 2.4.2, cette approche regroupe les méthodes qui attribuent à chaque image entrante une étiquette d'action en exploitant l'historique.
— **Discrimination de l'information pertinente (Information Discrimination)** Développée en sous-section 2.4.3 , cette approche améliore la classification image par image en intégrant des mécanismes de mémoire ou d'attention visant à sélectionner les informations les plus pertinentes issues du passé.
— **Détection du début d'action (Detection of Action Start)** Abordée en sous-section 2.4.4, cette approche vise à détecter en ligne les instants où une nouvelle action débute, permettant ainsi une segmentation temporelle plus explicite.
— **Anticipation pour la classification (Anticipation for Classification)** Décrite en sous-section 2.4.5, cette approche vise à prédire certaines caractéristiques futures afin d'améliorer la reconnaissance de l'action en cours, tout en respectant la contrainte causale.

Dans les sections suivantes, nous détaillerons ces catégories en mettant en évidence les architectures, stratégies d'apprentissage et métriques spécifiques associées à l'évaluation des performances de ces modèles en inférence en ligne.

2.4.1 Schéma général des approches en ligne

Malgré la diversité des architectures et des stratégies d'apprentissage, la majorité des méthodes de reconnaissance d'actions en ligne suivent un schéma de traitement commun organisé en trois étapes successives : **(1) l'extraction de caractéristiques**, **(2) l'agrégation temporelle** et **(3) la prédiction**. Ce pipeline constitue la base de la plupart des approches contemporaines, qui reposent sur des données visuelles, squelettiques ou multimodales.

2.4.1.1 Extraction de caractéristiques

Chaque frame I_t issue du flux vidéo est d'abord convertie en un vecteur de caractéristiques $\mathbf{x}_t = \phi(I_t)$ par un module d'encodage. Celui-ci peut être implémenté à l'aide d'un réseau convolutionnel préentraîné, tel que I3D [137] ou C3D [143], voire par des variantes plus légères adaptées à l'inférence en temps réel, telles que MobileNet [144] ou EfficientNet [145].

Pour les données squelettiques, les *Graph Convolutional Networks* (GCN) constituent le choix dominant, notamment ST-GCN [139], MS-G3D [146] ou CTR-GCN [147], qui modélisent explicitement la structure spatiale du corps humain à travers un graphe d'articulations.

L'objectif de cette étape est de transformer chaque frame en une représentation compacte, discriminante et adaptée à la modélisation temporelle ultérieure.

2.4.1.2 Agrégation temporelle

Une fois les représentations instantanées extraites, l'information doit être agrégée au fil du temps afin de capturer la dynamique du mouvement humain. Soit $\mathbf{X}_t \in \mathbb{R}^d$ la représentation des observations à l'instant t, extraite à partir des données visuelles ou squelettiques, et $\hat{\mathbf{Y}}_t$ la prédiction de l'action associée à cet instant. L'agrégation temporelle vise alors à exploiter une ou plusieurs observations passées pour produire une prédiction plus stable et plus cohérente. Cette agrégation temporelle peut être réalisée selon deux grands principes.

— **Méthodes à fenêtre glissante (sliding window)** : La prédiction à l'instant courant repose sur une séquence de L observations consécutives, formant une fenêtre temporelle de taille fixe. Un modèle noté $f(\cdot)$, prend en entrée la séquence $\mathbf{X}_{t-L+1:t} = \{\mathbf{X}_{t-L+1}, \ldots, \mathbf{X}_t\}$ et produit la prédiction :

$$\hat{\mathbf{Y}}_t = f(\mathbf{X}_{t-L+1:t}). \tag{2.1}$$

La fenêtre se déplace progressivement dans le temps, ce qui permet une analyse locale et réactive du flux vidéo, tout en limitant la dépendance à des observations trop anciennes.

— **Méthodes à buffer historique (history buffer)** : Ces approches reposent sur une représentation interne $\mathbf{H}_t$, qui résume l'historique des observations passées jusqu'à l'instant t. Cet état latent est mis à jour de manière incrémentale à partir de l'état précédent $\mathbf{H}_{t-1}$ et de l'observation courante $\mathbf{X}_t$ au moyen d'une fonction de transition $g(\cdot)$:

$$\mathbf{H}_t = g(\mathbf{H}_{t-1}, \mathbf{X}_t). \tag{2.2}$$

L'état ainsi obtenu est ensuite utilisé pour prédire l'action courante via une couche de sortie linéaire paramétrée par une matrice de poids $\mathbf{W}_o$ et un biais $\mathbf{b}_o$, suivie d'une fonction de normalisation de type Softmax :

$$\hat{\mathbf{Y}}_t = \text{Softmax}(\mathbf{W}_o\mathbf{H}_t + \mathbf{b}_o). \tag{2.3}$$

Les architectures récurrentes telles que les RNN, LSTM ou GRU, ainsi que certaines variantes de transformers unidirectionnels, sont couramment utilisées pour implémenter ce type de mémoire temporelle compressée.

Ainsi, qu'elle repose sur une fenêtre temporelle explicite ou sur une mémoire incrémentale, l'agrégation temporelle joue un rôle central dans la modélisation des dépendances temporelles et dans la stabilisation des prédictions produites en continu.

2.4.1.3 Prédiction

Enfin, le module de prédiction attribue une étiquette d'action à la frame courante à partir de la représentation temporelle agrégée. Dans un cadre de classification supervisée multi-classes, la prédiction à l'instant t est représentée par un vecteur de probabilités $\hat{\mathbf{Y}}_t$. L'apprentissage est généralement supervisé par une fonction de coût de type entropie croisée, calculée sur une séquence de longueur T :

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} \mathbf{Y}_t^{\top} \log(\hat{\mathbf{Y}}_t), \tag{2.4}$$

où $\mathbf{Y}_t$ désigne le vecteur de vérité terrain à l'instant t et $\hat{\mathbf{Y}}_t$ la distribution de probabilités prédite pour les différentes classes d'actions. Certaines variantes introduisent des mécanismes de régularisation temporelle, de pondération de la confiance ou d'adaptation dynamique du seuil de décision afin d'améliorer la stabilité des prédictions produites en continu.

En résumé, ce schéma en trois étapes constitue la base commune à l'ensemble des approches de détection d'actions en ligne. Les méthodes diffèrent principalement dans la manière dont elles implémentent l'agrégation temporelle et la gestion de la mémoire, ou dans l'intégration de modules supplémentaires d'attention, d'anticipation ou de détection explicite des transitions d'action. Les sections suivantes détaillent ces familles de méthodes et leurs spécificités respectives.

2.4.2 Détection d'actions image par image avec historique (Historical Per-frame Action Detection)

Les approches de *détection d'actions image par image avec historique* constituent la base conceptuelle la plus simple et la plus intuitive pour la reconnaissance d'actions en ligne. Leur principe repose sur la prédiction d'une étiquette d'action à chaque frame entrante, en exploitant uniquement les observations passées et la frame courante. Autrement dit, à l'instant t, le

modèle estime la probabilité de l'action en cours à partir d'un **historique temporel** ou *buffer* contenant les représentations récentes du flux (voir figure 2.8.(a)).Soit $\mathbf{X}_t$ la représentation extraite à l'instant t, la prédiction peut alors s'écrire :

$$\hat{\mathbf{Y}}_t = f(\mathbf{X}_{1:t}) \quad \text{ou, plus souvent,} \quad \hat{\mathbf{Y}}_t = f(\mathbf{X}_{t-L+1:t}), \tag{2.5}$$

où $\mathbf{X}_{1:t}$ désigne l'ensemble des observations passées jusqu'à l'instant t, et L représente la longueur effective de la mémoire temporelle utilisée (fixe ou variable). Le traitement est réalisé de manière séquentielle, la prédiction $\hat{\mathbf{Y}}_t$ étant mise à jour à chaque nouvelle frame.

Plusieurs travaux récents mettent en évidence l'efficacité de ce paradigme. E2E-LOAD (*End-to-End Long-form Online Action Detection*, [148]) s'attaque aux défis de la détection d'actions en ligne, notamment la compréhension à long terme des séquences de données et l'efficacité en termes de temps d'inférence. Pour cela, les auteurs proposent une architecture de bout en bout intégrant un module de fusion de représentations à court et à long terme, assurant une meilleure contextualisation temporelle tout en préservant l'inférence en temps réel. De plus, OnlineTAS (*Online Temporal Action Segmentation*, [149]) transpose ces principes au cadre de la segmentation en ligne, en intégrant une mémoire hiérarchique adaptative capable de gérer simultanément les contextes à court et à long terme. Cette méthode réduit la sur-segmentation en utilisant des MS-TCN intégrant des blocs GRU. Leurs expériences montrent une meilleure cohérence temporelle, mais révèlent encore deux limites majeures concernant le retard de détection du début des actions, dû aux ambiguïtés de la transition des actions, et, dans les cas où il y avait de la sur-segmentation , c'était lié à des prédictions incorrectes mais à un taux de confiance élevé.

2.4.3 Discrimination de l'information pertinente (Information Discrimination)

Cette approche vise à modéliser la relation entre les images actuelles et passées afin d'accumuler uniquement les informations pertinentes pour la prédiction en ligne.

À chaque instant t, le modèle évalue la contribution informationnelle des images passées contenues dans une **fenêtre temporelle glissante de taille** L, définie par $\{\mathbf{X}_{t-L+1}, \ldots, \mathbf{X}_{t-1}\}$. Soit $f(\cdot)$ une fonction d'extraction de caractéristiques appliquée aux images d'entrée. Pour chaque image passée $\mathbf{X}_i$ de la fenêtre ($i \in \{t-L+1, \ldots, t-1\}$), le score de pertinence est défini par :

$$\alpha_i^t = \sigma(\mathbf{W}(f(\mathbf{X}_t) \odot f(\mathbf{X}_i))), \tag{2.6}$$

où α_i^t mesure l'importance de l'image $\mathbf{X}_i$ relativement à l'image courante $\mathbf{X}_t$, $\odot$ désigne le produit élément par élément, $\mathbf{W}$ est une matrice de poids apprise, et $\sigma(\cdot)$ une fonction d'activation. La mémoire sélective est alors calculée par :

$$\mathbf{H}_t = \sum_{i=t-L+1}^{t-1} \alpha_i^t f(\mathbf{X}_i), \tag{2.7}$$

où $\mathbf{H}_t$ représente l'état de la mémoire agrégée à l'instant t. La taille de la mémoire reste constante : lorsque t augmente, la fenêtre glisse sans accumulation d'éléments supplémentaires. Ce qui permet de conserver uniquement les représentations les plus utiles (voir figure 2.8.(b)). Cette stratégie exploite efficacement un historique temporel tout en réduisant la redondance et la charge mémoire. Cependant, son coût computationnel demeure élevé, car à chaque instant t, l'image courante $\mathbf{X}_t$ est comparée explicitement à l'ensemble des images contenues dans la fenêtre glissante de taille L.

Eun *et al.* [150] introduisent l'*Information Discrimination Unit* (IDU), une évolution du *Gated Recurrent Unit* (GRU), qui apprend à filtrer les informations non pertinentes en adaptant dynamiquement les portes de mise à jour en fonction de la similarité entre l'image courante et la mémoire passée. Les auteurs soulignent toutefois que la qualité de la discrimination dépend fortement de la robustesse des représentations extraites et que le mécanisme demeure sensible au bruit visuel. Dans la continuité de cette idée, Chen *et al.* [151] proposent *GateHUB*, qui renforce la sélectivité informationnelle en combinant un mécanisme de *gated cross-attention* pour pondérer les images historiques en fonction de leur contribution à la prédiction actuelle, ainsi qu'un module de suppression de fond pour limiter les faux positifs dus à des arrière-plans similaires aux actions. Bien que cette approche améliore la pertinence du contexte temporel, elle demeure limitée par la brièveté de l'historique exploité, ce qui peut nuire à la détection d'actions rares ou espacées dans le temps, ainsi que par la difficulté à modéliser avec précision les actions à mouvement lent, dont la dynamique diffère de celle des gestes courants.

2.4.4 Détection du début d'action (Detection of Action Start)

La détection du début d'action (*Online Detection of Action Start*, ODAS) vise à identifier, dès qu'elle a lieu, la transition entre le fond et l'action, ainsi que la catégorie correspondante (voir figure 2.8.(c)). À chaque instant t, le modèle traite les images successives $\mathbf{X}_{1:t} = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}$ et estime, pour chaque classe d'action k, la probabilité que l'image

courante appartienne à cette classe :

$$s_t^{(k)} = P(\mathbf{Y}_t = k \mid \mathbf{X}_{1:t}), \tag{2.8}$$

où $s_t^{(k)}$ désigne la probabilité prédite pour la classe d'action k à l'instant t, et $\mathbf{Y}_t$ désigne l'étiquette d'action associée à l'instant t. En parallèle, le modèle évalue la probabilité qu'un début d'action survienne à cet instant :

$$p_t = P(\text{start@}t \mid \{s_\tau^{(k)}\}_{\tau \leq t,k}), \tag{2.9}$$

où p_t exprime la probabilité qu'une transition de la classe "Pas d'action" vers un état d'action se produise à l'instant t, en s'appuyant sur l'ensemble des scores de classification passés $s_\tau^{(k)}$. La décision finale combine ces deux estimations :

$$\pi_t^{(k)} = p_t \cdot s_t^{(k)}, \tag{2.10}$$

de sorte que $\pi_t^{(k)}$ représente la probabilité d'apparition d'un début d'action de catégorie k au temps t. Contrairement aux approches de reconnaissance ou de segmentation d'actions qui s'appuient sur de longues fenêtres temporelles pour modéliser le contexte global, la détection d'actions en ligne (ODAS) se concentre sur la transition immédiate entre la classe "Pas d'action" et une classe d'action, ce qui nécessite une fenêtre temporelle courte, généralement limitée à quelques images précédant l'instant t, et permet ainsi une latence de décision très faible.

Gao et al. [152] proposent StartNet. Leur méthode fusionne les sorties d'un réseau de classification d'action et d'un réseau de localisation entraîné par apprentissage par renforcement, afin d'optimiser une récompense favorisant la réactivité tout en limitant les fausses alertes. Les auteurs montrent que cette conception améliore la stabilité et la rapidité du modèle, mais qu'elle demeure fortement dépendante du réseau de classification, qui peine à distinguer les phases de transition entre le fond et le début de l'action. De leur côté, Hu et al. [153] estiment que les limites des approches existantes proviennent principalement du déséquilibre des données et des transitions floues entre les actions. Ils proposent une représentation (distribution-aware) des frontières d'action, modélisées comme une distribution continue plutôt qu'un point unique, ce qui permet au modèle d'apprendre une transition plus réaliste et progressive du fond à l'action.

2.4.5 Anticipation pour la classification (Anticipation for Classification)

Les approches d'anticipation d'action (*Anticipation for Classification*) visent à améliorer la reconnaissance d'actions en ligne en compensant l'absence d'informations futures inhérente à ce cadre. Elle repose sur la prédiction des représentations latentes des images à venir à partir de l'historique déjà observé, afin d'enrichir le contexte temporel utilisé pour classer l'action en cours (voir figure 2.8.(d)).

Formellement, à chaque instant t, le modèle encode les images passées $\mathbf{X}_{1:t} = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}$ en une représentation latente :

$$\mathbf{H}_t = f(\mathbf{X}_{1:t}), \tag{2.11}$$

où $f(\cdot)$ désigne le module d'encodage et $\mathbf{H}_t$ représente l'état latent à l'instant t. Le modèle anticipe ensuite les représentations futures pour les K instants suivants :

$$\hat{\mathbf{H}}_{t+i} = g(\mathbf{H}_t), \quad i = 1, \ldots, K, \tag{2.12}$$

où $g(\cdot)$ désigne le module d'anticipation et $\hat{\mathbf{H}}_{t+i}$ correspond à la représentation latente prédite à l'instant $t+i$. Les états futurs prédits sont ensuite combinés avec la mémoire courante afin de produire une représentation enrichie :

$$\mathbf{Z}_t = \text{Fuse}(\mathbf{H}_t, \hat{\mathbf{H}}_{t+1}, \ldots, \hat{\mathbf{H}}_{t+K}), \tag{2.13}$$

où Fuse$(\cdot)$ désigne une opération de fusion (par exemple concaténation ou somme pondérée). À partir de cette représentation enrichie, la probabilité de classe est calculée :

$$P(\mathbf{Y}_t = k \mid \mathbf{X}_{1:t}) = \text{Softmax}(\mathbf{W}\mathbf{Z}_t + b), \tag{2.14}$$

où W et b correspondent respectivement à la matrice de poids et au biais de la couche de classification finale, et $\mathbf{Y}_t$ désigne l'étiquette d'action prédite pour l'instant t. En anticipant les dynamiques d'action à court terme, cette approche permet d'atténuer la vision partielle propice au traitement en ligne et d'améliorer la précision de la classification, notamment dans les contextes où les signaux visuels sont faibles ou retardés. Toutefois, cette capacité prédictive s'accompagne d'un coût de calcul élevé lié à la génération et à la fusion de représentations futures, ce qui accroît la complexité du modèle.

Xu et al. [154] proposent les *Temporal Recurrent Networks* (TRN), basés sur un simple réseau récurrent afin de réduire la complexité du modèle et le temps de prédiction. En anticipant les caractéristiques futures afin d'affiner la prédiction actuelle, cette approche améliore la

cohérence temporelle dans des scénarios à faible variation temporelle, mais reste limitée dans des environnements à fortes variations, tels que les scènes industrielles. Afin de pallier cela, Yoon et al. [155] introduisent *Future-GAN*, qui s'appuie sur un générateur adversarial pour produire des représentations futures plus réalistes et plus diverses. Cette stratégie renforce la robustesse du modèle face à la variabilité des séquences, mais entraîne un coût de calcul élevé et une instabilité d'entraînement. Enfin, Lee et al. [156] proposent une approche composée de deux modules complémentaires, l'un, dédié au traitement des informations passées (IDU), et l'autre, appelé *Information Integration Unit* (IIU), chargé d'anticiper les caractéristiques futures. Les auteurs observent que l'ajout d'un module d'anticipation améliore significativement la détection précoce du début d'action par rapport aux méthodes reposant uniquement sur les observations passées.

2.4.6 Synthèse

La majorité des approches d'action en ligne présentées ont été développées et évaluées sur des bases de données centrées sur des activités quotidiennes, telles que *TVSeries* [157], *THUMOS14* [89] ou *HDD* [158] pour la conduite autonome. Ces séquences comportent généralement une seule action dominante par segment, séparée de longues périodes de fond relativement homogènes. Ce cadre diffère profondément des environnements industriels, et plus particulièrement des scénarios d'assemblage, où plusieurs actions sont réalisées successivement ou en parallèle au sein d'une même séquence, souvent avec des transitions rapides et des interactions complexes entre l'opérateur, les outils et les pièces manipulés.

Les approches telles que les *Temporal Recurrent Networks* [154] ou *StartNet* [152], testées dans des contextes tels que la conduite autonome ou des activités domestiques, profitent de la nature progressive et prévisible des séquences : les actions évoluent lentement (par exemple, approcher un carrefour ou s'arrêter à un feu rouge), et les transitions entre événements sont graduelles. Cette continuité temporelle facilite l'anticipation et la détection des actions, les modèles disposant de signaux visuels cohérents et réguliers pour prédire les événements à venir.

À l'inverse, dans les scénarios industriels, les activités d'assemblage se caractérisent par une forte variabilité temporelle et gestuelle : les transitions entre actions sont brusques, les gestes sont courts et précis, et les contextes visuels évoluent rapidement. Les modèles conçus pour des séquences lentes et homogènes montrent alors leurs limites, notamment pour la

détection des débuts d'action ou la modélisation des transitions rapides entre gestes. De plus, la plupart de ces méthodes reposent sur des hypothèses simplificatrices, telles que la présence d'un seul acteur ou l'absence d'interactions complexes avec des objets ou des outils, ce qui réduit leur capacité à généraliser aux environnements collaboratifs de l'industrie 5.0.

Ces limites ouvrent toutefois de nouvelles opportunités de recherche et d'application. L'adaptation des approches d'action en ligne au contexte industriel pourrait bénéficier de la prise en compte de signaux multimodaux (vision, capteurs inertiels, suivi d'objets). De plus, la disponibilité récente de jeux de données dédiés aux environnements industriels tels que *InHARD*[3], *HA4M*[65] et *IKEA-ASM*[4] (voir la section 1.4.1) rend désormais possible le développement de modèles d'apprentissage profond performants, spécifiquement adaptés à la détection d'actions en ligne dans des tâches d'assemblage. Néanmoins, malgré ces avancées, la détection d'action en ligne en milieu industriel demeure une tâche complexe, soumise à de fortes contraintes liées à la variabilité des gestes, à la dynamique des scènes et aux exigences de réactivité en temps réel. Les défis spécifiques à ce contexte, ainsi que la manière dont notre approche proposée y répond, sont présentés et discutés en détails dans le chapitre 5. Dans la continuité, nous présentons ci-après les métriques d'évaluation utilisées pour mesurer la performance et la réactivité des approches de détection d'actions en contexte industriel.

2.5 Métriques d'évaluation

L'évaluation des performances des approches d'analyse du mouvement humain (HMA) repose sur l'utilisation de métriques adaptées à la tâche à évaluer (reconnaissance, segmentation, détection, etc.). Dans un premier temps, la sous-section 2.5.1 propose un aperçu des principales métriques de l'état de l'art, regroupées par catégories et synthétisées dans un tableau récapitulatif. La sous-section 2.5.2 détaille ensuite les mesures que nous avons retenues pour nos expérimentations, en précisant leurs définitions, leurs formules de calcul et leurs illustrations associées. Enfin, la sous-section 2.5.3 présente une synthèse des points clés et justifie le choix final des métriques retenues dans notre étude.

2.5.1 Métriques dans l'état de l'art

L'évaluation des approches d'analyse du mouvement humain (HMA) s'appuie sur diverses métriques, chacune mettant en lumière des aspects spécifiques de la performance et facilitant

l'identification des axes d'amélioration. Le tableau 2.2 présente une synthèse des principales métriques rencontrées dans l'état de l'art. Les métriques les plus répandues (MoF, F1-score, mAP, FLOPs, Edit-score, IoU) sont détaillées ci-dessus, tandis que les autres sont rapportées à titre comparatif afin d'offrir une vue exhaustive des pratiques d'évaluation.

La *"Mean over Frame (MoF)"*, ou *"Accuracy"*, est la métrique la plus couramment utilisée dans la littérature étudiée. Cependant, de nombreux jeux de données souffrent d'un fort déséquilibre de classes, certaines catégories étant largement surreprésentées par rapport à d'autres. Dans ce contexte, un score de MoF élevé (parfois proche de 100%) ne reflète pas toujours fidèlement les performances réelles du modèle. C'est pourquoi l'utilisation du F1-score, qui combine la précision (*precision*) et le rappel (*recall*) en une seule mesure harmonisée, constitue une évaluation plus équilibrée, en particulier pour des données déséquilibrées.

Lorsqu'il s'agit d'évaluer des algorithmes de reconnaissance d'actions sur des séquences non segmentées (*untrimmed*), il est essentiel d'intégrer des métriques adaptées à la détection temporelle et spatio-temporelle. Pour la détection temporelle, la *Mean Average Precision* (mAP) est largement utilisée. Elle consiste à calculer la précision moyenne (*Average Precision*, AP) pour chaque classe d'action à différents seuils d'intersection sur union (IoU), souvent sous forme de multi-seuils (ex. mAP@$\{0.5 : 0.05 : 0.95\}$) pour une évaluation plus robuste. De plus, des métriques telles que l'Edit-score ou F1@$\{50\}$ sont particulièrement adaptées aux tâches de segmentation temporelle, car elles tiennent compte de l'ordre et de la durée des segments prédits, ce qui pénalise notamment la sur-segmentation définie dans la section 2.3.2.1.

En complément, des métriques de régression sont employées pour évaluer l'estimation de trajectoires humaines ou robotiques, réduisant ainsi les erreurs de positionnement et améliorant la sécurité dans des environnements collaboratifs [38]. Ces métriques jouent un rôle clé dans les applications liées à l'ergonomie et à la prévention des collisions entre opérateurs et robots.

Le coût de calcul constitue un autre critère essentiel d'évaluation. Il peut être mesuré par le nombre d'opérations en virgule flottante (*Floating Point Operations*, FLOPs), indicateur théorique de complexité, mais aussi par des mesures réelles telles que le temps de traitement, qui regroupe le temps d'inférence, le délai de détection et le post-traitement [25]. Il est important de noter que, contrairement aux FLOPs, le temps de traitement dépend fortement du matériel utilisé (CPU, GPU, optimisation logicielle).

En résumé, l'analyse du tableau 2.1 montre que les métriques les plus prédominantes pour l'évaluation sur les jeux de données de référence sont la *MoF*, le F1-score et la mAP à différents seuils d'IoU, tandis que des métriques comme l'Edit-score, F1@{50} et certaines mesures de régression sont privilégiées dans des contextes spécifiques tels que la segmentation et la détection d'actions en milieu industriel.

TABLEAU 2.2 – Métriques utilisées pour l'évaluation des approches d'analyse du mouvement humain.
TP : Vrai Positif, TN : Vrai Négatif, FP : Faux Positif, FN : Faux Négatif, P : étiquettes prédites, GT : Vérité Terrain, IoU : Intersection over Union, $D(.,.)$: distance de Levenshtein [159]

, L : durée de l'action, l : intervalle temporel entre le début de l'action réelle et la frame prédite, p : point clé prédit, q : position réelle du point clé, d : distance, $Recall_i$: rappel au seuil i, ADD : distance moyenne entre les sommets prédits et réels et le centroïde de la boîte englobante de l'objet, UWB : Ultra-Wide-Band, T_i : temps d'inférence, D_d : délai de détection, T_{pp} : temps de post-traitement, P^i_{ref} : sommet de référence, P^i_{pre} : sommet prédit, x : valeur réelle de la i-ème observation, y : valeur prédite de la i-ème observation, n : nombre total d'observations, k : nombre total de points clés, v : nombre total de sommets.

	Métrique	Formule	Objectif	Référence
Métriques générales communes	Mean-over-Frames (MoF)	$\frac{TP+TN}{TP+TN+FP+FN} \times 100$	Détermine à quelle fréquence les prédictions du modèle correspondent aux étiquettes réelles.	[3], [85], [86] [15], [45], [55], [59], [61], [62] [18], [54], [20], [21], [132] [119], [122]
	Precision	$TP/(TP+FP)$	Se concentre sur la précision des prédictions positives. Aide à évaluer dans quelle mesure le modèle est précis et fiable lorsqu'il prédit un résultat positif.	[15], [63] [25], [52]
	Recall	$TP/(TP+FN)$	Quantifie la capacité du modèle à trouver tous les échantillons pertinents sans en manquer, en minimisant le nombre de faux négatifs. Mesure combien d'éléments pertinents ont été détectés.	[59], [63] [25], [52]
	F1-score	$\frac{2*Precision*Recall}{Precision+Recall} \times 100$	Évaluer les modèles de classification comportant deux classes ou plus, en particulier lorsque les classes sont déséquilibrées.	[19], [63], [119] [25], [43], [84] [14], [60]
Untrimmed Métriques	IoU	$(P \cap GT)/(P \cup GT)$	Évaluer la qualité des algorithmes de localisation ou de segmentation d'objets. Mesure dans quelle mesure l'algorithme identifie et délimite avec précision les objets d'intérêt dans une image (spatial) ou un événement dans une séquence (temporel).	[3], [85], [119] [20], [86]
	Mean Average Precision (mAP)	$\frac{1}{n}\sum^{n}_{Recall_i} Precision(Recall_i)$	Représente les performances globales d'un modèle de détection d'objets ou de segmentation d'instances, en tenant compte à la fois de la précision et du rappel à différents seuils de confiance.	[58], [59]
	Edit-score	$\left(1 - \frac{D(P,Y)}{\max(\lvert P\rvert,\lvert Y\rvert)}\right) \times 100$	Mesure la fidélité de l'ordonnancement temporel des segments d'actions prédits P par rapport à la vérité terrain Y.	[60]
	Pourcentage de Prédictions Correctes (PCP)	$\frac{1}{k}\sum_{i=1}^{k} d(p_i, q_i) < t$	Indique la probabilité qu'un point clé détecté p se situe dans un seuil de distance t, donné en pixels, par rapport à sa position réelle q. Utilisé pour évaluer la précision.	[48]
	ADD	$\frac{1}{v}\sum_{i=1}^{v} \lVert P^i_{pre} - P^i_{ref}\rVert_2$	Estime la précision de la détection d'objets pour l'estimation de pose en 6D.	[38]
Métriques de Régression	Erreur de rotation (E_{rot}) Erreur de translation (E_{tra})	$\angle(\overrightarrow{P}, \overrightarrow{GT})$ $\lVert P - GT\rVert$	Estimer l'erreur d'orientation de l'objet. Estimer l'erreur de localisation de l'objet.	[38]
	Mean Absolute Error (MAE)	$(\frac{1}{n})\sum_{i=1}^{n} \lvert y_i - x_i\rvert$	Exprime l'erreur moyenne de prédiction du modèle en traitant toutes les erreurs de manière égale.	[29]
	Root Mean Squared Error (RMSE)	$\sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2}$	Exprime l'erreur moyenne de prédiction du modèle en pénalisant davantage les erreurs importantes.	[23], [29], [44]
	Geometric Dilution Of Precision (GDOP)	$RMSE_{loc}/RMSE_{range}$	Estime l'erreur causée par la position relative d'autres capteurs tels que le GPS ou l'UWB.	[23]
Time Cost	Latence	$Dd = l/L$	Évaluer le délai nécessaire pour détecter l'action.	[119]
	Temps de traitement	$T_i + D_d + T_{pp}$	Évaluer et optimiser l'efficacité et les performances du système.	[25], [55]

2.5.2 Métriques retenues

Pour l'évaluation, nous retenons trois métriques complémentaires : le *Mean-over-Frames* (MoF) pour quantifier la précision image par image, l'*Edit-score* pour évaluer la cohérence des segments prédits et le *F1-score* avec différents seuils d'IoU pour mesurer la qualité de leur recouvrement temporel. Ces trois métriques sont détaillées respectivement dans les sous-sections 2.5.2.1, 2.5.2.2 et 2.5.2.3.

2.5.2.1 Mean-over-Frames (MoF)

Le *Mean-over-Frames* (MoF) évalue la proportion d'images correctement classifiées sur l'ensemble de la séquence, sans tenir compte de la structure temporelle des actions. Cette métrique fournit une indication globale de la performance en classification image par image (*frame-wise*). L'équation ci-dessous en donne la définition formelle et une illustration de ce calcul est présentée en figure 2.9.

$$\text{MoF} = \frac{\text{Nombre d'images correctement prédites (TP+TN)}}{\text{Nombre total d'images (TP+TN+FP+FN)}} \times 100 \tag{2.15}$$

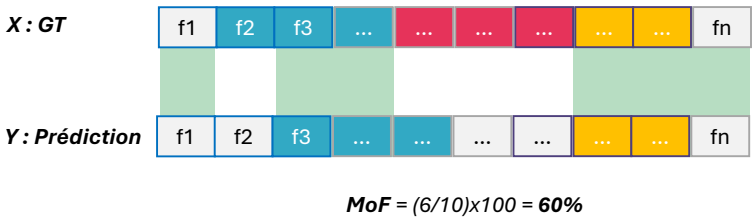


FIGURE 2.9 – Illustration du calcul du MoF. f_n : frame ou image d'indice n.

2.5.2.2 Edit-score

L'*Edit-score* mesure la fidélité de l'ordonnancement temporel des segments d'actions prédits par rapport à la vérité du terrain. Il est basé sur la distance de Levenshtein normalisée, qui quantifie le nombre minimal d'opérations (insertions, suppressions ou substitutions) nécessaires pour transformer la séquence prédite P en la séquence réelle Y et est calculée comme suit :

$$\text{Edit} = \left(1 - \frac{D(P,Y)}{\max(|P|,|Y|)}\right) \times 100 \tag{2.16}$$

où $D(P, Y)$ est la distance de Levenshtein calculée selon [159], et $|P|$, $|Y|$ sont les longueurs respectives des deux séquences. Comme illustré en figure 2.10, cette métrique permet d'évaluer la cohérence de l'ordre des actions ainsi que les erreurs structurelles, notamment la sur-segmentation (découpage excessif) et la sous-segmentation (fusion d'actions distinctes).

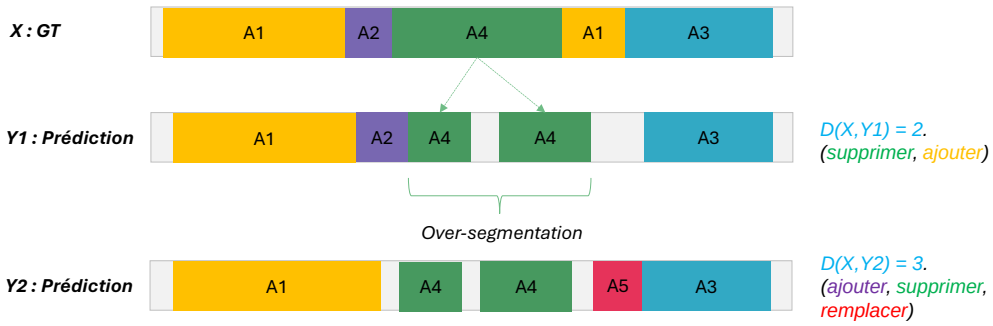


FIGURE 2.10 – Illustration de l'Edit-score : comparaison entre la séquence temporelle prédite et la vérité du terrain, mesurée par la distance de Levenshtein normalisée, afin de quantifier les erreurs d'ordre et de segmentation.

2.5.2.3 F1@IoU

Le *F1@IoU* est une métrique d'évaluation au niveau des blocs d'image (segments), utilisée pour mesurer la qualité globale de la segmentation des actions d'une séquence donnée, en combinant précision et rappel. Elle repose sur un seuil d'intersection sur union (IoU) qui définit le degré de recouvrement requis entre un segment prédit et la vérité du terrain pour qu'il soit considéré comme correct. Ce seuil peut être fixé, par exemple, à 0,50 (*F1@50*), mais d'autres valeurs peuvent être retenues en fonction du niveau de tolérance souhaité. La précision (*Precision*) et le rappel (*Recall*) sont calculés comme suit :

$$\text{Précision} = \frac{TP}{TP + FP}, \quad \text{Rappel} = \frac{TP}{TP + FN} \tag{2.17}$$

où TP (*vrais positifs*) correspond aux segments correctement détectés, FP (*faux positifs*) aux segments détectés à tort, et FN (*faux négatifs*) aux segments manqués. Le score F1 est ensuite obtenu comme la moyenne harmonique entre la précision et le rappel :

$$\text{F1@IoU} = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} \times 100 \tag{2.18}$$

Contrairement à une évaluation image par image, le *F1@IoU* prend en compte la qualité de la localisation des segments d'actions. Il pénalise les segments mal localisés (IoU inférieur

au seuil choisi) ; les segments manquants (FN) et la sur-segmentation (FP). La figure 2.11 illustre ces propos.



FIGURE 2.11 – Illustration du F1 score avec un IoU de 50% (la moitier de la taille des segments du GT) sur La séquence "Pred". GT : la sequence segmentée de verité terrain. Pred : la sequence segmentée predite par le model.

2.5.3 Synthèse

En résumé, cette section a présenté trois métriques principales utilisées pour l'évaluation expérimentale : le *MoF*, l'*Edit-score* et le *F1@IoU*, chacune offrant une perspective différente sur la performance des modèles. Elles sont particulièrement adaptées à notre problématique car elles permettent de combiner une mesure globale de précision avec une évaluation fine de la segmentation temporelle. Comme résumé dans le tableau 2.2, d'autres métriques sont également utilisées dans l'état de l'art pour des contextes ou des objectifs spécifiques (détection spatio-temporelle, estimation de la pose, analyse de trajectoires, etc.), mais notre choix se concentre sur celles offrant le meilleur compromis entre pertinence et comparabilité avec les travaux existants.

2.6 Discussion et tendances

L'étude des travaux de recherche consacrés à l'analyse du mouvement humain en contexte industriel met en évidence plusieurs observations et opportunités d'amélioration. Les principaux enseignements peuvent ainsi être regroupés selon les axes suivants.

2.6.1 Segmentation d'actions sur données non découpées

Alors que la segmentation d'actions hors ligne a démontré d'excellentes performances, comme dans [45], la segmentation en ligne présente encore des résultats plus modestes, soulignant la nécessité d'avancées supplémentaires dans la segmentation de séquences non découpées [60]. Avec la disponibilité croissante de jeux de données non découpés, il devient urgent de développer des méthodologies capables de segmenter efficacement et avec précision des

séquences de durées variables.

Les premiers travaux [160] s'appuyaient principalement sur le suivi d'objets et l'utilisation de fenêtres glissantes pour localiser les positions des actions et détecter les images de début et de fin d'une séquence. Pour relever ce défi, une solution possible consiste à développer des méthodes performantes capables de segmenter les séquences indépendamment des étiquettes d'actions considérées. Une telle approche peut s'appuyer sur l'utilisation de réseaux convolutifs temporels (TCN) [161], qui ont montré un fort potentiel dans ce domaine.

Par ailleurs, pour la localisation des séquences et une meilleure performance en reconnaissance continue d'actions, certains travaux, tels que [162], ont exploré des techniques d'apprentissage par transfert visant à adapter des modèles pré-entraînés sur des jeux de données généraux à des contextes industriels spécifiques. Ces méthodes exploitent des modèles pré-entraînés sur de grands jeux de données génériques, puis les affinent sur de plus petits jeux de données propres au domaine industriel. Cela permet de pallier la rareté des données et d'améliorer les capacités de généralisation. Parmi les exemples notables, on peut citer les modèles TSM et TSN [116], pré-entraînés sur le jeu de données Kinetics-400, qui ont obtenu des résultats prometteurs sur les ensembles InHARD et MECCANO, respectivement.

Dans ce contexte, nos travaux se positionnent d'abord sur la segmentation et la reconnaissance hors ligne de séquences longues et non découpées, puis s'étendent vers une approche en ligne. Cette démarche s'appuie sur des architectures de deep learning spatio-temporelles basées sur les données squelettiques, ainsi que sur des stratégies d'apprentissage par transfert adaptées au contexte industriel.

2.6.2 Approches

Dans le contexte industriel, l'objectif est de déployer les systèmes développés pour la reconnaissance d'actions au sein d'environnements de production embarqués. Pour être considérés comme efficaces dans de tels environnements, ces systèmes doivent répondre à plusieurs critères. Premièrement, ils doivent présenter de hautes performances, caractérisées par une bonne précision dans la reconnaissance des actions. Deuxièmement, ils doivent être rapides, capables d'assurer un traitement en temps réel et de fournir des résultats immédiats. De plus, ils doivent pouvoir traiter des données non découpées et fonctionner efficacement avec un nombre limité d'échantillons d'entraînement déjà annotés.

Un autre critère important est la capacité d'apprentissage continu, permettant l'intégration de nouvelles actions et la mise à jour progressive du modèle de reconnaissance. Ce type d'application implique un fort volet de modélisation spatio-temporelle, avec un accent particulier sur l'analyse de séries temporelles. L'interprétabilité des résultats produits par le système constitue également un enjeu majeur, garantissant une compréhension claire des mécanismes de décision.

Concernant les techniques existantes, plusieurs représentations méritent d'être considérées. Les approches basées sur les graphes offrent une modélisation adaptée des structures squelettiques [163], tandis que les *transformers* se sont révélés efficaces pour la représentation temporelle des données [164]. L'exploration d'une combinaison de graphes et de transformers pourrait constituer une piste de recherche prometteuse [135].

Un autre aspect clé concerne la réduction du temps d'exécution par l'optimisation de l'utilisation mémoire, par exemple via des techniques d'apprentissage incrémental basées sur les graphes [165]. Ces avancées ne se limitent pas au seul domaine industriel, mais peuvent également bénéficier à d'autres champs d'application.

Enfin, la fusion de données multimodales au sein d'architectures d'apprentissage profond peut être réalisée selon différentes stratégies, telles que la fusion précoce, la fusion tardive ou encore l'utilisation de mécanismes d'attention permettant de mettre l'accent sur les modalités les plus informatives. L'intégration efficace de ces approches de fusion et de l'apprentissage profond permet d'atteindre les performances requises, de garantir un traitement en temps réel et d'améliorer la précision des tâches de reconnaissance d'actions, faisant de ces solutions des approches particulièrement prometteuses pour une mise en œuvre pratique [81].

Nos travaux se distinguent par une adaptation au milieu industriel, en intégrant des stratégies d'apprentissage, notamment la distillation de connaissances en ligne, ainsi qu'une modélisation des gestes d'assemblage fondée sur les données squelettiques, les graphes et des mécanismes d'attention, afin de concevoir des systèmes performants, interprétables et robustes.

2.7 Conclusion

Dans ce chapitre, nous avons présenté un panorama détaillé des approches d'analyse du mouvement humain dans un contexte industriel. Nous avons décrit les principales familles

d'approches (offline et online) ainsi que les architectures mobilisées (réseaux convolutifs, récurrents, transformeurs, graphes, etc.). Nous avons également discuté des métriques d'évaluation les plus utilisées et des défis actuels, notamment la segmentation d'actions dans des séquences non découpées (*untrimmed*), la gestion des données déséquilibrées et les enjeux liés aux environnements industriels réels.

De cette analyse, il ressort que la segmentation temporelle hors ligne demeure un défi majeur pour les séquences longues et complexes, spécifiquement dans le contexte industriel. Les données sont souvent des images, vidéos RGB ou de profondeur ; cependant, les données squelettiques apparaissent comme un support pertinent pour réduire la complexité visuelle et se concentrer sur la dynamique du mouvement humain, mais elles restent encore peu exploitées dans ce contexte.

Ces constats nous conduisent à proposer, dans le chapitre suivant, une première contribution méthodologique : le modèle **GSK-C2F (Graph Skeleton Coarse-to-Fine)**. Cette approche vise à répondre au besoin identifié d'une segmentation temporelle précise et robuste dans des séquences industrielles non segmentées, en exploitant une représentation par squelettes humains et une stratégie d'analyse hiérarchique du temps (*coarse-to-fine*). Le chapitre 3 détaille cette méthodologie, ses fondements théoriques, ainsi que son évaluation à partir de bases de données de référence.

Chapitre 3

Segmentation d'actions dans des vidéos non découpées à partir de données squelettiques

3.1 Introduction

Dans le contexte de l'Industrie 5.0, l'analyse des mouvements et la segmentation d'actions d'assemblage sont essentielles, car elles ouvrent la voie à des applications à forte valeur ajoutée. En reconnaissance hors ligne, elles permettent notamment la génération de rapports ergonomiques post-activité, l'analyse des postures de l'opérateur, l'identification de contraintes articulaires, la proposition de retours adaptés, l'incitation à des pauses lorsque nécessaire et la prévention des troubles musculosquelettiques. Ces applications illustrent l'intérêt de développer des méthodes capables d'analyser, reconnaître et segmenter de manière fiable les gestes humains en situation réelle pour l'industrie du futur.

Dans le chapitre précédent, nous avons mené une revue approfondie des approches récentes de reconnaissance et de segmentation d'actions humaines, en nous concentrant particulièrement sur leur application dans des contextes industriels. Bien que ces méthodes produisent des résultats convaincants dans des contextes simples ou du quotidien, leurs performances se dégradent fréquemment lorsqu'elles sont déployées dans des environnements industriels complexes tels que les chaînes d'assemblage (voir section 1.5). Parmi les principaux défis rencontrés figurent la grande finesse des mouvements opératoires, la forte similarité visuelle entre certains gestes, ainsi que la nécessité de segmenter avec précision des séquences longues et non tronquées. À cela s'ajoutent la présence fréquente d'occultations et la complexité des mouvements des opérateurs, qui rendent difficile la capture fiable des motifs de mouvement lorsqu'on utilise des modalités telles que les images RGB. Toutefois, il convient de noter que les algorithmes d'estimation de pose ont récemment connu des avancées significatives, permettant d'obtenir des données squelettiques plus précises, ce qui en fait une alternative prometteuse aux données RGB pour l'analyse du mouvement.

Afin de répondre à ces enjeux et de poser les bases de systèmes de segmentation plus avancés, adaptés aux exigences de l'Industrie 5.0, ce chapitre présente notre première contribution : une méthode de segmentation d'actions dans des séquences vidéo non segmentées d'assemblage industriel, fondée exclusivement sur les données squelettiques. Nous nous concentrons ici sur une exploration approfondie et systématique des représentations basées sur le squelette, afin de comprendre leur pouvoir descriptif, de concevoir des descripteurs robustes et d'évaluer leur efficacité dans le cadre d'une segmentation temporelle précise.

Ce travail constitue une première étape de notre démarche de recherche. En nous ap-

puyant sur des caractéristiques squelettiques définies manuellement (telles que les positions articulaires, les orientations des membres et les vitesses articulaires), nous cherchons à établir une base de référence solide, révélatrice de la richesse expressive de cette modalité. Ces observations sont fondamentales, car elles guideront directement le développement de modèles plus sophistiqués présentés dans les chapitres suivants. En particulier, le chapitre 4 prolongera cette contribution en introduisant une modélisation du squelette basée sur les graphes, enrichie par des mécanismes d'attention, afin d'améliorer à la fois les performances et l'interprétabilité, des avancées qui rendront possible la reconnaissance en ligne.

Ce chapitre présente la méthodologie développée pour la segmentation d'actions humaines à partir de données squelettiques. Nous commençons, dans la section 3.2, par motiver le choix de cette modalité et décrire le pipeline d'extraction de caractéristiques mis en place, tout en justifiant les décisions de conception par une analyse empirique. La section 3.3 introduit ensuite l'architecture encodeur–décodeur proposée, structurée en blocs hiérarchiques et entraînée à partir de caractéristiques définies manuellement, afin de produire une segmentation image par image des actions. La section 3.4 détaille le protocole expérimental, les jeux de données utilisés ainsi que les paramètres d'entraînement retenus. Les performances du modèle sont ensuite confrontées à celles de l'état de l'art dans la section 3.5, tandis que la section 3.6 analyse, à l'aide d'études d'ablation, la contribution de chaque composant de l'architecture proposée. Enfin, la section 3.7 synthétise les résultats obtenus, discute des limites de cette approche et prépare la transition vers une modélisation plus avancée, basée sur les graphes, au chapitre suivant.

3.2 Le choix des données squelettiques

L'environnement des postes de travail industriels introduit une grande variabilité visuelle, avec des perturbations telles que les variations d'éclairage, les occultations partielles dues à des objets ou à d'autres opérateurs, ainsi que la présence d'éléments de fond encombrants (machines, postes de travail, éléments de stockage, ...). Ces facteurs rendent la détection et la reconnaissance d'actions particulièrement complexes à partir de données RGB classiques [11]. Face à ces défis, l'utilisation des données squelettiques s'impose comme une alternative robuste et efficace [3]. En effet, cette modalité repose sur une abstraction de la posture humaine sous forme de points clés articulaires, permettant de se concentrer sur la dynamique du mouvement sans être affectée par le contexte visuel environnant. De plus, les progrès récents

en estimation de pose, avec des algorithmes comme REG-Net [166] ou PW-YOLO-Pose [167], ont considérablement amélioré la fiabilité et la précision de l'extraction des squelettes, même en présence d'occultations ou de gestes complexes [168].

Inspirée par ces avancées et constatant un manque d'approches spécialisées pour la reconnaissance et la segmentation d'actions dans des contextes industriels, cette étude propose une méthode fondée sur la modalité squelette. Notre modèle encodeur-décodeur exploite les caractéristiques squelettiques pour reconnaître et segmenter les actions dans des séquences non découpées, en mettant l'accent sur les transitions entre actions durant des tâches d'assemblage.

3.3 Approche méthodologique proposée

Le modèle proposé, nommé "GSK-C2F" (Graph Skeleton Coarse-to-Fine) inspiré de [60] et adapté aux données de squelette, est illustré sur la figure 3.1. L'architecture se compose de trois modules principaux : un backbone à multiples canaux, un bloc encodeur-décodeur et un bloc de prédiction pour la reconnaissance et la segmentation d'actions. Le choix de cette architecture résulte d'une analyse approfondie des travaux existants, présentée dans la section 3.3.2.

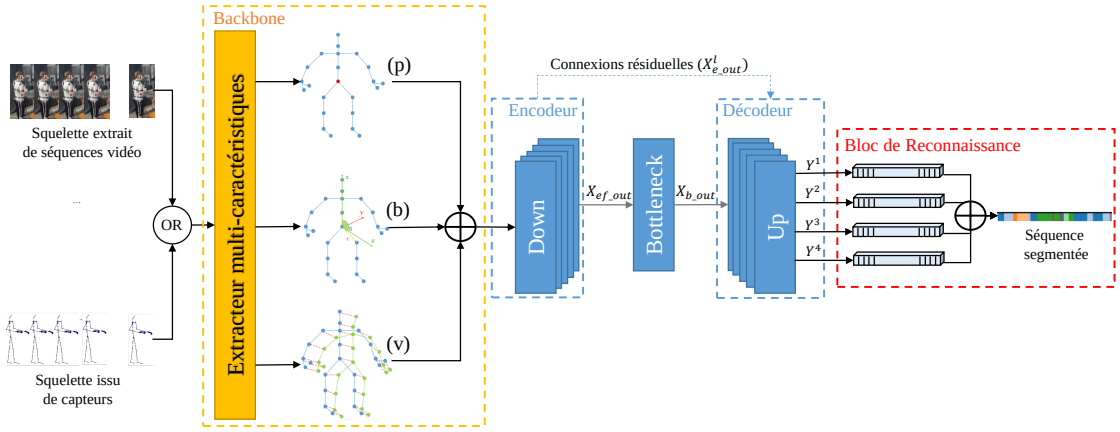


FIGURE 3.1 – Architecture globale de GSK-C2F. La séquence d'entrée illustrative provient du jeu de données InHARD [3]. p : positions articulaires, b : angles des membres, v : vitesses articulaires, Y_l : sortie de la couche l du décodeur.

Dans un premier temps, le backbone est chargé de pré-traiter les données brutes issues des squelettes 3D afin de les transformer en représentations plus expressives. Pour cela, nous

exploitons plusieurs descripteurs conçus à la main. Cette étape est cruciale, car elle permet d'extraire simultanément des informations spatiales, telles que la position des articulations (positions) et l'orientation des membres (bone angles), ainsi que des informations temporelles, telles que la vitesse des mouvements articulaires (velocity), souvent complémentaires, à partir des données squelettiques. Ensuite, le bloc encodeur-décodeur traite ces descripteurs à différentes échelles temporelles. Grâce à une première phase de compression dans l'encodeur suivie d'une reconstruction des motifs temporels (patterns) dans le décodeur, cette architecture permet de capturer aussi bien des dynamiques à court terme que des dépendances de plus longue durée. Cette capacité à modéliser le contexte temporel de manière hiérarchique est essentielle pour segmenter et reconnaître avec précision des actions humaines complexes, dans des séquences longues et non tronquées. Enfin, le module de prédiction exploite les sorties du décodeur à différents niveaux de résolution, en les combinant au moyen d'un mécanisme d'agrégation. Cette approche vise à atténuer le problème de sur-segmentation, fréquent dans les tâches de reconnaissance d'actions, en affinant les prédictions grâce à l'intégration de plusieurs niveaux d'information à différentes échelles temporelles. Les sous-sections suivantes décrivent en détails chacun de ces modules.

3.3.1 Le Backbone

Le backbone proposé vise à modéliser les données squelettiques dans l'espace et le temps. Plus précisément, la séquence d'entrée est d'abord traitée par un extracteur multi-caractéristiques, qui produit trois types de descripteurs : les positions articulaires 3D (positions), les angles des membres (Bone angles) et les vitesses articulaires (velocity), notés respectivement (p), (b) et (v) [169]-[171]. Ces modalités sont ensuite concaténées pour former un vecteur de caractéristiques par image. Dans ce qui suit, nous décrivons en détails la conception et le fonctionnement de cet extracteur multi-caractéristiques.

3.3.1.1 Extracteur multi-caractéristiques

Divers descripteurs de données squelettiques ont été examinés dans les travaux antérieurs [169], [170]. Dans le cadre de cette étude, nous distinguons trois types de descripteurs squelettiques : les positions articulaires 3D (p), les angles des membres (b) et les vitesses articulaires (v) (voir la figure3.2). L'extraction de ces descripteurs est assurée par le module d'extraction multi-caractéristiques (MFE : Multi-Feature Extractor) que nous proposons. Le processus

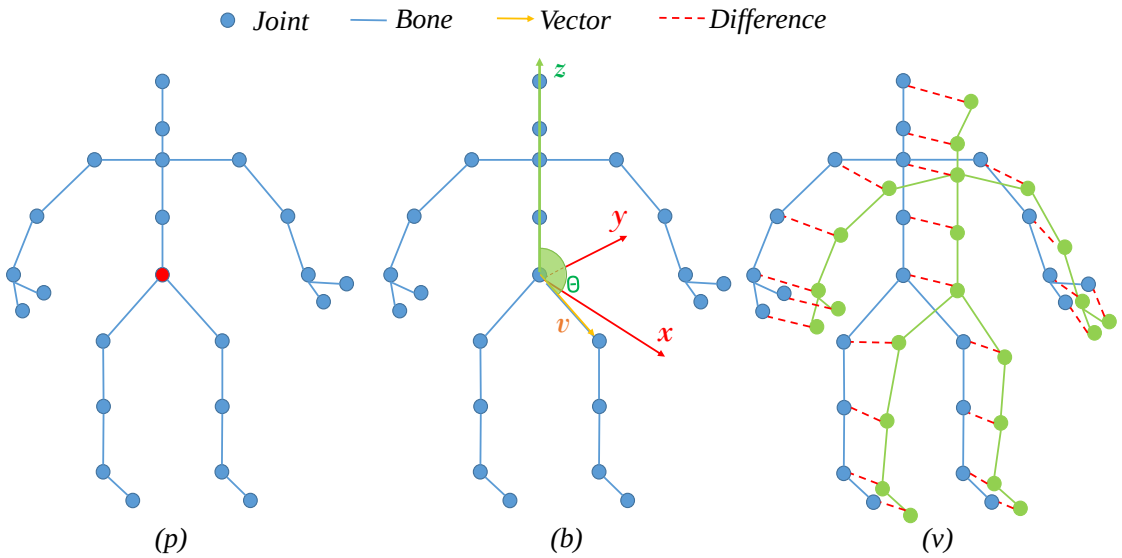


FIGURE 3.2 – Représentations squelettiques : (p) positions 3D des articulations en bleu (référencées par rapport au bassin en rouge), (b) angles des membres par rapport à l'axe z, et (v) vitesses articulaires.

d'extraction se déroule comme suit :

3.3.1.1.1 Positions articulaires (p) : Les données squelettiques brutes, notées p, sont représentées par le tenseur $\mathbf{X} \in \mathbb{R}^{C_{in} \times T_{in} \times N_{in}}$, où C_{in} est le nombre de canaux (typiquement 3 pour les coordonnées (x, y, z)), T_{in} est la durée de la séquence (en nombre d'images), et N_{in} est le nombre total d'articulations (noeuds) du squelette. Afin d'assurer l'invariance aux déplacements globaux du corps dans l'espace (par exemple dus au mouvement de la caméra ou à un changement de position absolue), les positions articulaires sont converties en coordonnées relatives par rapport à une articulation de référence, souvent appelée *articulation racine*. Dans cette approche, nous choisissons l'articulation du bassin (*Hip*) comme point de référence, désignée par l'indice c. Ce choix repose sur deux propriétés essentielles :

— sa position centrale dans la structure squelettique humaine, ce qui en fait un ancrage anatomique pertinent ;

— sa relative stabilité au cours de la plupart des actions, ce qui réduit les variations brusques de la référence spatiale.

Les positions relatives sont calculées pour chaque articulation i selon :

$$\mathbf{R}_i = \mathbf{X}_{:,:,i} - \mathbf{X}_{:,:,c}, \qquad \mathbf{X} \in \mathbb{R}^{C_{in} \times T_{in} \times N_{in}}. \tag{3.1}$$

où $\mathbf{X}_{:,:,i} \in \mathbb{R}^{C_{in} \times T_{in}}$ désigne l'ensemble des coordonnées (sur les C_{in} canaux et T_{in} instants) de l'articulation i, et c correspond à l'articulation de référence (le bassin). Ainsi, $\mathbf{R}_i \in \mathbb{R}^{C_{in} \times T_{in}}$ représente la position relative de l'articulation i par rapport au bassin. L'ensemble des positions relatives forme alors le tenseur $\mathbf{R} \in \mathbb{R}^{C_{in} \times T_{in} \times N_{in}}$. Ce processus permet de normaliser spatialement les données articulaires, renforçant la robustesse du modèle face aux variations de position globale du corps. Il contribue également à rendre les séquences comparables entre elles, même lorsqu'une même action est exécutée dans des contextes spatiaux différents. Par exemple, deux individus réalisant une action identique, telle que *tendre le bras*, peuvent produire des trajectoires très différentes en coordonnées absolues en raison de leur position initiale ou de la perspective du capteur. En revanche, l'utilisation de coordonnées relatives au bassin rend les trajectoires invariantes à la position globale et donc comparables entre séquences.

En résumé, cette représentation centrée sur la racine (bassin) simplifie l'apprentissage en supprimant les variations dues au mouvement global et recentre l'analyse sur la dynamique relative entre les articulations, plus pertinente pour la reconnaissance d'actions.

3.3.1.1.2 Angles des membres (b) : Les informations géométriques issues de la structure osseuse constituent un complément important aux positions articulaires dans la représentation des mouvements humains. Contrairement aux coordonnées brutes qui décrivent des positions absolues ou relatives, les caractéristiques basées sur les membres permettent de capturer directement la morphologie locale et la structure cinématique du squelette, offrant ainsi une perspective plus fine sur les dynamiques du mouvement.

Les angles des membres visent à modéliser l'orientation locale des segments corporels dans l'espace. Chaque membre est considéré comme un vecteur $\vec{v_i}$ défini entre deux articulations connectées. Pour quantifier l'orientation de ce vecteur, on mesure son angle par rapport à un vecteur de référence fixe, généralement l'axe vertical $z = [0, 0, 1]$. Dans notre cas, les coordonnées squelettiques sont acquises à l'aide du système inertiel Perception Neuron 32 v2 et exprimées dans un repère global cohérent aligné avec la gravité. Ainsi, l'axe vertical utilisé correspond à la verticale physique estimée, ce qui permet de capturer de manière stable l'orientation des segments corporels. L'angle θ_i entre le vecteur osseux $\vec{v_i}$ et l'axe de référence $\vec{v_{ref}}$ est obtenu par la formule suivante :

$$\theta_i = \arccos\left(\frac{\vec{v_{ref}} \cdot \vec{v_i}}{\|\vec{v_{ref}}\| \cdot \|\vec{v_i}\|}\right). \tag{3.2}$$

En empilant les angles de l'ensemble des membres au cours du temps, on obtient la modalité osseuse $b \in \mathbb{R}^{C_b \times T_{in} \times N_{in}}$, où $C_b = 1$ correspond au canal angulaire, T_{in} est la longueur temporelle de la séquence, et N_{in} le nombre de membres considérés. Cette représentation présente plusieurs avantages, elle encode l'orientation des segments indépendamment de la position globale et permet de capturer des configurations posturales spécifiques (ex. : bras levé, jambe fléchie). Elle apporte ainsi des informations géométriques et biomécaniques complémentaires. En combinaison avec la modélisation temporelle du réseau, ces descripteurs contribuent à mieux distinguer des actions visuellement similaires (comme "attraper" et "pousser"), qui diffèrent notamment par leur dynamique locale.

De plus, ces caractéristiques sont souvent plus stables temporellement que les coordonnées articulaires brutes, ce qui les rend particulièrement pertinentes pour des tâches de segmentation ou de reconnaissance d'actions sur des séquences longues.

3.3.1.1.3 Vitesses articulaires (v) : Les vitesses articulaires permettent de modéliser la dynamique du mouvement en mesurant la variation temporelle des positions des articulations squelettiques. Contrairement aux positions et aux angles, qui décrivent l'état statique ou la configuration d'une posture à un instant donné, la vitesse renseigne sur l'évolution du mouvement dans le temps. Elle est donc particulièrement utile pour capturer des informations dynamiques essentielles à la reconnaissance fine et à la segmentation des actions humaines.

Sur le plan méthodologique, la vitesse est estimée à partir de la dérivée temporelle discrète des positions articulaires. Plus précisément, pour chaque articulation i, on calcule la différence de position entre les instants t et $t+1$, correspondant à deux données consécutives de la séquence. La vitesse V_i est définie comme suit :

$$\mathbf{V}_t = \mathbf{X}_{:,t+1,:} - \mathbf{X}_{:,t,:}, \qquad \mathbf{X} \in \mathbb{R}^{C_{in} \times T_{in} \times N_{in}}, \quad t = 1, \ldots, T_{in} - 1. \tag{3.3}$$

où $\mathbf{X}$ est le vecteur représentant les positions 3D de toutes les articulations, avec C_{in} le nombre de coordonnées spatiales, T_{in} le nombre de frames, et N_{in} le nombre total d'articulations. $\mathbf{V}_t(i) \in \mathbb{R}^{C_{in}}$ correspond à la vitesse de l'articulation i, pour $i = 1, \ldots, N_{in}$. Par empilement temporel, on obtient le tenseur des vitesses $\mathbf{V} \in \mathbb{R}^{C_{in} \times (T_{in}-1) \times N_{in}}$.

L'intégration de cette modalité présente plusieurs intérêts. Elle permet tout d'abord de détecter les phases dynamiques et statiques d'une action : une vitesse nulle ou faible peut correspondre à un moment d'arrêt ou à une transition, tandis qu'une vitesse élevée indique

une phase d'activité. Cette propriété est particulièrement précieuse dans des séquences non découpées, où il est crucial de localiser les débuts et les fins d'action. La vitesse fournit une information dynamique complémentaire aux positions. En effet, des postures similaires peuvent correspondre à des phases de mouvement différentes (par exemple bras en train de monter ou de descendre). L'intégration de la vitesse permet ainsi au modèle de capturer la direction et la dynamique locale du geste, améliorant la discrimination entre actions visuellement proches.

3.3.2 Architecture encodeur-décodeur

Après avoir extrait des représentations squelettiques riches via le backbone (cf. section 3.3.1), l'étape suivante de notre architecture GSK-C2F est conçue pour tirer pleinement parti de ces représentations. Pour rappel, ce backbone produit des descripteurs riches du mouvement humain à partir de trois modalités complémentaires : (i) les positions articulaires brutes, qui capturent la configuration spatiale du corps ; (ii) les angles entre os, qui reflètent la posture et les interactions articulaires ; et (iii) les vitesses articulaires, qui encodent les dynamiques locales du mouvement.

Cependant, pour exploiter efficacement cette richesse descriptive dans le temps, un simple empilement de couches ou une modélisation linéaire ne suffit pas. Il est nécessaire de capturer à la fois les dépendances temporelles à court terme (ex. micro-gestes ou variations locales rapides) et à long terme (ex. phases d'une action prolongée ou structure temporelle globale). Les approches traditionnelles basées sur des réseaux de neurones récurrents (RNN), telles que les LSTM ou les GRU [11], [100], ont longtemps été utilisées pour ce type de modélisation séquentielle. Toutefois, leur structure itérative, leur mémoire limitée et leur coût de calcul élevé rendent leur application difficile sur de longues séquences industrielles à haute fréquence. Pour surmonter ces limites, les réseaux convolutifs temporels (Temporal Convolutional Networks, TCN) ont été introduits comme une alternative efficace. Par exemple, MS-TCN [101] propose une modélisation hiérarchique du temps à l'aide de convolutions dilatées, tout en conservant une capacité de parallélisation avantageuse. Ces modèles ont ensuite été améliorés par des variantes, telles que ED-TCN [138], qui introduit une structure encodeur–décodeur symétrique afin de mieux capter les dépendances temporelles à différentes échelles. Enfin, C2F-TCN [60] a proposé un raffinement progressif coarse-to-fine, où les prédictions sont affinées à plusieurs résolutions temporelles pour réduire la sur-segmentation et améliorer la localisation des transitions d'actions.

Dans cette lignée, notre approche GSK-C2F adopte une architecture encodeur-décodeur à plusieurs niveaux de résolution, inspirée de C2F-TCN, mais adaptée au traitement de représentations squelettiques enrichies. Contrairement aux versions classiques, notre encodeur-décodeur exploite les descripteurs multi-modaux générés par le backbone afin de modéliser les séquences à travers une hiérarchie temporelle. L'encodeur capture les dynamiques à différentes échelles, le goulot d'étranglement (bottleneck) intègre l'information contextuelle, et le décodeur reconstruit les prédictions d'actions avec une granularité croissante. Ce design permet de mieux exploiter les corrélations entre caractéristiques spatiales et temporelles, tout en localisant avec précision les actions pertinentes dans des séquences longues et non découpées typiques des environnements industriels. La figure 3.3 illustre la structure détaillée de chaque composant de ce bloc : (a) l'encodeur, (b) le goulot d'étranglement et (c) le décodeur.

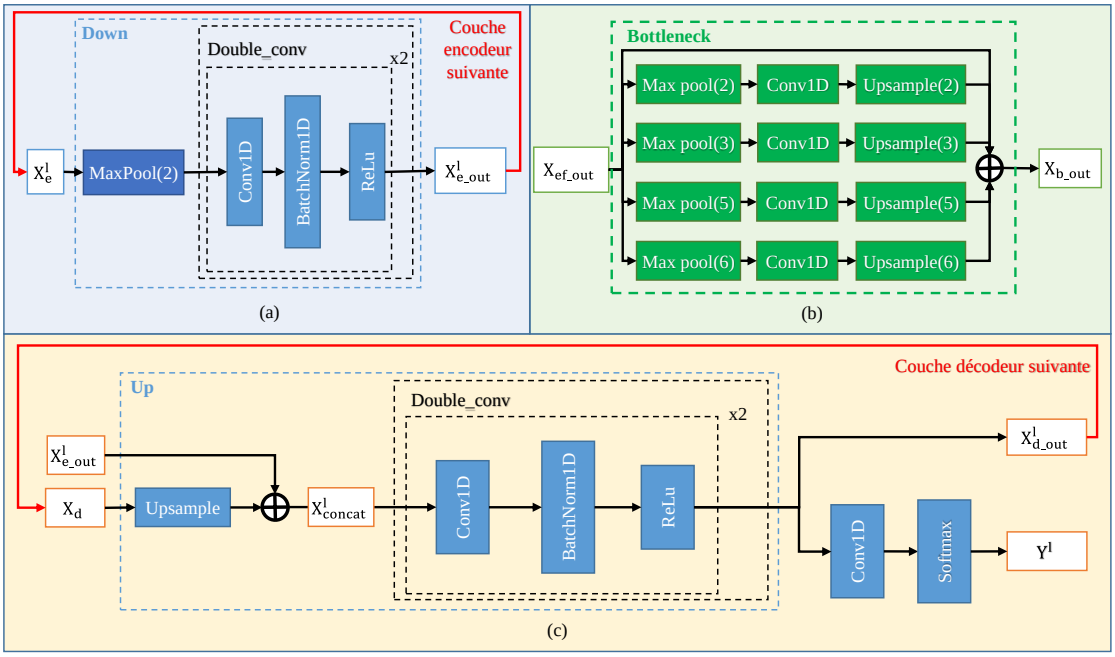


FIGURE 3.3 – Détails des modules : (a) encodeur, (b) goulot d'étranglement et (c) décodeur. Xe : entrée de l'encodeur, X_{ef_out} : sortie finale de l'encodeur, $X^l_{e_out}$: connexions résiduelles de la couche l, X_{b_out} : sortie du bottleneck (utilisée comme entrée du décodeur), X_d : entrée du décodeur, $X^l_{d_out}$: sortie du décodeur de la couche l, Y : prédiction.

3.3.2.1 Encodeur

À la suite de l'extraction des caractéristiques spatio-temporelles par le *backbone*, les trois types de caractéristiques d'entrée (positions p, angles osseux b et vitesses v) sont d'abord

fusionnées au sein du réseau afin de produire une représentation unifiée. Cette étape aboutit à un tenseur de caractéristiques noté $\mathbf{X}_e \in \mathbb{R}^{C_e \times T_{in}}$, où C_e désigne la dimension du descripteur latent, et T_{in} la dimension temporelle initiale. Comme illustré dans la Figure 3.3(a), la fonction principale de l'encodeur est de réduire progressivement la résolution temporelle de la séquence d'entrée tout en préservant les informations discriminantes nécessaires à la reconnaissance des actions complexes. Cette réduction s'effectue au sein d'une architecture hiérarchique composée de cinq niveaux (*Down*) successifs.

3.3.2.1.1 Bloc DoubleConv. Le bloc convolutif utilisé à chaque niveau est un bloc *DoubleConv*, constitué de deux convolutions 1D successives, chacune suivie d'une normalisation par lot et d'une activation ReLU. De manière compacte, un bloc *DoubleConv* appliqué à une entrée $\mathbf{X}$ s'écrit :

$$\text{DoubleConv}(\mathbf{X}) = \text{ReLU} \circ \text{BN} \circ \text{Conv1D} \circ \text{ReLU} \circ \text{BN} \circ \text{Conv1D}(\mathbf{X}). \tag{3.4}$$

La convolution 1D temporelle au niveau l est notée de façon compacte :

$$\text{Conv1D}(\mathbf{X}^l) = \mathbf{W} * \mathbf{X}^l + b, \tag{3.5}$$

où $\mathbf{X}^l \in \mathbb{R}^{C_e^l \times T^l}$ est l'entrée du niveau l, $\text{Conv1D}(\mathbf{X}^l) \in \mathbb{R}^{C_e^{l+1} \times T^l}$ la sortie, $\mathbf{W} \in \mathbb{R}^{C_e^{l+1} \times C_e^l \times (2K+1)}$ le banc de filtres (avec $2K+1$ la taille du noyau temporel), et $b \in \mathbb{R}^{C_e^{l+1}}$ le biais associé à chaque filtre de sortie. Le symbole $*$ désigne ici une convolution 1D appliquée le long de la dimension temporelle. La normalisation par lot standardise ensuite les activations selon :

$$\hat{\mathbf{X}}_{c,t} = \gamma_c \frac{\mathbf{X}_{c,t} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c, \tag{3.6}$$

où μ_c et σ_c^2 sont la moyenne et la variance calculées sur le mini-lot, et γ_c, β_c des paramètres appris. La fonction ReLU est définie par $\text{ReLU}(x) = \max(0, x)$, introduisant la non-linéarité nécessaire à la modélisation des dynamiques complexes.

3.3.2.1.2 Hiérarchie temporelle : MaxPool + DoubleConv. On note $\mathbf{X}^0 = \mathbf{X}_e$ l'entrée du premier niveau de l'encodeur. Chaque niveau l ($l = 0, \ldots, 4$) applique d'abord un sous-échantillonnage temporel par *max-pooling* d'un facteur 2, puis un bloc *DoubleConv*. Le sous-échantillonnage temporel est défini par :

$$\mathbf{X}_{c,t}^{l,\text{pool}} = \max\left(\mathbf{X}_{c,2t}^l,\ \mathbf{X}_{c,2t+1}^l\right), \quad t = 0, \ldots, \left\lfloor \frac{T^l}{2} \right\rfloor - 1, \quad c = 1, \ldots, C_e^l, \tag{3.7}$$

où $\mathbf{X}^l \in \mathbb{R}^{C_e^l \times T^l}$ est l'entrée du niveau l, et $\mathbf{X}^{l,\text{pool}} \in \mathbb{R}^{C_e^l \times T^{l+1}}$ la sortie sous-échantillonnée, avec

$$T^{l+1} = \left\lfloor \frac{T^l}{2} \right\rfloor. \tag{3.8}$$

Le traitement complet d'un niveau de l'encodeur s'écrit alors de façon compacte :

$$\mathbf{X}^{l+1} = \text{DoubleConv}\big(\text{MaxPool}(\mathbf{X}^l)\big), \quad l = 0, \ldots, 4, \tag{3.9}$$

où $\mathbf{X}^{l+1} \in \mathbb{R}^{C_e^{l+1} \times T^{l+1}}$. Après cinq niveaux, la dimension temporelle est ainsi réduite d'un facteur 32, soit

$$T^5 = \left\lfloor \frac{T_{in}}{32} \right\rfloor. \tag{3.10}$$

Cette architecture hiérarchique permet d'agréger l'information sur des fenêtres temporelles successivement plus larges, facilitant la capture des dépendances à long terme propres aux actions longues et complexes rencontrées dans les processus industriels d'assemblage. Par ailleurs, la réduction de la dimension temporelle diminue la charge computationnelle, rendant le modèle plus efficace, tout en conservant une richesse de représentation grâce aux convolutions successives et aux non-linéarités. Ce compromis entre expressivité, généralisation et efficacité computationnelle est essentiel pour assurer des performances robustes dans la reconnaissance d'actions à partir de séquences temporelles longues.

3.3.2.2 Goulot d'étranglement (Bottleneck)

Le module de goulot d'étranglement, ou *bottleneck*, constitue l'interface centrale de l'architecture entre les blocs d'encodage et de décodage. Comme illustré dans la figure 3.3(b), il reçoit en entrée la représentation encodée $\mathbf{X}^f \in \mathbb{R}^{C^f \times T^f}$, issue du dernier bloc de l'encodeur. Ici, $C^f = C_e^5$ désigne le nombre de canaux en sortie du cinquième (et dernier) bloc de l'encodeur, et $T^f = \lfloor T^{in}/32 \rfloor$ correspond à la dimension temporelle après les opérations successives de sous-échantillonnage. Cette représentation est donc fortement compressée dans le domaine temporel tout en étant riche en information latente le long des C^f canaux, ce qui en fait une base idéale pour des opérations de fusion multi-échelles. Le choix d'une structure de type *Temporal Pyramid Pooling* (TPP) dans ce goulot d'étranglement répond à plusieurs objectifs fondamentaux. D'une part, il s'agit de capturer des contextes temporels à différentes échelles de granularité en exploitant simultanément des fenêtres locales et globales. Ce principe permet de mieux modéliser des actions à temporalité variable, une caractéristique

typique des gestes humains dans les environnements industriels. D'autre part, cette approche offre une forme implicite de régularisation temporelle multi-échelle, en évitant que le modèle ne se spécialise sur des motifs d'une unique durée. Elle permet ainsi au modèle d'apprendre des représentations qui restent cohérentes malgré les différences d'échelle temporelle, en capturant simultanément des motifs courts, moyens et longs. Cela doit améliorer sa capacité à reconnaître des actions même lorsque la durée des sous-séquences varie.

Concrètement, quatre opérations de *max-pooling* unidimensionnelles sont appliquées en parallèle sur la séquence temporelle $\mathbf{X}^f$, chacune avec une taille de noyau $k_i \in \{2, 3, 5, 6\}$ et un pas $s_i = k_i$. Pour chaque niveau i, le sous-échantillonnage est défini par :

$$\mathbf{X}^{(i)}_{c,t} = \max\left(\mathbf{X}^f_{c,t\cdot s_i : t\cdot s_i + k_i - 1}\right), \quad t = 0, \ldots, \left\lfloor \frac{T^f - k_i}{s_i} \right\rfloor, \tag{3.11}$$

générant une représentation compressée $\mathbf{X}^{(i)} \in \mathbb{R}^{C^f \times T^{(i)}}$, avec $T^{(i)} = \left\lfloor \frac{T^f - k_i}{k_i} \right\rfloor$. Les quatre branches partagent ainsi le même nombre de canaux C^f, mais à des résolutions temporelles différentes. Chaque sortie $\mathbf{X}^{(i)}$ passe ensuite par une convolution 1D de noyau 1×1, notée ϕ, partagée entre les niveaux, qui projette les représentations dans un espace latent de taille d^f :

$$\mathbf{Z}^{(i)} = \phi(\mathbf{X}^{(i)}) \in \mathbb{R}^{d^f \times T^{(i)}}. \tag{3.12}$$

Cette projection réduit la dimension des canaux de C^f à d^f (avec $d^f < C^f$), ce qui diminue le coût computationnel des traitements ultérieurs, tout en imposant un espace latent commun entre les différentes échelles temporelles. Les représentations compressées $\mathbf{Z}^{(i)}$ sont ensuite rééchantillonnées (interpolation linéaire ou duplication de frames) afin de retrouver une dimension temporelle commune égale à T^f, soit :

$$\tilde{\mathbf{Z}}^{(i)} \in \mathbb{R}^{d^f \times T^f}, \quad \text{avec} \quad \tilde{\mathbf{Z}}^{(i)} = \text{Upsample}(\mathbf{Z}^{(i)}, T^f). \tag{3.13}$$

Une fois cette opération complétée pour les quatre branches, les représentations interpolées sont concaténées avec l'entrée originale $\mathbf{X}^f$ sur la dimension des canaux. Le tenseur final produit par le goulot d'étranglement est donc :

$$\mathbf{X}_{b_out} = \text{Concat}\left(\mathbf{X}^f, \tilde{\mathbf{Z}}^{(2)}, \tilde{\mathbf{Z}}^{(3)}, \tilde{\mathbf{Z}}^{(5)}, \tilde{\mathbf{Z}}^{(6)}\right) \in \mathbb{R}^{(C^f + 4\cdot d^f) \times T^f}. \tag{3.14}$$

Cette concaténation multi-niveaux enrichit la représentation centrale du réseau en y injectant des contextes temporels variés, à la fois locaux et globaux, tout en préservant l'alignement

temporel. Elle constitue une forme de fusion hiérarchique dense, favorable à la reconstruction détaillée de la dynamique d'action dans le module de décodage.

3.3.2.3 Décodeur

Le décodeur constitue la phase de reconstruction temporelle de l'architecture proposée. Il est conçu de manière symétrique à l'encodeur, afin de restaurer progressivement une résolution temporelle comparable à celle de la séquence d'entrée tout en exploitant les informations de haut niveau extraites par les couches profondes. Cette symétrie permet non seulement une architecture plus structurée, mais aussi une circulation plus efficace de l'information grâce à l'usage de connexions de saut (*skip connections*) inspirées des architectures de type U-Net. Le décodeur reçoit en entrée le tenseur produit par le goulot d'étranglement, noté $\mathbf{X}_d^1 = \mathbf{X}_{b_out} \in \mathbb{R}^{C_d^1 \times T^1}$, avec $T^1 = T^f$. De manière générale, on note $\mathbf{X}_d^l \in \mathbb{R}^{C_d^l \times T^l}$ la représentation en entrée du bloc l du décodeur, où C_d^l désigne le nombre de canaux à l'échelle l et T^l la longueur temporelle correspondante. Comme illustré dans la Figure 3.3(c), chaque bloc $l \in \{1, \ldots, 5\}$ applique d'abord un sur-échantillonnage temporel par interpolation linéaire, noté Upsample$(\cdot)$, qui a pour effet de doubler la résolution temporelle :

$$\tilde{\mathbf{X}}_d^l = \text{Upsample}(\mathbf{X}_d^l), \qquad \text{Upsample} : \mathbb{R}^{C_d^l \times T^l} \to \mathbb{R}^{C_d^l \times (2T^l)}. \tag{3.15}$$

Cette opération a pour but de rétablir une résolution temporelle plus fine sur laquelle les informations locales pourront être réinjectées. Pour cela, la représentation sur-échantillonnée $\tilde{\mathbf{X}}_d^l$ est concaténée avec la sortie correspondante de l'encodeur à la même échelle, notée $\mathbf{X}_{e_out}^l$, via une connexion de saut :

$$\mathbf{X}_{\text{concat}}^l = \text{Concat}(\tilde{\mathbf{X}}_d^l, \mathbf{X}_{e_out}^l). \tag{3.16}$$

Cette concaténation le long de la dimension des canaux fusionne les informations sémantiques profondes issues du décodeur avec les détails temporels locaux conservés dans les cartes de caractéristiques de l'encodeur. Les connexions de saut jouent ici un rôle central, ce sont elles qui réinjectent les détails fins potentiellement atténués lors de la compression hiérarchique, tout en facilitant l'optimisation du réseau en améliorant la propagation du gradient. La représentation $\mathbf{X}_{\text{concat}}^l$ est ensuite traitée par un bloc convolutionnel noté DoubleConv, identique à celui utilisé dans l'encodeur (deux convolutions 1D successives, chacune suivie d'une normalisation par lot et d'une activation ReLU) :

$$\mathbf{X}_{d_out}^l = \text{DoubleConv}(\mathbf{X}_{\text{concat}}^l), \tag{3.17}$$

et la sortie de ce bloc devient l'entrée du bloc suivant :

$$\mathbf{X}_d^{l+1} = \mathbf{X}_{d_out}^{l}. \tag{3.18}$$

En répétant ce processus de sur-échantillonnage, concaténation avec l'encodeur et raffinage convolutionnel à travers les cinq blocs du décodeur, on obtient en sortie une représentation de longueur temporelle T^{in}, enrichie par l'ensemble du traitement hiérarchique encodeur–bottleneck–décodeur, et mieux adaptée à la tâche finale de classification temporelle. Enfin, l'architecture symétrique encodeur–décodeur facilite l'apprentissage de représentations hiérarchiques, ce qui doit renforcer la capacité du modèle à généraliser efficacement à des séquences d'action nouvelles, jamais observées durant l'entraînement.

3.3.3 Bloc de Reconnaissance

Le bloc de reconnaissance joue un rôle crucial dans la génération des prédictions finales d'actions à chaque instant. Comme illustré à la figure 3.4, les sorties intermédiaires de chaque couche du décodeur, dont la longueur temporelle est notée T^l, sont systématiquement rééchantillonnées afin de retrouver la longueur initiale T de la séquence d'entrée. Ce rééchantillonnage est effectué par interpolation linéaire afin de garantir une correspondance temporelle directe entre les différentes échelles traitées.

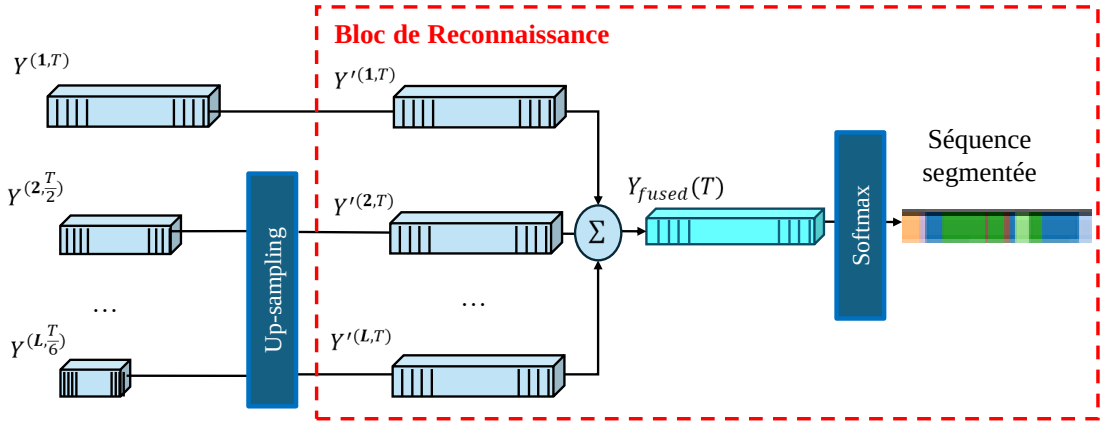


FIGURE 3.4 – Bloc de segmentation et de reconnaissance : pour chaque couche de décodeur l, la sortiea du décodeur $Y(l,t)$ (voir figure 3.3.c), est rééchantillonnée à la longueur temporelle initiale T afin de générer de nouvelles représentations $Y'(l,T)$. Ces caractéristiques sont ensuite fusionnées entre les couches, puis passées par une fonction softmax afin de produire des prédictions d'action, image par image.

Cette étape de sur-échantillonnage permet d'aligner les représentations extraites à différentes résolutions temporelles, ouvrant la voie à une fusion tardive (*late fusion*) de ces multiples niveaux de représentation. Plus précisément, pour chaque couche l du décodeur, la sortie $\mathbf{Y}(l,t) \in \mathbb{R}^{C \times T^l}$ est transformée en $\mathbf{Y}'(l,t) \in \mathbb{R}^{C \times T}$ par :

$$\mathbf{Y}'(l,t) = \text{Up-sampling}(\mathbf{Y}(l,t'), T) \quad \text{avec} \quad t' \in [1, T^l], \quad t \in [1, T] \tag{3.19}$$

où Up-sampling$(\cdot)$ représente l'opération d'interpolation linéaire temporelle. Cette démarche garantit que les informations extraites à différentes résolutions temporelles peuvent être mises en correspondance de manière temporelle précise. Ensuite, les représentations réechantillonnées $\mathbf{Y}'(l,t)$ issues des différentes couches du décodeur $l \in [1,5]$ sont fusionnées pour produire une représentation globale multi-échelle $\mathbf{Y}_{\text{fused}}(t)$. Cette fusion est réalisée au moyen d'une combinaison pondérée, où un poids w_l est associé à chaque couche. Les poids sont définis manuellement, ce qui permet d'inclure ou d'ignorer certaines résolutions temporelles en fonction de leur importance. comme suit :

$$\mathbf{Y}_{\text{fused}}(t) = \sum_{l=1}^{L} w_l \cdot \mathbf{Y}'(l,t), \tag{3.20}$$

où les poids $\{w_l\}_{l=1}^{L}$ sont définis pour sélectionner et combiner les résolutions temporelles jugées pertinentes. Cette stratégie de fusion tardive permet de réduire la sur-segmentation tout en améliorant la précision des prédictions. En agrégeant des représentations issues de plusieurs échelles temporelles, elle combine à la fois les détails locaux fournis par les couches superficielles et le contexte global capturé par les couches plus profondes. Cette complémentarité renforce la cohérence temporelle des actions prédites et permet une délimitation plus fiable des transitions [60].

Enfin, la représentation fusionnée $\mathbf{Y}_{fused}(t)$ est projetée dans l'espace des classes d'actions à l'aide d'une couche entièrement connectée (MLP à une couche linéaire), suivie d'une fonction softmax, qui produit la distribution de probabilité sur les différentes classes pour chaque instant t :

$$\hat{\mathbf{Y}}(t) = \text{Softmax}(W_c \cdot \mathbf{Y}_{fused}(t) + b_c) \tag{3.21}$$

où W_c et b_c sont respectivement les poids et biais appris de la couche de classification

finale, et $\hat{\mathbf{Y}}(t)$ est le vecteur de probabilité des classes d'action à l'instant t.

3.3.4 Synthèse

Afin de mieux comprendre les transformations opérées à travers les différentes composantes du modèle GSK-C2F, le tableau 3.1 présente un résumé des dimensions d'entrée et de sortie associées à chaque bloc fonctionnel de l'architecture. Cette vue d'ensemble illustre la progression du traitement, depuis les caractéristiques des données squelettiques brutes jusqu'aux prédictions d'actions, image par image.

TABLEAU 3.1 – Synthèse des dimensions à travers les blocs de GSK-C2F (voir Fig. 3.1). La dimension temporelle est notée T et la dimension des canaux C.

Bloc	Entrée	Sortie	Dimensions
Backbone	$X \in \mathbb{R}^{C_{in} \times T_{in} \times N_{in}}$	$X_e \in \mathbb{R}^{C_e \times T_{in}}$	$\mathbb{R}^{C_{in} \times T_{in} \times N_{in}} \rightarrow \mathbb{R}^{C_e \times T_{in}}$
Encodeur	$X_e \in \mathbb{R}^{C_e \times T_{in}}$	$X^f \equiv X_{ef_out} \in \mathbb{R}^{C^f \times T^f}$	$T^f = \lfloor T_{in}/32 \rfloor$
Goulot d'étranglement	$X^f \in \mathbb{R}^{C^f \times T^f}$	$X_{b_out} \in \mathbb{R}^{(C^f+4d^f) \times T^f}$	$\mathbb{R}^{C^f \times T^f} \rightarrow \mathbb{R}^{(C^f+4d^f) \times T^f}$
Décodeur	$X_d^1 = X_{b_out} + \{X_{e_out}^l\}$	$\{X_{d_out}^l\}_{l=1..5}$	$\mathbb{R}^{(C^f+4d^f) \times T^f} \rightarrow \mathbb{R}^{C_d^l \times T^l}$ avec $T^l = 2^{l-1}T^f$
Bloc de reconnaissance	$\{Y^{(l)}\}_{l=1..L}$	$\hat{Y}(t) \in \mathbb{R}^C$	$Y^{(l)} \in \mathbb{R}^{C \times T^l} \xrightarrow{\text{upsample}} \mathbb{R}^{C \times T_{in}} \Rightarrow \hat{Y} \in \mathbb{R}^{C \times T_{in}}$

3.4 Protocole expérimental

Cette section présente les éléments clés du protocole expérimental mis en place pour évaluer rigoureusement les performances du modèle GSK-C2F proposé. Nous détaillons d'abord les jeux de données utilisés dans la section 3.4.1, choisis pour leur pertinence pour la reconnaissance d'actions dans le contexte de l'assemblage industriel. Ensuite, le protocole d'entraînement est décrit dans la section 4.4.2, y compris les paramètres d'apprentissage du modèle et le prétraitement des données. Enfin, les métriques d'évaluation utilisées pour quantifier la qualité des prédictions sont présentées, ce qui permet une analyse complète et objective des résultats obtenus.

3.4.1 Jeux de données

Pour évaluer les performances de notre modèle **GSK-C2F**, et en nous basant sur l'état de l'art des jeux de données présentés dans la section 1.4.1, deux jeux de données pertinents ont été sélectionnés. Le premier, InHARD [3], est un jeu de données industriel capturant des séquences longues de gestes humains lors d'opérations d'assemblage dans un environnement complexe et réaliste. Le second, IKEA-ASM [4], est un jeu de données centré sur l'assemblage

de meubles, enregistré dans des environnements variés. Bien qu'il ne s'agisse pas d'un environnement industriel au sens strict, il reste directement lié aux tâches d'assemblage manuel, ce qui le rend pertinent pour évaluer la généralisabilité de notre modèle.

Le jeu InHARD a été annoté image par image avec treize classes d'actions, ainsi qu'une classe *"No action"*, ce qui reflète la richesse des scénarios industriels capturés. Quant au jeu IKEA-ASM, il comprend 33 classes d'actions et se distingue par des annotations fortement orientées objet. Certaines actions visuellement similaires sont étiquetées différemment selon l'objet manipulé. Ce type de données met en lumière les limites des approches basées uniquement sur les mouvements corporels et souligne l'intérêt d'une éventuelle intégration de la reconnaissance d'objets.

Ainsi, bien que des tests aient été menés sur IKEA-ASM pour évaluer la robustesse de notre modèle face à des variations contextuelles, le jeu de données InHARD demeure la référence principale de notre étude. Les expérimentations présentées dans les sections suivantes s'appuieront donc principalement sur ce dernier, mieux adapté à la reconnaissance d'actions fondée sur les dynamiques corporelles.

3.4.1.1 Le jeu de données InHARD

Le jeu de données **InHARD** (*Industrial Human Action Recognition Dataset*) [3] a été conçu pour la reconnaissance et l'analyse des gestes humains dans des environnements industriels collaboratifs. Il s'inscrit dans un contexte réel d'assemblage de tricycles, reproduisant fidèlement les contraintes et la complexité des tâches effectuées par des opérateurs humains. Ce jeu de données a été spécifiquement conçu pour le développement de méthodes d'apprentissage robustes dans des environnements industriels, où l'interaction homme–robot est essentielle.

InHARD offre une **double modalité d'acquisition** : des vidéos RGB multi-vues et des données squelettiques 3D. Trois caméras (Logitech C920) positionnées selon des vues *"haut"*, *"gauche"* et *"droite"* permettent de superviser largement les scènes et d'analyser les gestes sous différents angles. En parallèle, les **positions et rotations 3D de 21 articulations** du corps humain sont captées à l'aide du capteur *Perception Neuron 32 Edition v2*, avec une fréquence d'échantillonnage de **120 Hz**. Les données squelettiques sont stockées au format **BVH**. Un total de **16 participants** distincts (4 femmes et 12 hommes) ont contribué à la collecte de 38 séquences d'assemblage (voir la figure. 3.5), totalisant **4,804 échantillons** (segments d'actions) répartis en **14 classes** (dont une classe *No action*) annotées manuellement.

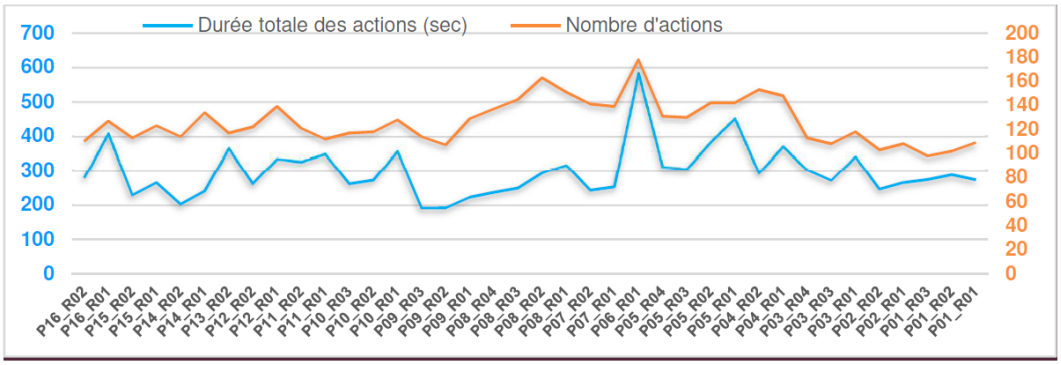


FIGURE 3.5 – Graphique croisé dynamique du nombre d'actions et leurs durées dans le jeu de données InHARD-13 [3].

Ces actions couvrent des manipulations précises et techniques, typiques des chaînes de production. Le jeu présente des **actions complexes**, de **durées variables (entre 0.5 et 26.9 secondes)** et une **distribution des classes déséquilibrée**, comme l'illustre la figure. 3.6, ce qui en fait un excellent terrain d'évaluation pour les approches d'apprentissage supervisé, notamment dans des contextes multi-classes difficiles. L'annotation temporelle a été réalisée à l'aide du logiciel *ANVIL* (voir la figure. 3.7), garantissant une segmentation précise des séquences en actions élémentaires.

Le jeu de données InHARD comprend plusieurs variantes : (1) **InHARD-13**, qui comprend 14 classes, dont 13 actions spécifiques et une classe "No action". (2) **InHARD-4**, un sous-ensemble d'InHARD-13 composé de quatre classes de bas niveau : "Pick in front", "Pick left", "Assemble system" et "No action". (3) **InHARD-3**, qui redéfinit trois actions de haut niveau issues d'InHARD-13. La classe "Assemble system" reste inchangée, tandis que les classes "Take" et "Put down" regroupent plusieurs actions existantes. La classe "Take" comprend notamment "Picking in front", "Picking left", "Take component" et "Take subsystem". La classe "Put down" inclut "Put down component" et "Put down subsystem".

Toutes ces variantes ont été utilisées pour comparer l'approche proposée aux méthodes de l'état de l'art, en mettant particulièrement l'accent sur InHARD-13 en raison de sa complexité et de sa richesse.

3.4.1.2 Jeu de données IKEA-ASM

IKEA-ASM [4] est un jeu de données conçu pour l'analyse fine des activités d'assemblage de **meubles**. Il se distingue par sa richesse **multi-vue** et **multi-modale**, incluant des séquences

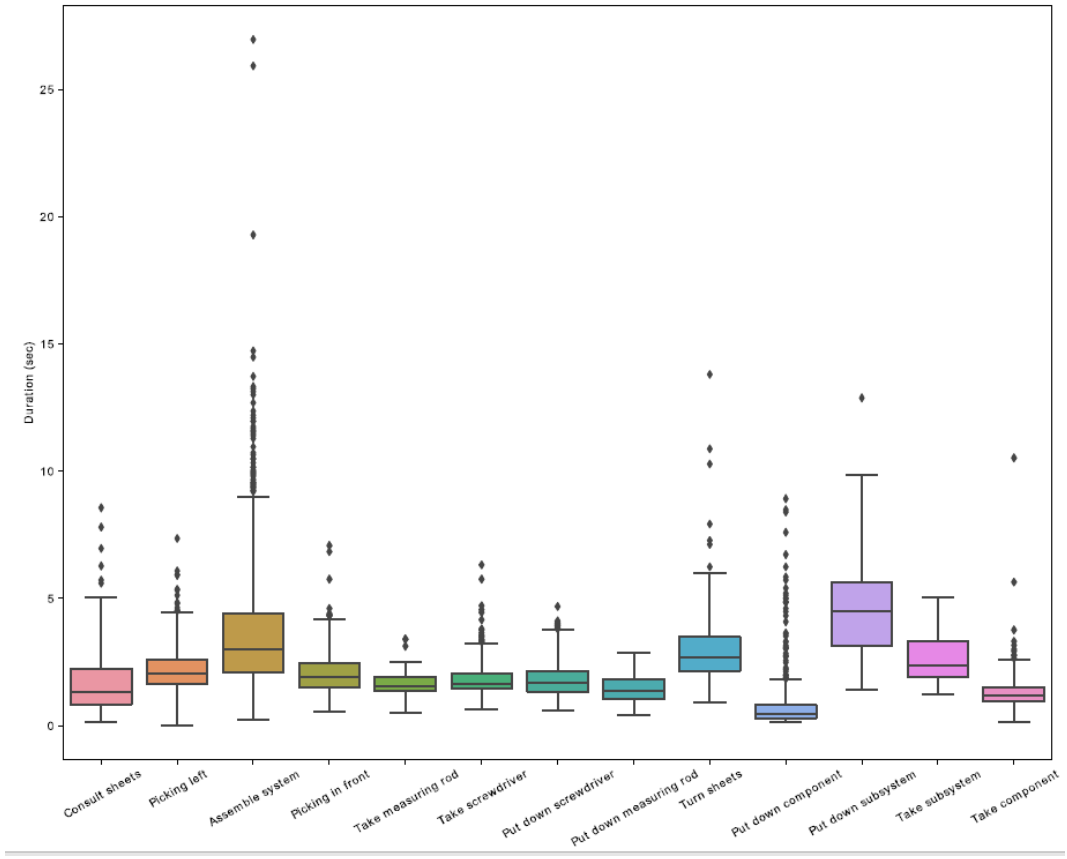


FIGURE 3.6 – Distribution de la durée des actions dans le jeu de données InHARD-13 [3]. Les classes "No action" et "Assemble system" comptent le plus grand nombre d'échantillons et la plus longue durée totale.

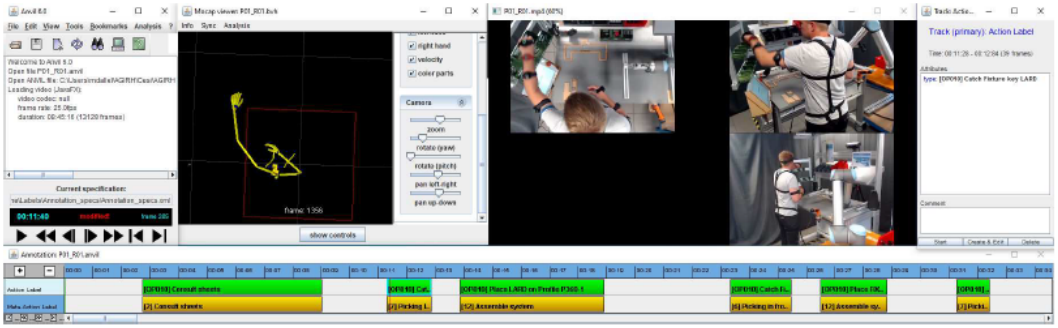


FIGURE 3.7 – Labélisation des actions dans le jeu de données InHARD-13 [3] avec l'outil ANVIL. En vert le label d'action (bas niveau), en jaune le label d'action (haut niveau).

RGB, des cartes de profondeur et des estimations de pose humaine. Il comprend **371 vidéos** documentant l'assemblage de **4 modèles de meubles IKEA** (tels que des tables ou des étagères), réalisées dans **5 environnements différents** par **48 assembleurs distincts**. Cette diversité

offre une forte variabilité inter-individuelle, contextuelle et comportementale. Les annotations couvrent **33 classes d'actions atomiques** (du type verbe-objet), segmentées temporellement.

La pose humaine est représentée par 12 articulations du corps et 5 points du visage, chacun étant accompagné d'un score de confiance. Ces données sont initialement annotées manuellement sur des images clés (environ 1% des images), puis interpolées automatiquement. La segmentation des objets suit un processus similaire.

3.4.2 Protocole d'entraînement

3.4.2.1 Prétraitement

Les entrées du modèle sont constituées de séquences de squelettes 3D décrivant les mouvements humains, obtenues soit à partir de systèmes de motion capture dédiés [84], soit estimées automatiquement à partir de vidéos RGB à l'aide d'algorithmes de pose lorsque les annotations squelettiques ne sont pas disponibles, comme dans le jeu de données IKEA-ASM. Dans le cadre de cette étude, nous utilisons principalement **MediaPipe**, mais d'autres outils de référence, tels qu'*OpenPose*, *YOLO* ou *PoseNet*, peuvent également être employés [172]. MediaPipe a été retenu dans cette étude pour sa rapidité d'exécution en temps réel, sa robustesse sur des données bruitées et sa facilité d'intégration dans des architectures légères. La figure 3.8 illustre quelques estimateurs de pose parmi les plus utilisés pour extraire automatiquement les squelettes à partir de flux vidéo RGB. Ces outils constituent une étape essentielle de l'analyse de l'activité humaine dans des environnements industriels complexes.

Afin d'assurer l'homogénéité temporelle et spatiale des données, chaque séquence est d'abord rééchantillonnée à **30 images par seconde** (fps). Les coordonnées articulaires sont ensuite normalisées à l'aide d'un *Min–Max scaling* [173]. Pour chaque dimension spatiale x, la normalisation est définie par :

$$x^{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.22}$$

Où $x_{\min}$ et $x_{\max}$ représentent respectivement les valeurs minimales et maximales observées dans la séquence pour cette dimension. Cette opération ramène toutes les valeurs dans l'intervalle $[0, 1]$, réduit les variations d'échelle entre les séquences et améliore la stabilité de l'apprentissage.

Ensuite, la durée minimale $T_{\min}$ observée sur l'ensemble des séquences du jeu de données est retenue comme longueur de référence. Chaque vidéo est alors divisée en segments de

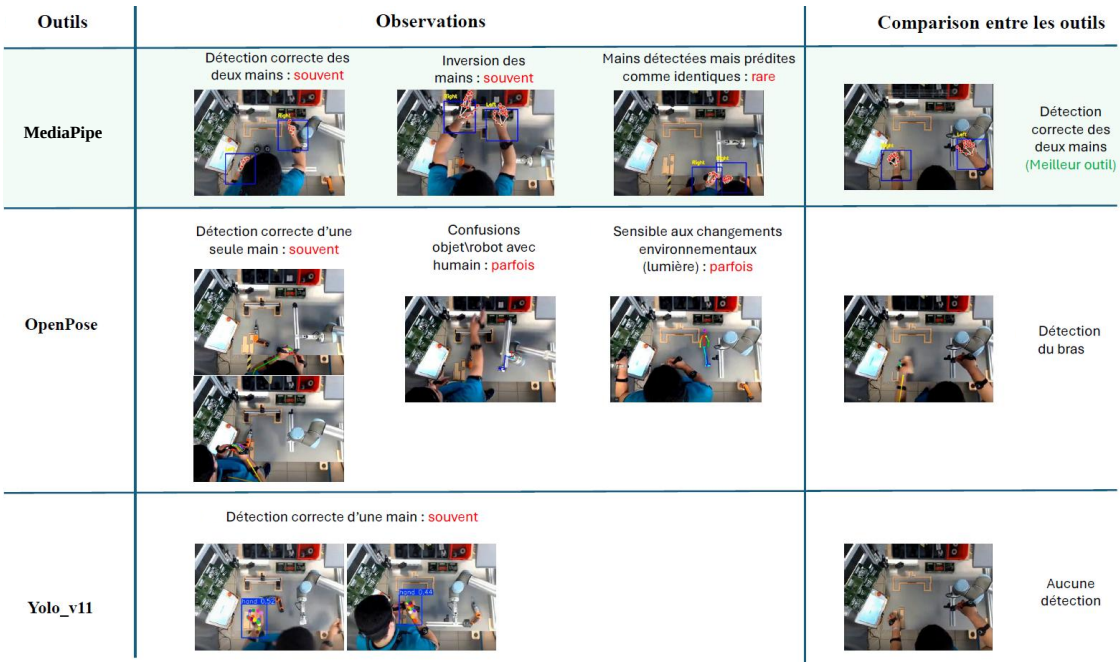


FIGURE 3.8 – Exemples d'outils de détection de la pose humaine utilisés pour extraire des squelettes à partir de séquences RGB.

taille $T_{\min}$, soit **7360 images** (soit environ **4 minutes** à 30 fps). Cette opération permet de standardiser les durées des séquences en entrée, ce qui est particulièrement adapté aux vidéos industrielles, dont les durées sont longues et peuvent atteindre quelques minutes [3].

3.4.2.2 Détails d'implémentation

Les expériences ont été réalisées à l'aide de l'optimiseur *Adam*, sur un total de **500 époques**, avec un taux d'apprentissage initial fixé à 10^{-4} et un coefficient de régularisation ℓ_2 (*weight decay*) de 3×10^{-3}. La taille des mini-lots (*batch size*) est de **5** séquences. Les vidéos ayant des durées variables, chaque séquence est ajustée à une longueur fixe T_{min} afin de garantir des dimensions homogènes dans les mini-lots. Lorsque la séquence est plus courte que T_{min}, un *padding* temporel est appliqué avec la valeur -100. Ces positions paddées sont ignorées lors du calcul de la fonction de perte, ce qui permet de préserver la cohérence de l'apprentissage. Les implémentations ont été réalisées à l'aide de la bibliothèque `PyTorch` (version 1.9.0+cu111), sur une station de travail équipée d'un GPU `NVIDIA Quadro RTX 6000` doté de `24 Go` de mémoire, d'un processeur `Intel Xeon W-2245` cadencé à `3,90 GHz` et de `64 Go` de mémoire vive (RAM).

Pour les méthodes basées sur des vidéos RGB, les caractéristiques visuelles sont extraites à l'aide d'un modèle **I3D** pré-entraîné sur le jeu de données *Kinetics*, largement utilisé pour l'analyse d'actions humaines. L'extraction des squelettes 2D à partir des séquences RGB (E-SK) est assurée par **MediaPipe**, un estimateur de pose rapide et fiable, adapté aux environnements industriels [172]. Afin de limiter les effets d'occlusion, les vues utilisées pour chaque dataset ont été sélectionnées empiriquement : la *vue latérale droite* a été retenue pour **InHARD**, tandis que la *vue du dessus* est privilégiée pour **IKEA-ASM** en raison de sa meilleure visibilité des articulations du haut du corps. Les méthodes exploitant les squelettes 3D utilisent directement les données articulaires tridimensionnelles fournies par le dataset **InHARD** (référencées sous l'appellation *3D-SK*).

Les ensembles d'apprentissage, de validation et de test sont définis selon les protocoles standards propres à chaque jeu de données. Pour **InHARD**, nous suivons un protocole **Cross-subject** [3], où les opérateurs des ensembles d'apprentissage et de validation sont exclus du test. Pour **IKEA-ASM**, qui comprend cinq environnements distincts, l'évaluation repose sur un protocole **Cross-environment** [4], dans lequel les environnements d'apprentissage et de validation diffèrent de ceux du test.

La fonction de coût combine une entropie croisée pour la classification image par image et une perte de transition afin d'assurer la cohérence temporelle, conformément à [60]. On note C le nombre de classes d'action (incluant éventuellement la classe *background*), $Y_t \in \{1, \ldots, C\}$ la classe réelle à l'instant t, $\mathbf{Z}_t \in \mathbb{R}^C$ les logits du modèle et $\hat{\mathbf{Y}}_t = \text{softmax}(\mathbf{Z}_t) \in [0,1]^C$ les probabilités prédites. L'entropie croisée masquée est définie par :

$$L_{ce} = -\frac{1}{\sum_{t=1}^{T} M_t} \sum_{t=1}^{T} M_t \, \log(\hat{\mathbf{Y}}_{t,Y_t}), \tag{3.23}$$

où $M_t \in \{0,1\}$ est un masque binaire valant 1 pour les instants valides (annotés) et 0 pour les instants ignorés car ils proviennent du padding utilisé pour uniformiser la longueur des séquences. Dans le cas où une stratégie de pondération des classes est activée, chaque terme est multiplié par un poids dépendant de la classe réelle Y_t :

$$L_{ce}^{w} = -\frac{1}{\sum_{t=1}^{T} M_t} \sum_{t=1}^{T} M_t \, w_{Y_t} \log(\hat{\mathbf{Y}}_{t,Y_t}), \tag{3.24}$$

où w_{Y_t} correspond au poids associé à la classe Y_t. Les différentes stratégies de pondération (HCW, LCW) ainsi que leur mode de calcul sont détaillés dans la section 3.6.1. La fonction

de perte de transition pénalise les variations brusques entre prédictions successives :

$$L_{tr} = \frac{1}{\sum_{t=2}^{T} M_t} \sum_{t=2}^{T} \min\left(\left\| \hat{\mathbf{Y}}_t - \hat{\mathbf{Y}}_{t-1} \right\|_2^2, \delta \right) \cdot M_t, \tag{3.25}$$

où δ borne l'influence des variations trop importantes (ici $\delta = 16$). La fonction de perte totale est alors :

$$L = L_{ce} + \lambda_{tr} L_{tr}, \tag{3.26}$$

où le coefficient de pondération est fixé à $\lambda_{tr} = 0.15$. Ce terme régularisateur vise à encourager la cohérence temporelle des prédictions en pénalisant les changements brusques et non justifiés de classe d'action entre deux images consécutives. En effet, sans cette contrainte, le modèle peut générer des prédictions instables, alternant rapidement entre différentes classes, même en l'absence de changement réel de l'activité observée. Cette régularisation contribue ainsi à réduire les transitions erratiques et à stabiliser la sortie du modèle sur toute la séquence temporelle. Le code est disponible sur gitlab pour plus de détails concernant l'implémentation du modèle et son architecture.

3.5 Comparaison avec l'état de l'art

Récemment, de plus en plus de travaux visent à résoudre le problème de la segmentation dans des séquences d'actions complexes. [90] ont mené des expériences visant à mesurer le temps d'assemblage à l'aide d'un ST-GCN pour la reconnaissance d'actions sur InHARD-3, en optimisant les couches du ST-GCN ainsi que la taille de la fenêtre glissante. Leur meilleure précision a été obtenue avec une fenêtre de 40 images. À des fins de comparaison, le modèle a également été réimplémenté sur les versions InHARD-13 et InHARD-4 du jeu de données. Par ailleurs, Orsag et al. [174] ont proposé un modèle basé sur les squelettes, conçu pour classifier les actions des opérateurs humains dans un environnement industriel. Leur approche exploite des couches LSTM cachées pour la classification spatiotemporelle des activités et a été évaluée sur le jeu de données InHARD-4. La méthode C2F-TCN [115] a également été reproduite à partir du code officiel et évaluée sur l'ensemble des versions du jeu de données InHARD, en utilisant la modalité RGB ainsi que la stratégie de pondération des classes décrite en Section 3.6.1.

La méthode proposée, GSK-C2F, a été implémentée et testée sur toutes les versions du jeu de données InHARD. Afin de permettre une comparaison équitable avec les méthodes

existantes de l'état de l'art, notre approche a été évaluée à l'aide des positions articulaires (p), qui constituent la caractéristique commune aux autres méthodes. De plus, ses performances ont été mesurées en utilisant la configuration de caractéristiques optimale pour chaque jeu de données, afin de mettre en évidence les avantages de la fusion de caractéristiques. Comme indiqué dans le tableau 3.2, GSK-C2F surpasse systématiquement les méthodes de l'état de l'art sur le jeu de données InHARD.

TABLEAU 3.2 – Comparaison des scores MoF (Mean-over-Frames) avec les travaux récents sur le jeu de données InHARD. * : méthodes réimplémentées. Les caractéristiques utilisées dans l'architecture de base sont indiquées entre parenthèses : p (positions articulaires), b (angles osseux), v (vitesses articulaires).

Méthode	Modalité	InHARD-3	InHARD-4	InHARD-13
C2F-TCN* [115]	RGB	50.13	32.57	34.28
LSTM-NN [174]	3D-SK	-	44.09	-
MTM* [90]	3D-SK	74.13	51.49	56.43
GSK-C2F(p)	E-SK	80.45	67.39	59.77
GSK-C2F(p)	3D-SK	84.82	70.28	<u>73.86</u>
GSK-C2F(pb)	**3D-SK**	**88.3**	<u>70.40</u>	72.35
GSK-C2F(pbv)	**3D-SK**	<u>85.47</u>	**72.98**	**73.90**

La méthode proposée présente de solides performances en segmentation d'actions sur toutes les variantes du jeu de données InHARD. La matrice de confusion pour InHARD-13 est présentée en figure 3.9 pour C2F-TCN et en figure 3.10 pour GSK-C2F. L'écart est manifeste. Le modèle C2F-TCN utilisé dans ce travail a été initialement implémenté et ajusté pour être testé sur InHARD ; les résultats présentés en 3.9 révèlent sa sensibilité au déséquilibre du jeu de données. Il classe à tort de nombreuses actions comme "No action" ou "Assemble system", qui représentent ensemble 60% des images du jeu de données. Il parvient néanmoins à bien reconnaître l'action "Put down subsystem", probablement en raison de ses caractéristiques distinctives, notamment sa longue durée et les mouvements marqués de l'opérateur lorsqu'il dépose le "sous-système", ce qui en facilite la détection. Toutefois, il atteint une précision image par image de seulement 34,28 %.

À l'inverse, l'utilisation de la modalité squelette améliore significativement les résultats de segmentation. Avec uniquement les positions articulaires comme caractéristiques, la méthode atteint une précision de 73.86% et un score F1@50 de 71.37%. En intégrant toutes les caractéristiques "positions articulaires, angles osseux et vitesses articulaires, la précision



FIGURE 3.9 – Matrices de confusion C2F-TCN (RGB) sur le jeu de données InHARD-13

image-par-image et le score F1@50 s'élèvent respectivement à 73.90% et 73.96%, comme présenté dans la section 3.6.3.

Avec cette approche, des actions telles que "Take screwdriver" et "Picking left" sont mieux reconnues, leurs mouvements étant mieux capturés et analysés. En revanche, l'action "Consult sheets" peut encore être mal classée comme "No action", en raison de sa forte similarité avec d'autres états inactifs.

Concernant le jeu de données IKEA-ASM, il convient de noter qu'il repose fortement sur

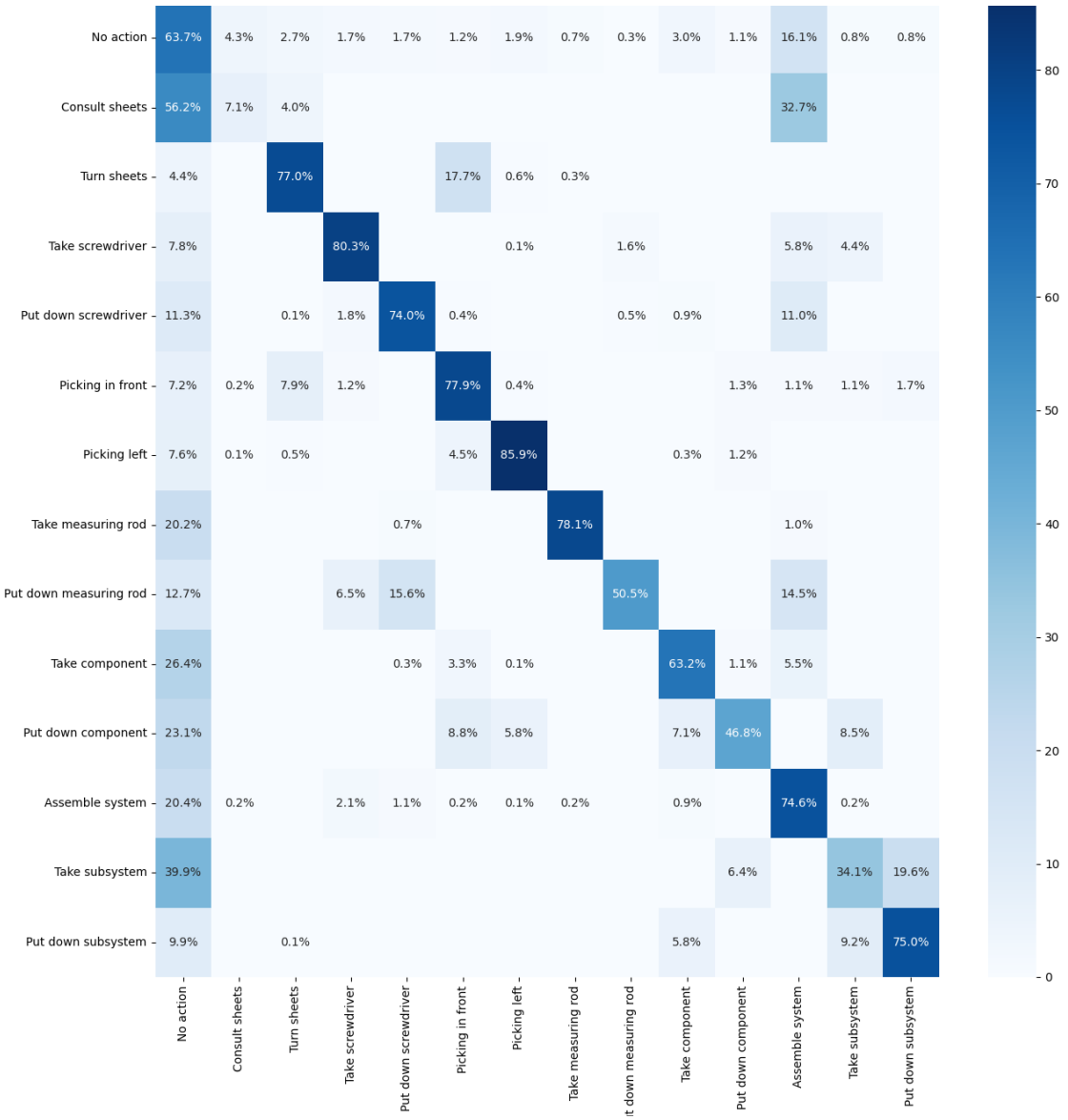


FIGURE 3.10 – Matrices de confusion GSK-C2F(pbv) sur le jeu de données InHARD-13

la présence et la manipulation d'objets ou d'outils pour annoter les actions. Cette dépendance rend la tâche de reconnaissance particulièrement complexe lorsqu'on se limite à l'analyse des seuls schémas de mouvement issus des articulations humaines. En d'autres termes, les caractéristiques squelettiques, bien qu'informatives, ne capturent pas toujours les indices visuels nécessaires pour distinguer certaines actions basées sur les interactions avec les objets. Malgré cette difficulté, notre méthode GSK-C2F a obtenu d'excellents résultats, comme l'illustre le tableau 3.3, atteignant la troisième meilleure précision de segmentation parmi les méthodes

existantes, et ce, sans aucun ajustement préalable ni stratégie de fine-tuning. De plus, l'extraction des caractéristiques squelettiques a été réalisée uniquement à l'aide de Mediapipe, ce qui atteste de la robustesse et de la capacité de généralisation de notre approche, même dans des contextes variés.

Tableau 3.3 – Comparaison des scores MoF (Mean-over-Frames) avec les travaux récents sur le jeu de données IKEA-ASM. Les caractéristiques utilisées dans l'architecture de base sont indiquées entre parenthèses : p (positions articulaires), b (angles osseux), v (vitesses articulaires).

Méthode	Modalité	IKEA-ASM
VAVA [175]	RGB	31.92
I3D [4]	RGB	57.57
CDFL [176]	RGB	60.3
P3D [4]	RGB	60.40
OODL [176]	RGB	<u>63.50</u>
EST [177]	**RGB**	**68.08**
GSK-C2F(p)	E-SK	62.43

Toutefois, au-delà de la précision de la segmentation, un aspect fondamental dans le contexte industriel demeure le coût d'inférence des méthodes déployées. En effet, pour une intégration en conditions réelles, notamment sur des systèmes embarqués ou à ressources limitées, il est crucial de maintenir un compromis optimal entre la performance et la complexité. Comme le montre le tableau 3.4, notre méthode GSK-C2F dépasse nettement les approches concurrentes tant en précision (MoF) qu'en efficacité computationnelle. Avec un score MoF de 73.86% et une complexité de seulement 15.75 GMACs, elle représente le meilleur compromis entre performance et coût d'inférence. Contrairement à ST-GCN ou MTM qui demandent plusieurs centaines, voire des milliers, de GMACs par inférence, GSK-C2F est plus de 80 fois moins coûteux que ST-GCN, tout en étant significativement plus précis. Ce résultat confirme la pertinence de notre approche pour une application en ligne dans des environnements contraints, tout en combinant précision, rapidité et légèreté du modèle.

3.6 Études d'ablation et analyse du modèle

Dans cette section, nous présentons une série d'études d'ablation ainsi qu'une analyse approfondie du comportement du modèle, menées sur les jeux de données InHARD-13, InHARD-4 et InHARD-3. Ces analyses visent à évaluer l'impact de différentes composantes du modèle

Method	MoF	Params (Million)	GMACs
ST-GCN* [119]	52.26	4.81	2549.17
MTM* [90]	<u>56.43</u>	**2.63**	<u>1393.69</u>
GSK-C2F	**73.86**	<u>4.77</u>	**15.75**

TABLEAU 3.4 – Comparaison du score *Mean-over-Frames* (MoF), du nombre de paramètres (en millions) et de la complexité computationnelle (en GMACs) pour différentes approches sur le jeu de données InHARD-13. * : méthodes réimplémentées.

en particulier dans des scénarios industriels complexes.

3.6.1 Ablation sur la pondération des classes

Les jeux de données industriels présentent de nombreux défis, notamment un fort déséquilibre dans la répartition des classes. Certaines actions sont fréquemment répétées lors des tâches d'assemblage, tandis que d'autres apparaissent occasionnellement. Ce déséquilibre affecte négativement la performance du modèle, qui prédit correctement les actions bien représentées, mais peine à apprendre les classes d'actions moins fréquentes.

Pour remédier à ce déséquilibre de données, une stratégie de pondération des classes a été mise en place. Cette pondération est appliquée directement dans la fonction de perte et garantit que les classes sous-représentées bénéficient d'une attention accrue lors de l'entraînement. L'évaluation a porté sur trois stratégies : aucune pondération, pondération avec un coefficient élevé (HCW) et pondération avec un coefficient faible (LCW). Les poids ont été déterminés en fonction de la durée totale des classes d'action, de sorte que les classes les moins fréquentes reçoivent des poids plus importants que les classes dominantes. Les poids de classe ont été calculés comme suit :

Pour chaque classe c dans l'ensemble des classes d'action C, la durée totale TD_c des segments de la classe S_c est calculée par :

$$TD_c = \left\{ \sum_{seg} durée(seg) \mid seg = 1, 2, \ldots, S_c, c \in C \right\} \tag{3.27}$$

Ensuite, les poids correspondants sont définis par :

$$HCW_c = \frac{\max(TD)}{TD_c}, \quad c \in C \tag{3.28}$$

$$LCW_c = \frac{1}{TD_c}, \quad c \in C \tag{3.29}$$

où $\max(TD)$ représente la durée totale de la classe la plus fréquente dans les séquences. Ces poids sont injectés dans la fonction de perte pondérée L_{ce}^{w} (voir section 3.4.2.2, équation 3.24) lors de l'entraînement.

TABLEAU 3.5 – Étude d'ablation de la stratégie de pondération (WS) dans GSK-C2F(p) sur InHARD-13.

WS	F1@10	F1@25	F1@50	Edit	MoF
Aucune	80.27	78.13	69.07	73.75	72.90
HCW	**82.47**	79.89	71.10	73.71	73.56
LCW	81.12	**80.86**	**71.37**	**74.37**	**73.86**

Comme l'illustre le tableau 3.5, l'ajout de poids améliore la précision de la segmentation. En particulier, la stratégie LCW donne de meilleurs résultats pour les caractéristiques de positions articulaires, car elle attribue des poids plus précis à chaque classe que la stratégie HCW. Cela permet de focaliser l'apprentissage sur les classes sous-représentées. L'amélioration est également visible dans la visualisation temporelle de la segmentation (figure 3.11). Dans un contexte industriel, où certaines actions critiques peuvent être sous-représentées, cette stratégie de pondération permet d'accroître la fiabilité des systèmes de reconnaissance.

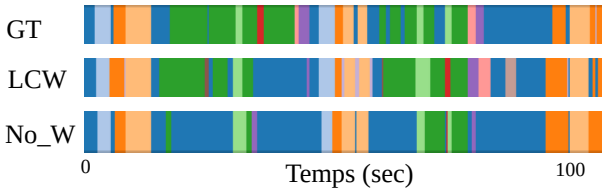


FIGURE 3.11 – Segmentation temporelle pour la vidéo "P13-R02" d'InHARD-13 avec GSK-C2F(p) : GT (vérité du terrain), LCW (pondération faible) et No_W (sans pondération).

3.6.2 Impact de la modalité squelette

La reconnaissance d'actions complexes dans des contextes industriels repose fortement sur la capacité à capturer avec précision les mouvements humains. Or, les environnements

d'assemblage présentent des arrière-plans encombrés et changeants qui compliquent l'analyse à partir de données RGB, bien que ces dernières soient couramment utilisées en raison de leur disponibilité et de leur richesse contextuelle. Toutefois, dans un cadre industriel, l'objectif principal reste la compréhension fine du geste opérateur, et non celle du contexte visuel global. Dans cette optique, les modalités squelettiques, centrées sur la seule structure corporelle, offrent une alternative robuste aux perturbations visuelles. Elles permettent de mieux isoler les dynamiques articulaires pertinentes pour la reconnaissance d'actions, notamment celles impliquant les bras et les mains [11], [178].

Nous comparons ici trois modalités d'entrée : (i) les données RGB brutes, (ii) le squelette 2D extrait automatiquement des vidéos (E-SK), et (iii) le squelette 3D issu d'un système de capture de mouvement (3D-SK). Cette étude vise notamment à valider l'intérêt d'extractions sans capteurs portés, telles que l'E-SK, pour un déploiement léger et non intrusif.

TABLEAU 3.6 – Étude comparative des modalités sur InHARD-13, InHARD-4 et InHARD-3.

Méthode	Jeu de données	Modalité	F1@10	F1@25	F1@50	Edit	MoF
C2F-TCN	InHARD-3	RGB	49,9	38,51	12,83	57,9	50,13
GSK-C2F(p)	InHARD-3	E-SK	68.31	65,23	49,23	**80.21**	80.45
GSK-C2F(p)	**InHARD-3**	**3D-SK**	**74.68**	**73.42**	**56.96**	81.47	**84.82**
C2F-TCN	InHARD-4	RGB	28,2	17,7	8,52	53,23	32,57
GSK-C2F(p)	InHARD-4	E-SK	80.13	70,19	49,06	74.67	67.39
GSK-C2F(p)	**InHARD-4**	**3D-SK**	**81.76**	**75.50**	**57.62**	**76.98**	**70.28**
C2F-TCN	InHARD-13	RGB	26,70	19,12	6,77	35,92	34,28
GSK-C2F(p)	InHARD-13	E-SK	67.83	62.47	43.16	62.84	59.77
GSK-C2F(p)	**InHARD-13**	**3D-SK**	**81.12**	**80.86**	**71.37**	**74.37**	**73.86**

Comme l'illustrent le tableau 3.6 et la figure 3.12, les méthodes basées sur le squelette (E-SK et 3D-SK) surpassent systématiquement celles basées sur le RGB sur l'ensemble des jeux de données InHARD. La modalité 3D-SK fournit des informations squelettiques plus précises et détaillées, ce qui en fait la meilleure option pour la reconnaissance d'actions complexes. La seule exception concerne InHARD-4 où la simplicité et l'équilibre du jeu réduisent l'écart de performance entre E-SK et 3D-SK.

Ces résultats mettent en évidence l'intérêt d'utiliser la modalité squelette dans un cadre industriel ; elle permet de centrer l'analyse sur les mouvements pertinents tout en minimisant les informations de fond non pertinentes [11]. Le modèle bénéficie ainsi d'une meilleure

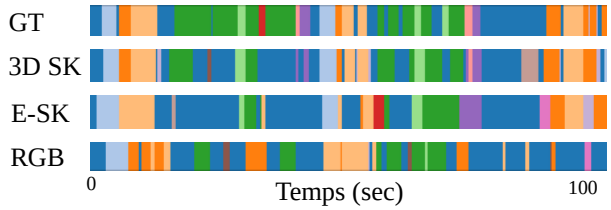


FIGURE 3.12 – Segmentation temporelle de la vidéo "P13-R02" d'InHARD-13 : vérité de terrain (GT) et résultats prédits pour chaque modalité : 3D-SK, E-SK et RGB.

capacité de segmentation temporelle, localisant plus efficacement les segments d'actions et réduisant les erreurs liées aux éléments hors du champ opératoire.

3.6.3 Ablation sur les caractéristiques du squelette

Dans les environnements industriels, la reconnaissance précise des actions humaines nécessite une modélisation fine du mouvement. La modalité squelette permet de capturer plusieurs types d'informations, dynamiques et structurelles, à travers différentes caractéristiques. Dans cette étude, nous explorons trois types de descripteurs squelettiques :

— **p** : les positions relatives des articulations, qui décrivent la posture spatiale instantanée,

— **b** : les angles des membres, qui modélisent les rotations et relations anatomiques,

— **v** : les vitesses articulaires, qui reflètent la dynamique du mouvement.

Chaque descripteur fournit une information partielle et complémentaire. Par exemple, les **positions** permettent une reconnaissance stable des postures, tandis que les **angles** ajoutent une robustesse aux variations inter-individuelles. Les **vitesses**, quant à elles, sont utiles pour détecter des transitions ou des actions rapides. L'objectif de cette étude est d'évaluer la contribution individuelle et combinée de ces caractéristiques dans un contexte d'assemblage industriel. Pour cela, une étude d'ablation a été menée sur le jeu de données InHARD-13.

TABLEAU 3.7 – Étude d'ablation des caractéristiques du squelette sur InHARD-13.

Caractéristiques	F1@{10}	F1@{25}	F1@{50}	Edit	MoF
p (Positions)	81.12	80.86	71.37	74.37	73.86
b (Angles)	74.45	70.10	58.62	68.05	69.08
v (Vitesses)	62.14	57.17	41.06	63.62	61.35
p + b	**82.19**	**81.11**	**73.04**	**75.04**	**74.54**
p + v	75.04	71.32	60.07	69.45	70.63
b + v	72.82	69.54	55.68	65.20	66.03
p + b + v	80.03	79.11	69.52	71.68	71.42

Comme le montre le tableau 3.7, les **positions articulaires (p)** offrent, à elles seules, la meilleure performance parmi les caractéristiques simples, confirmant leur rôle central dans la modélisation du mouvement. L'ajout des **angles des membres (b)** améliore encore les scores sur l'ensemble des indicateurs, révélant une complémentarité utile pour affiner la structure posturale.

En revanche, les **vitesses (v)**, bien qu'informatives en théorie pour détecter les changements rapides, se révèlent ici moins efficaces, probablement en raison de leur sensibilité au bruit et de l'approximation lors de l'extraction des trajectoires. Leur combinaison avec P ou B n'apporte pas de gain significatif, et l'ajout simultané de toutes les caractéristiques (p + b + v) n'atteint pas les performances de la combinaison P + B seule. Cela suggère une *redondance partielle* et une *dispersion du signal utile* lorsqu'on intègre la vitesse.

Ces résultats montrent que la **fusion judicieuse** des caractéristiques squelettiques est essentielle pour maximiser les performances, et que **la combinaison p + b** constitue un bon compromis entre précision, stabilité et robustesse.

3.7 Conclusion

Dans ce chapitre, nous avons proposé **GSK-C2F**, une approche dédiée à la segmentation d'actions dans des vidéos non découpées issues d'environnements industriels. En combinant plusieurs descripteurs squelettiques (positions articulaires et angles des membres), la méthode exploite la richesse et la complémentarité de ces descripteurs pour construire des représentations expressives, particulièrement adaptées aux gestes fins des bras et des mains.

Les différentes études d'ablation ont mis en évidence l'importance du choix et de la combinaison des modalités et ont confirmé l'apport spécifique des angles pour renforcer la structure posturale. GSK-C2F montre ainsi une excellente capacité de généralisation, avec des résultats compétitifs sur plusieurs configurations du jeu de données InHARD et un bon classement sur IKEA-ASM sans ajustement spécifique.

Toutefois, bien que ces représentations capturent efficacement la dynamique des mouvements, elles ne modélisent pas explicitement les relations topologiques entre les articulations. Or, dans un contexte d'assemblage industriel, les gestes sont fortement structurés par des dépendances spatiales (liaisons articulaires) et temporelles (enchaînements d'actions).

Pour franchir ce cap, le prochain chapitre s'intéressera à une modélisation explicite de ces

relations via des **représentations graphiques** du squelette humain. Cette structuration permettra d'exploiter la connectivité naturelle du corps à travers des *graph convolutional networks* (GCNs). Par ailleurs, l'intégration de **mécanismes d'attention** permettra de pondérer dynamiquement l'importance des articulations et des connexions dans le temps, ouvrant la voie à une meilleure *interprétabilité* du modèle, en identifiant par exemple les zones clés activées lors de chaque action. Cette nouvelle perspective vise à renforcer la capacité des modèles à capturer la complexité et la variabilité des gestes humains dans les tâches industrielles.

Chapitre 4

Segmentation d'actions basées sur des graphes avec un mécanisme d'attention

4.1 Introduction

Dans le prolongement du chapitre précédent, qui présentait les bases de la segmentation d'actions à partir de descripteurs squelettiques construits manuellement, ce chapitre introduit notre seconde contribution : un modèle de segmentation temporelle d'actions plus avancé qui exploite directement la structure topologique du squelette humain, représentée sous forme de graphe, au moyen de réseaux de convolution sur graphes (*Graph Convolutional Networks* – GCN) intégrés à une architecture d'apprentissage profond adaptée aux longues séquences industrielles. Pour cela, nous nous appuyons sur l'architecture encodeur-décodeur introduite dans le chapitre précédent, que nous enrichissons ici afin d'y intégrer trois éléments majeurs :

- une modélisation explicite de la connectivité articulaire via les GCN, pour permettre une meilleure capture des relations spatiales,
- un mécanisme d'attention pour offrir au modèle la capacité de se focaliser dynamiquement sur les régions spatio-temporelles les plus pertinentes,
- une stratégie de *differential pooling* est introduite afin de permettre une hiérarchisation progressive et structurée de l'information spatiale tout au long de l'apprentissage.

Nous validons l'efficacité de cette approche à travers des expérimentations approfondies menées sur deux jeux de données publics représentatifs des tâches d'assemblage industriel, avec à l'appui des résultats quantitatifs et des visualisations qualitatives illustrant le comportement du modèle.

Le reste de ce chapitre est structuré comme suit. La section 4.2 introduit la représentation du squelette humain sous forme de graphe ainsi que les principes fondamentaux des réseaux de convolution sur graphes (*Graph Convolutional Networks* – GCNs). La Section 4.3 décrit ensuite en détails l'architecture proposée, y compris le mécanisme d'attention et la stratégie de *differential pooling*. La section 4.4 présente le protocole expérimental ainsi que les détails de l'implémentation. Dans la section 4.5, les performances de notre méthode sont comparées à celles des approches de l'état de l'art. La section 4.6 analyse l'impact des différents modules à travers des études d'ablation, tandis que la section 4.7 propose des visualisations qualitatives pour illustrer ces résultats. Enfin, la section 4.8 conclut en résumant les principaux résultats et en mettant en perspective les apports de cette contribution.

4.2 Modélisation Graphique du Mouvement Humain

Dans le prolongement de notre précédente approche fondée sur des descripteurs squelettiques manuels et une architecture encodeur-décodeur, nous cherchons à exploiter davantage la structure intrinsèque du corps humain. En effet, le squelette peut être naturellement représenté sous forme de graphe, où les articulations constituent des nœuds reliés par des arêtes reflétant les connexions anatomiques. Cette modélisation graphique permet de capturer à la fois les relations spatiales entre les articulations et les dépendances temporelles à travers les séquences de mouvements. Pour tirer parti de cette structure non euclidienne des données squelettiques, les GCNs s'avèrent une solution particulièrement adaptée et puissante comme expliqué en section 4.2.1. Ils permettent d'apprendre des représentations discriminantes en agrégeant l'information locale au sein du graphe, tout en respectant sa topologie. Dans cette section, nous introduisons d'abord les principes essentiels des GCN appliqués à la modélisation du mouvement humain, puis détaillons la topologie spécifique du graphe squelettique retenue dans notre travail.

4.2.1 Réseaux de Convolution Graphique (GCN)

La généralisation des réseaux de neurones aux données structurées sous forme de graphes constitue un domaine en plein essor dans la recherche en apprentissage profond [139]. Ces architectures sont particulièrement adaptées à de nombreux contextes où les données sont naturellement représentées par des graphes, tels que les réseaux sociaux, la modélisation moléculaire ou encore l'analyse du squelette humain. Dans ce dernier cas, le squelette peut être naturellement représenté comme un graphe, les articulations jouant le rôle de nœuds et les connexions anatomiques, celles-ci, d'arêtes. Une telle topologie permet une modélisation efficace des dynamiques corporelles [179].

Au cours des dernières années, un large éventail de modèles de réseaux de neurones graphiques (GNN) a été proposé, incluant les Graph Convolutional Networks (GCN), les Graph Attention Networks (GAT) ou encore les modèles de type GraphSAGE[180]. Ces différentes architectures peuvent être interprétées comme des algorithmes de propagation de messages, où les représentations des nœuds sont calculées de manière itérative à partir de leurs voisins via des fonctions d'agrégation différentiables [181]. Les GNNs ont trouvé des applications dans de nombreux domaines : classification de nœuds[182], [183], prédiction de liens[184],

classification de graphes[122], [185], ou encore en chimio-informatique[186], [187].

Dans notre contexte, où l'objectif est la reconnaissance et la segmentation d'actions humaines à partir de données squelettiques, les GCN représentent une solution particulièrement adaptée[188]. Le squelette humain est ici traité comme un graphe spatio-temporel : les convolutions sont appliquées à la fois sur la structure spatiale (connexions anatomiques entre les articulations à un instant donné) et sur la dimension temporelle (évolution des articulations au fil du temps) [119]. Ce mécanisme, connu sous le nom de Spatio-Temporal Graph Convolutional Network (ST-GCN), permet de capturer simultanément les dépendances spatiales et temporelles entre les séquences d'actions, offrant ainsi une modélisation fine et robuste des mouvements.

Plus récemment, les modèles à attention, tels que les GAT, ont gagné en popularité pour la reconnaissance d'actions. L'intégration de mécanismes d'attention permet en effet d'extraire des représentations plus discriminantes en pondérant l'importance relative des nœuds voisins. Par exemple, Shi et al.[140] ont proposé un réseau GCN adaptatif capable d'apprendre dynamiquement la topologie du graphe squelettique. Sur cette base, Ye et al.[189] ont introduit ADD-GCN, intégrant un module d'attention sémantique afin de renforcer les capacités discriminantes et contextuelles du modèle. Plus récemment encore, Chi et al.[190] ont développé InfoGCN, un framework permettant de modéliser la topologie intrinsèque et contextuelle à l'aide d'un encodeur graphique. Malgré ces progrès, un défi fondamental persiste : comment transformer efficacement les représentations de nœuds obtenues en sortie du réseau en une représentation globale cohérente de l'action ?

Les approches classiques se contentent souvent d'opérations simples comme la somme ou la moyenne des embeddings (espace latent) de nœuds[191], ce qui limite leur capacité à capturer des représentations hiérarchiques riches. Or, dans les contextes industriels, les actions présentent souvent des motifs complexes, marqués à la fois par une forte similarité entre différentes classes (inter-classe) et par une variabilité importante des durées d'exécution au sein d'une même classe (intra-classe). Les méthodes d'agrégation naïves tendent alors à négliger les propriétés structurelles profondes du graphe, ce qui nuit à la qualité de la segmentation [181].

Par ailleurs, la majorité des travaux existants portent sur des actions humaines générales (sportives ou quotidiennes) et ne tiennent pas compte des spécificités des gestes industriels : fins, répétitifs et souvent associés à la manipulation d'outils. Ces caractéristiques rendent la

tâche plus complexe et exigent des architectures capables de capturer des variations subtiles.

Enfin, bien que les GCN reposent sur une topologie fixe efficace pour représenter le squelette, leur capacité à s'adapter aux mouvements dynamiques et variés rencontrés en milieu industriel demeure limitée [192]. Les GAT, grâce à leur mécanisme d'attention contextuelle, offrent une plus grande flexibilité pour gérer cette variabilité, mais peuvent être vulnérables au bruit ainsi qu'aux gestes répétitifs fréquents lors des tâches d'assemblage [193]. Ainsi, malgré leurs succès dans la reconnaissance d'actions générales, ces architectures montrent leurs limites dans des environnements industriels complexes où il est indispensable de modéliser des représentations hiérarchiques. Cela implique de capturer à la fois les relations locales entre articulations proches et les interactions globales entre différentes parties du corps à divers niveaux d'abstraction. Par ailleurs, la segmentation doit intégrer le contexte global de la séquence tout en distinguant les variations subtiles propres aux gestes industriels. Enfin, la robustesse face au bruit et aux mouvements similaires est essentielle pour assurer une segmentation précise et fiable. Les modèles classiques peinent à concilier ces exigences, ce qui en réduit l'efficacité dans ce cadre spécifique.

Face à ces défis, il devient essentiel de développer des architectures capables de modéliser à la fois la complexité spatiale et temporelle des mouvements industriels, tout en intégrant des mécanismes d'attention et d'agrégation hiérarchique pour mieux capturer les subtilités des gestes. C'est dans cette optique que nous proposons, dans ce chapitre, une nouvelle méthode spécifiquement adaptée à la segmentation fine d'actions en contexte industriel, notamment les actions d'assemblage. En nous appuyant sur les travaux récents [60], [115], [181], [189], [190], nous intégrons une représentation graphique du squelette au sein d'un cadre d'apprentissage profond avec encodeur-décodeur, enrichi par un mécanisme d'attention et un module d'agrégation hiérarchique. Cette architecture conserve la structure spatiale à plusieurs niveaux, ce qui permet au modèle de mieux saisir les relations locales et globales entre les articulations. Cette compréhension hiérarchique améliore ainsi la précision de la segmentation des actions à partir des données squelettiques.

4.2.2 Topologie du graphe squelettique

Dans notre approche, nous modélisons l'opérateur en mouvement sous la forme d'un graphe squelettique, qui constitue la base du traitement par réseaux de neurones sur graphes. Chaque image d'une vidéo est ainsi convertie en un ensemble de coordonnées articulaires,

extraites automatiquement à partir de la détection du squelette humain. Ce squelette est ensuite formalisé comme un graphe $G = (V, E)$, où V représente les articulations (nœuds) et E les connexions anatomiques (arêtes) entre elles. La qualité de cette modélisation a un impact direct sur la capacité du modèle à apprendre des représentations discriminantes pour les gestes complexes et souvent subtils rencontrés dans les environnements industriels.

La structure du graphe est décrite par une matrice d'adjacence $\mathbf{A} \in \mathbb{R}^{N \times N}$, où $A_{ij} = 1$ si les articulations i et j sont connectées par une arête anatomique, et 0 sinon. Sur le plan temporel, une séquence d'actions est représentée par un tenseur $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$, où T est le nombre d'images (frames), N le nombre d'articulations, et C le nombre de canaux caractéristiques (par exemple, coordonnées 3D, vitesses articulaires). Cette représentation spatio-temporelle forme l'entrée du réseau de neurones, et permet de capturer conjointement les relations structurelles entre articulations et leur évolution dynamique au cours du temps.

Afin de structurer efficacement les relations spatiales entre les nœuds du graphe (articulations), nous adoptons une décomposition hiérarchique de la matrice d'adjacence, inspirée de Yan et al.[139]. Cette approche distingue plusieurs types de connexions entre articulations, permettant au modèle de traiter des dépendances à différentes échelles :

— $\mathbf{A}^{(0)}$: connexions internes (self-links), reliant chaque articulation à elle-même pour préserver son information propre.
— $\mathbf{A}^{(1)}$: connexions proches (voisinage spatial), reliant les articulations directement connectées par des os anatomiques.
— $\mathbf{A}^{(2)}$: connexions éloignées, reliant les articulations situées à une distance plus grande dans la structure du squelette, afin de modéliser des interactions à plus grande portée.

Cette modélisation à trois branches permet de capturer simultanément les relations locales, intermédiaires et globales entre les articulations, enrichissant ainsi la capacité expressive du modèle. Pour garantir une propagation stable de l'information à travers le graphe, chaque matrice $A^{(k)}$ est normalisée de manière symétrique selon la formule suivante :

$$\hat{\mathbf{A}}^{(k)} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A}^{(k)} \mathbf{D}^{-\frac{1}{2}}, \quad \text{pour } k = 0, 1, 2 \tag{4.1}$$

où $\mathbf{D}$ est la matrice diagonale des degrés associée à $\mathbf{A}^{(k)}$. Cette normalisation assure que les contributions des nœuds voisins sont pondérées de manière équilibrée, évitant ainsi des effets de sur- ou de sous-propagation d'information. Les matrices normalisées obtenues sont

ensuite empilées pour former un tenseur d'adjacence multi-branche, intégré dans le module de convolution spatio-temporelle :

$$\hat{\mathbf{A}} = \langle \hat{\mathbf{A}}^{(0)}, \hat{\mathbf{A}}^{(1)}, \hat{\mathbf{A}}^{(2)} \rangle \in \mathbb{R}^{3\times N\times N} \tag{4.2}$$

Cette structure permet au réseau d'extraire simultanément des informations issues de différentes échelles spatiales à chaque étape de convolution.

Afin de confirmer l'intérêt de cette topologie hiérarchique, nous avons mené une étude comparative de différentes configurations, que nous détaillons dans la section 4.6.2. Cette analyse nous a permis d'identifier la topologie la plus adaptée aux particularités des gestes industriels, caractérisés par des transitions rapides, une symétrie partielle et une coordination fine des membres.

4.3 Approche méthodologique proposée

4.3.1 Formulation du problème

Dans cette étude, l'objectif est de prédire, pour chaque image d'une séquence vidéo, la classe d'action réalisée par un opérateur lors d'une tâche d'assemblage industriel. Comme détaillé dans la section 4.2.2, le corps humain est représenté sous la forme d'un graphe squelettique $G = (V, E)$, dont la topologie encode les relations anatomiques entre les articulations. La séquence temporelle correspondante est décrite par un tenseur $\mathbf{X} \in \mathbb{R}^{T\times N\times C}$, où T, N et C désignent respectivement le nombre d'images, d'articulations et de canaux caractéristiques. Ainsi, le problème de reconnaissance et de segmentation d'actions peut être formulé comme une fonction d'apprentissage f_θ paramétrée par les poids du modèle, qui associe à chaque séquence squelettique $\mathbf{X}$ une séquence de prédictions de classes d'actions $\hat{\mathbf{Y}}$:

$$\hat{\mathbf{Y}} = f_\theta(\mathbf{X}, \mathbf{A}), \tag{4.3}$$

où $\mathbf{A}$ est la matrice d'adjacence décrivant la structure du graphe squelettique (cf. section 4.2.2).

Contrairement aux réseaux de convolution classiques (CNN) opérant sur des grilles euclidiennes fixes, les réseaux de convolution sur graphes (GCN) (voir section 4.2.1) permettent de modéliser efficacement les interactions spatiales et temporelles entre les articulations. Cette

capacité à exploiter la structure non euclidienne du corps humain constitue un atout majeur pour la segmentation fine d'actions complexes et souvent subtiles dans des contextes industriels.

Dans la section suivante, nous présentons l'architecture proposée, **G-EAD**, qui met en œuvre ce principe de modélisation spatio-temporelle à l'aide de convolutions sur graphes et de mécanismes d'attention hiérarchiques.

4.3.2 Architecture proposée

Dans la continuité de notre méthodologie décrite précédemment en chapitre 3.3, nous proposons une nouvelle architecture baptisée **G-EAD** (*Graph Encoder-Attention-Decoder*), qui étend et améliore les fondations posées par le modèle **GSK-C2F** [115]. Ce dernier reposait sur une stratégie *coarse-to-fine* multi-échelle efficace pour affiner la segmentation des actions à différents niveaux de granularité temporelle. Toutefois, il n'exploitait que des caractéristiques globales sans tirer pleinement parti de la structure spatiale du squelette.

Pour surmonter ces limitations, G-EAD intègre explicitement une représentation graphique du squelette, exploitant ainsi les dépendances spatio-temporelles entre les articulations. L'architecture introduit également plusieurs modules novateurs, dont un encodeur basé sur des graphes, un mécanisme d'attention multi-niveau et un décodeur hiérarchique, afin d'extraire, enrichir et exploiter les motifs dynamiques les plus pertinents.

Comme illustré à la figure 4.1, l'approche G-EAD traite une séquence de données squelettiques 3D en entrée, encode les relations complexes entre les articulations au fil du temps, puis produit une sortie segmentée image par image. Chaque segment est associé à une classe d'action (par exemple, "Prendre à gauche" (Pick left), "Tourner la feuille" (Turn sheet)), avec une précision renforcée par l'utilisation conjointe de la topologie squelettique et du raffinement temporel multi-échelle.

Les sections suivantes sont organisées comme suit : nous présentons d'abord une vue d'ensemble de l'approche proposée, puis détaillons ses composants principaux, notamment les convolutions spatio-temporelles sur graphes, l'encodeur, le goulot d'étranglement, le mécanisme d'attention, les blocs de décodage et, pour finir, le bloc d'agrégation.

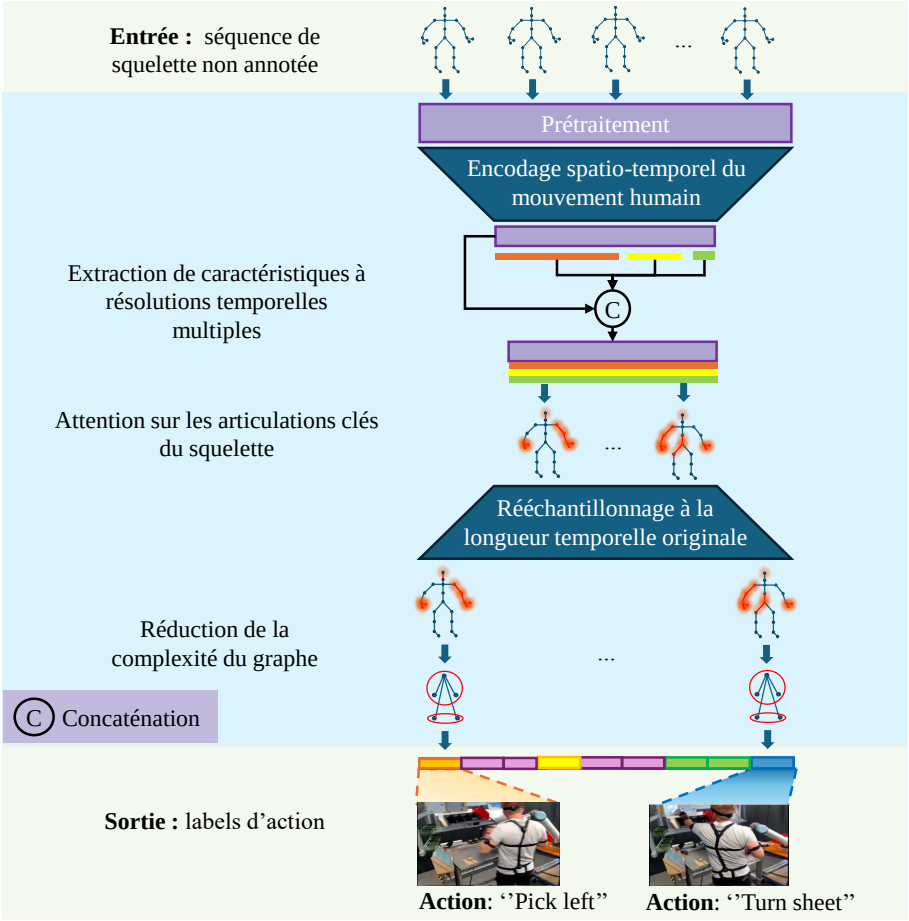


FIGURE 4.1 – Schéma global de l'approche proposée : du signal d'entrée, constitué d'une séquence 3D de squelettes (en haut), jusqu'à la sortie segmentée des actions (en bas). Chaque segment correspond à une classe d'action (par exemple, "Prendre à gauche (Pick left)", "Tourner la feuille (Turn sheet)").

4.3.3 Vue d'ensemble de l'approche

L'approche **G-EAD**, illustrée dans la figure 4.2, prend en entrée une séquence vidéo représentant une activité d'assemblage, où chaque image est représentée sous forme de données de squelette, et produit en sortie une séquence segmentée, dans laquelle chaque image est associée à une classe d'action.

Les données squelettiques, qu'elles soient capturées à partir de systèmes de capture de mouvement ou extraites d'images RGB, sont d'abord prétraitées, puis introduites dans l'architecture G-EAD. Le modèle se compose de trois éléments clés :

— une architecture encodeur-décodeur basée sur des graphes ;

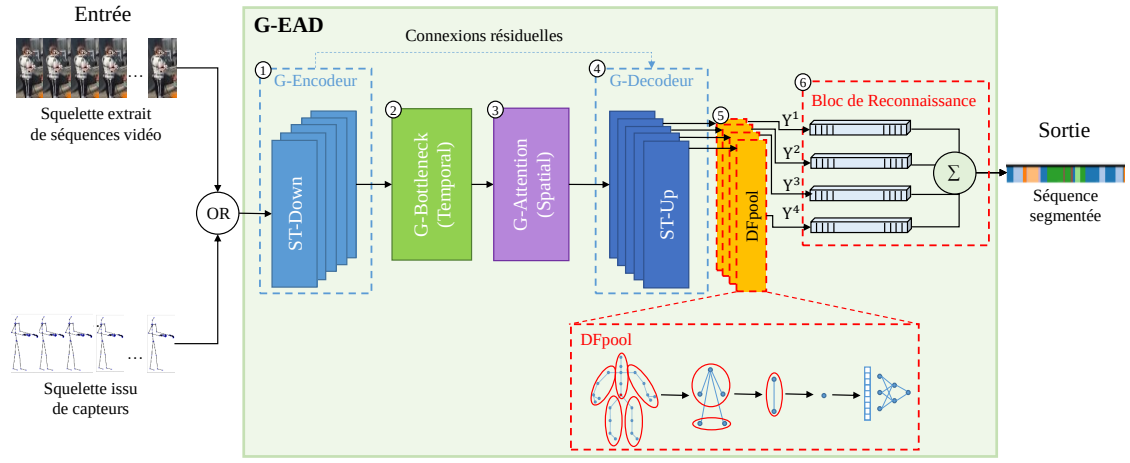


FIGURE 4.2 – Vue d'ensemble de l'architecture proposée, nommée G-EAD : d'abord, les séquences vidéo de squelettes sont traitées par un encodeur de graphes (1), suivi d'un goulot d'étranglement (2) pour capturer des caractéristiques multi-résolutions. Ensuite, un mécanisme d'attention (3) est appliqué afin de se concentrer sur les articulations pertinentes. Les séquences sont ensuite traitées par un décodeur de graphes (4), dont la sortie de chaque couche est transmise à un module de regroupement différentiel (5) afin de réduire hiérarchiquement le graphe. Enfin, la segmentation des actions est réalisée par un bloc d'agrégation à prédiction ensembliste (6).

— un bloc d'agrégation utilisant le regroupement différentiel ;

— un bloc de reconnaissance et de segmentation.

Le bloc encodeur de graphes transforme les séquences de données squelettiques en représentations graphiques en appliquant des convolutions sur les articulations, en tenant compte des connexions spatiales (entre les articulations) et de l'évolution temporelle (au fil des images) à différentes échelles (différentes couches d'encodage). Cela permet de capturer efficacement les relations entre les mouvements dans l'espace et dans le temps (voir la figure 4.2-1).

Au niveau du goulot d'étranglement, le modèle extrait des représentations internes à plusieurs résolutions temporelles, ce qui lui permet de tenir compte de la dynamique des actions à différentes échelles, un aspect clé pour une segmentation fine des séquences vidéo [115] (voir la figure 4.2-2).

Ensuite, un mécanisme d'attention spatiale est appliqué pour permettre au modèle de se focaliser sur les régions les plus informatives du squelette. Il évalue l'importance relative de chaque articulation à chaque instant, en apprenant des poids d'attention, ce qui permet au modèle de mieux représenter les configurations articulaires impliquées dans des gestes complexes (voir la figure 4.2-3).

Les représentations enrichies par attention sont ensuite transmises au décodeur de graphes, où un regroupement différentiel est appliqué. Cette stratégie hiérarchique permet de réduire la taille du graphe tout en conservant les structures discriminantes essentielles (voir la figure 4.2-(4 et 5)).

Enfin, le bloc de reconnaissance et de segmentation fusionne les prédictions issues de plusieurs niveaux du décodeur à l'aide d'une méthode de combinaison pondérée. Cette combinaison multi-niveaux permet non seulement de réduire les effets de sur-segmentation, mais surtout d'assurer une classification temporelle plus cohérente, précise et adaptée aux gestes répétitifs et subtils observés dans les tâches d'assemblage (voir la figure 4.2-6).

4.3.4 Réseau de convolution spatio-temporel sur graphe

Pour modéliser à la fois la structure spatiale du squelette humain et son évolution temporelle, nous combinons les réseaux de convolution sur graphes (GCN), adaptés aux relations topologiques entre articulations, avec les réseaux de convolution temporelle (TCN), capables de modéliser les dépendances temporelles. L'architecture résultante, appelée ST-GCN (Spatial Temporal Graph Convolutional Network), est particulièrement efficace pour la reconnaissance et la segmentation d'actions humaines à partir de données squelettiques [139]. Elle permet d'extraire conjointement les motifs spatiaux et temporels pertinents à la compréhension du mouvement.

4.3.4.1 Convolution spatiale sur graphe.

Un réseau de convolution sur un graphe est utilisé pour modéliser les dépendances spatiales entre les articulations connectées. À chaque couche, les caractéristiques du nœud h_i sont mises à jour en agrégeant celles de ses voisins $j \in N(i)$, pondérées par la matrice d'adjacence normalisée $\hat{A}$ (voir la section 4.2.2). Pour capturer des relations de voisinage à différents niveaux, la matrice d'adjacence est décomposée en trois composantes distinctes, selon [139] :

$$\mathbf{A}^{(0)} = \mathbf{A}_{\text{root}}, \quad \mathbf{A}^{(1)} = \mathbf{A}_{\text{root}} + \mathbf{A}_{\text{close}}, \quad \mathbf{A}^{(2)} = \mathbf{A}_{\text{further}} \tag{4.4}$$

où :

— $\mathbf{A}_{\text{root}}$ correspond aux auto-connections à distance de saut zéro, préservant les caractéristiques individuelles des articulations,

— $\mathbf{A}_{\text{close}}$ les connexions du voisinage local en combinant les nœuds équidistants du centre avec ceux spatialement plus proches,

— $\mathbf{A}_{\text{further}}$ encode les dépendances à plus longue portée où le nœud source est plus éloigné du centre que le nœud cible.

Chacune de ces matrices est ensuite normalisée pour assurer la stabilité de l'apprentissage, via une normalisation symétrique :

$$\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \tag{4.5}$$

où $\mathbf{D}$ est la matrice de degré associée à $\mathbf{A}$. On applique cette normalisation aux composantes décomposées, donnant :

$$\hat{\mathbf{A}}^{(0)} = \hat{\mathbf{A}}_{\text{root}}, \quad \hat{\mathbf{A}}^{(1)} = \hat{\mathbf{A}}_{\text{root}} + \hat{\mathbf{A}}_{\text{close}}, \quad \hat{\mathbf{A}}^{(2)} = \hat{\mathbf{A}}_{\text{further}} \tag{4.6}$$

Ces matrices sont ensuite empilées pour former un tenseur d'adjacence normalisé final :

$$\hat{\mathbf{A}} = \left[\hat{\mathbf{A}}^{(0)}, \hat{\mathbf{A}}^{(1)}, \hat{\mathbf{A}}^{(2)}\right], \quad \hat{\mathbf{A}} \in \mathbb{R}^{3\times N\times N} \tag{4.7}$$

Cette structure multi-niveaux permet de modéliser efficacement la dynamique spatiale, en tenant compte à la fois des relations locales et globales entre articulations. La mise à jour des caractéristiques d'un nœud i à la couche $l+1$ s'écrit :

$$\mathbf{H}_i^{(l+1)} = \text{ReLU}\left(\sum_{k=0}^{2}\sum_{j=1}^{N} \hat{\mathbf{a}}_{ij}^{(k)}\, \mathbf{H}_j^{(l)}\, \mathbf{W}_k^{(l)}\right) \tag{4.8}$$

où $\text{ReLU}(\cdot)$ est la fonction d'activation non linéaire, $\hat{a}_{ij}^{(k)}$ représente le coefficient normalisé décrivant l'influence du nœud j sur le nœud i dans la partition k, $\mathbf{H}_j^{(l)} \in \mathbb{R}^{C_l}$ représente la caractéristique du nœud j à la couche l, et $\mathbf{W}_k^{(l)} \in \mathbb{R}^{C_l \times C_{l+1}}$ est la matrice de poids associée à la partition k. En notation matricielle compacte, l'opération de convolution sur graphe peut être exprimée comme :

$$\text{GCN}(\mathbf{H}^{(l)}) = \text{ReLU}\left(\sum_{k=0}^{2} \hat{\mathbf{A}}^{(k)}\mathbf{H}^{(l)}\mathbf{W}_k^{(l)}\right) \tag{4.9}$$

où $\hat{\mathbf{A}}^{(k)} \in \mathbb{R}^{N\times N}$ désigne la matrice d'adjacence normalisée correspondant à la partition k,, $\mathbf{H}^{(l)} \in \mathbb{R}^{N\times C_l}$ est la matrice des caractéristiques des nœuds à la couche l, et la sortie appartient à $\mathbb{R}^{N\times C_{l+1}}$.

4.3.4.2 Convolution temporelle.

En complément de la convolution spatiale sur graphe, un bloc de convolution temporelle (TCN) est utilisé afin de modéliser les dépendances dynamiques le long de la dimension temporelle T. Les dépendances entre instants successifs sont capturées par des convolutions appliquées le long de l'axe temporel. Le bloc TCN est composé de deux blocs *DoubleConv*, chacun comprenant une convolution 2D de noyau $(K_t, 1)$, suivie d'une normalisation par lot et d'une activation ReLU. Cette opération relie chaque articulation à ses voisins temporels $(t-k, \ldots, t, \ldots, t+k)$, sans introduire de connexions supplémentaires entre articulations.

4.3.4.3 Intégration dans le G-Encodeur

Basé sur cette combinaison, le ST-GCN a été intégré dans le composant encodeur-décodeur de notre approche GSK-C2F, présenté au chapitre 3.3.2. Cette intégration vise à enrichir l'architecture initiale en exploitant conjointement la structure spatiale du squelette et la dynamique temporelle des actions. Contrairement aux approches purement convolutives ou séquentielles, le ST-GCN permet de modéliser explicitement les dépendances entre articulations à l'aide d'une représentation en graphe, tout en capturant l'évolution des mouvements dans le temps [139]. Cela améliore la qualité des descripteurs extraits et renforce la capacité du modèle à distinguer des actions complexes dans des environnements industriels contraints.

4.3.5 G-Encodeur

Pour analyser efficacement les séquences d'actions humaines dans un contexte industriel, il est essentiel de capturer simultanément les relations spatiales entre les articulations et l'évolution temporelle des mouvements. C'est dans cette optique qu'un module encodeur, appelé **G-Encodeur**, a été conçu. Celui-ci constitue la première étape du traitement dans notre architecture et joue un rôle central dans l'extraction de représentations discriminantes à partir des données squelettiques.

Cet encodeur repose sur l'architecture **ST-GCN**, présentée dans la section 4.3.4, et l'enrichit par une structure hiérarchique à plusieurs couches. Cette structure permet un *sous-échantillonnage progressif* dans le temps, crucial pour modéliser les dépendances à long terme sans perdre les motifs de mouvement fins. L'objectif est double : (i) compresser efficacement la séquence d'entrée tout en conservant ses propriétés discriminantes, et (ii) apprendre des

représentations spatio-temporelles riches et hiérarchisées adaptées à la reconnaissance et segmentation d'actions complexes.

Concrètement, l'entrée du G-Encodeur est une séquence squelettique représentée par un tenseur $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$, où T est la longueur temporelle, N le nombre d'articulations, et C la dimension des descripteurs. Cette entrée est ensuite traitée par une série de cinq couches séquentielles, notées `ST-Down`, dont la structure est illustrée dans la figure 4.3.

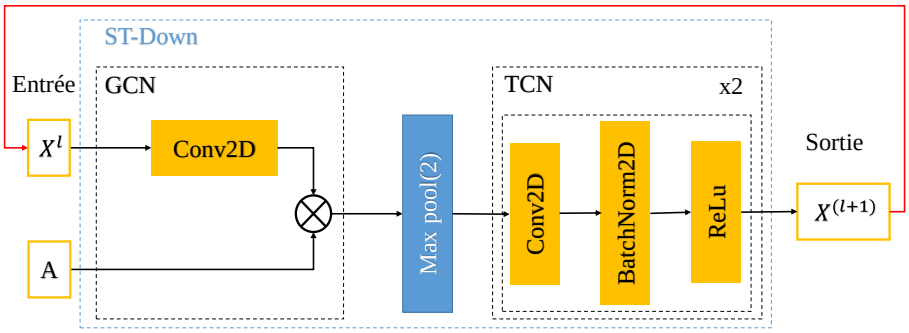


FIGURE 4.3 – Structure d'une couche d'encodeur (`ST-Down`) : les représentations squelettiques passent d'abord par une convolution graphique (GCN), suivie d'un max-pooling temporel, puis d'une convolution temporelle (TCN). Cette combinaison permet de capter des motifs d'action tout en réduisant la durée de la séquence. A : matrice d'adjacence, $\otimes$: agrégation spatiale par multiplication matricielle.

Chaque couche comprend trois étapes principales : (i) **Une convolution graphique (GCN)** permettant de modéliser les relations spatiales entre les articulations, en s'appuyant sur la structure du graphe squelettique, codée par la matrice d'adjacence A (voir la section 4.3.4.1). (ii) **Une opération de max-pooling temporel** qui réduit la résolution temporelle de moitié, favorisant la capture de dépendances à plus long terme à chaque couche. (iii) **Une convolution temporelle (TCN)** qui extrait des motifs de mouvement à différentes échelles temporelles (voir la section 4.3.4.2). Au niveau de la couche l de l'encodeur, la transformation s'écrit formellement comme suit :

$$\mathbf{X}^{(l+1)} = \mathrm{TCN}\left(\mathrm{MaxPool}\left(\mathrm{GCN}(\mathbf{X}^l, \mathbf{A})\right)\right) \tag{4.10}$$

avec $\mathbf{X}^{(l)} \in \mathbb{R}^{T^l \times N \times C}$ et $\mathbf{X}^{(l+1)} \in \mathbb{R}^{\frac{T^l}{2} \times N \times C}$, où la longueur temporelle est réduite d'un facteur 2 à chaque niveau de l'encodeur (*i.e.*, $T^{(l+1)} = \frac{T^{(l)}}{2}$).

Cette structure hiérarchique du G-Encodeur permet non seulement de compresser l'information de manière contrôlée, mais aussi d'extraire des descripteurs robustes à différents

niveaux de granularité. Elle est particulièrement adaptée au traitement de séquences d'actions longues, comme c'est souvent le cas dans les environnements industriels, notamment pour les activités d'assemblage.

4.3.6 G-Bottleneck

Le module G-Bottleneck constitue une amélioration directe du goulot d'étranglement initialement introduit dans l'architecture GSK-C2F (voir le chapitre 3.3.2.2). Dans cette version étendue, nous intégrons un *temporal pyramid pooling* (TPP), inspiré de ses succès en reconnaissance d'images et en analyse vidéo [104], [115]. Contrairement au bottleneck original de GSK-C2F, qui visait essentiellement à compresser les données sous la forme d'un vecteur fixe, le module proposé exploite le TPP comme un mécanisme d'**augmentation temporelle multi-échelle**, spécifiquement adapté aux données squelettiques tout en respectant la topologie du graphe. Par ailleurs, un bloc d'**attention spatiale** (voir la section 4.3.7) est intégré afin de pondérer dynamiquement les articulations en fonction de leur importance dans le contexte de l'action. Ce mécanisme doit améliorer non seulement la performance du modèle en focalisant l'attention sur les articulations les plus pertinentes, mais apporter également une meilleure interprétabilité. Il devient possible d'identifier quelles parties du squelette contribuent le plus à la reconnaissance de chaque action, facilitant ainsi l'interprétation des décisions du modèle. La figure 4.4 illustre le fonctionnement du module bottleneck, dont l'entrée correspond à la sortie finale de l'encodeur $\mathbf{X} \in \mathbb{R}^{T_f \times N \times C}$, où T_f représente la longueur temporelle obtenue après les opérations de sous-échantillonnage de l'encodeur, N le nombre d'articulations, et C la dimension des descripteurs.

Ce module constitue une adaptation du bottleneck présenté au Chapitre 3.3.2.2, où la principale différence réside dans la prise en compte explicite de la structure squelettique, grâce à des convolutions 2D plutôt qu'à des convolutions 1D. Quatre opérations de max-pooling parallèles sont appliquées sur la dimension temporelle avec des noyaux de taille k_i, suivies chacune d'une convolution 2D pour réduire la dimension des canaux, avant un ré-échantillonnage à la longueur temporelle originale :

$$\mathbf{X}_i \quad = \quad \mathrm{Up}(\mathrm{Conv}(\mathrm{MaxPool}(\mathbf{X}, k_i)), T), \quad k_i \in \{2, 3, 5, 6\} \tag{4.11}$$

où *Conv* est une convolution partagée 2D et *Up* une opération d'upsampling temporelle.

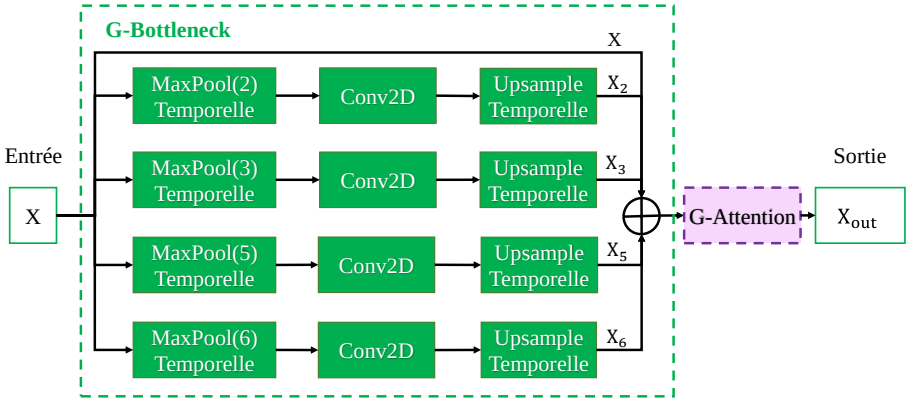


FIGURE 4.4 – Illustration du module G-Bottleneck : des opérations de max-pooling parallèles sont appliquées sur l'axe temporel à l'aide de différents noyaux. Chaque branche subit une compression de canaux par convolution, puis un rééchantillonnage à la longueur initiale. L'ensemble est concaténé à la représentation d'origine.

Enfin, les quatre représentations multi-échelles ainsi obtenues sont concaténées avec la représentation originale $\mathbf{X}$ le long de la dimension des canaux, puis raffinées à l'aide du module G-Attention présenté en section 4.3.7, formant la sortie $\mathbf{X}_{\text{out}} \in \mathbb{R}^{T \times N \times (4 \times d^x + C)}$:

$$\mathbf{X}_{\text{out}} = \text{G-Attention}(\text{Concat}(\mathbf{X}, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_5, \mathbf{X}_6)) \tag{4.12}$$

Cette adaptation du bottleneck rend le modèle plus sensible aux variations d'échelle temporelle, ce qui doit renforcer son aptitude à distinguer des actions similaires mais de durées différentes, un défi récurrent dans les environnements industriels complexes. Ci-après, nous détaillons le mécanisme d'attention spatiale intégré à notre architecture.

4.3.7 G-Attention

Après l'agrégation temporelle multi-échelle, nous appliquons un bloc d'attention graphique afin d'améliorer la modélisation spatiale. Alors que le pooling pyramidal se concentre sur des résolutions temporelles variées, le module d'attention proposé est intégré pour renforcer la capacité du modèle à se focaliser sur les articulations les plus pertinentes dans les séquences d'actions basées sur le squelette. En pondérant dynamiquement l'importance de chaque articulation à différents instants, ce mécanisme doit aider le modèle à intégrer plus précisément les dépendances spatiales.

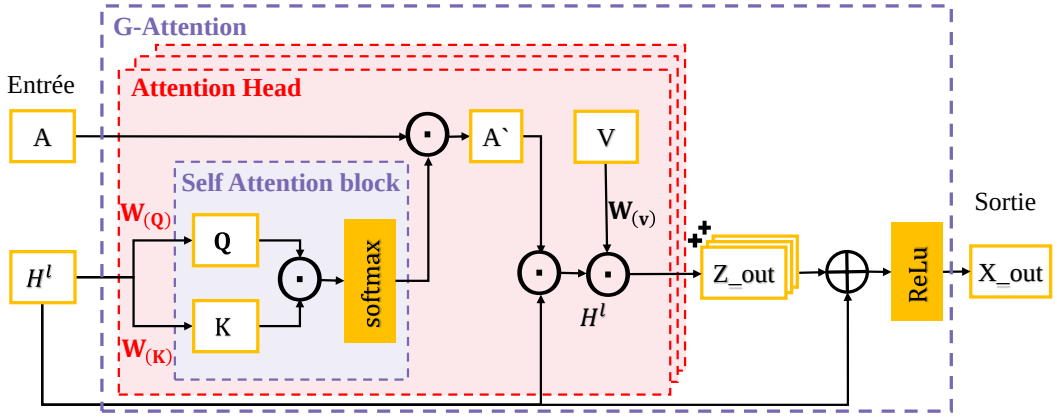


FIGURE 4.5 – Illustration de G-Attention : d'abord, les requêtes et les clés sont extraites des caractéristiques d'entrée H pour produire une matrice de similarité spatiale. Puis, les scores d'attention sont calculés via une opération softmax et modulés par la matrice d'adjacence A par multiplication élément par élément. Enfin, plusieurs têtes d'attention sont concaténées, suivies d'une connexion résiduelle et d'une activation ReLU. $\odot$: multiplication élément par élément.

Soit une carte de caractéristiques d'entrée $\mathbf{f}^{(\text{in})} \in \mathbb{R}^{T\times N\times C}$, où T désigne la longueur temporelle, N le nombre d'articulations, et C la dimension des canaux. Les caractéristiques sont projetées dans un espace d'embedding de dimension C_e, produisant $\tilde{\mathbf{f}} \in \mathbb{R}^{T\times N\times C_e}$. La représentation est ensuite réorganisée afin de regrouper l'information temporelle pour chaque articulation $\bar{\mathbf{H}}^{(l)} \in \mathbb{R}^{N\times(TC_e)}$. Un encodage positionnel est ajouté afin d'injecter explicitement une information sur la position temporelle dans la séquence. Pour chaque position temporelle $t \in \{1,\dots,T\}$ et chaque dimension k, l'encodage peut être défini sous forme sinusoïdale :

$$\mathbf{PE}(t,2k) = \sin\left(\frac{t}{10000^{\frac{2k}{C_e}}}\right), \quad \mathbf{PE}(t,2k+1) = \cos\left(\frac{t}{10000^{\frac{2k}{C_e}}}\right).$$

On obtient ainsi une matrice d'encodage positionnel $\mathbf{PE} \in \mathbb{R}^{T\times C_e}$, où chaque pas de temps possède une représentation unique. L'encodage positionnel $\mathbf{PE}$ est ensuite ajouté à la représentation avant le calcul de l'attention :

$$\mathbf{H}^{(l)} = \tilde{\mathbf{H}}^{(l)} + \mathbf{PE}.$$

L'embedding d'entrée $\mathbf{H}^{(l)}$ est projeté linéairement en requêtes $\mathbf{Q}$, clés $\mathbf{K}$ et valeurs $\mathbf{V}$ via des matrices de poids apprenables $\mathbf{W}_Q$, $\mathbf{W}_K$, et $\mathbf{W}_V$ respectivement :

$$\mathbf{Q} = \mathbf{H}^{(l)}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{H}^{(l)}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{H}^{(l)}\mathbf{W}_V \tag{4.13}$$

Ces projections sont implémentées par des couches convolutionnelles 1×1. Pour calculer les scores d'attention, une matrice de similarité paire à paire $C_k \in \mathbb{R}^{N \times N}$ est obtenue par produit scalaire entre requêtes et clés :

$$\mathbf{C}_k = \mathbf{Q}\mathbf{K}^\top \tag{4.14}$$

Les scores d'attention $\mathbf{SA} \in \mathbb{R}^{N \times N}$ sont ensuite calculés via un softmax appliqué par ligne :

$$\mathbf{SA} = \text{softmax}(\mathbf{C}_k) = \text{softmax}(\mathbf{Q}\mathbf{K}^\top) \tag{4.15}$$

Les scores d'attention sont alors modulés par la structure du graphe grâce à une multiplication élément par élément avec la matrice d'adjacence $\mathbf{A}$, ce qui concentre l'attention sur la connectivité du graphe. Les scores modulés servent à pondérer les valeurs $\mathbf{V}$ modulées par l'embedding d'entrée $\mathbf{H}$ afin de produire la matrice de sortie $\mathbf{Z}$ pour chaque tête d'attention. Les sorties des différentes têtes sont concaténées, suivies d'une connexion résiduelle avec une activation ReLU afin de conserver l'information d'entrée originale et de générer la sortie finale du bloc d'attention graphique. Ce processus est illustré en figure 4.5 et formulé comme suit :

$$\mathbf{X}_{\text{out}} = \text{ReLU}\left(\text{Concat}\left((\mathbf{A} \odot \mathbf{SA})\right)\mathbf{H}\mathbf{V}_i\right)_{i=1}^{h} + \mathbf{H} \tag{4.16}$$

où $\mathbf{H}$ désigne l'embedding d'entrée initial, h est le nombre total de têtes d'attention, et $\odot$ désigne la multiplication élément par élément. Cette architecture doit permettre au modèle d'apprendre dynamiquement la pertinence entre les articulations, ce qui facilite la capture des interactions complexes au sein du squelette graphique.

4.3.8 G-Decoder

Le module décodeur est conçu pour compléter l'encodeur et restaurer la résolution temporelle de la séquence d'entrée. Alors que l'encodeur effectue un échantillonnage progressif descendant (downsampling) pour extraire des caractéristiques temporelles de haut niveau, le décodeur travaille à reconstruire la dimension temporelle et à affiner la sortie, garantissant

ainsi que le modèle puisse générer une segmentation détaillée et précise de la séquence d'actions. Il comprend également un **bloc d'agrégation différentiel**, présenté dans la section 4.3.9. Ce mécanisme apprend à regrouper les nœuds du graphe en clusters pertinents à l'aide d'un processus différentiable, ce qui permet une compression structurée de l'information.

L'architecture du décodeur est symétrique à celle de l'encodeur, à l'exception du bloc GCN, qui n'y figure pas (voir la figure 4.6). Il est composé de cinq couches, chacune comprenant une unité de suréchantillonnage (up-sampling) et un bloc convolutionnel constitué d'une convolution 2D, d'une normalisation de batch 2D (BatchNorm) et d'une fonction d'activation ReLU. À chaque couche, l'unité de suréchantillonnage double la longueur temporelle de l'entrée $\mathbf{X}_{gd}^{l}$ par interpolation linéaire, puis fusionne la sortie correspondante de l'encodeur via des connexions résiduelles (skip connections) $\mathbf{X}_{ge_out}^{l}$.

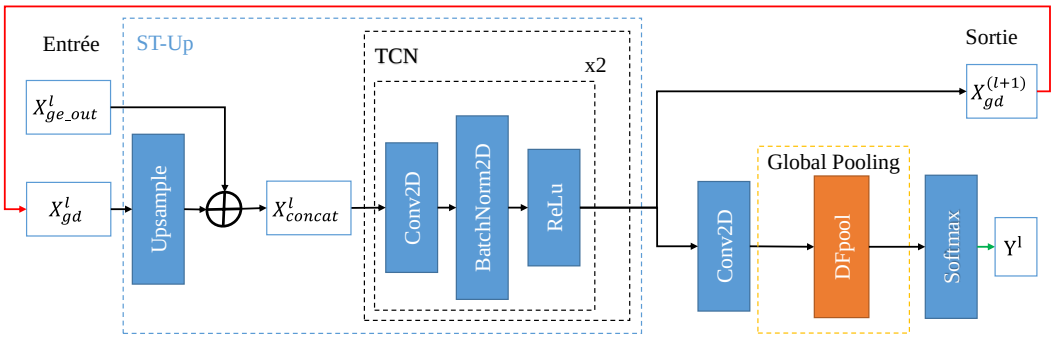


FIGURE 4.6 – Illustration d'une couche du décodeur (ST-Up) : d'abord, l'entrée $\mathbf{X}_{gd}^{l}$ est suréchantillonnée, puis combinée avec la sortie de l'encodeur ($\mathbf{X}^{(l+1)}$ en figure 4.3) via les connexions résiduelles $\mathbf{X}_{ge_out}^{l}$, et enfin traitée par le bloc TCN. La sortie $\mathbf{X}_{gd}^{(l+1)}$ est utilisée comme entrée pour la couche suivante et pour le pooling global lors de la prédiction $\mathbf{Y}^{l}$.

Ces connexions permettent d'intégrer les informations globales du décodeur aux détails locaux extraits par l'encodeur.

$$\begin{aligned} \mathbf{X}_{gd}^{(l+1)} &= \text{Conv}(\text{ReLU}(\text{BN}(\text{Conv}(\text{Up}(\mathbf{X}_{gd}^{l}) + \mathbf{X}_{ge_out}^{l})))) \\ &\quad \text{pour} \quad l = 1, 2, \ldots, n \end{aligned} \tag{4.17}$$

où BN représente la couche de BatchNorm 2D. Cette architecture permet au décodeur graphique d'améliorer les caractéristiques abstraites issues du bloc d'attention graphique tout en préservant les détails spatio-temporels cruciaux des sorties de l'encodeur grâce aux connexions résiduelles.

Enfin, en sortie du décodeur, un **bloc d'agrégation hiérarchique basé sur le differential pooling** est appliqué pour convertir le graphe enrichi en un vecteur latent fixe. Ce mécanisme, détaillé dans la section suivante, permet de synthétiser la structure globale tout en conservant les informations critiques extraites à différents niveaux de granularité.

4.3.9 Agrégation différentielle

Le bloc d'agrégation vise à résumer les représentations graphiques produites par le décodeur en un graphe plus compact et plus informatif. Pour cela, nous utilisons un *pooling différentiel* (*DiffPool*) [181], qui réduit progressivement le nombre de nœuds tout en préservant les relations structurelles importantes.

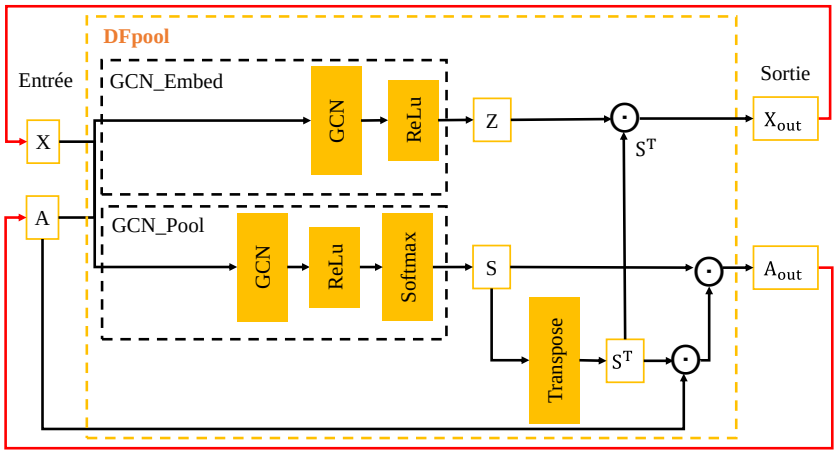


FIGURE 4.7 – Illustration d'une couche de pooling différentiel (DFpool) : à chaque couche, les embeddings de caractéristiques et les matrices d'adjacence sont générés, réduisant progressivement le nombre d'articulations du squelette (par exemple, de 21 à 13 puis à 1). La sortie X_{out} ainsi que la matrice d'adjacence mise à jour A_{out} sont transmises à la couche suivante.

Comme illustré en figure 4.7, une couche DFpool repose sur deux GCN distincts : (i) **GCN_Embed** : génère les embeddings des nœuds, produisant une représentation raffinée $\mathbf{Z}$; (ii) **GCN_Pool** : apprend une matrice d'assignation $\mathbf{S}$ indiquant comment regrouper les nœuds en clusters. Ces deux sorties permettent d'agréger à la fois les caractéristiques et la structure du graphe :

$$\mathbf{Z} = \text{ReLU}\left(\text{GCN_Embed}(\mathbf{X}, \mathbf{A})\right) \tag{4.18}$$

$$\mathbf{S} = \text{ReLU}\left(\text{GCN_Pool}(\mathbf{X}, \mathbf{A})\right) \tag{4.19}$$

$$\mathbf{X}_{\text{out}} = \mathbf{S}^T\mathbf{Z} \tag{4.20}$$

$$\mathbf{A}_{\text{out}} = \mathbf{S}^T\mathbf{A}\mathbf{S} \tag{4.21}$$

Dans notre modèle, chaque sortie du décodeur subit un DFpool, ce qui réduit le graphe (par exemple de 21 à 13, puis à 1 nœud) tout en produisant une représentation agrégée $\mathbf{X}_{\text{out}}$ transmise à la couche suivante.

4.3.10 Bloc de reconnaissance et de segmentation

Après l'agrégation des embeddings graphiques, l'étape finale consiste en la reconnaissance et la segmentation des actions. Ce bloc est essentiel car il permet au modèle de surmonter les problèmes de sur-segmentation en tirant parti des informations provenant à la fois des couches de haut niveau et de bas niveau du décodeur, conduisant ainsi à une segmentation d'action plus précise.

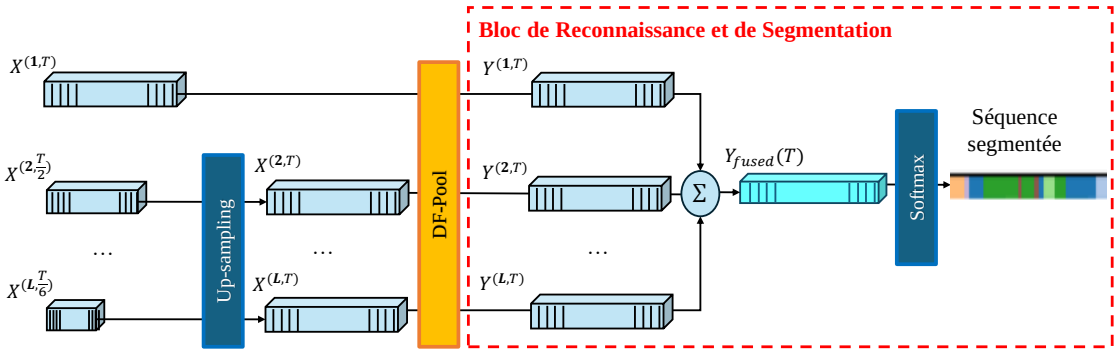


FIGURE 4.8 – Bloc de segmentation et reconnaissance : les sorties du décodeur $X(l,t)$ sont rééchantillonnées à T, agrégées via un pooling différentiel, puis fusionnées entre couches avant la fonction softmax finale.

Ce bloc reprend la structure introduite dans GSK-C2F (voir section 3.3.3), à l'exception d'un ajout majeur : nous insérons un **pooling global différentiel** (*differential pooling*) immédiatement avant la prédiction finale (voir figure 4.8). Ce module permet d'agréger les représentations graphiques finales du décodeur avant la fusion inter-couches. Les sorties agrégées $\mathbf{X}$ de chaque couche l, de longueur temporelle T^l, sont d'abord rééchantillonnées (par upsampling) à la longueur originale T. Elles sont ensuite transformées par le pooling différentiel, Puis fusionnées pour produire les prédictions image par image. Le processus est défini par :

$$\mathbf{Y}(l,t) = \text{DF-Pool}(\text{Up}(\mathbf{X}(l,t'),T)), \qquad t' \in [1,T^l],\ t \in [1,T] \tag{4.22}$$

$$\mathbf{Y}_{\text{fused}}(t) = \sum_{l=1}^{L} w_l \cdot \mathbf{Y}(l,t) \tag{4.23}$$

où Up désigne la fonction d'upsampling et w_l les poids de fusion tardive. Le modèle est ensuite entraîné avec la même fonction de perte que dans GSK-C2F (équation 4.24), combinant la classification par image et une pénalité de transition.

4.4 Protocole Expérimental

Cette section présente les éléments clés du protocole expérimental mis en place pour évaluer rigoureusement les performances du modèle G-EAD proposé. Nous détaillons d'abord les jeux de données utilisés en Section 4.4.1, choisis pour leur pertinence pour la reconnaissance d'actions dans le contexte de l'assemblage industriel. Ensuite, le protocole d'entraînement est décrit dans la Section 4.4.2, y compris les paramètres d'apprentissage du modèle et le prétraitement des données. Enfin, les métriques d'évaluation utilisées pour quantifier la qualité des prédictions sont présentées, ce qui permet une analyse complète et objective des résultats obtenus.

4.4.1 Jeux de données

Pour évaluer les performances de notre modèle GEAD, nous utilisons deux jeux de données industriels axés sur des scénarios d'assemblage. Le jeu de données **InHARD** [3] constitue l'un des principaux benchmarks de notre étude et fournit des séquences longues et variées enregistrées dans un environnement industriel réel. Dans le chapitre précédent, nous avions utilisé le jeu de données IKEA-ASM [4]. Cependant, comme expliqué dans la section 3.5, ce jeu de données repose fortement sur la présence et la manipulation d'objets pour définir et annoter les actions. De nombreuses classes d'action y sont distinguées principalement par les interactions opérateur–outil, ce qui complique considérablement la reconnaissance lorsque l'on se limite aux seules informations squelettiques. Dans ce contexte, un modèle fondé uniquement sur les trajectoires articulaires ne dispose pas des indices visuels nécessaires pour capturer l'ensemble des actions d'IKEA-ASM.

Afin de disposer d'un protocole d'évaluation plus adapté à la reconnaissance basée exclusivement sur les mouvements corporels, nous nous tournons vers le jeu de données **HA4M** [65], qui propose un protocole d'assemblage plus structuré et inclut un plus grand nombre de sujets, ce qui permet d'évaluer la capacité du modèle à généraliser à différents styles opératoires. Les expérimentations décrites dans les sections suivantes reposent sur InHARD et HA4M, et seul ce dernier est présenté ici, InHARD ayant déjà été introduit dans la section 3.4.1.1.

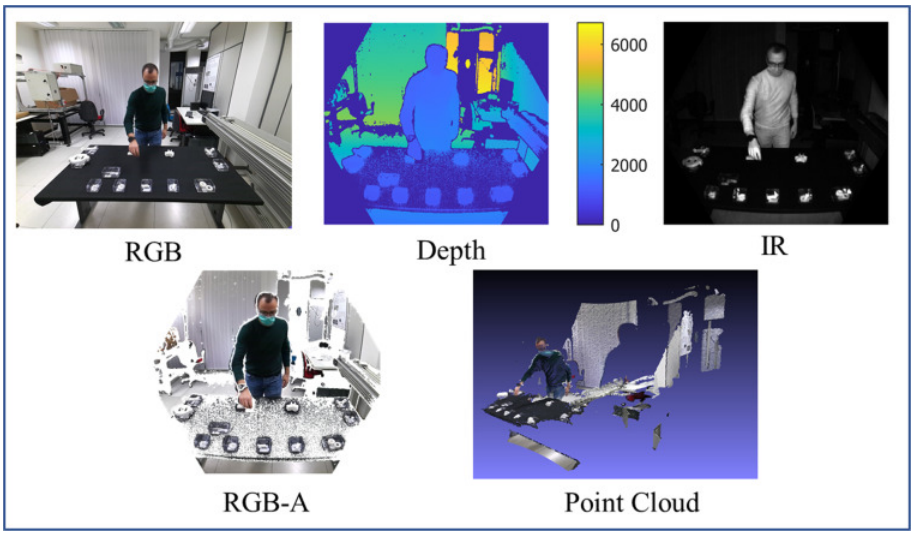


FIGURE 4.9 – Modalités présentes dans le jeu de données HA4M [65].

Le jeu de données HA4M (*Human Action Multi-modal Monitoring in Manufacturing*) porte sur une tâche d'assemblage réelle (le montage manuel d'un engrenage épicycloïdal) réalisée par **41 opérateurs**. Chaque sujet exécute la procédure selon son propre style, ce qui entraîne une variabilité notable des gestes et des durées d'exécution. L'ensemble comprend **217 séquences** capturées avec une caméra Microsoft Azure Kinect positionnée face à l'opérateur, garantissant une vue frontale stable. Le jeu de données comprend notamment des images RGB et en profondeur (voir la figure 4.9), ainsi que des données squelettiques extraites à l'aide du SDK de l'Azure Kinect. Dans le cadre de cette étude, seules les données squelettiques sont exploitées. Chaque séquence est annotée image par image selon **12 classes d'actions** industrielles, auxquelles s'ajoute une classe *"do not care"*. Les annotations sont fournies sous forme de fichiers contenant l'étiquette temporelle associée à chaque image. Ce jeu de données présente plusieurs défis typiques des environnements industriels, tels que des gestes visuellement proches, un déséquilibre entre les classes et une diversité de styles opéra-

toires, ce qui en fait un cadre pertinent pour l'évaluation des modèles de reconnaissance et de segmentation d'actions.

4.4.2 Protocole d'entraînement

Comme détaillé dans la section 3.4.2.2, nous avons utilisé l'optimiseur Adam pendant 500 époques avec un coefficient de régularisation (weight decay) de 3×10^{-3} et des tailles de batch de 5 et 32 pour les jeux de données InHARD et HA4M, respectivement.

Pour la méthode G-EAD basée sur l'attention, une taille de batch de 2 et warmup_steps = 5 ont été utilisées. Les paramètres de l'optimiseur Adam ont été fixés comme suit : coefficient de momentum $\beta_1 = 0.9$, coefficient de RMSprop $\beta_2 = 0.98$, et constante de stabilité numérique $\epsilon = 10^{-9}$. Le schéma d'apprentissage du taux d'apprentissage suit la stratégie proposée par Vaswani et al. [142], largement adoptée dans les architectures à base d'attention.

Les plages d'hyperparamètres suivantes ont été explorées : les tailles d'embedding d'augmentation dans le goulot d'étranglement $(d^x) \in \{1, 4, 8, 16, 32\}$ et le nombre de têtes d'attention $\in \{1, 2, 4\}$. Comme illustré à la figure 4.10, la meilleure performance en MoF a été obtenue avec deux têtes d'attention et une taille d'embedding (d^x) de 4. Une augmentation de d^x au-delà de 4 n'a pas systématiquement amélioré les performances et a souvent introduit une plus grande variance. Les configurations avec deux têtes d'attention ont généralement offert une meilleure stabilité et précision.

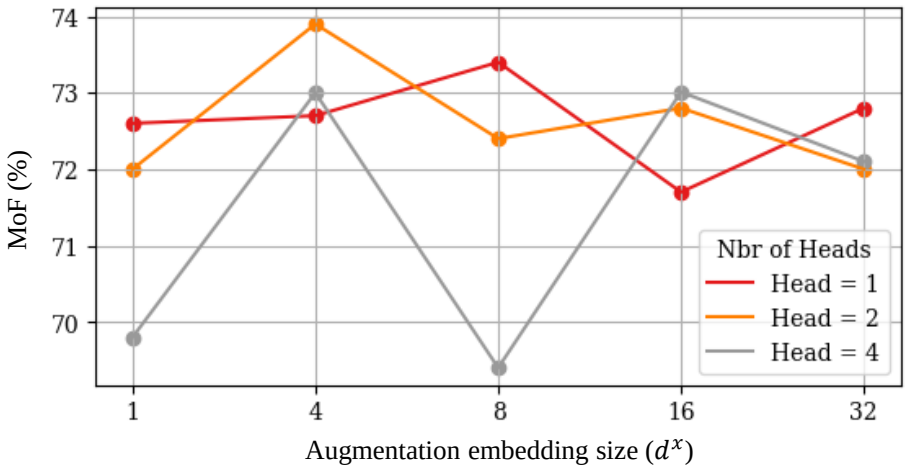


FIGURE 4.10 – Impact de la taille d'embedding d'augmentation dans le goulot d'étranglement (d^x) et du nombre de têtes d'attention sur la performance en MoF de l'approche G-EAD évaluée sur le jeu de données InHARD.

Dans nos expérimentations, différentes topologies squelettiques ont été utilisées en fonction des jeux de données, en fonction des conditions de capture et de la fiabilité des articulations. Pour InHARD, nous avons testé une version complète à 72 articulations (incluant les doigts) et une version réduite à 21 articulations, car les données fines des mains issues de la combinaison de motion capture étaient souvent bruitées ; il était donc pertinent de comparer l'impact de leur utilisation ou de leur non-utilisation. Pour HA4M, nous avons utilisé une topologie à 23 articulations correspondant au haut du corps, ainsi qu'une version plus compacte à 14 articulations centrée uniquement sur les bras, ces derniers étant les segments les plus impliqués dans les actions d'assemblage et offrant des signaux plus pertinents que ceux des membres inférieurs, souvent occultés par la table de travail. Ces variantes assurent ainsi une représentation squelettique adaptée à la visibilité et à la pertinence des articulations dans chaque contexte. Une stratégie de partitionnement spatial a été appliquée afin de définir des matrices d'adjacence à trois niveaux : le nœud racine, le groupe centripète (nœuds proches de la racine) et le groupe centrifuge (nœuds éloignés), comme décrit dans [139]. Une seule fonction de perte est appliquée à la sortie combinée, selon [60] :

$$L = L_{ce} + \lambda_{tr} L_{tr} \tag{4.24}$$

où L_{ce} est la perte d'entropie croisée par frame, L_{tr} est une perte de transition favorisant la continuité des étiquettes d'action entre les frames voisines, et $\lambda_{tr} = 0.15$, conformément à [60]. Pour pallier le déséquilibre des classes, une stratégie de pondération des classes, inspirée de [115], a été adoptée, attribuant des poids plus élevés aux classes sous-représentées.

Pour le jeu de données InHARD, nous avons suivi le protocole officiel décrit dans [3] pour les séparations d'apprentissage/validation/test. Concernant HA4M, en l'absence d'un protocole standard, deux stratégies ont été mises en œuvre : (1) **Cross-subject** : apprentissage sur les 32 premiers sujets et test sur les sujets restants ; (2) **Cross-sample** : division 80/20 pour l'apprentissage et le test avec une validation croisée à 5 sous-ensembles (splits). Le code est disponible sur gitlab pour plus de détails concernant l'implémentation du modèle et son architecture.

4.5 Comparaison avec l'état de l'art

4.5.1 Résultats sur le jeu de données InHARD

Le jeu de données InHARD [3] constitue un défi majeur pour la reconnaissance d'actions en raison de la complexité de ses séquences, de leur longueur (jusqu'à 14 minutes) et de la présence de transitions subtiles entre actions. Plusieurs versions de segmentation de ce jeu sont disponibles, allant de la version InHARD-3 et InHARD-4 à une segmentation plus fine comprenant 13 classes d'actions spécifiques (InHARD-13).

Le tableau 4.1 présente une comparaison des scores moyens de MoF (Mean over Frames) obtenus par les approches de l'état de l'art sur ces différentes versions. Afin de garantir une comparaison équitable, nous avons réimplémenté certaines méthodes en utilisant les données squelettiques, notamment *C2F-TCN* [60], ce qui aboutit à *SK-TCN*. Ces versions prennent uniquement les séquences squelettiques en entrée, plutôt que les flux RGB utilisés dans les versions originales. Parmi les méthodes squelettiques 3D ré-implémentées, ST-GCN et MTM [90] offrent de meilleures performances que C2F-TCN, mais leurs résultats restent nettement inférieurs aux architectures à base de graphe proposées.

On observe également que SK-TCN a connu une amélioration substantielle des performances par rapport à C2F-TCN sur l'ensemble des sous-ensembles d'InHARD. Cette progression met en évidence l'intérêt d'exploiter les informations squelettiques plutôt que les données RGB pour des tâches fines de reconnaissance et de segmentation d'actions.

Enfin, parmi les architectures proposées, notre premier modèle *GSK-C2F* [115] intègre une stratégie de pondération par classe afin de traiter efficacement les déséquilibres de classes, ce qui se traduit par des gains de performance constants sur toutes les versions d'InHARD (notamment +16,68 % sur la version v13). Notre second modèle **G-EAD** démontre quant à lui les meilleures performances sur toutes les versions, en particulier sur la version la plus fine (InHARD-13), où il atteint un MoF de **74.59%**. Cette amélioration est attribuable à l'architecture profonde de type encodeur-décodeur basé sur les graphes, à la hiérarchie multi-échelle temporelle, mais aussi à l'introduction d'un mécanisme d'**attention spatiale** au niveau du goulot d'étranglement. Ce dernier permet de se concentrer sélectivement sur les articulations les plus discriminantes selon le contexte temporel, ce qui améliore la précision de la segmentation tout en renforçant l'explicabilité du modèle.

Par ailleurs, bien que GSK-C2F obtienne de bons résultats, nous montrons dans notre étude d'ablation en section 4.6.1 qu'il se révèle moins robuste lorsque l'on fait varier le nombre de jointures, ce qui limite sa généralisation aux versions de granularité différente (21 et 72 jointures sur InHARD). À l'inverse, G-EAD présente une meilleure stabilité et apporte une nette amélioration des performances sur les actions rares, comme exposé dans la section 4.7.1, ce qui confirme sa capacité à mieux gérer les classes faiblement représentées.

TABLEAU 4.1 – Comparaison des scores MoF (%) avec les meilleurs résultats rapportés par chaque méthode de l'état de l'art sur le jeu de données InHARD (toutes topologies confondues). * : méthodes ré-implémentées. NA : information non disponible.

Méthode	Modalités	**InHARD-3**	**InHARD-4**	**InHARD-13**
C2F-TCN* [60]	RGB	50.13	32.57	34.28
LSTM-NN [174]	3D SK	NA	44.09	NA
ST-GCN* [119]	3D SK	75.94	59.63	52.26
MTM* [90]	3D SK	74.13	51.49	56.43
SK-TCN	3D SK	84.31	61.75	57.18
GSK-C2F	3D SK	84.82	70.28	73.86
G-EAD	**3D SK**	**85.40**	**70.55**	**74.59**

4.5.2 Résultats sur le jeu de données HA4M

Le jeu de données HA4M [65] propose un environnement d'assemblage plus structuré, avec des enchaînements d'actions pré-définis et des transitions temporelles plus régulières. Ce jeu comporte également un plus grand nombre de sujets, ce qui permet de tester la capacité du modèle à généraliser à des styles de mouvement variés.

Dans le tableau 4.2, la méthode proposée, G-EAD, est comparée à d'autres approches de l'état de l'art sur le jeu de données HA4M. Bien que certaines méthodes telles que Att-C3D-AssNet [194] et MS-TCN++ [195] présentent de bonnes performances, il est important de noter que ces approches n'ont été testées que sur un seul jeu de données (HA4M), ce qui limite leur portée comparative. De plus, les découpages d'entraînement/test utilisés dans ces travaux ne sont pas publiés. Également, l'absence de code source pour ces approches empêche toute reproduction fidèle des résultats et complique une comparaison équitable. Malgré ces limites, l'approche proposée, G-EAD, avec son architecture encodeur-décodeur et ses mécanismes d'attention, dépasse les performances de notre méthode GSK-C2F présentée au chapitre 3.3. Cela montre également que G-EAD fonctionne efficacement sur deux jeux

de données représentant des scénarios d'assemblage différents (InHARD et HA4M), ce qui démontre sa capacité de généralisation. Dans notre publication [196], les expérimentations ont été entièrement reproductibles grâce au code ainsi qu'aux protocoles d'entraînement et de test fournis. Cela garantit que les travaux futurs pourront reproduire la méthode et comparer les résultats, ce qui constitue une base de référence fiable pour les recherches à venir sur le jeu de données HA4M.

Méthode	nb. articulations	Edit	MoF
GSK-C2F	23 (Haut)	91.79	91.81
MS-TCN++* [195]	14 (Bras)	**96.38**	94.90
MS-TCN++* [195]	23 (Haut)	<u>95.90</u>	<u>94.92</u>
Att-C3D-AssNet* [194]	RGB-A	-	**97.10**
G-EAD	23 (Haut)	92.22	92.81
G-EAD	14 (Bras)	92.78	93.11

TABLEAU 4.2 – Comparaison avec l'état de l'art sur HA4M en utilisant G-EAD. Haut : partie supérieure du corps, RGB-A : alignement RGB-Depth. * : résultats non reproductibles, car la distribution d'entraînement/test et le code non publiés.

4.6 Études d'ablation

Cette section présente les études d'ablation et l'analyse du modèle réalisées sur le jeu de données InHARD-13. Elle examine l'impact de l'intégration des représentations graphiques dans le modèle de segmentation à différents niveaux de l'architecture. De plus, elle évalue l'intérêt du mécanisme d'attention dans l'espace latent, ainsi que l'effet de la stratégie de regroupement (pooling) utilisée au sein du bloc d'agrégation.

4.6.1 Influence de la représentation graphique et du mécanisme d'attention

Dans cette étude, nous proposons une série de modèles visant à améliorer la reconnaissance et la segmentation d'actions en exploitant des architectures graphiques spatio-temporelles. Chaque modèle prend en entrée des données squelettiques et produit en sortie une séquence segmentée d'actions.

Une première version du modèle, appelée B_G-C2F, a été conçue en intégrant un réseau de convolution spatio-temporel sur graphes comme backbone afin de capturer la structure spatiale ainsi que la dynamique temporelle des données squelettiques. En s'appuyant sur

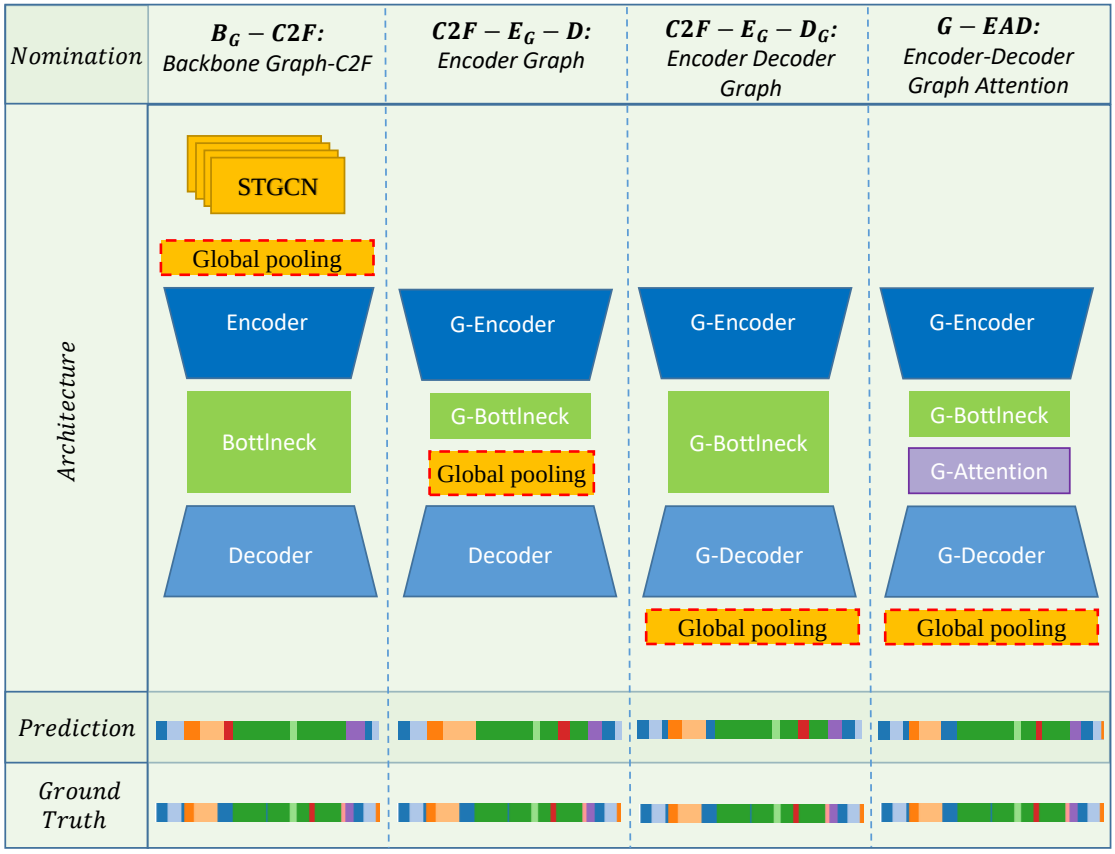


FIGURE 4.11 – Évolution de l'architecture. Étape 1 (B_G-C2F) : Intégration du graphe en tant que backbone pour l'extraction des caractéristiques. Étape 2 (C2F-E_G-D) : Ajout de la structure graphique dans l'encodeur afin d'affiner l'extraction des caractéristiques avant le décodage. Étape 3 (C2F-E_G-D_G) : Intégration du graphe également au niveau du décodeur pour enrichir la représentation. Étape 4 (G-EAD) : Introduction d'un mécanisme d'attention spatiale pour capturer plus efficacement les dépendances entre les articulations les plus pertinentes dans les séquences d'actions squelettiques.

cette base, le modèle C2F-E_G-D introduit une structure graphique spatio-temporelle au sein de l'encodeur, ce qui permet un raffinement plus poussé de l'extraction des caractéristiques avant l'étape de décodage. Cette approche est ensuite étendue dans C2F-E_G-D_G, où les graphes spatio-temporels sont exploités tout au long du pipeline encodeur-décodeur, offrant ainsi une analyse plus complète des relations temporelles et spatiales dans les séquences en entrée. Enfin, le modèle G-EAD est proposé ; il intègre un mécanisme d'attention spatiale permettant au réseau de se concentrer sélectivement sur les articulations les plus pertinentes. Cette attention renforce la capacité du modèle à reconnaître des actions subtiles ou complexes, souvent associées à des postures ergonomiques contraignantes ou à des exécutions inefficaces

des tâches. Les différentes variantes proposées sont illustrées dans la Fig. 4.11. Cette étude analyse ainsi l'impact de l'introduction de graphes spatio-temporels à différents niveaux de l'architecture : en tant que backbone (B_G-C2F), au sein de l'encodeur (C2F-E_G-D), dans l'ensemble du schéma encodeur-décodeur (C2F-E_G-D_G), ainsi qu'avec l'ajout d'un mécanisme d'attention (G-EAD).

Méthode	Articulations	F1@50	Edit	MoF
GSK-C2F	21	56.74	61.84	65.38
B_G-C2F	21	65.36	70.56	72.20
C2F-E_G-D	21	69.05	72.79	72.62
C2F-E_G-D_G	21	**72.45**	<u>75.94</u>	<u>73.98</u>
G-EAD	21	<u>71.33</u>	**76.10**	**74.59**
GSK-C2F	72	**71.37**	**74.37**	<u>73.86</u>
B_G-C2F	72	32.23	52.38	43.73
C2F-E_G-D	72	55.00	60.60	60.30
C2F-E_G-D_G	72	45.72	49.98	56.38
G-EAD	72	<u>71.26</u>	<u>72.52</u>	**74.42**

TABLEAU 4.3 – Étude d'ablation des différentes architectures basées sur les graphes sur le jeu de données InHARD-13. B_G-C2F : ST-GCN utilisé comme backbone. C2F-E_G-D : ST-GCN intégré à l'encodeur. C2F-E_G-D_G : ST-GCN intégré à l'encodeur, au goulot d'étranglement et au décodeur. G-EAD : architecture encodeur-décodeur avec ST-GCN et mécanisme d'attention.

Une étude d'ablation a été menée afin d'évaluer l'impact de l'intégration de réseaux de convolutions spatio-temporelles sur graphes (ST-GCN) à différents niveaux de l'architecture pour la reconnaissance et la segmentation d'actions, en utilisant le jeu de données InHARD-13 [3]. Deux topologies du squelette ont été explorées : une première avec 21 articulations, représentant les articulations principales du corps humain, et une seconde avec 72 articulations, incluant les doigts. Toutefois, ces articulations supplémentaires sont statiques dans la majorité des séquences et n'apportent que peu d'information dynamique. Leur présence peut donc introduire du bruit, rendant la configuration à 21 articulations plus réaliste et adaptée au contexte industriel. Les résultats détaillés sont présentés dans le tableau 4.3.

Pour la topologie à 21 articulations, on observe que l'ajout progressif de graphes spatio-temporels à différents blocs de l'architecture améliore systématiquement les performances. Le modèle C2F-E_G-D_G, intégrant des graphes dans l'encodeur, le goulot d'étranglement et le décodeur, atteint 73.98% de MoF, soit un gain de 8.60% par rapport à l'architecture de

base. Cette amélioration est attribuée à une meilleure modélisation des relations entre les articulations, notamment les coordinations complexes entre les épaules, le dos et les bras. Ces dépendances sont essentielles pour identifier des gestes inefficaces ou répétitifs, souvent à l'origine de troubles musculosquelettiques (TMS). Le modèle G-EAD, en intégrant un mécanisme d'attention spatio-temporelle en plus des ST-GCN, permet de se concentrer sur les articulations les plus pertinentes. Il obtient les meilleures performances globales avec 74.59% de MoF, mettant en avant sa capacité à capter les subtilités des mouvements complexes.

En revanche, avec la topologie à 72 articulations (incluant les doigts), les performances chutent drastiquement (ex. : 56.38% de MoF pour C2F-E_G-D_G), en raison du bruit introduit par les articulations non informatives. Toutefois, l'ajout du mécanisme d'attention dans G-EAD permet de compenser ce désavantage, atteignant 74.42% de MoF, soit la meilleure performance MoF toutes configurations confondues.

Il est également important de noter que GSK-C2F (notre première contribution) présente une tendance inverse : il obtient de meilleures performances avec la topologie à 72 articulations (73.86 % de MoF) qu'avec celle à 21 articulations (65.38 %). Cette observation suggère que GSK-C2F est plus sensible aux variations de granularité du squelette et peut, dans ce cas précis, bénéficier de la richesse supplémentaire des articulations fines. À l'inverse, G-EAD demeure plus robuste et maintient des performances élevées dans les deux configurations, vraisemblablement grâce à son mécanisme d'attention qui filtre sélectivement les articulations non pertinentes et se focalise sur les zones critiques du corps, ce qui est fondamental dans des contextes d'analyse ergonomique

Ainsi, ces résultats mettent en évidence l'importance du choix de la granularité du squelette et du design de l'architecture pour les tâches de reconnaissance et segmentation d'actions dans un cadre industriel centré sur l'humain.

Pour mieux comprendre le comportement et la capacité d'interprétation du modèle à base d'attention proposé, G-EAD, la figure 4.12 illustre les cartes d'attention correspondant à trois instants clés (6 s, 7 s et 8 s) lors de l'exécution de l'action « Prendre à gauche ». Chaque carte de chaleur représente la distribution de l'attention entre les articulations du squelette à un moment donné, où une intensité plus élevée (couleurs claires) indique une attention plus forte. Ces matrices d'attention sont accompagnées des images correspondantes de l'opérateur en train d'effectuer l'action.

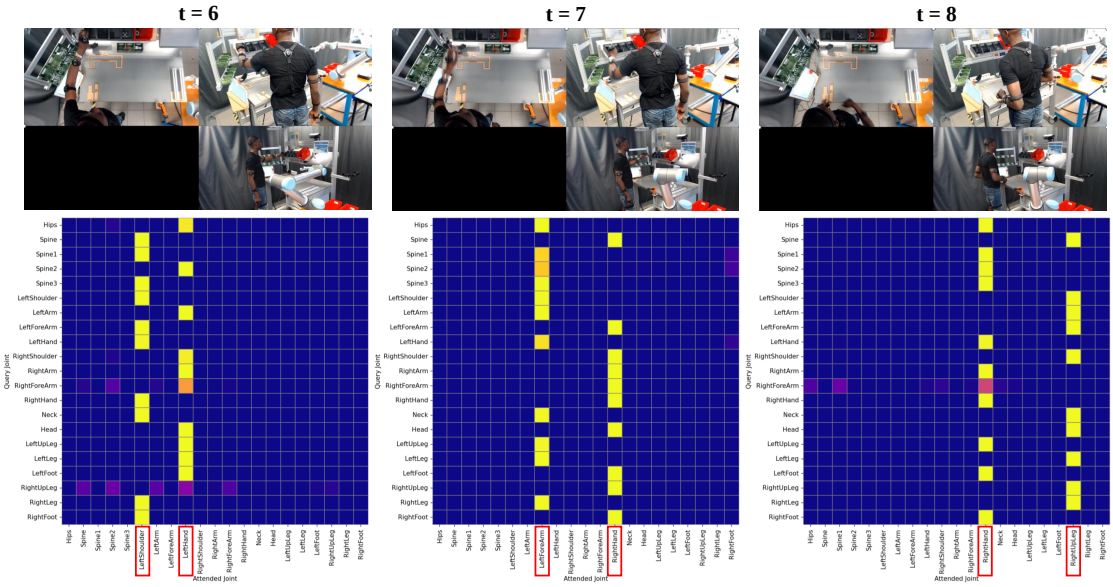


FIGURE 4.12 – Représentation des cartes d'attention pour chaque seconde dans l'intervalle de temps (secondes 6 à 8) de l'action « Prendre à gauche ». Images extraites de la vidéo « P13-R02 » du jeu de données InHARD-13.

À la première seconde (6 s), l'opérateur commence par tendre le « bras gauche » vers l'avant. La carte d'attention montre une activation significative entre l'« épaule gauche » et la « main gauche », illustrant la capacité du modèle à se concentrer sur les parties du corps impliquées dans l'initiation du mouvement. À (7 s), l'opérateur saisit l'outil avec la « main gauche » tout en repositionnant légèrement la « main droite ». L'attention se déplace en conséquence : l'« avant-bras gauche » reçoit une attention plus marquée, reflétant le mouvement de retour, tandis que la « main droite » devient active en préparation de la tâche suivante. Enfin, à (8 s), lorsque l'opérateur amorce une nouvelle action « Prendre devant », l'attention du modèle se déplace vers la « main droite » et la « cuisse droite ». Cela suggère que le modèle capture les légers changements de posture corporelle, tels que le mouvement du buste vers l'avant, indiquant une anticipation du mouvement suivant.

Cette illustration améliore non seulement l'interprétabilité du modèle, mais révèle également la manière dont il apprend les dynamiques temporelles et les dépendances physiques entre les articulations. De plus, un tel niveau de visualisation est essentiel dans les applications liées à l'ergonomie et à la sécurité, où le suivi du stress articulaire et des schémas de mouvement peut alimenter des systèmes de rétroaction destinés à prévenir les troubles musculosquelettiques. Au-delà de la précision, cette interprétabilité visuelle permet de valider

le modèle, d'analyser ses erreurs et de l'affiner, en révélant quand et où il peut se tromper, ouvrant ainsi la voie à des systèmes industriels plus transparents et fiables, centrés sur l'humain.

Afin d'évaluer le coût d'inférence de l'architecture proposée, une analyse de complexité est présentée en comparaison avec des méthodes récentes à base de graphes évaluées sur le jeu de données InHARD. Le tableau 4.4 compare le modèle proposé (G-EAD) avec les baselines réimplémentées ST-GCN et MTM en termes de précision MoF, de nombre de paramètres entraînables et de complexité de calcul en GMACs (opérations Giga-Multiply-Accumulate).

Méthode	**MoF (%)**	**Paramètres (Millions)**	**Complexité (GMACs)**
ST-GCN* [119]	52.26	<u>4.81</u>	2549.17
MTM* [90]	<u>56.43</u>	**2.63**	**1393.69**
G-EAD (proposé)	**74.59**	21.31	<u>1446.77</u>

TABLEAU 4.4 – Comparaison du score moyen par image (MoF), du nombre de paramètres (en millions) et de la complexité de calcul (GMACs) entre des approches à base de graphes évaluées sur InHARD-13 à 21 jointures. * : méthodes ré-implémentées.

L'augmentation du nombre de paramètres dans G-EAD provient principalement de l'utilisation de couches de convolution 2D au sein d'une architecture encodeur-décodeur. En particulier, les premières couches de l'encodeur et les dernières du décodeur contribuent fortement à cette complexité, car elles traitent de longues séquences temporelles tout en conservant la topologie du graphe à travers le réseau.

Un mécanisme d'attention est également intégré afin d'améliorer l'interprétabilité, en fournissant des indications sur les décisions prises par le modèle (voir la figure 4.12). Pour limiter son impact sur la complexité du modèle, ce module est placé intentionnellement au niveau du goulot d'étranglement de l'architecture, où les dimensions des caractéristiques sont réduites, ce qui rend les calculs plus efficaces.

Malgré cette augmentation du nombre de paramètres, la complexité de calcul (GMACs) reste modérée, comparable à celle de MTM et nettement inférieure à celle de ST-GCN. Cela suggère que bien que G-EAD nécessite davantage de mémoire pour stocker ses paramètres, son temps d'inférence et sa consommation énergétique restent compatibles avec un déploiement pratique. De plus, G-EAD atteint une amélioration significative de la précision, avec environ +18 % de MoF par rapport à ST-GCN. En résumé, G-EAD améliore à la fois la précision et l'interprétabilité du modèle, tout en conservant un temps d'inférence compétitif.

4.6.2 Impact de la topologie du graphe squelettique

Nous avons mené une étude d'ablation sur la structure de la matrice d'adjacence afin d'évaluer l'impact des différentes topologies sur la qualité de la segmentation d'actions. Comme présenté précédemment, la matrice d'adjacence $\hat{\mathbf{A}}$ peut être décomposée en trois composantes : $\hat{\mathbf{A}}^{(0)}$ (auto-connexions), $\hat{\mathbf{A}}^{(1)}$ (connexions proches de la jointure concernée) et $\hat{\mathbf{A}}^{(2)}$ (connexions plus éloignées). Dans notre implémentation, chaque connexion valide reçoit initialement un coefficient unitaire (valeur 1). Ces matrices sont ensuite normalisées de manière à ce que la somme des coefficients de chaque ligne soit égale à 1, conformément à la méthode de normalisation utilisée dans ST-GCN [139]. Ainsi, aucune pondération spécifique n'est imposée a priori ; le modèle apprend directement l'importance relative de chaque type de connexion au cours de l'entraînement.

Nous avons évalué plusieurs variantes :

— $\mathbf{A}^{(0)}$ seul : ne considère que les connexions internes (self-links).
— $\hat{\mathbf{A}}^{(0)} + \hat{\mathbf{A}}^{(1)}$: ajoute les voisins immédiats.
— $\hat{\mathbf{A}}^{(0)} + \hat{\mathbf{A}}^{(2)}$: combine les extrémités et le centre, mais ignore les voisins proches.
— $\hat{\mathbf{A}}^{(0)} + \hat{\mathbf{A}}^{(1)} + \hat{\mathbf{A}}^{(2)}$: topologie complète à trois branches.

Les résultats, présentés dans le tableau 4.5, indiquent que la configuration complète à trois composantes permet d'obtenir les meilleures performances, en particulier pour la précision des transitions d'actions, souvent critiques dans les scénarios industriels.

Tableau 4.5 – Impact des différentes topologies de graphe sur les performances de segmentation (MoF en %) sur le jeu de données InHARD-13 à 21 jointures.

Topologie	**MoF (%)**
$\hat{\mathbf{A}}^{(0)}$	68.2
$\hat{\mathbf{A}}^{(0)} + \hat{\mathbf{A}}^{(2)}$	69.5
$\hat{\mathbf{A}}^{(0)} + \hat{\mathbf{A}}^{(1)}$	71.6
$\hat{\mathbf{A}}^{(0)} + \hat{\mathbf{A}}^{(1)} + \hat{\mathbf{A}}^{(2)}$	**74.59**

Ces résultats confirment que l'inclusion simultanée des trois types de connexions (self, proche, lointaine) améliore significativement la qualité de représentation du graphe, ce qui permet de mieux capter à la fois les mouvements fins (autour des articulations clés impliquées dans l'action) et les articulations distantes, souvent impliquées dans les tâches industrielles.

4.6.3 Différentes méthodes d'agrégation

Le pooling global joue un rôle essentiel dans les modèles basés sur les graphes en permettant de capturer le contexte global et les relations entre les nœuds, ce qui est crucial pour comprendre la structure entière du graphe au-delà des simples caractéristiques isolées des nœuds. Comme illustré à la figure 4.11, cette opération est appliquée à chaque étape où les embeddings du graphe doivent être transformés en un vecteur de taille fixe, ce qui garantit la conservation de l'information structurelle à chaque instant.

Les méthodes traditionnelles utilisent souvent un pooling moyen (AvgPool), qui se contente de calculer la moyenne des caractéristiques des nœuds en un seul vecteur, au risque de perdre les relations hiérarchiques importantes du graphe. À l'inverse, l'approche proposée exploite le pooling différentiel (DFpool) [181], qui agrège hiérarchiquement l'information du graphe en apprenant des assignations douces de clusters, ce qui permet de capturer la structure complexe à plusieurs niveaux.

Cette agrégation hiérarchique est particulièrement pertinente pour les applications en ergonomie. De nombreuses tâches physiques impliquent des mouvements coordonnés entre groupes d'articulations, par exemple la coordination "épaule, coude, main" lors de la manipulation d'outils, ou "bas du dos, genou, cheville" lors du levage de charges. Les écarts par rapport à ces schémas coordonnés peuvent signaler une fatigue, des postures contraignantes ou des risques ergonomiques. En capturant efficacement ces interdépendances, DFpool fournit une représentation plus riche que celle des méthodes de pooling classiques. Comme illustré dans le tableau 4.6, on observe une amélioration notable des performances de segmentation, avec un gain de **+3.28 au score F1@50** et de **+2.53 au score Edit**. Ces résultats reflètent la capacité renforcée du modèle à produire des segmentations précises tout en limitant les erreurs de sur-segmentation.

TABLEAU 4.6 – Impact de la stratégie de pooling (PS) sur les performances de segmentation du modèle G-EAD sur le jeu de données InHARD-13 à 21 jointures.

PS	F1@50	Edit	MoF
AvgPool	68.05	73.57	73.81
DFpool	**71.33 (+3.28)**	**76.10 (+2.53)**	**74.59 (+0.78)**

4.7 Visualisation et discussion

Cette section illustre les défis liés à la segmentation des actions et met en évidence la robustesse de l'approche proposée. Elle comprend une analyse détaillée des résultats obtenus sur les jeux de données InHARD-13 et HA4M.

4.7.1 Répartition des actions

Les jeux de données industriels posent souvent des défis majeurs, notamment en raison d'un fort déséquilibre entre les classes. Dans les tâches d'assemblage, certaines actions sont fréquemment effectuées, tandis que d'autres sont rares, comme l'illustre la figure 3.5 pour le jeu de données InHARD. Ce déséquilibre affecte la performance des modèles, car la précision de reconnaissance est généralement corrélée à la fréquence des échantillons d'actions par classe. En effet, les modèles parviennent généralement à mieux prédire les actions surreprésentées, mais obtiennent des résultats moins satisfaisants pour les classes minoritaires.

Pour atténuer cet effet, notre première contribution, GSK-C2F, utilise une stratégie de pondération par classe afin d'accorder une plus grande importance aux classes sous-représentées durant l'apprentissage. La nouvelle approche proposée, G-EAD, va plus loin en intégrant une représentation graphique qui capture plus efficacement les schémas de mouvement, ce qui permet une meilleure différenciation entre les actions, même avec un nombre limité d'échantillons. Cette amélioration se traduit par une performance de reconnaissance accrue, en particulier pour les classes minoritaires.

L'efficacité de cette approche est démontrée par la matrice de confusion (figure 4.13), qui met en évidence des gains de reconnaissance pour les actions peu représentées. En particulier, G-EAD permet une augmentation moyenne de 20 % de la précision pour les actions suivantes : "3 : Prendre tournevis", "7 : Prendre réglette", "8 : Poser réglette", "10 : Poser composant" et "12 : Prendre sous-ensemble".

Cette amélioration démontre la capacité du modèle à mieux distinguer les actions en apprenant des schémas de mouvement propres à chaque classe. Elle met également en lumière l'intérêt de l'utilisation des données squelettiques dans des environnements complexes tels que les postes de travail industriels. En effet, les données squelettiques sont robustes aux occultations et aux variations de luminosité, car elles se concentrent uniquement sur le squelette humain et ignorent le bruit de fond, offrant ainsi une information précise sur le mouvement

humain.

Cependant, une performance plus faible est observée pour la classe d'action "1 : Consulter fiche". Cela s'explique par la difficulté intrinsèque à la distinguer de la classe "Aucune action", ces deux classes présentant des schémas squelettiques similaires, impliquant peu de mouvements et essentiellement des postures du haut du corps, sans gestes distinctifs ni interactions avec des outils. Ce type de confusion met en évidence une limite fondamentale des approches basées uniquement sur le mouvement squelettique pour la reconnaissance fine d'activités, notamment dans le cas de tâches passives ou peu dynamiques.

Dans ce cas particulier, l'intégration d'une estimation du regard (gaze estimation) comme modalité supplémentaire pourrait considérablement améliorer la précision de reconnaissance. Les schémas de regard fournissent un contexte sémantique précieux, permettant au modèle d'inférer la zone d'attention de l'utilisateur, même en l'absence de mouvements corporels expressifs.

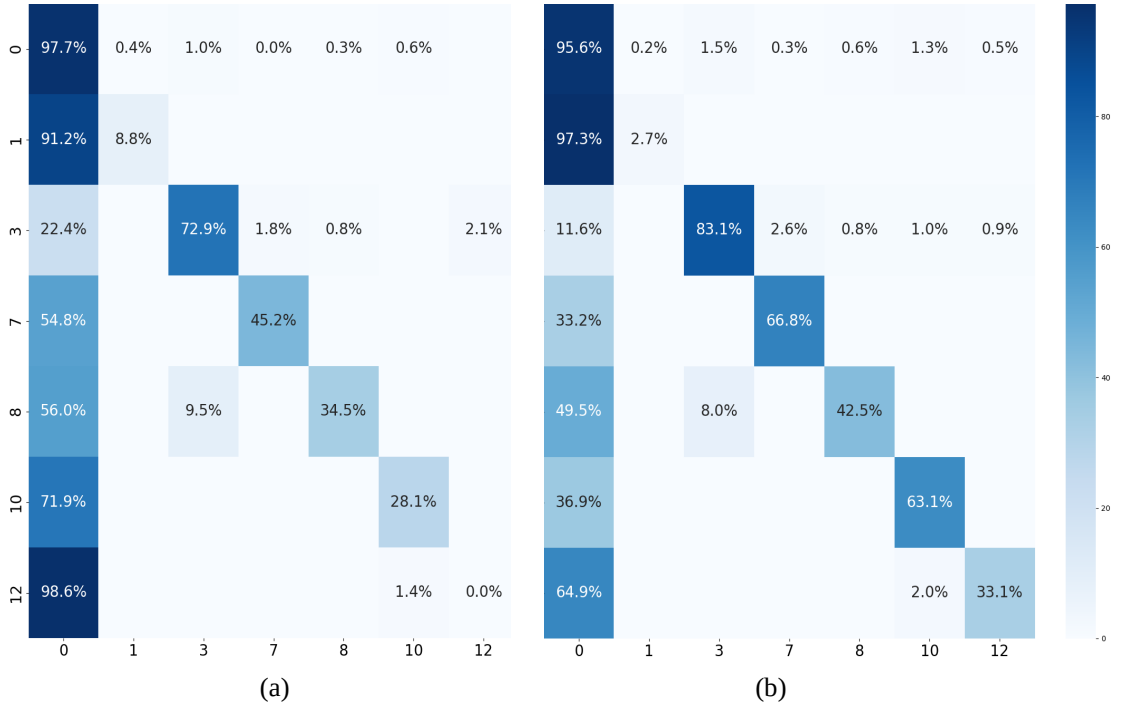


FIGURE 4.13 – Matrice de confusion pour les classes d'actions sous-représentées sur InHARD-13, avec (a) GSK-C2F et (b) le modèle proposé G-EAD utilisant 21 articulations. Id d'action, Classe (Nombre d'échantillons) : 0 : "Aucune action" (6606) ; 1 : "Consulter fiche" (132) ; 3 : "Prendre tournevis" (420) ; 7 : "Prendre réglette" (76) ; 8 : "Poser réglette" (74) ; 10 : "Poser composant" (385) ; 12 : "Prendre sous-ensemble" (39).

4.7.2 Variabilité de la durée des actions

La segmentation des actions en contexte industriel pose un défi majeur en raison de la forte variabilité des durées d'action. Certaines durent plusieurs secondes, tandis que d'autres se produisent très rapidement, comme l'illustre la figure 3.5, ce qui rend leur détection difficile pour les systèmes de reconnaissance. Ce problème est particulièrement critique dans les activités d'assemblage industriel, où les séquences sont composées d'actions diverses aux durées variées. De plus, les transitions entre actions surviennent souvent de manière rapide et sont parfois mal annotées, car les frontières entre les actions peuvent être ambiguës et difficiles à définir. Par conséquent, détecter avec précision les instants de début et de fin des actions devient une tâche complexe.

Pour faire face à ces défis, le modèle G-EAD proposé s'appuie sur une architecture encodeur-décodeur par graphes, permettant de capturer plusieurs résolutions temporelles et d'extraire des motifs de mouvement informatifs. Le nombre de couches d'encodage est soigneusement sélectionné pour renforcer cette capacité. Le goulot d'étranglement génère des séquences à différentes échelles temporelles, ce qui permet au modèle d'apprendre efficacement les variations des durées d'action. La force du modèle réside dans sa capacité à exploiter une modélisation par graphes pour représenter simultanément les dynamiques temporelles multi-résolution tout en conservant l'intégrité spatiale et structurelle des données de mouvement.

La figure 4.14 présente les chronologies de segmentation d'actions obtenues par le modèle G-EAD basé sur des représentations squelettiques, comparées au modèle C2F-TCN [60] qui repose sur des données RGB. Les améliorations introduites par G-EAD sont mises en évidence par des rectangles rouges, dans lesquels les instants de début d'action sont détectés avec une plus grande précision. De plus, les actions brèves, souvent ignorées ou mal reconnues par C2F-TCN, sont plus correctement détectées et segmentées avec G-EAD. Ces gains résultent de la grande précision de l'analyse du mouvement, rendue possible par la modalité squelettique, ainsi que de la capacité du modèle à préserver les dépendances spatiales et temporelles.

L'analyse des erreurs montre que les mauvaises classifications surviennent principalement autour des frontières d'action, en particulier pour les actions de courte durée, souvent imbriquées dans des séquences plus longues. Par exemple, l'action "Visser composant" est parfois confondue avec le geste adjacent "Tenir composant", lorsque la transition est fluide et mal segmentée dans les annotations. Ces erreurs proviennent à la fois d'annotations ambiguës et

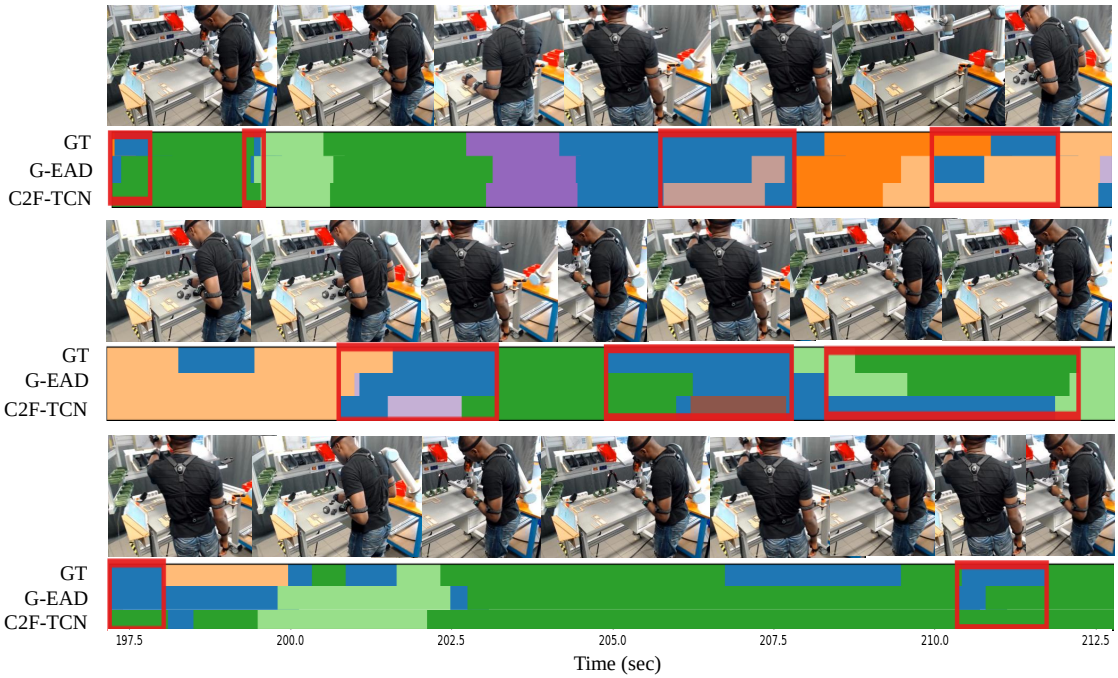


FIGURE 4.14 – Chronologies de segmentation d'actions sur le jeu de données InHARD pour les modèles G-EAD et C2F-TCN (approche RGB) [60] (vidéo P13R02). Chaque barre colorée correspond à une action parmi les 14 classes d'action. GT : Vérité terrain (Ground Truth).

du contexte limité pris en compte par le modèle à un instant donné.

Ces écarts soulignent un enjeu important pour le déploiement en conditions réelles : bien que le modèle offre de bonnes performances globales, le raffinement des frontières temporelles reste un point difficile, notamment dans les contextes liés à la sécurité, où reconnaître précisément le début d'une action (par exemple saisir un outil) est essentiel pour l'analyse ergonomique ou la détection de risques.

Pour remédier à cela, les travaux futurs pourraient explorer l'intégration d'informations contextuelles complémentaires, au-delà des caractéristiques squelettiques, telles que les indices liés aux objets présents sur le poste de travail ou aux outils utilisés lors de l'assemblage. Cela permettrait d'améliorer la capacité du modèle à distinguer des actions visuellement similaires. Cette intégration peut être réalisée par l'ajout de représentations d'objets ou d'outils, comme proposé dans [19], ou à l'aide d'algorithmes de détection d'objets à partir d'images RGB ou d'autres modalités.

4.7.3 Variabilité de l'ordre des actions et des gestes

Les opérateurs présentent une variabilité notable dans l'exécution des tâches, tant dans la manière d'effectuer les actions que dans le temps nécessaire à leur accomplissement. Comme discuté dans la section 1.4.2.4 concernant les défis propres aux jeux de données industriels, les séquences d'actions présentent à la fois une forte variabilité inter-classe et une variabilité intra-classe. La figure 1.5 illustre notamment que certaines actions distinctes peuvent partager des schémas gestuels très similaires, ce qui complique leur distinction. À l'inverse, la figure 1.6 montre que des actions d'une même classe peuvent être exécutées différemment selon les opérateurs, ce qui accroît la difficulté de généralisation des modèles de reconnaissance et de segmentation.

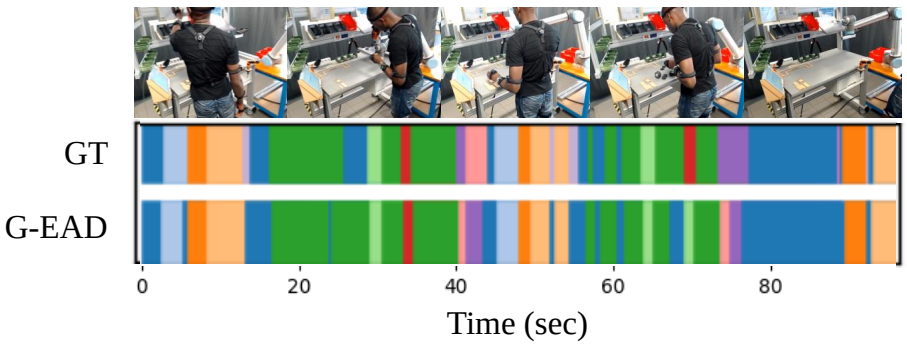


FIGURE 4.15 – Chronologie de la segmentation dans le jeu de données InHARD (vidéo P13R02 – 100 premières secondes) obtenue par le modèle G-EAD. Chaque barre colorée représente l'une des 14 actions. GT : Vérité Terrain.

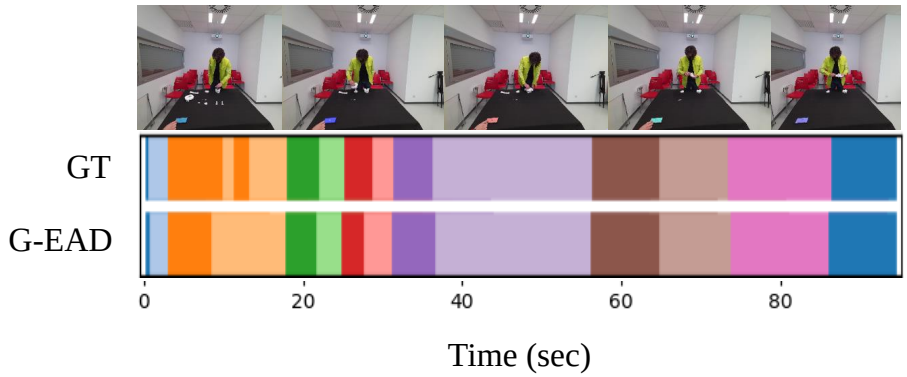


FIGURE 4.16 – Chronologie de la segmentation dans le jeu de données HA4M (vidéo ID035U001) obtenue par le modèle G-EAD. Chaque barre colorée représente l'une des 12 actions. GT : Vérité Terrain.

Par ailleurs, l'ordre d'exécution des actions dans un processus d'assemblage peut varier.

Certains scénarios, comme le benchmark InHARD-13, autorisent une certaine flexibilité dans la réalisation des tâches, tandis que d'autres, comme HA4M, imposent un ordre strictement défini. Cette distinction est clairement représentée dans la figure 4.15, qui montre la complexité accrue de InHARD-13, comparée à HA4M illustrée en figure 4.16, ainsi que son impact sur la performance de reconnaissance. La liberté d'exécution et l'imprévisibilité des séquences d'actions ajoutent une complexité importante au benchmark InHARD-13, ce qui rend la détection des segments d'action plus difficile.

Bien que la modalité squelettique améliore la reconnaissance d'actions grâce à une analyse précise du mouvement, elle reste limitée dans les scénarios où le mouvement seul ne suffit pas à différencier les classes. L'approche proposée, reposant uniquement sur les données squelettiques, ne capture que les mouvements de l'opérateur sans intégrer les éléments contextuels tels que les objets ou l'environnement de travail. Malgré ces limites, l'approche proposée permet de répondre aux principaux défis industriels. En traitant les séquences d'actions à différentes résolutions temporelles, elle s'adapte à la variabilité de l'ordre des actions et capture un contexte plus large d'exécution. De plus, son analyse précise du mouvement améliore l'apprentissage des schémas gestuels, renforçant ainsi la capacité du modèle à gérer la variabilité gestuelle et à produire une segmentation plus précise des actions.

4.8 Conclusion

Dans ce chapitre, nous avons proposé G-EAD, un modèle de segmentation d'actions basé sur une architecture Graphe encodeur-Graphe Attention-Graphe Décodeur, permettant ainsi d'exploiter, à chaque couche, la représentation squelettique sous forme de graphe. Cette approche permet de convertir les données de pose humaine en représentations graphiques exploitables et d'extraire des caractéristiques de mouvement précises grâce à un encodeur multi-niveaux. Le module d'attention intégré au goulot d'étranglement renforce la capacité du modèle à se concentrer sur les articulations les plus pertinentes pour la reconnaissance d'action. Par ailleurs, l'agrégation des prédictions issues de plusieurs niveaux du décodeur permet d'améliorer la robustesse globale du modèle, en réduisant notamment les erreurs de sur-segmentation.

Les évaluations menées sur deux bases de données industrielles, InHARD et HA4M, confirment l'efficacité du modèle. G-EAD atteint un score MoF de 74,59 % sur InHARD,

se positionnant au niveau de l'état de l'art, et affiche une performance compétitive sur HA4M avec un MoF de 93,11 %. En outre, bien que GSK-C2F, qui constituait notre première contribution, présente déjà de bonnes performances globales, G-EAD s'en distingue par une amélioration plus marquée sur les classes peu représentées. Les analyses montrent en effet des gains pouvant dépasser 20% de MoF pour certaines de ces catégories, attestant d'une meilleure capacité du modèle à gérer les situations de fort déséquilibre inter-classes.

Au-delà de la performance en segmentation, ces résultats ouvrent la voie à des applications concrètes dans l'industrie, notamment en analyse ergonomique. La capacité du modèle à identifier précisément les gestes effectués par les opérateurs permet d'envisager des analyses post-activité axées sur la détection de postures contraignantes ou de gestes répétitifs. En combinant ces informations avec des outils d'évaluation ergonomique comme RULA, il devient envisageable d'estimer la charge physique et la fatigue des opérateurs et de proposer des recommandations pour améliorer les conditions de travail telles que des ajustements des postes, une réorganisation des séquences opératoires, ou encore l'instauration de pauses actives.

Toutefois, la méthode actuelle repose sur le traitement intégral des vidéos, ce qui la limite à une application hors ligne (*offline*). Or, les environnements industriels exigent de plus en plus de solutions réactives, capables d'intervenir en temps réel. Dans la suite de ce manuscrit, nous étudierons cette nouvelle problématique en présentant l'évolution des modèles précédemment proposés, GSK-C2F et G-EAD, conçus pour fonctionner en temps réel *online*. L'objectif est de permettre la détection d'actions à la volée, condition nécessaire à la mise en place de systèmes d'assistance intelligents dans le cadre de l'Industrie 5.0.

Chapitre 5

Détection d'actions en ligne

5.1 Introduction

Au cours des chapitres précédents, nous avons étudié la segmentation d'actions humaines dans un cadre hors ligne, où l'intégralité de la séquence est disponible avant toute phase de traitement. Cette approche offre un contexte favorable à l'extraction de représentations riches et précises et demeure particulièrement utile pour de nombreuses applications industrielles telles que l'analyse ergonomique, l'établissement de rapports post-activité ou encore l'identification des gestes répétitifs au sein d'une activité. Cependant, malgré ces avantages, elle reste limitée dès lors que l'on vise des applications nécessitant des prises de décision rapides ou des interactions dynamiques avec l'opérateur. En effet, l'Industrie 5.0, fondée sur la collaboration étroite entre l'humain et la machine, impose de nouvelles exigences en matière de perception et d'intelligence artificielle, notamment la capacité à analyser et à interpréter les actions en temps réel.

La détection d'actions en ligne constitue ainsi une étape cruciale pour le déploiement de systèmes intelligents capables d'intervenir immédiatement dans le processus de production. Contrairement à la segmentation hors ligne, elle repose sur l'analyse séquentielle et incrémentale des données, sans accès au futur, ce qui accentue la difficulté à anticiper des gestes partiellement observés ou ambigus. Cette contrainte rend l'identification plus délicate mais ouvre en contrepartie la voie à des applications industrielles à fort impact, telles que l'assistance ergonomique aux opérateurs, la prévention des postures à risque, l'adaptation dynamique des robots collaboratifs aux intentions humaines ou encore l'optimisation en continu des processus d'assemblage. La détection en ligne devient ainsi un vecteur essentiel de sécurité, de qualité et de productivité, en phase avec les objectifs de l'Industrie 5.0.

Néanmoins, cette problématique soulève plusieurs défis scientifiques majeurs. L'absence d'informations futures contraint le modèle à raisonner uniquement à partir des observations présentes et passées, ce qui exige une capacité accrue de généralisation et de robustesse. Par ailleurs, les contraintes liées au temps réel imposent le recours à des architectures à la fois efficaces en calcul et suffisamment expressives pour modéliser la complexité des gestes humains dans un environnement industriel. Enfin, les mêmes contraintes que pour les architectures hors ligne, telles que la variabilité inter-individuelle, les occultations partielles dues aux outils ou aux machines, ainsi que le bruit inhérent aux capteurs, représentent autant d'obstacles supplémentaires à surmonter pour garantir la fiabilité des prédictions.

Pour relever ces défis, nous proposons, dans ce chapitre, une approche basée sur la distillation de connaissances, qui s'appuie sur un paradigme enseignant-étudiant, où un modèle enseignant, entraîné hors ligne, transmet ses connaissances à un modèle étudiant conçu pour fonctionner en ligne. Cette distillation de connaissances permet de tirer parti de la richesse d'un apprentissage hors ligne tout en respectant les contraintes temporelles et computationnelles imposées par l'inférence en temps réel. Cette stratégie s'inscrit dans la continuité des travaux présentés dans les chapitres précédents, tout en marquant une étape décisive vers le déploiement effectif de systèmes de détection d'actions adaptés aux besoins de l'Industrie 5.0.

La structure du chapitre est la suivante. Nous commençons par présenter, en section 5.2, les principaux enjeux liés à la détection d'actions en ligne. La section 5.3 présente ensuite un panorama des méthodes de distillation de connaissances. La section 5.4 introduit la méthode de distillation proposée dans cette étude, tandis que la section 5.5 présente les différentes leçons transmises par le modèle enseignant ainsi que leurs configurations possibles. La section 5.6 expose le protocole expérimental, analyse en détails les résultats obtenus et propose une discussion comparative de l'impact des différentes leçons, permettant d'en dégager les conclusions principales. Enfin, la section 5.7 conclut ce chapitre et ouvre sur les perspectives qui découlent de ce travail.

5.2 Enjeux de la détection d'actions en ligne

La détection d'actions en ligne, appliquée aux environnements industriels, se distingue par une série de défis spécifiques qui conditionnent sa performance et son applicabilité. Contrairement à la segmentation hors ligne, elle repose sur un traitement incrémental des données, ce qui accentue la difficulté d'anticiper correctement les actions en cours. Deux grands axes se dégagent : la contrainte temporelle liée au temps réel et la gestion de l'incertitude inhérente à l'absence d'accès au futur. Ces aspects, auxquels s'ajoutent les perspectives applicatives propres à l'Industrie 5.0, sont détaillés ci-dessous.

5.2.1 Contrainte de temps réel

La détection en ligne impose un traitement continu des flux de données, avec une latence minimale entre l'acquisition et la prédiction. Les systèmes doivent être capables d'analyser des séquences complexes rapidement, sans sacrifier la précision des résultats. Dans un environne-

ment industriel, ce critère est déterminant : un retard dans la détection peut compromettre la sécurité d'un opérateur, retarder un processus d'assemblage ou perturber la fluidité de la collaboration homain-machine. L'enjeu consiste donc à développer des modèles alliant efficacité computationnelle et richesse des représentations, afin de répondre aux contraintes du temps réel.

5.2.2 Gestion de l'incertitude et des ambiguïtés

La détection en ligne est confrontée à une incertitude structurelle : dans les premiers instants d'une action, les indices observables sont souvent insuffisants pour distinguer clairement plusieurs classes proches. L'ambiguïté est d'autant plus forte que le système ne dispose d'aucune information future pour confirmer ou infirmer son hypothèse. Le défi consiste à concilier réactivité et fiabilité, en tirant parti des indices partiels disponibles sans multiplier les erreurs de prédiction. Des stratégies basées sur l'attention sélective, la modélisation probabiliste ou l'incorporation de connaissances a priori constituent des pistes prometteuses pour répondre à cette difficulté [197].

5.2.3 Perspectives applicatives

Au-delà de ces considérations techniques, la détection des actions en ligne revêt une importance stratégique dans le cadre de l'Industrie 5.0. Elle permet d'envisager des systèmes intelligents capables d'offrir un retour immédiat aux opérateurs afin de prévenir les erreurs lors des tâches d'assemblage, de détecter en temps réel des postures inadaptées pour assurer une assistance ergonomique, d'améliorer la formation en guidant l'exécution des gestes. Dans un cadre plus large, les prédictions en ligne peuvent également constituer un signal d'entrée pour des systèmes collaboratifs homain–robot, en fournissant une estimation continue de l'activité en cours, pouvant contribuer la coordination et l'adaptation dynamique des robots collaboratifs. Ces perspectives renforcent l'idée que la détection en ligne n'est pas seulement un défi scientifique, mais également un levier essentiel pour la sécurité, la qualité et la productivité des environnements industriels du futur [11].

5.3 Méthodes de distillation de connaissances (enseignant-étudiant)

Comme nous l'avons souligné dans la section précédente, la détection d'actions en ligne se heurte à plusieurs défis majeurs : la contrainte de traitement en temps réel, la variabilité des gestes et des contextes industriels, ainsi que l'incertitude liée à l'absence d'informations futures. Ces limitations restreignent l'usage direct de modèles profonds de grande capacité, qui, bien qu'efficaces pour la segmentation hors ligne, demeurent trop coûteux en calcul et peu adaptés aux exigences d'un déploiement opérationnel. Une alternative consisterait à simplifier l'architecture pour satisfaire les contraintes temporelles, mais cela se traduit généralement par une perte significative de précision et de robustesse, ce qui compromet l'intérêt de la solution en pratique.

C'est précisément dans ce contexte que la distillation de connaissances (Knowledge Distillation, KD) s'impose comme une approche particulièrement pertinente. Introduite par Hinton, Vinyals et Dean [198], cette technique vise à concilier les performances des modèles complexes avec l'efficacité requise pour une exécution en ligne. Le principe repose sur un paradigme enseignant–étudiant, où un modèle enseignant, entraîné sans contrainte de calcul hors ligne, transmet son savoir à un modèle étudiant plus compact, spécifiquement conçu pour répondre aux exigences du temps réel. Ce transfert permet de doter l'étudiant d'une capacité de généralisation et d'une expressivité accrues, tout en respectant les contraintes de latence et de ressources computationnelles.

L'intérêt de la KD réside dans sa capacité à fournir un signal d'apprentissage enrichi, qui dépasse la supervision classique fondée uniquement sur les étiquettes de vérité de terrain. Les prédictions d'un modèle enseignant contiennent en effet des informations plus fines, telles que les relations entre classes voisines, la structure temporelle des séquences ou encore les dépendances spatiales entre articulations. Ces éléments, difficilement accessibles par un apprentissage supervisé standard, constituent un levier essentiel pour améliorer la qualité des modèles étudiants. En transmettant ces connaissances implicites, la KD permet de rapprocher les performances d'un étudiant léger de celles d'un enseignant complexe, tout en garantissant l'efficacité requise pour la détection d'actions en ligne.

Dans la suite, nous présentons les principales familles de méthodes de distillation, en fonction de la nature des signaux transférés : distillation au niveau des logits en section 5.3.1, au niveau des caractéristiques intermédiaires en section 5.3.2, distillation des relations en

section 5.3.3, et, enfin, une analyse de ces différents niveaux de distillation en section 5.3.4.

5.3.1 Distillation au niveau des logits

La distillation au niveau des logits constitue la forme la plus classique et la plus répandue du transfert de connaissances. Elle repose sur l'idée d'utiliser les distributions de probabilité produites par le modèle enseignant comme signal d'apprentissage pour l'étudiant [199], comme l'illustre la figure 5.1. Plus précisément, au lieu de fournir une supervision "dure" (hard labels) qui attribue une probabilité nulle aux classes incorrectes et une valeur unitaire à la classe correcte, l'enseignant génère une supervision "douce" (soft labels) qui reflète la proximité relative entre les classes.

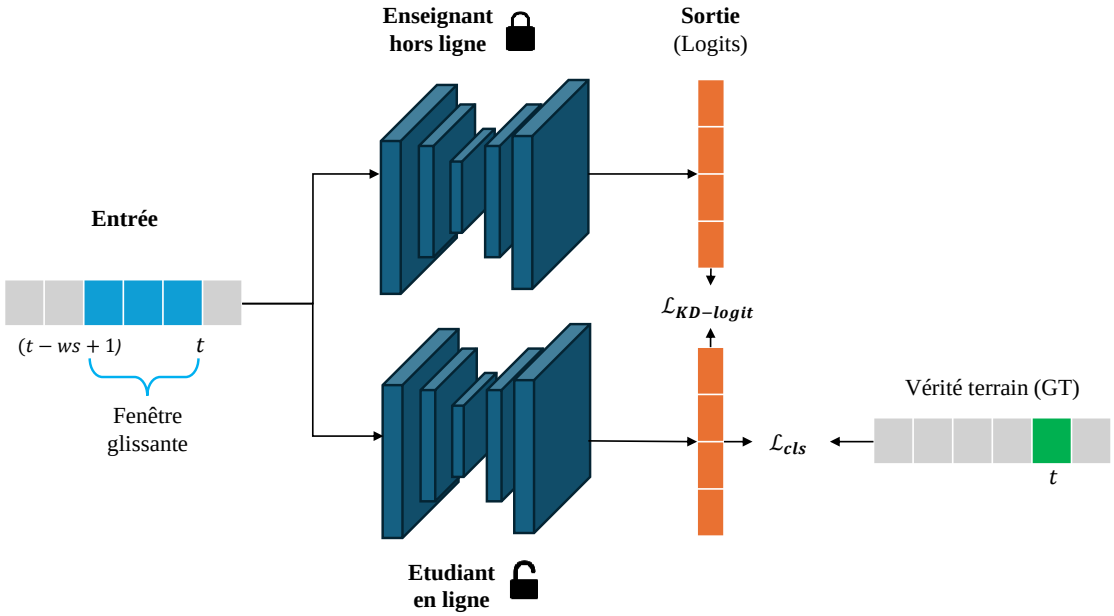


FIGURE 5.1 – Vue d'ensemble de l'architecture Enseignant–Étudiant proposée avec distillation par logits. ws : taille de la fenêtre glissante. $\mathcal{L}_{\text{cls}}$: la fonction de perte de classification (voir l'équation 3.23). $\mathcal{L}_{\text{KD-logit}}$: la fonction de perte de distillation de logits (voir l'équation 5.2).

Ce principe a été formalisé par l'introduction d'un paramètre de température dans la fonction softmax, ce qui permet d'obtenir une distribution de probabilités plus lissée et de mieux refléter les relations entre les classes [200]. La fonction de perte utilisée pour entraîner le modèle étudiant s'écrit :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\, \mathcal{L}_{\text{KD-logit}} \tag{5.1}$$

avec $\mathcal{L}_{\text{cls}}$ la fonction de perte de classification (voir Équation 3.23). La composante de distillation $\mathcal{L}_{\text{KD-logit}}$ est définie par :

$$\mathcal{L}_{\text{KD-logit}} = \text{KL}\left(\text{softmax}\left(\frac{\mathbf{Z}^T}{T}\right) \,\middle\|\, \text{softmax}\left(\frac{\mathbf{Z}^S}{T}\right)\right) \tag{5.2}$$

Où $\mathbf{Z}^T$ et $\mathbf{Z}^S$ les logits du modèle enseignant et du modèle étudiant, T la température, α le coefficient d'équilibrage, et $\text{KL}(\cdot\|\cdot)$ la divergence de Kullback–Leibler.

Dans le cadre de la reconnaissance des actions, notamment dans des environnements industriels, cette approche présente un intérêt majeur. Les actions peuvent en effet être très proches les unes des autres dans leurs phases initiales (par exemple, tendre la main pour saisir un tournevis ou manipuler une vis), ce qui rend difficile une discrimination immédiate lorsqu'on se limite à des étiquettes binaires. En transmettant les distributions de probabilité du modèle enseignant, l'étudiant bénéficie d'un signal plus riche qui capture ces similarités inter-classes et l'aide à différencier plus finement les gestes partiellement observés. Ce mécanisme est particulièrement précieux dans un contexte en ligne, où les décisions doivent être prises avant même que l'action ne soit entièrement réalisée.

Toutefois, la distillation au niveau des logits présente certaines limites. Elle se concentre exclusivement sur la sortie finale du modèle enseignant et néglige l'information contenue dans ses représentations internes [201]. Or, pour des tâches complexes telles que la détection d'actions, les couches intermédiaires du réseau encodent des structures temporelles et spatiales essentielles (par exemple, les relations entre articulations ou la dynamique des segments temporels). Ces considérations ont conduit à l'exploration de méthodes complémentaires, telles que la distillation au niveau des caractéristiques, que nous présentons dans la section suivante.

5.3.2 Distillation au niveau des caractéristiques

Dans les réseaux profonds, une part essentielle du savoir appris par le modèle réside dans les représentations intermédiaires extraites des différentes couches de l'architecture. Ces représentations encodent des informations de haut niveau sur la structure des données : relations spatiales entre entités, régularités temporelles, ou encore descripteurs abstraits discriminants pour la classification. La distillation au niveau des caractéristiques (feature-based distillation)

vise précisément à exploiter cette richesse. Le principe consiste à forcer le modèle étudiant à imiter les représentations internes de l'enseignant en minimisant la distance entre leurs cartes de caractéristiques [202], comme illustré à la figure 5.2. La fonction de perte utilisée pour entraîner le modèle étudiant s'écrit :

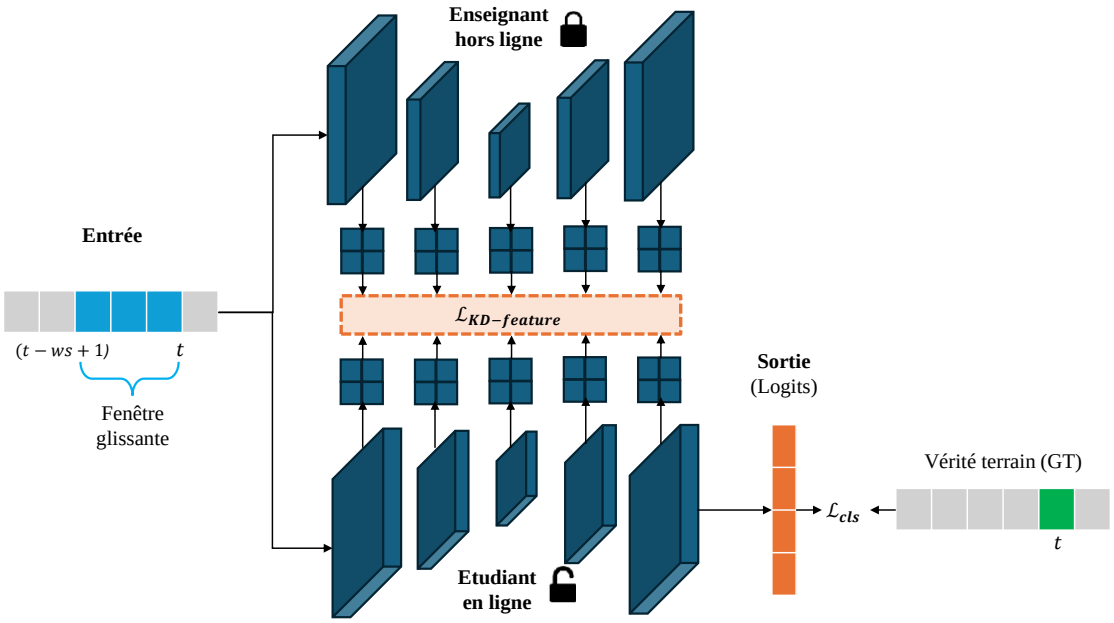


FIGURE 5.2 – Vue d'ensemble de l'architecture Enseignant–Étudiant proposée avec distillation par caractéristiques. ws : taille de la fenêtre glissante. $\mathcal{L}_{\text{cls}}$: la fonction de perte de classification (voir l'équation 3.24). $\mathcal{L}_{\text{KD-feature}}$: la fonction de perte de distillation des caractéristiques internes (voir l'équation 5.4).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\, \mathcal{L}_{\text{KD-feature}} \tag{5.3}$$

avec $\mathcal{L}_{\text{cls}}$ la fonction de perte de classification (voir Équation 3.23). La composante de distillation $\mathcal{L}_{\text{KD-feature}}$ est définie par :

$$\mathcal{L}_{\text{KD-feature}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{F}_i^{(s)} - \mathbf{F}_i^{(t)}\|_2^2 \tag{5.4}$$

Où $\mathbf{F}_i^{(s)}$ et $\mathbf{F}_i^{(t)}$ désignent les cartes de caractéristiques correspondantes de l'étudiant et de l'enseignant, et N le nombre total de blocs considérés.

Dans le cadre de la reconnaissance des actions humaines. Les couches intermédiaires du modèle enseignant capturent souvent des descripteurs spatio-temporels complexes, tels que la

corrélation entre les articulations du squelette, les trajectoires locales des segments corporels, ou encore des motifs temporels caractéristiques de certaines actions. En transmettant ces représentations, l'enseignant fournit à l'étudiant une base plus structurée pour interpréter les séquences, ce qui se traduit par une meilleure robustesse face à la variabilité des gestes des opérateurs. De plus, dans un contexte industriel, les données peuvent être incomplètes ou bruitées, en raison d'occlusions dues aux outils ou aux machines [15]. Les représentations internes de l'enseignant, plus abstraites et robustes que les données brutes, constituent alors un signal d'apprentissage particulièrement utile. L'étudiant, contraint par une architecture plus légère, bénéficie ainsi d'une supervision qui lui permet de mieux généraliser aux situations réelles.

Néanmoins, cette approche comporte également certaines limites. La correspondance entre les couches de l'enseignant et de l'étudiant n'est pas évidente, notamment lorsque leurs architectures diffèrent fortement [203]. De plus, imiter directement des caractéristiques d'un modèle peut entraîner une dépendance excessive à la structure de ce dernier, ce qui réduit la capacité d'adaptation du modèle étudiant à son propre espace de représentation [204]. Ces considérations ont conduit à l'exploration de méthodes complémentaires visant non seulement à transférer le contenu des représentations, mais aussi les relations qu'elles encodent, ce que nous aborderons dans la section suivante.

5.3.3 Distillation des relations

La distillation relationnelle vise à transmettre la structure des dépendances entre ces représentations plutôt que leur contenu brut [205]. Autrement dit, au lieu de forcer l'étudiant à imiter les cartes de caractéristiques couche par couche, on lui enseigne comment ces activations s'organisent et interagissent entre elles.

Dans le cadre de la reconnaissance d'actions humaines, cette approche est particulièrement adaptée car les gestes industriels ne se définissent pas uniquement par l'état isolé d'une articulation, mais par la cohérence des relations spatio-temporelles qui les relient. Ainsi, le mouvement coordonné de la main et de l'avant-bras, ou la séquence d'interactions entre un opérateur et un outil, constitue des signaux cruciaux que l'étudiant peut apprendre grâce à cette forme de distillation.

Cependant, cette richesse informationnelle a un coût. La construction et la comparaison de matrices relationnelles (par exemple de similarité ou de corrélation) introduisent une com-

plexité de calcul plus élevée [200]. Pour cette raison, la distillation relationnelle est souvent envisagée comme un complément aux approches plus légères, ouvrant la voie à des stratégies hybrides plus efficaces.

5.3.4 Analyse des méthodes de distillation

Les stratégies de distillation étudiées précédemment, qui reposent sur les logits, les caractéristiques internes ou les relations spatio-temporelles, offrent des informations complémentaires, mais chacune présente des limites lorsqu'elles sont utilisées séparément. La distillation au niveau des logits est légère, directement liée à la tâche de classification et simple à intégrer, mais elle ne transmet qu'une information peu expressive sur les représentations apprises par le modèle enseignant. À l'inverse, la distillation des caractéristiques permet de guider l'étudiant vers des représentations intermédiaires plus riches et plus discriminantes, tout en conservant un coût de calcul raisonnable. Cependant, elle ne suffit pas à elle seule à capturer toute la diversité des signaux utiles pour la reconnaissance d'actions. Par ailleurs, bien que la distillation relationnelle puisse fournir une structure spatio-temporelle plus complète, son coût de calcul la rend peu adaptée à un cadre strictement en ligne. Afin de concilier expressivité, efficacité et contraintes de temps réel, nous proposons, dans ce travail, une méthode de distillation hybride basée exclusivement sur les logits et les caractéristiques, combinant leurs avantages respectifs tout en évitant la complexité de la distillation relationnelle.

5.4 Méthode de distillation proposée

Après avoir analysé les forces et les limites des différentes stratégies de distillation en section 5.3, nous introduisons ici une approche hybride conçue pour exploiter de manière progressive les signaux les plus adaptés au contexte en ligne. L'objectif est de pallier les insuffisances des approches de distillation de connaissances individuelles mises en évidence en section 5.3.4, en combinant une distillation basée sur les prédictions du modèle enseignant (logits) avec une supervision plus riche issue des caractéristiques internes, afin d'assurer à la fois un apprentissage efficace dans un contexte en temps réel (online). La fonction de perte utilisée pour l'entraînement du modèle étudiant est formulée comme suit :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\, \mathcal{L}_{\text{KD-logit}} + \beta\, \mathcal{L}_{\text{KD-feature}}. \tag{5.5}$$

Le terme $\mathcal{L}_{\text{cls}}$ correspond à la fonction de perte de classification supervisée (voir l'équation 3.23). Le coefficient α pondère la fonction de perte $\mathcal{L}_{\text{KD-logit}}$, qui transfère les connaissances décisionnelles (logits) de l'enseignant à l'étudiant (voir l'équation 5.2). De même, β pondère la fonction de perte $\mathcal{L}_{\text{KD-feature}}$, responsable du transfert des représentations intermédiaires apprises par l'enseignant (voir l'équation 5.4).

L'un des aspects essentiels de notre approche réside dans son **apprentissage progressif**. Plutôt que d'appliquer simultanément toutes les fonctions de pertes, la distillation est effectuée par étapes successives afin de stabiliser l'apprentissage du modèle étudiant. Dans un premier temps, une phase de préchauffage (3 à 7 époques) permet à l'étudiant d'apprendre uniquement à partir de la supervision directe $\mathcal{L}_{\text{cls}}$. Ensuite, la distillation par logits est progressivement introduite sur une période de 15 à 30 époques appelé *logit_kd_epochs*, en augmentant linéairement le poids α selon un coefficient de montée appelé *logit_ramp*. Enfin, la distillation par caractéristiques est activée pendant une durée plus courte de 5 à 15 époques appelé *feat_kd_epochs* avec un coefficient β croissant, contrôlé par le paramètre *feat_ramp*. Ce processus hiérarchique assure une intégration stable et graduelle des connaissances, allant des sorties globales (logits) aux représentations internes (features).

L'ensemble des hyperparamètres a été ajusté à l'aide d'une recherche bayésienne réalisée avec *Optuna* [206]. Les plages explorées incluent : un taux d'apprentissage (*learning rate*) entre 3×10^{-4} et 3×10^{-3}, un coefficient de régularisation (*weight decay*) entre 10^{-6} et 10^{-3}, une température de distillation (T) choisie parmi $\{6, 8, 10, 12, 16\}$, un α compris entre 0.3 et 0.9, et un β entre 0.05 et 0.3. Les valeurs optimales des paramètres varient selon la leçon et la configuration étudiante considérée. À titre illustratif, pour la meilleure configuration retenue (*Étudiant P1*), dont les résultats sont présentés en Section 5.6.6, les valeurs optimales obtenues sont :

$$T = 12, \quad \alpha = 0.637, \quad \beta = 0.201,$$

$$\texttt{learning_rate} = 2.40 \times 10^{-3}, \quad \texttt{weight_decay} = 2.58 \times 10^{-6},$$

$$\texttt{warmup_epochs} = 3, \quad \texttt{logit_kd_epochs} = 20, \quad \texttt{feat_kd_epochs} = 5,$$

$$\texttt{logit_ramp} = 5, \quad \texttt{feat_ramp} = 3.$$

Cette optimisation automatique permet d'identifier les combinaisons de paramètres garantissant le meilleur compromis entre la stabilité d'entraînement, la précision temporelle et la capacité de généralisation du modèle étudiant. Ainsi, notre cadre de distillation progressive

combine de manière cohérente les deux formes de transfert, décisionnel (logits) et représentationnel (features), en exploitant les informations issues à différents niveaux du modèle enseignant. Cette méthode constitue la base commune des leçons présentées dans les sections suivantes, où chaque type de connaissance transférée est analysé individuellement et évalué expérimentalement.

5.5 Leçons proposées transmises par le modèle enseignant

Au-delà des aspects méthodologiques de la distillation présentés précédemment, il est essentiel de s'interroger sur la nature des connaissances que le modèle "Enseignant" peut transmettre au modèle "Étudiant". Dans le cadre industriel, et plus particulièrement dans les tâches d'assemblage, plusieurs dimensions clés influencent directement la qualité de la détection des actions en ligne.

Nous proposons trois formes principales de connaissances à distiller. La première concerne la **dimension temporelle**, essentielle pour modéliser la continuité et la dépendance entre les gestes successifs dans un flux d'action continu. La seconde porte sur la **relation entre les actions et les objets manipulés**, car les outils et les pièces d'assemblage constituent des indices sémantiques déterminants pour comprendre l'intention du geste. Enfin, la troisième concerne la **gestion de l'occultation**, problématique récurrente en milieu industriel où les articulations de l'opérateur peuvent être partiellement invisibles ou perturbées par l'environnement.

Ces trois formes de savoir constituent les piliers de notre stratégie de distillation en ligne. Elles sont détaillées dans les sous-sections suivantes, chacune proposant une méthode de transfert spécifique, à savoir la leçon temporelle (section 5.5.1), la leçon liée aux objets (section 5.5.2) et la leçon liée à l'occultation (section 5.5.3). Enfin, la section 5.5.4 complète cette présentation en introduisant les variantes cumulatives, où plusieurs leçons sont combinées soit de manière directe, soit de manière progressive afin d'évaluer leur complémentarité et leur impact sur la performance du modèle en ligne.

5.5.1 Leçon temporelle

Au fil de nos expérimentations, le passage d'un apprentissage **hors ligne** à un apprentissage **en ligne** a mis en évidence une limite majeure des modèles en temps réel, la **perte de cohérence temporelle**. Notre modèle hors ligne, entraîné sur la séquence complète d'as-

semblage, exploitait efficacement le contexte global des actions et affichait de très bonnes performances. Cependant, lors de son adaptation au cadre en ligne, où le modèle ne dispose que d'une fenêtre d'observation restreinte centrée sur le présent, nous avons constaté une baisse significative de ses performances (voir tableau 5.1). Cette dégradation ne provenait pas d'une faiblesse architecturale, mais du fait que le modèle, privé de la vision d'ensemble de la séquence, perdait la capacité à interpréter les dépendances entre les gestes successifs. C'est cette constatation qui nous a conduits à formuler la **leçon temporelle** : transférer, par distillation de connaissances, la compréhension acquise par le modèle enseignant hors ligne vers le modèle étudiant en ligne, afin de restaurer une continuité temporelle implicite malgré la contrainte de causalité.

Principe de la distillation temporelle : Dans les séquences industrielles, les actions d'assemblage suivent un ordre globalement structuré mais flexible. Un opérateur peut, par exemple, (consulter le manuel), puis (préparer un outil), ou (positionner une pièce) avant de (visser), tout en adaptant ponctuellement cet ordre en fonction de la tâche ou du contexte. Cette organisation semi-séquentielle reflète une **structure temporelle récurrente et contextuelle** que le modèle enseignant, entraîné sur des séquences complètes (passées et futures), apprend implicitement à représenter.

Comme illustré à la figure 5.3, l'objectif de la distillation temporelle est donc de transférer cette compréhension séquentielle globale, identifiée comme (leçon temporelle), au modèle étudiant en ligne, contraint par une fenêtre d'observation limitée. Ainsi, l'étudiant peut maintenir une représentation cohérente de la dynamique des gestes et anticiper plus efficacement les transitions d'action.

5.5.2 Leçon liée aux objets

Après avoir présenté, dans la section 5.5.1, l'intérêt de la **leçon temporelle** pour renforcer la cohérence séquentielle du modèle en ligne, nous avons cherché à enrichir davantage le contexte du modèle enseignant. En observant les séquences industrielles, il est apparu que les gestes de l'opérateur ne peuvent être interprétés isolément de leur environnement matériel, car chaque action s'inscrit dans une interaction fonctionnelle avec un objet tel qu'un composant, une pièce ou un outil. Ainsi, alors que la leçon temporelle capture la dynamique du mouvement humain dans le temps, la **leçon liée aux objets** vise à modéliser le lien entre le geste et l'objet manipulé.

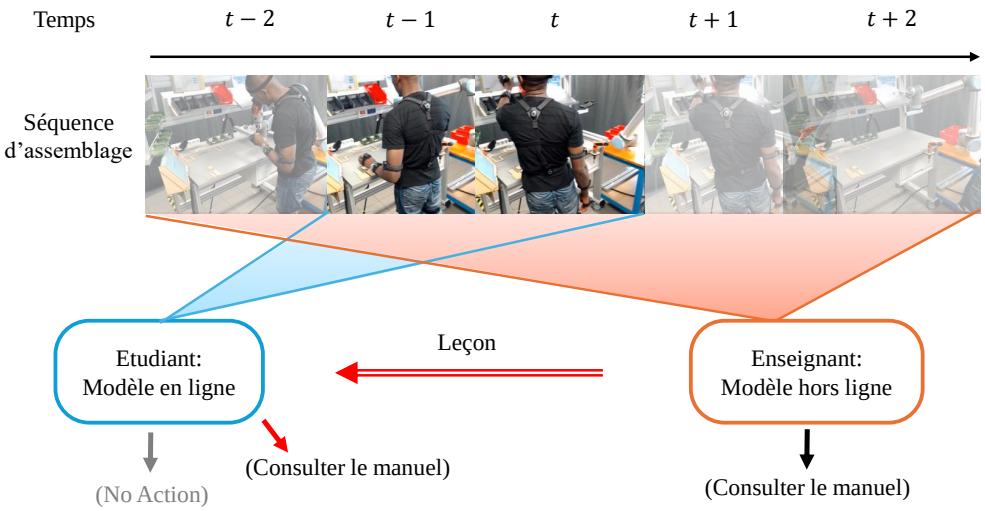


FIGURE 5.3 – Schéma du processus de distillation de la leçon temporelle : Le modèle enseignant, disposant de l'ensemble de la séquence, fournit un signal temporel cohérent, qui est transmis au modèle étudiant, disposant uniquement d'une fenêtre temporelle limitée, afin d'améliorer la détection d'action en ligne.

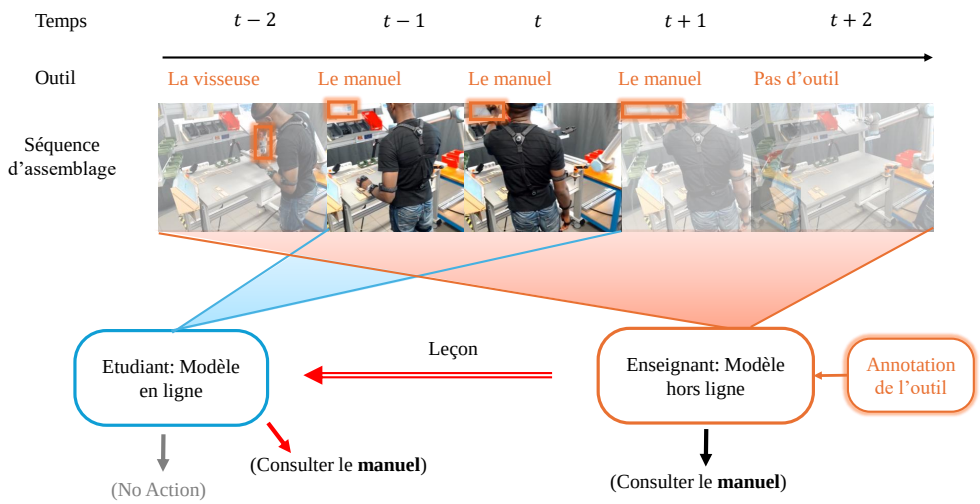


FIGURE 5.4 – Schéma du processus de distillation de la leçon objet : Le modèle enseignant exploite à la fois le contexte temporel global et l'information de l'outil manipulé, ici "Le manuel", qui oriente directement vers la bonne action. Ce signal enrichi est ensuite transmis au modèle étudiant afin d'améliorer la détection d'action en ligne.

Dans le cadre industriel, (Saisir un tournevis), (Positionner une pièce) ou (Visser) ne se résume pas à des configurations articulaires, mais à des interactions spécifiques entre l'opérateur et l'outil. Afin de capturer cette dimension fonctionnelle, nous avons enrichi le modèle

hors ligne en y intégrant explicitement le **contexte outil**, extrait directement des annotations du jeu de données (Voir la figure 5.4). Ces annotations indiquent l'outil que l'opérateur utilise à chaque instant t. Le terme outil renvoie donc aux outillages utilisés (tournevis, clés, etc.) incluant la gamme d'assemblage (manuel), à l'exclusion des composants et pièces manipulés. Ce choix se justifie par le fait que l'outil employé constitue un indicateur fort de l'action réalisée, tandis que les pièces présentent une variabilité élevée et contribuent moins à la discrimination des actions dans notre cadre. Ce modèle intègre ainsi à la fois la dynamique temporelle complète des séquences et la sémantique des outils manipulés. La combinaison de ces deux sources de connaissance (temporelle et d'objet) constitue la **leçon liée aux objets**, qui est ensuite transmise, par distillation, au modèle en ligne, dont la fenêtre temporelle et le champ perceptif sont limités.

Principe de la distillation liée aux objets : La leçon liée aux objets repose sur une double supervision : la dynamique temporelle, déjà présentée en section 5.5.1, et le contexte outil, que nous intégrons explicitement au modèle enseignant. Nous extrayons ainsi l'ensemble des outils distincts présents dans les labels d'action et associons à chacun d'eux une représentation continue via un **embedding** appris lors de l'entraînement du modèle.

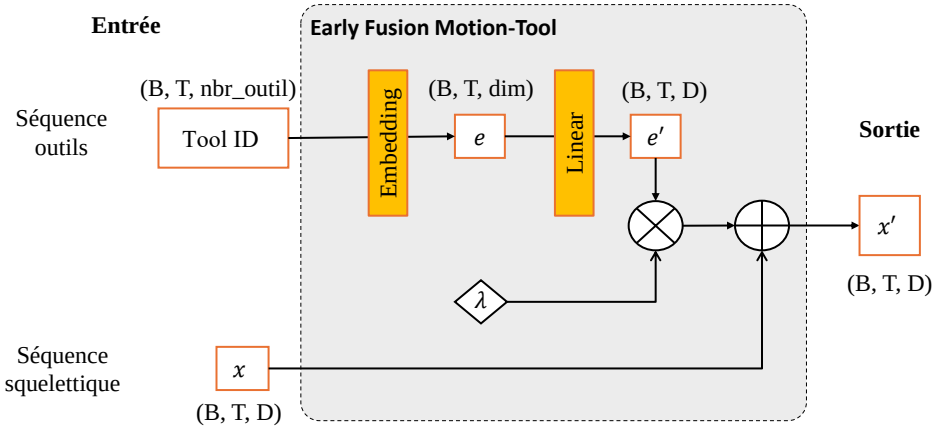


FIGURE 5.5 – Processus de fusion résiduelle (Early Fusion Motion–Tool) entre les embeddings d'objets et les caractéristiques de mouvement. Les tenseurs suivent la forme $(B, T, \cdot)$, où B désigne la taille du batch, T la longueur de la séquence et $\cdot \in \{nbr_{outil}, dim, D\}$ la taille des embeddings. x : caractéristiques squelettiques, e embedding de l'outil manipulé, W_o : matrice projetant l'espace des outils vers celui du mouvement, λ : coefficient de l'outil, x' : caractéristiques de sorite. $\oplus$: addition. $\otimes$: multiplication.

Pour chaque instant t, l'outil manipulé est représenté par un vecteur latent $\mathbf{e}_t$, qui est

projeté dans l'espace de mouvement par la matrice $\mathbf{W}_{dim}$. La fusion entre mouvement et information outil est réalisée par une addition résiduelle (voir la figure 5.5 illustrant le module "Early Fusion Motion-Tool" responsable de cette procédure), ce qui donne :

$$\mathbf{x}'_t = \mathbf{x}_t + \lambda \cdot \tanh(\mathbf{W}_{dim}\mathbf{e}_t) \tag{5.6}$$

où $\mathbf{x}_t \in \mathbb{R}^{B\times T\times D}$ correspond aux caractéristiques squelettiques au temps t, $\mathbf{e}_t \in \mathbb{R}^{B\times T\times dim}$ désigne l'embedding de l'outil, $\mathbf{W}_{dim} \in \mathbb{R}^{B\times T\times dim}$ est la matrice qui projette l'espace des outils vers l'espace des mouvements, λ contrôle l'importance de l'information outil dans la fusion, et $\mathbf{x}'_t$ représente les caractéristiques enrichies résultant de cette combinaison. Cette opération permet d'intégrer le contexte lié à l'outil sans modifier fortement les caractéristiques de mouvement. Le modèle étudiant, qui ne dispose que d'une fenêtre glissante et d'aucun contexte sur l'outil $\mathbf{x}_{t-l:t}$, apprend ensuite à imiter ces représentations et les prédictions associées à l'aide d'une distillation progressive, présentée en section 5.4.

5.5.3 Leçon liée à l'occultation

Après avoir exploré la dimension temporelle et le contexte des objets manipulés, un autre défi récurrent dans les environnements industriels concerne la **gestion des occultations**. Dans les postes d'assemblage, certaines parties du corps, notamment les jambes ou les bras, peuvent être masquées par les outils, les objets présents dans l'espace de travail (machines...), ou par le corps de l'opérateur lui-même. Cette situation complique la modélisation du mouvement et limite les performances des modèles d'action en ligne, qui doivent produire des prédictions fiables à partir d'informations partielles.

Principe de la leçon liée à l'occultation : Plusieurs approches pourraient être envisagées pour simuler le problème d'occultation, telles que l'injection de bruit dans les coordonnées articulaires, le masquage ponctuel de certaines articulations, ou encore la suppression de segments temporels afin de reproduire des pertes intermittentes de suivi. Chacune de ces stratégies permettrait de modéliser un défaut spécifique des technologies de capture de mouvement. Toutefois, pour établir une première validation claire et contrôlée du concept, nous avons retenu une forme d'occultation simple, dans laquelle le modèle étudiant ne reçoit qu'une partie du squelette, par exemple les articulations du haut du corps, tandis que le modèle enseignant exploite la structure complète, comme illustré à la figure 5.6. Cette configuration vise à transmettre au modèle étudiant la capacité du modèle enseignant à interpréter correc-

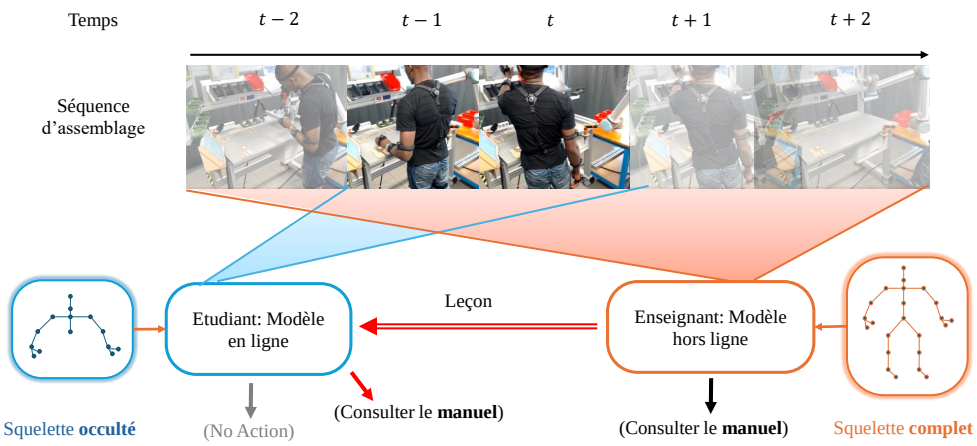


FIGURE 5.6 – Schéma du processus de distillation de la leçon d'occultation. Le modèle enseignant utilise le contexte temporel complet et un squelette non occulté pour guider un modèle étudiant doté d'un squelette partiellement occulté, améliorant ainsi la détection d'actions en ligne.

tement les gestes, même lorsque certaines articulations sont occultées ou absentes des données d'entrée.

5.5.4 Leçons cumulées

Dans les sections précédentes, nous avons présenté l'approche de combinaison de deux leçons (temporelle et objet) en section 5.5.2 et (temporelle et occlusion) en section 5.5.3. Ces approches visent à améliorer les performances du modèle étudiant par rapport à la distillation uniquement temporelle décrite à la section 5.5.1. Pour aller plus loin, nous nous intéressons également à cumuler les trois leçons identifiées : temporelle, d'objet et d'occultation.

5.5.4.1 Leçons combinées

Une première approche naturelle consiste à regrouper (combiner) directement ces trois sources d'information au sein d'un même modèle d'enseignement, puis à transférer cette connaissance cumulée vers le modèle étudiant. Cependant, comme nous le verrons dans les résultats présentés dans le tableau 5.7 cette stratégie s'avère peu efficace. La fusion simultanée de signaux issus de contextes très différents (dynamique temporelle, informations sur les outils et variations topologiques liées aux occlusions) introduit un niveau de bruit trop élevé, ce qui perturbe la capacité de l'enseignant à produire des représentations cohérentes et exploitables pour l'apprentissage de l'étudiant.

5.5.4.2 Leçons progressives

Pour dépasser les limites des leçons combinées, nous nous sommes appuyés sur une autre idée : **la distillation progressive**. Un travail récent [197] a proposé d'utiliser une série d'enseignants hiérarchisés pour la distillation temporelle, chaque enseignant couvrant une portion plus large du contexte temporel, du plus simple (proche de la fenêtre du modèle en ligne) au plus complexe (proche d'un modèle hors ligne). Le modèle étudiant apprend alors successivement auprès de ces enseignants, de manière graduelle *du facile au difficile*, ce qui a conduit à des performances supérieures.

Cette approche nous a inspiré une question centrale : au lieu d'appliquer la distillation progressive à un seul type de connaissance (par exemple le contexte temporel), serait-il possible d'en tirer parti pour cumuler des connaissances issues de contextes différents ? Autrement dit, l'étudiant pourrait-il apprendre successivement auprès de plusieurs enseignants spécialisés ? L'un pour la dynamique temporelle, un autre pour le contexte outil, un troisième pour la robustesse aux occlusions, afin d'intégrer ces trois dimensions sans créer les interférences observées lors d'une fusion directe comme présenté en section 5.5.4.1.

Principe de la distillation progressive des connaissances (Leçons) : Les leçons progressives visent à structurer le transfert de connaissances en séquençant l'apprentissage de l'étudiant. L'idée est de décomposer la supervision en plusieurs étapes cohérentes, chacune portée par un enseignant spécialisé dans un aspect particulier. L'étudiant est ainsi exposé à des connaissances plus simples dans un premier temps, puis à des connaissances de plus en plus complexes. Ce principe permet d'organiser la montée en difficulté et d'éviter les interférences observées lors de la combinaison directe des trois leçons. Cette approche a déjà montré son efficacité, notamment pour la dynamique temporelle [197] comme expliqué précédemment. Différents enseignants peuvent être construits en variant l'horizon temporel utilisé : un premier se limitant à un contexte futur court, un second s'appuyant sur un contexte intermédiaire, puis un troisième couvrant un horizon étendu, proche d'un modèle hors ligne. L'étudiant apprend alors progressivement à saisir des dépendances temporelles de plus en plus longues, en suivant un parcours (du facile au difficile).

Ce même principe peut être transposé aux autres leçons étudiées dans ce chapitre. Par exemple, pour l'occlusion, une progression naturelle consisterait à passer d'un enseignant basé sur une vision partielle du squelette (les bras), à un enseignant exploitant le haut du corps,

puis à un dernier enseignant utilisant l'ensemble du squelette. Pour la leçon liée aux objets, un enseignant simple pourrait utiliser un embedding d'outil faiblement pondéré, tandis qu'un enseignant plus avancé intégrerait un contexte outil riche, fusionné avec les caractéristiques de mouvement.

Au-delà des progressions internes à une même dimension, il est également possible de définir une progression inter-dimensionnelle. L'étudiant pourrait, par exemple, commencer par apprendre une structure temporelle stable, puis intégrer le contexte d'outil, avant de s'adapter à des conditions plus difficiles, telles que l'occultation. Cette organisation doit permettre de construire une hiérarchie cohérente entre les trois types de connaissances proposées dans ce chapitre : dynamique temporelle, information liée à l'outil et robustesse face aux articulations manquantes (occultations).

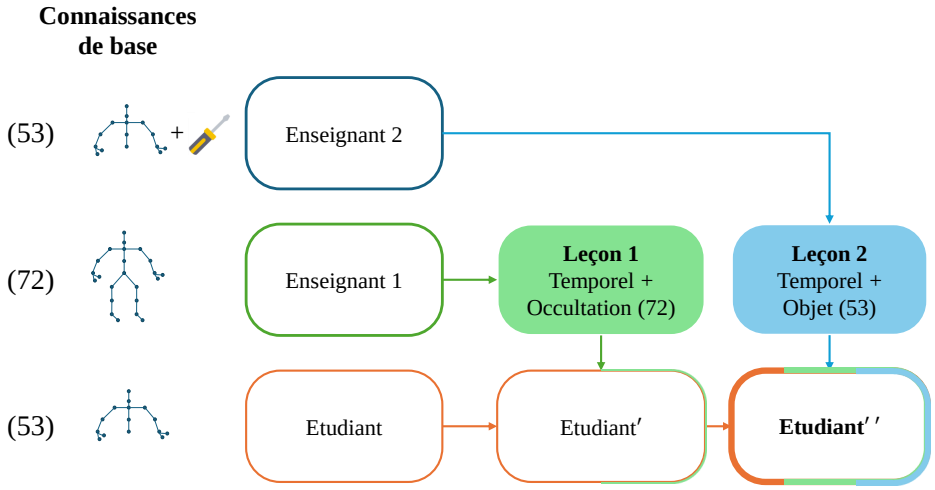


FIGURE 5.7 – Exemple d'un processus de distillation progressive. Cet exemple correspond au modèle (Étudiant P2) présent dans le tableau 5.7.

La figure 5.7 illustre ce principe par un exemple : le modèle étudiant, doté d'une connaissance de base de la topologie à 53 articulations (squelette occulté), apprend auprès d'un 1er enseignant, qui possède une topologie plus riche à 72 articulations (squelette complet). Ensuite, ce modèle **étudiant pré-entraîné** apprend une seconde fois avec un enseignant, cette fois dans un contexte d'outils plus riches (annotation d'outils). Chaque enseignant ne transmet qu'un aspect ciblé de la connaissance sous forme de leçons, et l'étudiant les assimile l'un après l'autre. Cette structuration offre un cadre d'apprentissage plus contrôlé et potentiellement plus efficace que la combinaison directe, qui expose l'étudiant à une surcharge d'informations hétérogènes dès les premières étapes de l'entraînement.

5.6 Résultats expérimentaux et discussion

5.6.1 Choix des modèles enseignant–étudiant

Toutes les expériences de distillation présentées dans ce chapitre utilisent comme base commune l'architecture GSK-C2F introduite au chapitre 3.3. Ce choix résulte d'une analyse comparative préalable entre GSK-C2F et GEAD, deux modèles que nous avons développés et évalués dans les mêmes conditions expérimentales en mode en ligne. Le tableau 5.1 résume cette comparaison.

Tableau 5.1 – Performances des modèles de détection d'action en ligne : Comparaison entre GSK-C2F et GEAD sur la base InHARD-13 à 21 articulations (complexité, coûts de calcul et performances).

Modèle	Params	FLOPs	Temps/ép.	Entraînement total	Inférence	MoF test
GSK-C2F	**4.2 M**	**0.8079 Mmac / 0.4076 Gmac**	**3.23 min**	**397.28 min**	**0.0229 s**	43.99%
GEAD	21.35 M	100.66 Gmac / 47.18 Gmac	44.93 min	3414.35 min	0.0382 s	**61.91%**

Bien que GEAD obtienne de meilleures performances en MoF (61.91 % contre 43.99 % pour GSK-C2F), plusieurs raisons nous ont conduits à privilégier GSK-C2F comme modèle de référence pour l'ensemble des leçons de distillation.

La première raison tient aux contraintes de calcul liées aux scénarios en ligne. GSK-C2F est un modèle léger (4.2 M paramètres, soit cinq fois moins que GEAD) et repose uniquement sur des opérations de convolution 1D. Cette structure le rend particulièrement adapté à l'inférence continue, avec un temps de prédiction moyen de 0.0229 s par frame, contre 0.0382 s pour GEAD. Cette différence, bien que modérée à l'échelle d'une image, devient significative lorsqu'il s'agit de traiter des séquences entières en temps réel.

La seconde raison concerne le coût d'entraînement. Comme le montre le tableau, GSK-C2F requiert 3.23 minutes par époque et un total de 397 minutes d'apprentissage, alors que GEAD atteint 44.93 minutes par époque pour un total de 3414 minutes. Ce rapport d'un ordre de grandeur est principalement dû au recours intensif de GEAD aux convolutions 2D et aux mécanismes d'attention, nécessaires pour manipuler explicitement la topologie du graphe squelettique tout au long du réseau. À l'inverse, GSK-C2F privilégie une architecture plus simple et plus rapide, ce qui facilite l'exploration de multiples configurations (hyperparamètres, topologies, combinaisons de leçons et stratégies de distillation progressive).

Enfin, la faible consommation en FLOPs de GSK-C2F (0.8079 Mmac en apprentissage contre 100.66 GMac pour GEAD) en fait un choix naturel pour un protocole expérimental impliquant de nombreux entraînements successifs, comme c'est le cas dans les différentes leçons présentées dans ce chapitre. Notons toutefois que GEAD, en raison de sa capacité à exploiter efficacement l'ensemble du contexte temporel et spatial, pourrait constituer un excellent candidat pour jouer le rôle de *modèle enseignant hors ligne*, tandis que GSK-C2F serait utilisé comme étudiant en ligne. Cette direction, particulièrement pertinente pour des scénarios de distillation asymétrique, constitue toutefois une piste de recherche que nous évoquons uniquement dans les perspectives en fin de chapitre.

5.6.2 Protocole expérimental

Leçon temporelle : Pour évaluer l'impact de la distillation temporelle, trois modèles ont été considérés : (i) un modèle enseignant hors ligne, entraîné sur des séquences complètes d'assemblage ; (ii) un modèle en ligne de base (baseline), ne disposant d'aucune distillation ; et (iii) un modèle étudiant en ligne, entraîné par distillation temporelle à partir du modèle enseignant. Les trois modèles partagent la même architecture que celle de notre première contribution, GSK-C2F.

Nous avons considéré quatre topologies squelettiques, celles qui seront utilisées dans la leçon temporelle ainsi que les leçons étudiées par la suite : (i) un squelette complet avec doigts (72 nœuds), (ii) un squelette du haut du corps avec doigts (53 nœuds), (iii) un squelette complet sans doigts (21 nœuds), et (iv) un squelette du haut du corps sans doigts (15 nœuds). Ces différentes topologies permettent d'analyser l'effet de la distillation en fonction du niveau de granularité articulaire.

Afin de rendre l'**entraînement** réalisable dans un délai raisonnable, un pas (`stride`) de 3 images (frames) a été utilisé durant la phase d'apprentissage. Ce choix, également employé dans [207], réduit le nombre de fenêtres temporelles d'un facteur trois tout en préservant une diversité suffisante d'exemples. En phase d'**inférence**, le `stride` est fixé à 1 afin d'évaluer les performances en conditions réelles (en ligne). Les méthodes de détection en ligne, définies en section 2.4.1, reposent sur l'utilisation d'une fenêtre glissante de 8 secondes (soit 240 images à 30 fps). Cette configuration est motivée par notre architecture GSK de type encodeur–décodeur (voir Section 3.3.2), qui requiert une séquence minimale de 240 frames en entrée. Conformément à l'approche basée sur l'historique présentée en section 2.4.2, le modèle prédit,

à chaque instant, l'étiquette de la dernière frame de la fenêtre, correspondant à la frame courante, à partir des 240 frames passés, soit 8 secondes. Cette stratégie respecte la contrainte d'inférence en ligne : seule la phase initiale nécessite l'accumulation des 8 premières secondes, après quoi la fenêtre glissante est déplacée au fil du temps et les prédictions sont mises à jour en continu, comme illustré dans la Figure 5.8.

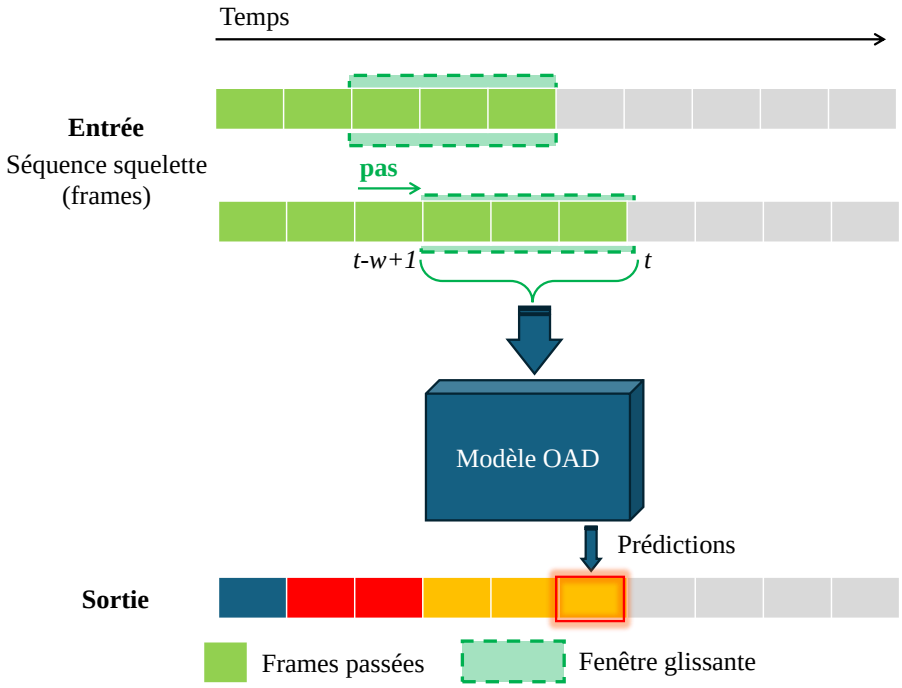


FIGURE 5.8 – Schéma du processus de détection d'action en ligne par fenêtre glissante. Chaque couleur dans la séquence de sortie représente une classe d'action. w : la taille de la fenêtre glissante.

Leçon liée aux objets : La procédure expérimentale suivie pour la leçon liée aux objets reprend le même protocole que celui de la leçon temporelle. La différence vient du fait que, dans cette leçon, nous ajoutons au modèle enseignant un module supplémentaire destiné à intégrer le contexte outil, en plus de la dynamique temporelle (voir la figure 5.5), ce qui permet à l'enseignant de produire des signaux plus complets, en tenant compte de l'outil effectivement manipulé.

L'ajout de ce module implique l'introduction de nouveaux hyperparamètres propres à la fusion "mouvement–outil". Le modèle enseignant est ainsi optimisé en explorant l'espace de recherche suivant :

— `embedding_dim` $\in \{8, 12, 16, 32\}$: taille du vecteur représentant l'outil (embedding).

— `dropout` $\in [0, 0.3]$: taux de dropout appliqué pour éviter le surapprentissage.
— `mask_p` $\in [0.1, 0.4]$: probabilité de masquer l'information "outil" pour rendre l'apprentissage plus robuste.
— `max_tool_scale` $\in [0.05, 0.45]$: facteur contrôlant l'amplitude maximale de la contribution de l'outil dans la fusion résiduelle.

Leçon liée aux occlustations : La procédure expérimentale suivie pour la leçon d'occultations reprend le même protocole que celui de la leçon temporelle. La particularité de cette leçon tient au fait que le modèle enseignant et le modèle étudiant n'utilisent pas la même topologie squelettique. L'enseignant opère sur un squelette complet (72 ou 21 nœuds), tandis que l'étudiant reçoit une version occultée correspondant uniquement au haut du corps (53 ou 15 nœuds). Cette configuration simule une situation d'occultation persistante, où certaines articulations ne sont pas visibles ou ne peuvent pas être capturées.

L'évaluation des performances de détection en ligne d'actions a été réalisée sur le jeu de données InHARD-13 [3] décrit en section 3.4.1.1. L'apprentissage, au cours des trois leçons, suit la fonction de perte progressive décrite dans la section 5.4. Les coefficients de pondération α (distillation des logits) et β (distillation des caractéristiques) sont optimisés pour chaque leçon et chaque configuration étudiante. À titre illustratif, pour notre meilleure configuration *Étudiant P1* (présentée dans la section 5.6.6), les valeurs optimales obtenues sont $\alpha = 0.637$ et $\beta = 0.201$, avec une température de distillation $T = 12$. Le code est disponible sur gitlab pour plus de détails concernant l'implémentation du modèle et son architecture.

5.6.3 Analyse de la leçon temporelle

Les résultats du tableau 5.2 mettent en évidence l'apport de la distillation temporelle à l'apprentissage en ligne. Pour les trois premières topologies (72, 53 et 21 nœuds), le modèle étudiant dépasse le modèle de base (sans distillation) avec des gains de MoF allant de +2.09 % à +4.42 %. Ces améliorations confirment que l'information temporelle globale fournie par le modèle enseignant, qui offre le contexte complet de l'activité et une vue d'ensemble de la séquence, aide le modèle étudiant à mieux anticiper les transitions d'actions au sein de la fenêtre glissante.

On observe toutefois une tendance différentielle selon la granularité du squelette : les topologies comportant le squelette complet (72 et 21 nœuds) tirent un bénéfice particulièrement marqué de la distillation temporelle, tandis que la topologie réduite à 15 nœuds ne montre

TABLEAU 5.2 – Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 entre le modèle de base (sans distillation) et le modèle étudiant (avec distillation temporelle).

Modèle	Mode	KD transfer	MoF (%)
Squelette à 72 nœuds			
Enseignant	Offline	-	73.86
Base (sans distillation)	Online	-	48.55
Étudiant	**Online**	**Temporel**	**52.97 (+4.42)**
Squelette à 53 nœuds			
Enseignant	Offline	-	68.35
Base (sans distillation)	Online	-	45.72
Étudiant	**Online**	**Temporel**	**47.81 (+2.09)**
Squelette à 21 nœuds			
Enseignant	Offline	-	65.38
Base (sans distillation)	Online	-	43.99
Étudiant	**Online**	**Temporel**	**48.01 (+4.02)**
Squelette à 15 nœuds			
Enseignant	Offline	-	64.57
Base (sans distillation)	**Online**	-	**55.22**
Étudiant	Online	Temporel	54.47 (-0.75)

pas de gain. Ce comportement suggère que le squelette à 15 nœuds atteint déjà un niveau de performance proche du maximum permis par sa structure simplifiée. La réduction drastique de la granularité articulaire limite l'information exploitable, de sorte que la distillation temporelle ne dispose pas d'un signal suffisant pour dépasser ce plafond de performance. De manière générale, ces résultats démontrent que la distillation temporelle améliore la robustesse du modèle en ligne, en particulier lorsque la structure squelettique conserve un niveau suffisant de détails.

5.6.4 Analyse de la leçon liée aux objets

Les résultats du tableau 5.3 montrent que la distillation liée aux objets apporte un gain de performance sur l'ensemble des topologies, confirmant l'importance du contexte outil en complément des seuls signaux de mouvement. L'amélioration est particulièrement marquée pour les squelettes de 72, 21 et 53 nœuds, où l'information d'interaction humain-outil enrichit efficacement la représentation des mouvements squelettiques. Même pour la configuration la

TABLEAU 5.3 – Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 entre le modèle de base (sans distillation) et le modèle étudiant (avec distillation liée aux objets).

Modèle	**Mode**	**Transfert KD**	**MoF (%)**
Squelette à 72 nœuds			
Enseignant	Offline + Object	-	68.55
Base (sans distillation)	Online	-	48.55
Étudiant	**Online**	**Temporel + Object**	**56.07 (+7.52)**
Squelette à 53 nœuds			
Enseignant	Offline + Object	-	66.88
Base (sans distillation)	Online	-	45.72
Étudiant	**Online**	**Temporel + Object**	**58.07 (+12.35)**
Squelette à 21 nœuds			
Enseignant	Offline + Object	-	63.96
Base (sans distillation)	Online	-	43.99
Étudiant	**Online**	**Temporel + Object**	**53.65 (+9.66)**
Squelette à 15 nœuds			
Enseignant	Offline + Object	-	64.34
Base (sans distillation)	Online	-	55.22
Étudiant	**Online**	**Temporel + Object**	**56.12 (+0.90)**

plus compacte à 15 nœuds, un bénéfice mesurable subsiste, ce qui souligne la robustesse de cette leçon pour renforcer la détection d'actions en ligne.

Un point à noter en comparant le modèle enseignant (offline + objet) avec le modèle enseignant offline présenté précédemment dans le tableau 5.2. Bien que l'ajout du contexte outil entraîne une légère baisse du MoF, cette dégradation provient surtout du fait que cette métrique ne comptabilise que les frames correctement prédits, sans prendre en compte la cohérence temporelle des segments.

Comme l'illustre la Figure 5.9, l'enseignant avec le contexte "objet" génère des prédictions plus cohérentes sur le plan fonctionnel (intervalles "2 et 3"), mais introduit davantage de micro-erreurs locales, qui affectent fortement le MoF (intervalle "1"). À l'inverse, l'enseignant de base produit des segments plus continus, même lorsqu'ils sont incorrects (intervalles "1" et "3"), ce qui limite ces petites fluctuations et conduit à un MoF plus élevé. Ainsi, la baisse

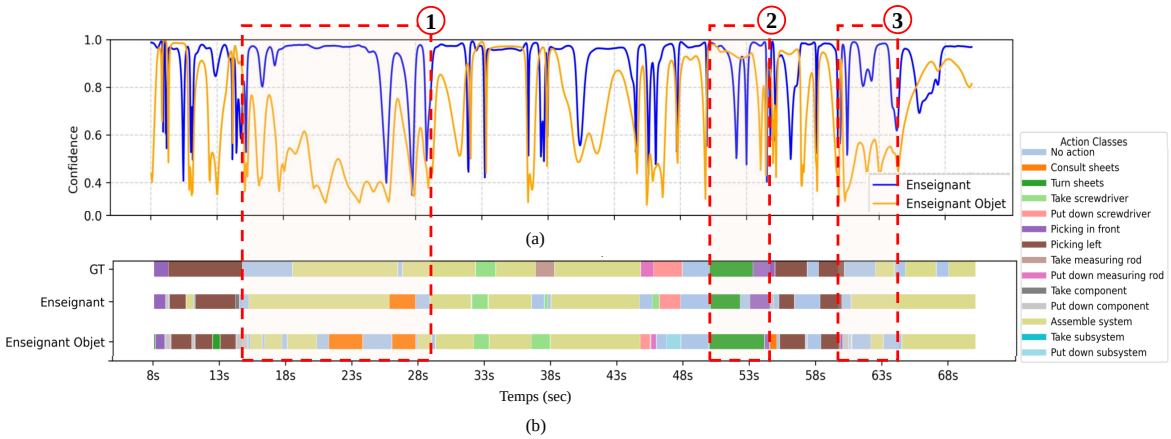


FIGURE 5.9 – Prédictions des modèles hors ligne sur l'intervalle [8 sec - 70 sec] de la séquence (P13-R02) du jeu de données InHARD. (a) Confiance des modèles au fil du temps. (b) Segments temporels prédits (chaque couleur correspond à une action). Enseignant : modèle hors ligne (contexte offline), Enseignant Objet : modèle hors ligne (contexte offline + objet), GT : vérité du terrain.

observée ne reflète pas une réelle perte de qualité, mais la sensibilité du MoF aux erreurs locales, alors que l'information d'objet demeure bénéfique pour la distillation et contribue aux gains obtenus par le modèle étudiant.

5.6.5 Analyse de la leçon liée aux occultations

TABLEAU 5.4 – Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 entre le modèle de base (sans distillation) et le modèle étudiant (avec distillation liée aux occultations : fb (full-body) → up (upper-body)).

Modèle	Mode	Transfert KD	MoF (%)
Squelette : 72→53 nœuds			
Enseignant	Offline (fb)	-	73.86
Base (sans distillation)	Online (up)	-	45.72
Étudiant	**Online (up)**	**Temporel (fb)**	**55.70 (+9.98)**
Squelette : 21→15 nœuds			
Enseignant	Offline (fb)	-	67.48
Base (sans distillation)	Online (up)	-	55.22
Étudiant	**Online (up)**	**Temporel (fb)**	**55.67 (+0.45)**

Les résultats du tableau 5.4 montrent que la distillation inter-topologique liée à l'occultation, où l'enseignant exploite un squelette complet et l'étudiant une version occultée (limitée au haut du corps), apporte un gain mesurable sur les deux topologies évaluées. L'amélioration est particulièrement marquée pour la configuration "72→53 nœuds", ce qui indique que l'étudiant bénéficie largement des relations spatiales globales apprises par l'enseignant. Le gain reste plus modeste pour la configuration "21→15 nœuds", en raison de la forte réduction de la richesse du squelette, mais dépasse néanmoins le modèle de base (sans distillation), confirmant que l'accès indirect au squelette complet demeure utile même dans une topologie très compacte. Ces résultats valident l'intérêt de transférer les connaissances issues du squelette complet pour renforcer la détection d'actions en ligne lorsque le modèle étudiant opère sur des entrées occultées.

5.6.6 Analyse des leçons cumulées

Tableau 5.5 – Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 à 53 articulations entre le modèle de base (sans distillation) et les modèles étudiants (avec distillation cumulée : Leçons combinées vs progressives).

Modèle étudiant	**Connaissances étudiant**	**Connaissances enseignant**	**MoF (%)**
Base (sans distillation)	53	-	45.72
Leçons combinées			
Étudiant C1	53	Temporel (72)	55.70 (+9.98)
Étudiant C2	**53**	**Temporel + objet (53)**	**58.07 (+12.35)**
Étudiant C3	53	Temporel + objet (72)	50.78 (+5.06)
Leçons progressives			
Étudiant P1	**(53)→Temporal (53)**	**Temporel + objet (53)**	**58.31 (+12.59)**
Étudiant P2	(53)→Temporel (72)	Temporel + objet (72)	56.22 (+ 10.50)

Les résultats présentés dans le tableau 5.7 évaluent l'impact de la distillation cumulée sur la topologie à 53 articulations. Ce choix, plutôt que celui du squelette à 15 articulations, repose sur les observations précédentes montrant que le modèle à 53 nœuds offrait les gains les plus constants et les plus importants, en faisant une base pertinente pour étudier des stratégies de distillation plus avancées. Les colonnes *Connaissances étudiant* et *Connaissances enseignant* précisent respectivement les données utilisées pour entraîner le modèle étudiant

TABLEAU 5.6 – Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 à 53 articulations entre le modèle de base (sans distillation) et les modèles étudiants (avec distillation cumulée : Leçons combinées vs progressives).

Modèle étudiant	**Connaissances de base**	**Connaissances Enseignant 1**	**Connaissances Enseignant 2**	**MoF (%)**
Base (sans distillation)	53	-		45.72
Étudiant S1	53	Temporel (53)	-	**47.81 (+2.09)**
Leçons combinées				
Étudiant C1	53	Temporel (72)	-	55.70 (+9.98)
Étudiant C2	**53**	**Temporel + objet (53)**	-	**58.07 (+12.35)**
Étudiant C3	53	**Temporel + objet (72)**	-	50.78 (+5.06)
Leçons progressives				
Étudiant P1	53	Temporel (53)	**Temporel + objet (53)**	**58.31 (+12.59)**
Étudiant P2	53	Temporel (72)	**Temporel + objet (72)**	56.22 (+ 10.50)
Modèle Enseignant	Temporel + objet (53)	-	-	68.55

et celles fournies par le modèle enseignant lors de la distillation. D'autres configurations progressives auraient pu être explorées en variant l'ordre des leçons ou le nombre de modèles enseignants impliqués ; toutefois, par manque de temps, nous nous sommes concentrés sur les configurations proposées (notamment les modèles Étudiant P1 et Étudiant P2), qui servent ici de preuves de concept.

Dans la configuration à leçons combinées, les trois modèles étudiants (C1, C2 et C3) améliorent le modèle de base (sans distillation), mais avec des effets distincts. **Étudiant C2** obtient la meilleure performance de ce groupe en intégrant simultanément les leçons temporelles et d'objet issues d'un squelette identique à 53 nœuds (leçon d'objet). **Étudiant C1**, qui bénéficie uniquement de la leçon temporelle mais issue cette fois du squelette complet à 72 nœuds (leçon d'occultation), présente également une amélioration notable. **Étudiant C3**, qui combine les trois leçons temporelles, d'objet et d'occultation issues du squelette complet à 72 nœuds, obtient le plus faible gain dans ce groupe. Cette observation suggère qu'une combinaison dense de connaissances hétérogènes, appliquée en une seule étape, ne garantit

pas toujours une meilleure assimilation.

La stratégie progressive offre les meilleurs résultats et se démarque nettement par sa cohérence. **Étudiant P1** réalise la meilleure performance globale en apprenant d'abord des signaux temporels simples (issus de la même topologie) avant d'intégrer la leçon "objet" (temporel+objet). **Étudiant P2** confirme cette tendance, même lorsqu'une partie de l'apprentissage progressif est réalisée avec le squelette complet à 72 articulations (topologie différente du modèle étudiant à 53 nœuds). Dans les deux cas, l'apprentissage progressif, allant d'une supervision simple à une supervision plus riche, se révèle plus efficace que d'imposer immédiatement un ensemble complexe de leçons. Ces observations soulignent que, au-delà du choix de la topologie, la manière d'organiser la distillation des connaissances joue un rôle déterminant dans l'efficacité du modèle étudiant.

5.6.7 Discussion et visualisation

Pour mieux comprendre l'apport réel de chaque leçon distillée et la manière dont elle influence la détection d'actions en ligne, nous analysons dans cette section les performances des modèles en ligne présents dans le tableau 5.7. Cette discussion permet non seulement de comparer l'efficacité des différentes leçons (temporelle, occultation, objet), mais aussi d'évaluer leurs combinaisons directes ou progressives afin d'identifier les stratégies les plus adaptées au contexte en ligne. La figure 5.10 associée illustre les séquences de sortie prédites (timelines) de chaque modèle en ligne, ainsi que l'évolution de leur confiance dans le temps au cours d'une activité d'assemblage du jeu de données InHARD. Cela permet de déterminer précisément comment chaque type de connaissance modifie le comportement du modèle étudiant en ligne, ce qui nous permet d'interpréter plus finement les gains observés et d'esquisser des pistes d'amélioration. Les modèles en ligne choisis sont les suivants :

— Base (sans distillation)
— Étudiant S1 (distillation temporel)
— Étudiant C1 (distillation temporel + occlusion)
— Étudiant C2 (distillation temporel + objet)
— Étudiant P1 (distillation progressive temporel puis objet)
— Étudiant P2 (distillation progressive occultation puis temporel puis objet)

La vérité de terrain (GT) constitue un repère de comparaison (séquence d'actions labellisée). Les segments sont colorés selon les classes d'action présentes dans le jeu de données InHARD,

ce qui permet d'observer visuellement la succession réelle des actions.

TABLEAU 5.7 – Comparaison des performances (MoF) de détection d'actions en ligne sur InHARD-13 à 53 articulations entre le modèle de base (sans distillation) et les modèles étudiants (avec distillation cumulée : Leçons combinées vs progressives).

Modèle étudiant	**Connaissances étudiant**	**Connaissances enseignant**	**MoF (%)**
Base (sans distillation)	53	-	45.72
Étudiant S1	53	Temporel (53)	47.81 (+2.09)
Étudiant C1	72	Temporel (53)	55.70 (+9.98)
Étudiant C2	53	Temporel + objet (53)	58.07 (+12.35)
Étudiant P1	**(53)→Temporal (53)**	**Temporel + objet (53)**	**58.31 (+12.59)**
Étudiant P2	(53)→Temporel (72)	Temporel + objet (72)	56.22 (+ 10.50)

Comme l'illustre la figure 5.10-b-Base, on constate que le modèle de base (sans distillation) s'éloigne nettement du déroulement réel (GT). Dans les intervalles "1, 2, 3 , 4 et 5", où le modèle n'arrive pas à détecter la bonne action malgré un taux de confiance élevé. La même instabilité apparaît dans l'intervalle "4", où de nombreux micro-segments incorrects traduisent une difficulté à détecter clairement le début et la fin d'une action.

La **leçon temporelle**, bien qu'efficace pour stabiliser la prédiction, n'apporte qu'un gain limité (+2.09 %). Ce résultat s'explique par la nature même du cadre en ligne : le modèle Étudiant S1 dispose déjà d'un tampon temporel de 8 secondes qui lui fournit un contexte local suffisant pour capturer la dynamique de la plupart des actions. La distillation temporelle seule n'apporte donc qu'un complément modéré. Cela est illustré dans la figure 5.10-b-S1, où l'on observe que les segments de certaines classes d'action deviennent plus continus et mieux alignés avec la vérité de terrain que celle du modèle de base (sans distillation), notamment dans les intervalles "1, 2 et 6". Cependant, une sur-segmentation est évidente dans l'intervalle "4".

La **leçon liée à l'occultation** (Étudiant C1), qui transfère des connaissances temporelles depuis un squelette complet vers un modèle étudiant avec un squelette occulté (limité au haut du corps), montre une amélioration notable (+9,98 %). Cette stratégie permet à l'étudiant d'hériter indirectement de relations spatiales globales que sa propre topologie ne peut pas représenter. Cela est visible en figure 5.10-b-C1, où la détection se stabilise mieux dans les intervalles "3 et 4".

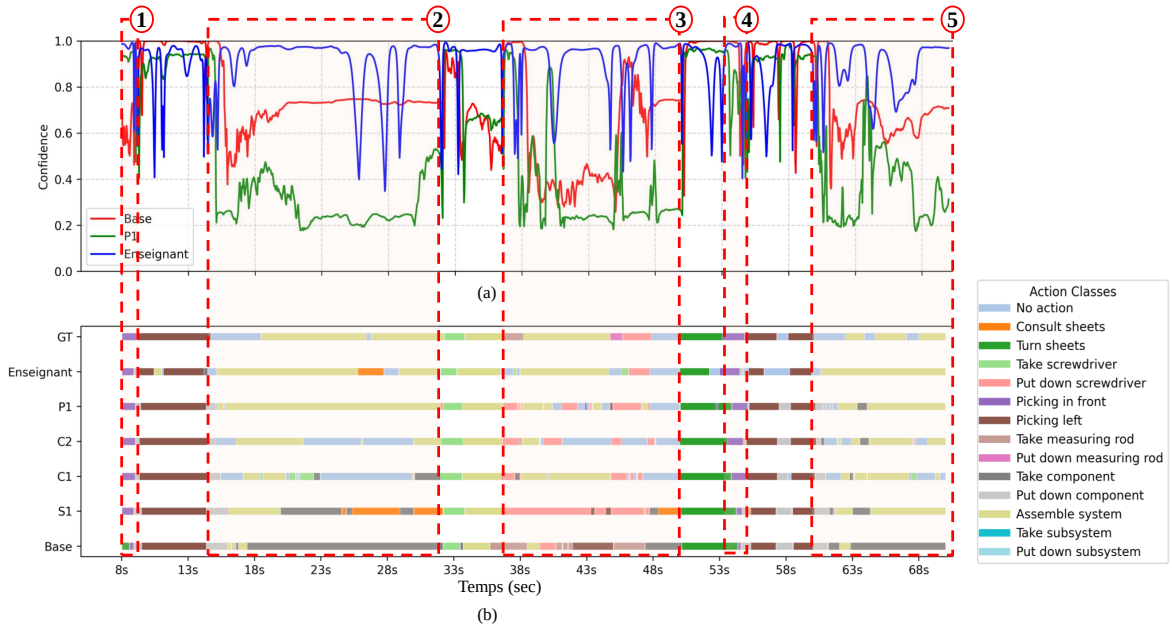


FIGURE 5.10 – Prédictions des approches en ligne utilisant différentes configurations de distillation à partir du modèle Enseignant sur l'intervalle [8 sec - 70 sec] de la séquence (P13-R02) du jeu de données InHARD. (a) Confiance des modèles au fil du temps. (b) Segments temporels prédits (chaque couleur correspond à une action). Base : modèle de base (sans distillation), S1 : étudiant avec distillation temporelle, C1 : étudiant avec distillation par occultation, C2 : étudiant avec distillation d'objet, P1 : étudiant avec distillation objet-temporelle progressive, Enseignant : modèle hors ligne, GT : vérité du terrain.

La **leçon liée aux objets** se distingue nettement par son impact. L'Étudiant C2 atteint 58.07 % (+12.35 points). Ce gain élevé s'explique par la nature des actions du jeu de données InHARD et, plus globalement, des jeux de données industriels, où la manipulation d'un outil constitue souvent l'information la plus discriminante. Lorsque deux actions présentent des mouvements articulaires similaires, c'est l'usage de l'objet qui permet de lever l'ambiguïté. La distillation de cette information fournit ainsi au modèle étudiant un signal que les seules coordonnées squelettiques ne suffisent pas à capturer. Son efficacité est supérieure à celle de la leçon d'occultation, ce qui suggère que le contexte outil a un effet plus structurant que la seule information posturale complète dans le cadre de tâches d'assemblage. Dans la figure 5.10-b-C2, on observe que le modèle Étudiant C2 parvient à détecter correctement l'action dans l'intervalle "4". La détection des actions "Assemble system" et "Picking in front" apparaît plus stable, ce qui témoigne d'une meilleure prise en compte du contexte lié aux objets manipulés.

Enfin, les **leçons cumulées** révèlent une tendance forte : l'apprentissage progressif surpasse

systématiquement la combinaison directe de leçons. L'Étudiant P1, entraîné d'abord sur la dynamique temporelle puis sur la combinaison temporel+objet, atteint la meilleure performance en ligne 58,31% (+12,59 points). Cela montre qu'un séquencement des connaissances, du plus simple au plus riche, facilite l'assimilation et évite de submerger l'étudiant par un signal trop complexe dès le départ. Cette observation rejoint les principes du curriculum learning, où l'ordre d'introduction des connaissances joue un rôle déterminant dans la stabilité et l'efficacité de l'apprentissage. Comme illustré en figure 5.10-b-P1. La sur-segmentation observée dans l'intervalle "2" disparaît totalement. Le modèle en ligne maintient une détection plus précise et plus cohérente, aussi bien pour les actions de "Picking in front" que pour l'action "Assemble system". Grâce à la combinaison graduelle des leçons, il exploite mieux le contexte objet et produit une segmentation nettement plus stable et plus proche de la vérité du terrain (GT).

Cependant, on observe que le modèle Étudiant P1, qui cumule progressivement les leçons "temporelle" et "objet", obtient de meilleures performances que l'Étudiant P2, qui intègre successivement les leçons "occultation", "temporelle" et "objet". Cette différence semble moins liée à l'occultation elle-même qu'à l'hétérogénéité des connaissances transférées. Le modèle Étudiant P2 reçoit trois signaux de nature différente, dont la fusion séquentielle peut déstabiliser son apprentissage. Pour exploiter efficacement ce triplet de leçons, des mécanismes de fusion plus adaptés, tels que les approches Mixture of Experts, où chaque type de connaissance serait activé ou pondéré en fonction de sa pertinence contextuelle, pourraient constituer une alternative prometteuse [208].

D'autre part, on constate sur la figure 5.10-a que le modèle hors ligne (Enseignant), bien qu'étant le plus performant, est également celui qui affiche la confiance la plus élevée, même lorsqu'il commet des erreurs. Le modèle en ligne de base (sans distillation) reproduit partiellement ce comportement car il montre lui aussi une confiance importante, notamment lors de ses erreurs dans les intervalles "2" et "3", mais avec une performance nettement dégradée. À l'inverse, la variante P1 (distillation progressive) présente une confiance plus modérée que le modèle de base, mais cette confiance augmente nettement lorsqu'elle produit des prédictions correctes comme dans les intervalles "1, 4, 5". Un zoom de l'intervalle 3 est présenté en figure 5.11 et permet de visualiser plus précisément ces dynamiques de confiance. On y observe clairement que l'étudiant P1 reste principalement confiant lorsqu'il détecte correctement l'action (intervalles "1" : "Take screw driver" et "3" : "Turn sheet"), alors qu'il adopte une confiance

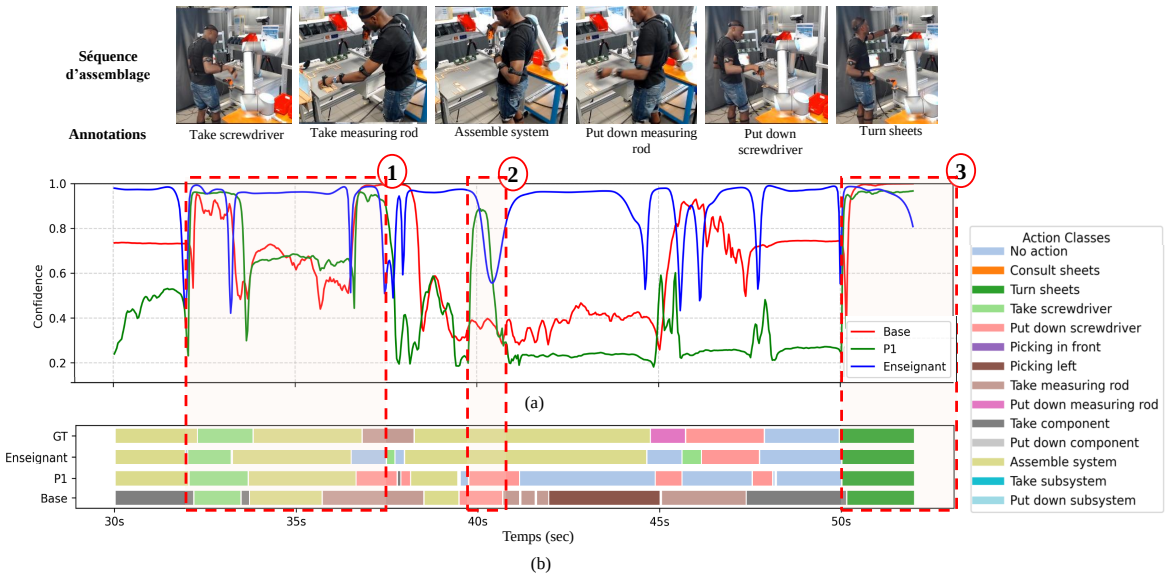


FIGURE 5.11 – Prédictions des approches de détection d'action en ligne sur l'intervalle [8 sec - 70 sec] de la séquence (P13-R02) du jeu de données InHARD. (a) Confiance des modèles au fil du temps. (b) Segments temporels prédits (chaque couleur correspond à une action). Base : modèle en ligne de base (sans distillation). P1 : modèle étudiant en ligne avec distillation, objet-temporelle progressive. Enseignant : modèle hors ligne. GT : vérité du terrain.

faible dès que la prédiction tend à être erronée, ce qui traduit un comportement plus fiable et mieux calibré. Ce comportement contraste avec celui du modèle hors ligne et du modèle de base, dont la confiance demeure élevée même en cas de prédictions erronées.

Cette visualisation détaillée met également en évidence une limite non négligeable du jeu de données lui-même. Certaines erreurs apparentes du modèle proviennent en réalité d'ambiguïtés ou d'imprécisions dans l'annotation. Par exemple, entre 38 et 45 secondes, l'action est annotée comme "Assemble system", une classe englobant l'utilisation d'outils tels que le "screwdriver" ou la "measuring rod". Cette ambiguïté peut pousser le modèle à prédire à tort des actions liées à l'outil (modèle P1 : intervalle "2" : "Put down screwdriver"), ce qui entraîne une sur-segmentation visible dans cette zone de la séquence. De plus, à certains moments ([42 sec - 45 sec]), l'opérateur reste quasiment immobile pour maintenir des pièces lors de l'assemblage ; cette absence de mouvement peut conduire le modèle à prédire "No Action". Ces observations soulignent que, au-delà des limites du modèle, une amélioration de la qualité et de la granularité des annotations pourrait réduire significativement les erreurs, notamment lors des phases de transition d'une action à la suivante.

5.7 Conclusion

Dans ce chapitre, nous avons abordé la problématique de la détection d'actions en ligne dans un contexte industriel exigeant. Pour répondre aux limites imposées par ce cadre, nous avons exploré l'usage de la distillation des connaissances et proposé des leçons spécifiquement adaptées à ces environnements : la distillation temporelle, la distillation objet et la distillation par occultation.

Nous avons également étudié différentes stratégies de distillation portant à la fois sur les logits et sur les représentations intermédiaires, et introduit une contribution centrale : une distillation progressive, déclinée à la fois dans les méthodes de transfert et dans les leçons elles-mêmes. Plusieurs configurations ont été expérimentées et comparées afin d'évaluer l'impact de chaque stratégie, ce qui nous a permis de valider l'intérêt d'une distillation progressive pour améliorer les performances en ligne et renforcer la robustesse du modèle. L'évaluation réalisée sur le jeu de données InHARD a mis en évidence les avantages de notre approche, tout en révélant certaines limites persistantes, notamment celles liées à l'ambiguïté ou à l'imprécision de certaines annotations, qui influencent directement la détection d'actions en ligne.

Enfin, ces travaux ouvrent plusieurs perspectives. La première concerne l'intégration d'une détection explicite des objets, plutôt que de s'appuyer uniquement sur des annotations disponibles. La localisation des outils à l'aide de boîtes englobantes offrirait un signal visuel plus riche, permettant de mieux caractériser les interactions humain–objet, essentielles dans les tâches d'assemblage. Bien que prometteuse, cette approche impliquerait un protocole d'annotation image par image très coûteux, ce qui dépasse le cadre de ce travail.

Les différentes leçons proposées (temporelle, occultation et objet) se sont révélées complémentaires, notamment grâce à la stratégie de distillation progressive mise en œuvre. D'autres configurations auraient pu être explorées, en variant l'ordre de distillation des leçons ou le nombre de modèles enseignants. Une perspective intéressante serait d'étendre cette idée à un modèle de type Mixture of Experts (MoE), capable d'exploiter cette complémentarité en sélectionnant dynamiquement, en fonction de l'incertitude du modèle, la stratégie de distillation la plus pertinente, ouvrant ainsi la voie à une approche adaptative et plus efficace.

Enfin, une piste importante concerne une distillation multi-niveaux entre GEAD (hors ligne) et GSK-C2F (en ligne). Grâce à son architecture profonde, basée sur les graphes, et à sa puissance discriminative grâce au module d'attention, GEAD constitue une source

particulièrement riche en signaux temporels et structurels. Transférer ces connaissances vers GSK-C2F, un modèle léger, permettrait d'améliorer les capacités de prédiction en temps réel tout en respectant les contraintes de l'inférence en ligne.

Conclusion générale

Cette thèse s'inscrit dans le contexte de l'industrie 5.0, où la place centrale de l'humain dans les systèmes de production impose de nouvelles exigences en matière de perception, d'assistance et de collaboration avec les agents robotiques. Comprendre automatiquement les actions humaines devient alors un enjeu essentiel, tant pour améliorer l'ergonomie des postes de travail que pour renforcer la sécurité, l'efficacité opérationnelle et la mise à jour continue des jumeaux numériques. Toutefois, les environnements industriels présentent des défis majeurs : gestes hétérogènes, interactions fréquentes avec des outils, grande variabilité entre opérateurs, séquences longues et déséquilibrées, occultations, ainsi que la nécessité de fournir des prédictions fiables en temps réel. Dans ce cadre exigeant, la reconnaissance d'actions demeure un problème complexe, nécessitant des modèles capables de modéliser finement les dynamiques spatiales et temporelles tout en restant interprétables et adaptés aux contraintes du terrain.

Afin d'aborder ces défis, le travail de thèse a débuté par une étude approfondie de l'état de l'art. Cette revue a couvert les principaux domaines constitutifs de l'analyse du mouvement humain en milieu industriel : les technologies d'acquisition utilisées pour enregistrer le mouvement, les jeux de données générés par ces dispositifs, ainsi que les approches d'apprentissage profond dédiées à la reconnaissance et la segmetation d'actions ainsi que leurs métriques d'évaluation. Cette étape a permis d'établir une base solide pour situer les verrous scientifiques actuels et définir les axes de recherche de la thèse.

5.8 Résumé des contributions

La première contribution a reposé sur l'exploitation de la modalité squelettique, choisie pour sa précision, sa compacité et sa robustesse dans des environnements complexes. Nous avons proposé une approche coarse-to-fine capable de modéliser les dynamiques temporelles à plusieurs échelles, améliorant nettement la détection des actions de courte durée et dépassant les performances des méthodes existantes. Cette méthodologie a ensuite été étendue par l'intégration d'une représentation en graphe, préservant la topologie du squelette tout au long du modèle. Couplée à un mécanisme d'attention, cette architecture a permis d'obtenir une représentation des mouvements plus expressive, tout en offrant une meilleure interprétabilité des décisions du modèle.

Les approches développées ont été évaluées sur plusieurs jeux de données industriels, no-

tamment InHARD, IKEA-ASM et HA4M. Les résultats obtenus sur InHARD ont dépassé l'état de l'art, tandis que ceux obtenus sur IKEA-ASM et HA4M se sont révélés très prometteurs, confirmant la pertinence des contributions proposées et leur capacité à généraliser à différents contextes industriels.

Enfin, une troisième contribution a porté sur la détection d'actions en ligne, indispensable aux applications industrielles en temps réel. Nous avons proposé une stratégie de distillation progressive de connaissances, articulée autour de leçons adaptées aux contextes industriels : temporel, d'occultation et d'objet. Ce transfert de connaissances d'un modèle hors ligne vers un modèle étudiant léger a permis d'améliorer la stabilité temporelle, de renforcer la robustesse face aux pertes de données et de mieux distinguer les actions impliquant des outils lors des tâches d'assemblage.

5.9 Défis et perspectives futures

Au-delà des contributions méthodologiques et expérimentales développées dans ce mémoire, plusieurs voies de recherche et des perspectives applicatives peuvent être explorées pour prolonger et renforcer ce travail. Ces perspectives se déclinent en deux catégories : scientifiques, visant à améliorer les fondements techniques de la reconnaissance d'actions, et applicatives, orientées vers l'intégration concrète de ces approches dans des environnements industriels réels.

5.9.1 perspectives scientifiques

Adaptation aux évolutions de processus industriels : Un enjeu essentiel dans les environnements industriels est la capacité d'un modèle à s'adapter lorsque le processus évolue par l'introduction de nouvelles actions, la modification de la séquence de montage, l'ajout de nouveaux outils ou le changement de stratégie opératoire. Une perspective majeure consiste donc à étudier des mécanismes d'apprentissage incrémental, permettant d'intégrer de nouvelles connaissances sans nécessiter un réapprentissage complet depuis le début. Des approches telles que la distillation de connaissances étudiée dans ce travail, ou le one-shot et le zero-shot learning pourraient offrir des solutions prometteuses.

Distillation entre architectures différentes : Une autre perspective consiste à explorer plus en profondeur la distillation entre modèles hétérogènes, en tirant parti de la complémentarité

entre des architectures complexes et des modèles légers. Par exemple, l'utilisation d'un modèle enseignant plus performant et expressif, tel que GEAD présenté en chapitre 4, pourrait fournir un signal de supervision particulièrement riche, tandis qu'un modèle étudiant plus rapide comme GSK-C2F présenté en chapitre 3 serait chargé d'assurer l'inférence en temps réel. L'étude de ce transfert de connaissances entre différentes architectures pourrait ouvrir la voie à des systèmes à la fois performants et opérationnels dans des environnements complexes.

Optimisation et réduction des coûts de calcul : Les systèmes en ligne doivent fonctionner en temps réel, parfois sur des plateformes embarquées limitées en ressources. Plusieurs pistes peuvent être envisagées : adaptation dynamique de la taille de la fenêtre temporelle, filtrage intelligent des frames non pertinents, avec des mécanismes d'attention sélective, pour concentrer les ressources de calcul sur les moments critiques. Ces optimisations contribueraient à des systèmes plus légers, plus rapides et plus économes en énergie.

5.9.2 perspectives applicatives

Génération de données et mise à jour de jumeaux numériques : La détection d'actions en ligne permettrait d'alimenter automatiquement un jumeau numérique avec des informations précises sur l'état de la tâche, les outils utilisés, les transitions ou les phases d'assemblage. L'utilisation de jumeaux numériques interfacés en XR pour, par exemple, des opérations de formations sur les opérations pourrait servir à créer de nouveaux jeux de données à partir des interactions dans ces environnements virtuels, permettant l'enrichissement continu des modèles.

Intégration de démonstrateur ergonomique en temps réel : Les informations extraites par les modèles (postures, gestes répétitifs, transitions...) pourraient être exploitées pour fournir un retour immédiat à l'opérateur ou aux ergonomes. Un tel système constituerait un outil précieux pour analyser objectivement la charge physique, détecter les situations à risque et proposer des recommandations d'amélioration des postes de travail.

Vers une collaboration humain–robot plus fluide : Le système développé ouvre la voie à des applications de cobotique, où un robot collaboratif adapte son comportement en fonction de l'activité humaine en cours. Une piste importante consiste à connecter la détection d'actions en ligne à des modules décisionnels ou de planification afin de permettre au robot d'ajuster son assistance (vitesse, anticipation, force, positionnement...) en fonction de l'état de la tâche et de l'intention de l'opérateur.

Ouverture vers d'autres domaines et activités : Au-delà de ses applications industrielles, ce travail ouvre des perspectives plus larges pour l'analyse et la compréhension des actions humaines dans divers contextes. Les approches développées au cours de cette thèse pourraient contribuer à des avancées dans des domaines tels que la logistique ou la maintenance, mais aussi dans des secteurs non industriels tels que le domaine médical (analyse des gestes cliniques, suivi de la rééducation), le sport (évaluation de la performance, prévention des blessures), ou encore la formation professionnelle et technique. Ces extensions témoignent de la portée plus large de ce travail, qui dépasse le cadre strict de l'industrie et peut servir de base à des systèmes intelligents capables d'interpréter l'activité humaine dans des environnements variés.

Bibliographie

[1] J. Alves, T. M. Lima et P. D. Gaspar, "Is Industry 5.0 a Human-Centred Approach ? A Systematic Review," *Processes*, t. 11, n° 1, p. 193, 2023. doi : doi.org/10.3390/pr11010193.

[2] M. G. Morshed, T. Sultana, A. Alam et Y.-K. Lee, "Human Action Recognition : A Taxonomy-Based Survey, Updates, and Opportunities," *Sensors*, t. 23, n° 4, p. 2182, 2023. doi : 10.3390/s23042182.

[3] M. Dallel, V. Havard, D. Baudry et X. Savatier, "Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics," p. 1-6, 2020. doi : 10.1109/ICHMS49158.2020.9209531.

[4] Y. Ben-Shabat, X. Yu, F. Saleh et al., "The ikea asm dataset : Understanding people assembling furniture through actions, objects and pose," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, p. 847-859. doi : 10.1109/WACV48630.2021.00089.

[5] P. Zheng, Y. Sang et J. Wang, "Industrial applications of the Internet of Things : Challenges, approaches, and opportunities," *Engineering Applications of Artificial Intelligence*, t. 76, p. 25-37, 2018. doi : 10.1016/j.procs.2022.12.282.

[6] P. K. R. Maddikunta et al., "Industry 5.0 : A survey on enabling technologies and potential applications," *Journal of Industrial Information Integration*, t. 26, p. 100 257, 2022. doi : 10.1016/j.jii.2021.100257.

[7] H. Bagheri et al., "Resilient human-centered cyber-physical systems : Concept, design and implementation," *Computers in Industry*, t. 131, p. 103 503, 2021. doi : 10.1007/s10846-018-0881-x.

[8] V. Havard, B. Jeanne, M. Lacomblez et D. Baudry, "Digital twin and virtual reality : a co-simulation environment for design and assessment of industrial workstations," *Production & Manufacturing Research*, t. 7, n° 1, p. 472-489, 2019. doi : 10.1080/21693277.2019.1660283.

[9] V. Havard, D. Baudry, B. Jeanne, A. Louis et X. Savatier, "A use case study comparing augmented reality (AR) and electronic document-based maintenance instructions considering tasks complexity and operator competency level," *Virtual Reality*, t. 25, n° 4, p. 999-1014, 2021. doi : 10.1007/s10055-020-00493-z.

[10] Y. Feddoul, N. Ragot, F. Duval, V. Havard, D. Baudry et A. Assila, "Exploring human-machine collaboration in industry : A systematic literature review of digital twin and robotics interfaced with extended reality technologies," *The international journal of advanced manufacturing technology*, t. 129, n° 5, p. 1917-1932, 2023. doi : 10.1007/s00170-023-12291-3.

[11] T. Benmessabih, R. Slama, V. Havard et D. Baudry, "Online human motion analysis in industrial context : A review," *Engineering Applications of Artificial Intelligence*, t. 131, p. 107 850, 2024. doi : 10.1016/j.engappai.2024.107850.

[12] R. Poppe, "Vision-based human motion analysis : An overview," *Computer Vision and Image Understanding*, t. 108, n° 1-2, p. 4-18, 2007. doi : 10.1016/j.cviu.2006.10.016.

[13] M. Menolotto, D.-S. Komaris, S. Tedesco, B. O'Flynn et M. Walsh, "Motion capture technology in industrial applications : A systematic review," *Sensors*, t. 20, n° 19, p. 5687, 2020. doi : 10.3390/s20195687.

[14] M. Rana, A. Rahman et D. Smith, "A Semi-supervised Approach for Activity Recognition from Indoor Trajectory Data," *arXiv preprint arXiv :2301.03134*, 2023. doi : 10.48550/arXiv.2301.03134.

[15] F. Sener, D. Chatterjee, D. Shelepov et al., "Assembly101 : A large-scale multi-view video dataset for understanding procedural activities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 21 096-21 106. doi : 10.1109/CVPR52688.2022.02042.

[16] A. Singh, B. Choudhury, S. Roy et al., "A comprehensive review on vision based human activity recognition in industry 4.0 and beyond," *Journal of Manufacturing Systems*, t. 59, p. 513-535, 2021. doi : 10.1007/s11042-020-09004-3.

[17] M. Bortolini, M. Faccio, M. Gamberi et F. Pilati, "Motion Analysis System (MAS) for production and ergonomics assessment in the manufacturing processes," *Computers & Industrial Engineering*, t. 139, p. 105 485, 2020. doi : 10.1016/j.cie.2018.10.046.

[18] M. Al-Amin, W. Tao, D. Doell et al., "Action recognition in manufacturing assembly using multimodal sensor fusion," *Procedia Manufacturing*, t. 39, p. 158-167, 2019. doi : 10.1016/j.promfg.2020.01.288.

[19] F. Niemann, H. Bas, J. Steffens, N. Nair et M. ten Hompel, *Context-Aware Activity Recognition in Logistics (CAARL)–A optical marker-based Motion Capture Dataset*, 2021. doi : 10.3390/s22010134.

[20] D. Moutinho, L. F. Rocha, C. M. Costa, L. F. Teixeira et G. Veiga, "Deep learning-based human action recognition to leverage context awareness in collaborative assembly," *Robotics and Computer-Integrated Manufacturing*, t. 80, p. 102 449, 2023. doi : 10.1016/j.rcim.2022.102449.

[21] A. F. Mohammadi, M. Rezayati, H. W. van de Venn et H. Karimpour, "A mixed-perception approach for safe human–robot collaboration in industrial automation," *Sensors*, t. 20, n° 21, p. 6347, 2020. doi : 10.3390/s20216347.

[22] F. Vicentini, "Collaborative robotics : a survey," *Journal of Mechanical Design*, t. 143, n° 4, 2021. doi : 10.1007/s10846-024-02141-z.

[23] M. Delamare, F. Duval et R. Boutteau, "A new dataset of people flow in an industrial site with uwb and motion capture systems," *Sensors*, t. 20, n° 16, p. 4511, 2020. doi : 10.3390/s20164511.

[24] Q. Abu Al-Haija et J. Al-Saraireh, "Asymmetric identification model for human-robot contacts via supervised learning," *Symmetry*, t. 14, n° 3, p. 591, 2022. doi : 10.3390/sym14030591.

[25] W. Kwon, Y. Jin et S. J. Lee, "Uncertainty-aware knowledge distillation for collision identification of collaborative robots," *Sensors*, t. 21, n° 19, p. 6674, 2021. doi : 10.3390/s21196674.

[26] S. A. B. Birjandi, J. Kühn et S. Haddadin, "Observer-extended direct method for collision monitoring in robot manipulators using proprioception and imu sensing," *IEEE Robotics and Automation Letters*, t. 5, n° 2, p. 954-961, 2020. doi : 10.1109/LRA.2020.2967287.

[27] J. Zhang, H. Liu, Q. Chang, L. Wang et R. X. Gao, "Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly," *CIRP annals*, t. 69, n° 1, p. 9-12, 2020. doi : 10.1016/j.cirp.2020.04.077.

[28] S. Grushko, A. Vysockỳ, D. Heczko et Z. Bobovskỳ, "Intuitive spatial tactile feedback for better awareness about robot trajectory during human–robot collaboration," *Sensors*, t. 21, n° 17, p. 5748, 2021. doi : 10.3390/s21175748.

[29] F. Pastor, D.-h. Lin-Yang, J. M. Gómez-de-Gabriel et A. J. Garcia-Cerezo, "Dataset with Tactile and Kinesthetic Information from a Human Forearm and Its Application to Deep Learning," *Sensors*, t. 22, n° 22, p. 8752, 2022. doi : 10.3390/s22228752.

[30] P. Maurice, A. Malaisé, C. Amiot et al., "Human movement and ergonomics : An industry-oriented dataset for collaborative robotics," *The International Journal of Robotics Research*, t. 38, n° 14, p. 1529-1537, 2019. doi : 10.1177/0278364919882089.

[31] A. G. da Silva, M. V. Mendes Gomes et I. Winkler, "Virtual reality and digital human modeling for ergonomic assessment in industrial product development : a patent and literature review," *Applied Sciences*, t. 12, n° 3, p. 1084, 2022. doi : 10.3390/app12031084.

[32] P. Chiabert et K. Aliev, "Analyses and study of human operator monotonous tasks in small enterprises in the era of industry 4.0," in *Product Lifecycle Management Enabling Smart X : 17th IFIP WG 5.1 International Conference, PLM 2020, Rapperswil, Switzerland, July 5–8, 2020, Revised Selected Papers 17*, Springer, 2020, p. 83-97. doi : 10.1007/978-3-030-62807-9_8.

[33] A. K. Inkulu, M. R. Bahubalendruni et A. Dara, "Challenges and opportunities in human robot collaboration context of Industry 4.0-a state of the art review," *Industrial Robot : the international journal of robotics research and application*, t. 49, n° 2, p. 226-239, 2022. doi : 10.1108/IR-04-2021-0077.

[34] M. Joshi et V. Deshpande, "A systematic review of comparative studies on ergonomic assessment techniques," *International Journal of Industrial Ergonomics*, t. 74, p. 102 865, 2019. doi : 10.1016/j.ergon.2019.102865.

[35] V. Havard, B. Jeanne, M. Lacomblez et D. Baudry, "Digital twin and virtual reality : a co-simulation environment for design and assessment of industrial workstations," *Production & Manufacturing Research*, t. 7, n° 1, p. 472-489, 2019. doi : 10.1080/21693277.2019.1660283.

[36] W. Kim, L. Peternel, M. Lorenzini, J. Babič et A. Ajoudani, "A human-robot collaboration framework for improving ergonomics during dexterous operation of power tools," *Robotics and Computer-Integrated Manufacturing*, t. 68, p. 102 084, 2021. doi : 10.1016/j.rcim.2020.102084.

[37] V. M. Manghisi, A. Evangelista et A. E. Uva, "A Virtual Reality Approach for Assisting Sustainable Human-Centered Ergonomic Design : The ErgoVR tool," *Procedia Computer Science*, t. 200, p. 1338-1346, 2022. doi : 10.1016/j.procs.2022.01.335.

[38] J. Sedlar, K. Stepanova, R. Skoviera et al., "Imitrob : Imitation Learning Dataset for Training and Evaluating 6D Object Pose Estimators," *IEEE Robotics and Automation Letters*, 2023. doi : 10.48550/ARXIV.2209.07976.

[39] M. Eisenbach, D. Aganian, M. Köhler, B. Stephan, C. Schröter et H.-M. Gross, *Visual scene understanding for enabling situation-aware cobots*. Universitätsbibliothek, 2022. doi : 10.22032/dbt.51471.

[40] Y. Zhang, K. Ding, J. Hui, J. Lv, X. Zhou et P. Zheng, "Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly," *Advanced Engineering Informatics*, t. 54, p. 101 792, 2022. doi : 10.1016/j.aei.2022.101792.

[41] A. Castro, F. Silva et V. Santos, "Trends of human-robot collaboration in industry contexts : Handover, learning, and metrics," *Sensors*, t. 21, n° 12, p. 4113, 2021. doi : 10.3390/s21124113.

[42] E. Matheson, R. Minto, E. G. Zampieri, M. Faccio et G. Rosati, "Human–robot collaboration in manufacturing applications : A review," *Robotics*, t. 8, n° 4, p. 100, 2019. doi : 10.3390/robotics8040100.

[43] R. Zhang, J. Li, P. Zheng, Y. Lu, J. Bao et X. Sun, "A fusion-based spiking neural network approach for predicting collaboration request in human-robot collaboration," *Robotics and Computer-Integrated Manufacturing*, t. 78, p. 102 383, 2022. doi : 10.1016/j.rcim.2022.102383.

[44] H. Zhou, G. Yang, B. Wang et al., "An attention-based deep learning approach for inertial motion recognition and estimation in human-robot collaboration," *Journal of Manufacturing Systems*, t. 67, p. 97-110, 2023. doi : 10.1016/j.jmsy.2023.01.007.

[45] F. Tassi, F. Iodice, E. De Momi et A. Ajoudani, "Sociable and ergonomic human-robot collaboration through action recognition and augmented hierarchical quadratic programming," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022, p. 10 712-10 719. doi : 10.48550/arXiv.2207.03435.

[46] A. Toichoa Eyam, W. M. Mohammed et J. L. Martinez Lastra, "Emotion-driven analysis and control of human-robot interactions in collaborative applications," *Sensors*, t. 21, n° 14, p. 4626, 2021. DOI : 10.3390/s21144626.

[47] K. Darvish, F. Wanderlingh, B. Bruno, E. Simetti, F. Mastrogiovanni et G. Casalino, "Flexible human–robot cooperation models for assisted shop-floor tasks," *Mechatronics*, t. 51, p. 97-114, 2018. DOI : 10.1016/j.mechatronics.2018.03.006.

[48] D. Lagamtzis, F. Schmidt, J. R. Seyler et T. Dang, "CoAx : Collaborative Action Dataset for Human Motion Forecasting in an Industrial Workspace.," in *ICAART (3)*, 2022, p. 98-105. DOI : 10.5220/0010775600003116.

[49] Z. Liu, Q. Liu, W. Xu, L. Wang et Z. Ji, "Adaptive real-time similar repetitive manual procedure prediction and robotic procedure generation for human-robot collaboration," *Advanced Engineering Informatics*, t. 58, p. 102 129, 2023. DOI : 10.1016/j.aei.2023.102129.

[50] W. Xu, S. Feng, B. Yao, Z. Ji et Z. Liu, "Turn-Taking Prediction for Human–Robot Collaborative Assembly Considering Human Uncertainty," *Journal of Manufacturing Science and Engineering*, t. 145, n° 12, p. 121 007, 2023. DOI : 10.1115/1.4063231.

[51] R. Zhang, J. Lv, J. Li, J. Bao, P. Zheng et T. Peng, "A graph-based reinforcement learning-enabled approach for adaptive human-robot collaborative assembly operations," *Journal of Manufacturing Systems*, t. 63, p. 491-503, 2022. DOI : 10.1016/j.jmsy.2022.05.006.

[52] H. Zamora, Mauricio-Andrés, J. A. Castro-Vargas, J. Azorin-Lopez et J. Garcia-Rodriguez, "Deep learning-based visual control assistant for assembly in Industry 4.0," *Computers in Industry*, t. 131, p. 103 485, 2021. DOI : 10.1016/j.compind.2021.103485.

[53] C. Chen, T. Wang, D. Li et J. Hong, "Repetitive assembly action recognition based on object detection and pose estimation," *Journal of Manufacturing Systems*, t. 55, p. 325-333, 2020. DOI : 10.1016/j.jmsy.2020.04.018.

[54] J. Hernandez, G. Valarezo, R. Cobos, J. W. Kim, R. Palacios et A. G. Abad, "Hierarchical Human Action Recognition to Measure the Performance of Manual Labor," *IEEE Access*, t. 9, p. 103 110-103 119, 2021. DOI : 10.1109/ACCESS.2021.3095934.

[55] Z. Wang, R. Qin, J. Yan et C. Guo, "Vision sensor based action recognition for improving efficiency and quality under the environment of industry 4.0," *Procedia CIRP*, t. 80, p. 711-716, 2019. DOI : 10.1016/j.procir.2019.01.106.

[56] T. Schreiter, T. R. de Almeida, Y. Zhu et al., "The Magni Human Motion Dataset : Accurate, Complex, Multi-Modal, Natural, Semantically-Rich and Contextualized," *arXiv preprint arXiv :2208.14925*, 2022. DOI : 10.48550/ARXIV.2208.14925.

[57] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras et A. J. Lilienthal, "Thör : Human-robot navigation data collection and accurate motion trajectories dataset," *IEEE Robotics and Automation Letters*, t. 5, n° 2, p. 676-682, 2020. DOI : 10.1109/LRA.2020.2965416.

[58] S. Li, J. Fan, P. Zheng et L. Wang, "Transfer learning-enabled action recognition for human-robot collaborative assembly," *Procedia CIRP*, t. 104, p. 1795-1800, 2021. doi : 10.1016/j.procir.2021.11.303.

[59] A. Upadhyay, P. A. Barua, A. Dubey, S. Sengupta, S. M. Kuriakose et P. Goenka, "P3DAttnNet : Automated Assembly Plan Generation from Video Demonstration,"

[60] D. Singhania, R. Rahaman et A. Yao, "C2F-TCN : A Framework for Semi and Fully Supervised Temporal Action Segmentation," *arXiv preprint arXiv :2212.11078*, 2022. doi : 10.1109/TPAMI.2023.3284080.

[61] M. Liu, F. Meng, C. Chen et S. Wu, "Novel motion patterns matter for practical skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2023. doi : 10.1609/aaai.v37i2.25258.

[62] J. Zhang, P. Wang et R. X. Gao, "Hybrid machine learning for human action recognition and prediction in assembly," *Robotics and Computer-Integrated Manufacturing*, t. 72, p. 102 184, 2021. doi : 10.1016/j.rcim.2021.102184.

[63] F. Niemann, C. Reining, F. Moya Rueda et al., "Lara : Creating a dataset for human activity recognition in logistics using semantic attributes," *Sensors*, t. 20, n° 15, p. 4083, 2020. doi : 10.3390/s20154083.

[64] Y. Yun, P. Agarwal et A. D. Deshpande, "Accurate, robust, and real-time estimation of finger pose with a motion capture system," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2013, p. 1626-1631. doi : 10.1109/IROS.2013.6696567.

[65] G. Cicirelli, R. Marani, L. Romeo et al., "The HA4M dataset : Multi-Modal Monitoring of an assembly task for Human Action recognition in Manufacturing," *Scientific Data*, t. 9, n° 1, p. 745, 2022. doi : 10.57760/sciencedb.01872.

[66] S. L. Colyer, M. Evans, D. P. Cosker et A. I. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports medicine-open*, t. 4, n° 1, p. 1-15, 2018. doi : 10.1186/s40798-018-0139-y.

[67] F. Iodice, E. De Momi et A. Ajoudani, "Hri30 : An action recognition dataset for industrial human-robot interaction," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, p. 4941-4947. doi : 10.1109/ICPR56361.2022.9956300.

[68] P. M. S. Ribeiro, A. C. Matos, P. H. Santos et J. S. Cardoso, "Machine learning improvements to human motion tracking with imus," *Sensors*, t. 20, n° 21, p. 6383, 2020. doi : 10.3390/s20216383.

[69] C. Tamantini, F. Cordella, C. Lauretti et L. Zollo, "The WGD—A dataset of assembly line working gestures for ergonomic analysis and work-related injuries prevention," *Sensors*, t. 21, n° 22, p. 7600, 2021. doi : 10.3390/s21227600.

[70] G. Cicirelli et et al., "A deep learning approach for real-time human action recognition in collaborative robotics," *Information Fusion*, 2022.

[71] P. Maurice et et al., "Assessment of IMU-based human motion capture systems for ergonomics applications," *Applied Ergonomics*, 2019.

[72] U. Zakia et C. Menon, "Dataset on Force Myography for Human–Robot Interactions," *Data*, t. 7, n° 11, p. 154, 2022. doi : 10.3390/data7110154.

[73] A. Zakia et C. Menon, "Force myography-based hand posture recognition using low-cost sensors," *IEEE Sensors Journal*, 2022.

[74] Z. Xiao et et al., "Hand gesture recognition using force myography : A comparative study," *Biomedical Signal Processing and Control*, 2019. doi : 10.3389/felec.2024.1503424.

[75] G. Bassani, A. Filippeschi et C. A. Avizzano, "A Dataset of Human Motion and Muscular Activities in Manual Material Handling Tasks for Biomechanical and Ergonomic Analyses," *IEEE Sensors Journal*, t. 21, n° 21, p. 24 731-24 739, 2021. doi : 10.1109/JSEN.2021.3113123.

[76] S. Jiang, Q. Gao, H. Liu et P. B. Shull, "A novel, co-located EMG-FMG-sensing wearable armband for hand gesture recognition," *Sensors and Actuators A : Physical*, t. 301, p. 111 738, 2020. doi : 10.1016/j.sna.2019.111738.

[77] Z. Zheng, Z. Wu, R. Zhao, Y. Ni, X. Jing et S. Gao, "A review of EMG-, FMG-, and EIT-based biosensors and relevant human–machine interactivities and biomedical applications," *Biosensors*, t. 12, n° 7, p. 516, 2022. doi : 10.3390/bios12070516.

[78] Z. G. Xiao et C. Menon, "A review of force myography research and development," *Sensors*, t. 19, n° 20, p. 4557, 2019. doi : 10.3390/s19204557.

[79] A. Prakash, N. Sharma et S. Sharma, "An affordable transradial prosthesis based on force myography sensor," *Sensors and Actuators A : Physical*, t. 325, p. 112 699, 2021. doi : 10.1016/j.sna.2021.112699.

[80] A. Ke, J. Huang, L. Chen, Z. Gao et J. He, "An Ultra-Sensitive Modular Hybrid EMG–FMG Sensor with Floating Electrodes," *Sensors*, t. 20, n° 17, p. 4775, 2020. doi : 10.3390/s20174775.

[81] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang et J. Liu, "Human action recognition from various data modalities : A review," *IEEE transactions on pattern analysis and machine intelligence*, 2022. doi : 10.1109/TPAMI.2022.3183112.

[82] N. Yoshimura, J. Morales, T. Maekawa et T. Hara, "OpenPack : A Large-scale Dataset for Recognizing Packaging Works in IoT-enabled Logistic Environments," *arXiv preprint arXiv :2212.11152*, 2022. doi : 10.1109/PerCom59722.2024.10494448.

[83] F. Sturm, E. Hergenroether, J. ReC2Ft, P. S. Vojnovikj et M. Siegel, "Challenges of the Creation of a Dataset for Vision Based Human Hand Action Recognition in Industrial Assembly," *arXiv preprint arXiv :2303.03716*, 2023. doi : 10.48550/ARXIV.2303.03716.

[84] M. Dallel, V. Havard, Y. Dupuis et D. Baudry, "Digital twin of an industrial workstation : A novel method of an auto-labeled data generator using virtual reality for human action recognition in the context of human–robot collaboration," *Engineering applications of artificial intelligence*, t. 118, p. 105 655, 2023. doi : 10.1016/j.engappai.2022.105655.

[85] C. Munasinghe, F. M. Amin, D. Scaramuzza et H. W. van de Venn, "COVERED, CollabOratiVE Robot Environment Dataset for 3D Semantic segmentation," in *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2022, p. 1-4. DOI : 10.1109/ETFA52439.2022.9921525.

[86] F. Ragusa, A. Furnari, S. Livatino et G. M. Farinella, "The meccano dataset : Understanding human-object interactions from egocentric videos in an industrial-like domain," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, p. 1569-1578. DOI : 10.1109/WACV48630.2021.00161.

[87] M. Al-Amin et et al., "A wearable EMG-based system for movement recognition in industrial environments," *Procedia Computer Science*, 2019. DOI : 10.1109/EMBC44109.2020.9176615.

[88] F. Caba Heilbron, V. Escorcia, B. Ghanem et J. Carlos Niebles, "Activitynet : A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, p. 961-970. DOI : 10.1109/CVPR.2015.7298698.

[89] Y.-G. Jiang, J. Liu, A. Roshan Zamir et al., *THUMOS Challenge : Action Recognition with a Large Number of Classes*, http://crcv.ucf.edu/THUMOS14/, 2014.

[90] J. Koch, L. Buesch, M. Gomse et T. Schueppstuhl, "A methods-time-measurement based approach to enable action recognition for multi-variant assembly in human-robot collaboration," *Procedia CIRP*, t. 106, p. 233-238, 2022. DOI : 10.1016/j.procir.2022.02.184.

[91] A. Vysocky, S. Grushko, T. Spurny, R. Pastor et T. Kot, "Generating Synthetic Depth Image Dataset for Industrial Applications of Hand Localization," *IEEE Access*, t. 10, p. 99 734-99 744, 2022. DOI : 10.1109/ACCESS.2022.3206948.

[92] H. Nguyen Ngoc, G. Lasa et I. Iriarte, "Human-centred design in industry 4.0 : case study review and opportunities for future research," *Journal of Intelligent Manufacturing*, t. 33, n° 1, p. 35-76, 2022. DOI : 10.1007/s10845-021-01796-x.

[93] M. C. Zizic, M. Mladineo, N. Gjeldum et L. Celent, "From industry 4.0 towards industry 5.0 : A review and analysis of paradigm shift for the people, organization and technology," *Energies*, t. 15, n° 14, p. 5221, 2022. DOI : 10.3390/en15145221.

[94] G.-Y. Kim, D. Kim, S. Do Noh et al., "Human Digital Twin System for Operator Safety and Work Management," in *IFIP International Conference on Advances in Production Management Systems*, Springer, 2022, p. 529-536. DOI : 10.1007/978-3-031-16411-8_61.

[95] A. Parashar et R. S. Shekhawat, "Protection of gait data set for preserving its privacy in deep learning pipeline," *IET Biometrics*, t. 11, n° 6, p. 557-569, 2022. DOI : 10.1049/bme2.12093.

[96] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan et A. C. Kot, "Ntu rgb+ d 120 : A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, t. 42, n° 10, p. 2684-2701, 2019. DOI : 10.1109/TPAMI.2019.2916873.

[97] N. Trivedi, A. Thatipelli et R. K. Sarvadevabhatla, "Ntu-x : An enhanced large-scale dataset for improving pose-based recognition of subtle human actions," in *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, 2021, p. 1-9. doi : 10.1145/3490035.3490270.

[98] H. Ali, C. Grönlund et Z. Shah, "Leveraging GANs for data scarcity of COVID-19 : Beyond the hype," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, p. 659-667. doi : 10.1109/CVPRW59228.2023.00073.

[99] M. Ba, Z. Ji, Z. Liu, B. Yao, W. Xu et Y. Zhong, "Human Action Detection Based on Multimodal Feature Fusion for Human-Robot Collaborative Assembly," in *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2023, p. 1-6. doi : 10.1109/CASE56687.2023.10260516.

[100] G. Ding, F. Sener et A. Yao, "Temporal action segmentation : An analysis of modern techniques," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. doi : 10.1109/TPAMI.2023.3327284.

[101] H. Gammulle, D. Ahmedt-Aristizabal, S. Denman, L. Tychsen-Smith, L. Petersson et C. Fookes, "Continuous human action recognition for human-machine interaction : a review," *ACM Computing Surveys*, t. 55, n° 13s, p. 1-38, 2023. doi : 10.1145/3587931.

[102] R. Ghoddoosian, S. Sayed et V. Athitsos, "Hierarchical modeling for task recognition and action segmentation in weakly-labeled instructional videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, p. 1922-1932. doi : 10.1109/WACV51458.2022.00020.

[103] X. Liu, Q. Wang, Y. Hu et al., "End-to-End Temporal Action Detection With Transformer," *IEEE Transactions on Image Processing*, t. 31, p. 5427-5441, 2022. doi : 10.1109/TIP.2022.3195321.

[104] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li et D. Tao, "Tridet : Temporal action detection with relative boundary modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, p. 18 857-18 866. doi : 10.1109/CVPR52729.2023.01808.

[105] Z. Zhao, D. Wang et X. Zhao, "Movement enhancement toward multi-scale video feature representation for temporal action detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, p. 13 555-13 564. doi : 10.1109/ICCV51070.2023.01247.

[106] Z. Liu, L. Wang, D. Zhou et al., "Temporal segment transformer for action segmentation," *arXiv preprint arXiv :2302.13074*, 2023. doi : 10.48550/arXiv.2302.13074.

[107] A. J. Rana et Y. S. Rawat, "We don't need thousand proposals : Single shot actor-action detection in videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, p. 2960-2969. doi : 10.1109/WACV48630.2021.00300.

[108] Y. Liu, X. Zhong, S. Zhai et al., "Prompt-enhanced hierarchical transformer elevating cardiopulmonary resuscitation instruction via temporal action segmentation," *Computers in Biology and Medicine*, t. 167, p. 107 672, 2023, ISSN : 0010-4825. DOI : https://doi.org/10.1016/j.compbiomed.2023.107672. adresse : https://www.sciencedirect.com/science/article/pii/S001048252301137X.

[109] B. van Amsterdam, A. Kadkhodamohammadi, I. Luengo et D. Stoyanov, "Aspnet : Action segmentation with shared-private representation of multiple data sources," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, p. 2384-2393. DOI : 10.1109/CVPR52729.2023.00236.

[110] D. Damen, H. Doughty, G. M. Farinella et al., "Scaling Egocentric Vision : The EPIC-KITCHENS Dataset," in *European Conference on Computer Vision (ECCV)*, 2018. DOI : 10.48550/arXiv.1804.02748.

[111] D. Liu, Q. Li, A.-D. Dinh, T. Jiang, M. Shah et C. Xu, "Diffusion action segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, p. 10 139-10 149. DOI : 10.1109/TPAMI.2024.3509434.

[112] X. Chen et Y. Yu, "Automatic repetitive action counting for construction worker ergonomic assessment," *Automation in Construction*, t. 167, p. 105 726, 2024. DOI : 10.1016/j.autcon.2024.105726.

[113] S. Chand, H. Zheng et Y. Lu, "A vision-enabled fatigue-sensitive human digital twin towards human-centric human-robot collaboration," *Journal of Manufacturing Systems*, t. 77, p. 432-445, 2024. DOI : 10.1016/j.jmsy.2024.10.002.

[114] S. Gaugel et M. Reichert, "PrecTime : A deep learning architecture for precise time series segmentation in industrial manufacturing operations," *Engineering Applications of Artificial Intelligence*, t. 122, p. 106 078, 2023. DOI : 10.1016/j.engappai.2023.106078.

[115] T. Benmessabih, R. Slama, V. Havard et D. Baudry, "GSK-C2F : Graph Skeleton Modelization for Action Segmentation and Recognition Using a Coarse-to-Fine Strategy," in *European Conference on Computer Vision*, Springer, 2024, p. 169-185. DOI : 10.1007/978-3-031-92591-7_11.

[116] A. Deng, T. Yang et C. Chen, "A Large-scale Study of Spatiotemporal Representation Learning with a New Benchmark on Action Recognition," *arXiv preprint arXiv :2303.13505*, 2023. DOI : 10.1109/ICCV51070.2023.01876.

[117] P. Wang, H. Liu, L. Wang et R. X. Gao, "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration," *CIRP annals*, t. 67, n° 1, p. 17-20, 2018. DOI : 10.1016/j.cirp.2018.04.066.

[118] L. Orsag, T. Stipancic et L. Koren, "Towards a Safe Human–Robot Collaboration Using Information on Human Worker Activity," *Sensors*, t. 23, n° 3, p. 1283, 2023. DOI : doi.org/10.3390/s23031283.

[119] M. Dallel, V. Havard, Y. Dupuis et D. Baudry, "A sliding window based approach with majority voting for online human action recognition using spatial temporal graph convolutional neural networks," in *2022 7th International Conference on Machine Learning Technologies (ICMLT)*, 2022, p. 155-163. doi : doi.org/10.1145/3529399.3529425.

[120] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan et J. Liu, "Online Human Action Detection using Joint Classification-Regression Recurrent Neural Networks," *European Conference on Computer Vision*, 2016. doi : 10.1007/978-3-319-46478-7_13.

[121] C. Tang, W. Li, P. Wang et L. Wang, "Online human action recognition based on incremental learning of weighted covariance descriptors," *Information Sciences*, t. 467, p. 219-237, 2018. doi : 10.1016/j.ins.2018.08.003.

[122] J. Koch, L. Büsch, M. Gomse et T. Schüppstuhl, "A Methods-Time-Measurement based Approach to enable Action Recognition for Multi-Variant Assembly in Human-Robot Collaboration," *Procedia CIRP*, t. 106, p. 233-238, 2022. doi : 10.1016/j.procir.2022.02.184.

[123] M. M. Islam et T. Iqbal, "Hamlet : A hierarchical multimodal attention-based human activity recognition algorithm," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, p. 10 285-10 292. doi : 10.1109/IROS45743.2020.9340987.

[124] C. Chen, R. Jafari et N. Kehtarnavaz, "UTD-MHAD : A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*, IEEE, 2015, p. 168-172. doi : 10.1109/ICIP.2015.7350781.

[125] L. Xia, C. Chen et J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, IEEE, 2012, p. 20-27. doi : 10.1109/CVPRW.2012.6239233.

[126] A. Kubota, T. Iqbal, J. A. Shah et L. D. Riek, "Activity recognition in manufacturing : The roles of motion capture and sEMG+ inertial wearables in detecting fine vs. gross motion," in *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, 2019, p. 6533-6539. doi : 10.1109/ICRA.2019.8793954.

[127] S. Mahmud, M. Tonmoy, K. K. Bhaumik et al., "Human activity recognition from wearable sensor data using self-attention," *arXiv preprint arXiv :2003.09018*, 2020. doi : 10.48550/arXiv.2003.09018.

[128] A. Reiss et D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *2012 16th international symposium on wearable computers*, IEEE, 2012, p. 108-109. doi : 10.1109/ISWC.2012.13.

[129] D. Roggen, A. Calatroni, M. Rossi et al., "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh international conference on networked sensing systems (INSS)*, IEEE, 2010, p. 233-240. doi : 10.1109/INSS.2010.5573462.

[130] M. Zhang et A. A. Sawchuk, "USC-HAD : A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, p. 1036-1043. doi : 10.1145/2370216.2370438.

[131] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz et G. Tröster, "Wearable activity tracking in car manufacturing," *IEEE Pervasive Computing*, t. 7, n° 2, p. 42-50, 2008. doi : 10.1109/MPRV.2008.40.

[132] T. Belay Tuli, V. M. Patel et M. Manns, "Industrial Human Activity Prediction and Detection Using Sequential Memory Networks," in *Proceedings of the Conference on Production Systems and Logistics : CPSL 2022*, Hannover : publish-Ing., 2022, p. 62-72. doi : 10.15488/12144.

[133] Y. Zhang, L. Wang, H. Chen, A. Tian, S. Zhou et Y. Guo, "IF-ConvTransformer : A framework for human activity recognition using IMU fusion and ConvTransformer," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, t. 6, n° 2, p. 1-26, 2022. doi : 10.1145/3534584.

[134] R. Yao, G. Lin, Q. Shi et D. C. Ranasinghe, "Efficient dense labelling of human activity sequences from wearables using fully convolutional networks," *Pattern Recognition*, t. 78, p. 252-266, 2018. doi : 10.1016/j.patcog.2017.12.024.

[135] Z. Jiao, G. Jia et Y. Cai, "Ensuring computers understand manual operations in production : Deep-learning-based action recognition in industrial workflows," *Applied Sciences*, t. 10, n° 3, p. 966, 2020. doi : 10.3390/app10030966.

[136] L. Büsch, J. Koch, D. Schoepflin, M. Schulze et T. Schüppstuhl, "Towards Recognition of Human Actions in Collaborative Tasks with Robots : Extending Action Recognition with Tool Recognition Methods," *Sensors*, t. 23, n° 12, p. 5718, 2023. doi : 10.3390/s23125718.

[137] J. Carreira et A. Zisserman, "Quo Vadis, Action Recognition ? A New Model and the Kinetics Dataset," in *CVPR*, 2017.

[138] C. Lea, M. D. Flynn, R. Vidal, A. Reiter et G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 156-165. doi : 10.1109/CVPR.2017.113.

[139] S. Yan, Y. Xiong et D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, t. 32, 2018. doi : 10.5555/3504035.3504947.

[140] L. Shi, Y. Zhang, J. Cheng et H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, p. 12 026-12 035. doi : 10.1109/CVPR.2019.01230.

[141] D. Ahn, S. Kim, H. Hong et B. C. Ko, "STAR-Transformer : A Spatio-temporal Cross Attention Transformer for Human Action Recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, p. 3330-3339. doi : 10.1109/WACV56688.2023.00333.

[142] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in neural information processing systems*, t. 30, 2017. doi : 10.48550/arXiv.1706.03762.

[143] D. Tran, L. Bourdev, R. Fergus, L. Torresani et M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *ICCV*, 2015. doi : 10.1109/ICCV.2015.510.

[144] A. G. Howard et al., "MobileNets : Efficient Convolutional Neural Networks for Mobile Vision Applications," in *arXiv preprint arXiv :1704.04861*, 2017. doi : 10.48550/arXiv.1704.04861.

[145] M. Tan et Q. V. Le, "EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks," in *ICML*, 2019. doi : 10.48550/arXiv.1905.11946.

[146] Z. Liu, H. Zhang, Z. Chen et Z. Wang, "Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition," in *CVPR*, 2020. doi : 10.1109/CVPR42600.2020.00022.

[147] Y. Chen, Z. Zhang, W. Xie, S. Pu et Q. Tian, "Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition," in *ICCV*, 2021. doi : 10.1109/ICCV48922.2021.01311.

[148] S. Cao, W. Luo, B. Wang, W. Zhang et L. Ma, "E2e-load : end-to-end long-form online action detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, p. 10 422-10 432. doi : 10.1109/ICCV51070.2023.00956.

[149] Q. Zhong, G. Ding et A. Yao, "Onlinetas : An online baseline for temporal action segmentation," *Advances in Neural Information Processing Systems*, t. 37, p. 58 984-59 005, 2024. doi : 10.52202/079017-1881.

[150] H. Eun, J. Moon, J. Park, C. Jung et C. Kim, "Learning to discriminate information for online action detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, p. 809-818. doi : 10.1109/CVPR42600.2020.00089.

[151] J. Chen, G. Mittal, Y. Yu, Y. Kong et M. Chen, "Gatehub : Gated history unit with background suppression for online action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 19 925-19 934. doi : 10.1109/CVPR52688.2022.01930.

[152] M. Gao, M. Xu, L. S. Davis, R. Socher et C. Xiong, "Startnet : Online detection of action start in untrimmed videos," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, p. 5542-5551. doi : 10.1109/ICCV.2019.00564.

[153] X. Hu, S. Wang, M. Li, Y. Li et S. Du, "Distribution-aware activity boundary representation for online detection of action start in untrimmed videos," *IEEE Signal Processing Letters*, t. 31, p. 765-769, 2024. doi : 10.1109/LSP.2024.3352503.

[154] M. Xu, M. Gao, Y.-T. Chen, L. S. Davis et D. J. Crandall, "Temporal recurrent networks for online action detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, p. 5532-5541. doi : 10.1109/ICCV.2019.00563.

[155] D.-H. Yoon, N.-G. Cho et S.-W. Lee, "A novel online action detection framework from untrimmed video streams," *Pattern Recognition*, t. 106, p. 107 396, 2020. doi : 10.1016/j.patcog.2020.107396.

[156] S. Lee, H. Eun, J. Moon et al., "Learning to discriminate information for online action detection : Analysis and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, t. 45, n° 5, p. 5918-5934, 2022. doi : 10.1109/CVPR42600.2020.00089.

[157] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek et T. Tuytelaars, "Online action detection," in *European Conference on Computer Vision*, Springer, 2016, p. 269-284.

[158] V. Ramanishka, Y.-T. Chen, T. Misu et K. Saenko, "Toward driving scene understanding : A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, p. 7699-7707. doi : 10.1109/CVPR.2018.00803.

[159] L. Yujian et L. Bo, "A normalized Levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, t. 29, n° 6, p. 1091-1095, 2007. doi : 10.1109/TPAMI.2007.1078.

[160] C. Reining, F. Niemann, F. Moya Rueda, G. A. Fink et M. ten Hompel, "Human activity recognition for production and logistics—a systematic literature review," *Information*, t. 10, n° 8, p. 245, 2019. doi : 10.3390/info10080245.

[161] G. Ding, F. Sener et A. Yao, "Temporal action segmentation : An analysis of modern techniques," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, t. 46, n° 2, p. 1011-1030, 2023. doi : 10.1109/TPAMI.2023.3327284.

[162] X.-Y. Zhang, C. Li, H. Shi, X. Zhu, P. Li et J. Dong, "AdapNet : Adaptability Decomposing Encoder–Decoder Network for Weakly Supervised Action Recognition and Localization," *IEEE Transactions on Neural Networks and Learning Systems*, t. 34, n° 4, p. 1852-1863, 2023. doi : 10.1109/TNNLS.2019.2962815.

[163] Y. Xing et J. Zhu, *Deep learning-based action recognition with 3D skeleton : a survey*, 2021. doi : 10.1049/cit2.12014.

[164] Q. Wen, T. Zhou, C. Zhang et al., "Transformers in time series : A survey," *arXiv preprint arXiv :2202.07125*, 2022. doi : 10.24963/ijcai.2023/75.

[165] L. Hedegaard, N. Heidari et A. Iosifidis, "Continual spatio-temporal graph convolutional networks," *Pattern Recognition*, t. 140, p. 109 528, 2023. doi : 10.1016/j.patcog.2023.109528.

[166] F. Zuo, J. Liu, Y. Ren, L. Wang et Z. Wen, "A Reliable Bolt Key-Points Detection Method in Industrial Magnetic Separator Systems," *IEEE Transactions on Instrumentation and Measurement*, 2025. doi : 10.1109/TIM.2025.3547466.

[167] Q. Su, J. Zhang, M. Chen et H. Peng, "PW-YOLO-Pose : A novel algorithm for pose estimation of power workers," *IEEE Access*, 2024. doi : 10.1109/ACCESS.2024.3437359.

[168] J. Kim, H. Kim, C. Oh, C. Kim et H. J. Kim, "Real-time Moving Object Detection for 3D LiDAR using Occlusion Accumulation in Range Image," *IEEE Transactions on Instrumentation and Measurement*, 2025. DOI : 10.1109/TIM.2025.3545840.

[169] Y.-F. Song, Z. Zhang, C. Shan et L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE transactions on pattern analysis and machine intelligence*, t. 45, n° 2, p. 1474-1488, 2022. DOI : 10.1109/TPAMI.2022.3157033.

[170] N. Trivedi et R. K. Sarvadevabhatla, "Psumnet : Unified modality part streams are all you need for efficient pose-based action recognition," in *European Conference on Computer Vision*, Springer, 2022, p. 211-227. DOI : 10.1007/978-3-031-25072-9_14.

[171] Z. Qin, Y. Liu, P. Ji et al., "Fusing Higher-Order Features in Graph Neural Networks for Skeleton-Based Action Recognition," *IEEE Transactions on Neural Networks and Learning Systems*, t. 35, n° 4, p. 4783-4797, 2024. DOI : 10.1109/TNNLS.2022.3201518.

[172] B. Kwolek, "Continuous Hand Gesture Recognition for Human-Robot Collaborative Assembly," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, p. 2000-2007. DOI : 10.1109/ICCVW60793.2023.00214.

[173] V. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi et V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *2020 third international conference on smart systems and inventive technology (icssit)*, IEEE, 2020, p. 729-735. DOI : 10.1109/ICSSIT48917.2020.9214160.

[174] L. Orsag, T. Stipancic et L. Koren, "Towards a Safe Human–Robot Collaboration Using Information on Human Worker Activity," *Sensors*, t. 23, n° 3, 2023, ISSN : 1424-8220. DOI : 10.3390/s23031283. adresse : https://www.mdpi.com/1424-8220/23/3/1283.

[175] W. Liu, B. Tekin, H. Coskun, V. Vineet, P. Fua et M. Pollefeys, "Learning to align sequential actions in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 2181-2191. DOI : 10.1109/CVPR52688.2022.00222.

[176] R. Ghoddoosian, I. Dwivedi, N. Agarwal, C. Choi et B. Dariush, "Weakly-supervised online action segmentation in multi-view instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 13 780-13 790. DOI : 10.1109/CVPR52688.2022.01341.

[177] M. Peven et G. D. Hager, "Embedding Task Structure for Action Detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, p. 6604-6613. DOI : 10.1109/WACV57701.2024.00647.

[178] R. Slama, W. Rabah et H. Wannous, "STr-GCN : Dual Spatial Graph Convolutional Network and Transformer Graph Encoder for 3D Hand Gesture Recognition," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023, p. 1-6. DOI : 10.1109/FG57933.2023.10042643.

[179] J. Zhu, Z. Ye, M. Ren et G. Ma, "Transformative skeletal motion analysis : optimization of exercise training and injury prevention through graph neural networks," *Frontiers in neuroscience*, t. 18, p. 1 353 257, 2024. DOI : 10.3389/fnins.2024.1353257.

[180] B. Khemani, S. Patil, K. Kotecha et S. Tanwar, "A review of graph neural networks : concepts, architectures, techniques, challenges, datasets, applications, and future directions," *Journal of Big Data*, t. 11, n° 1, p. 18, 2024. DOI : 10.1186/s40537-023-00876-4.

[181] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton et J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," *Advances in neural information processing systems*, t. 31, 2018. DOI : 10.48550/arXiv.1806.08804.

[182] A. Begga, M. Á. Lozano et F. Escolano, "HEX-GNN : Hierarchical EXpanders for Node Classification," in *Conference of the Spanish Association for Artificial Intelligence*, Springer, 2024, p. 233-242. DOI : 10.1007/978-3-031-62799-6_24.

[183] A. J. Fofanah, D. Chen, L. Wen et S. Zhang, "Addressing imbalance in graph datasets : Introducing gate-gnn with graph ensemble weight attention and transfer learning for enhanced node classification," *Expert Systems with Applications*, t. 255, p. 124 602, 2024. DOI : 10.1016/j.eswa.2024.124602.

[184] Y. Wang, X. Hu, Q. Gan, X. Huang, X. Qiu et D. Wipf, "Efficient link prediction via gnn layers induced by negative sampling," *IEEE Transactions on Knowledge and Data Engineering*, 2024. DOI : 10.1109/TKDE.2024.3481015.

[185] P. Pradhyumna, G. Shreya et al., "Graph neural network (GNN) in image and video understanding using deep learning for computer vision applications," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, 2021, p. 1183-1189. DOI : 10.1109/ICESC51422.2021.9532631.

[186] Y. Kong, X. Zhao, R. Liu et al., "Integrating concept of pharmacophore with graph neural networks for chemical property prediction and interpretation," *Journal of Cheminformatics*, t. 14, n° 1, p. 52, 2022. DOI : doi.org/10.1186/s13321-022-00634-3.

[187] A. S. Setlur, V. Niranjan, A. Balaji et C. Karunakaran, "ProteoDockNet : Novel GNN-based ligand binding affinities prediction architecture via SMILES to key liver, kidney and brain proteins using QSAR data," *Computational and Structural Biotechnology Reports*, t. 1, p. 100 011, 2024. DOI : 10.1016/j.csbr.2024.100011.

[188] P. Yang, Q. Wang, H. Chen et Z. Wu, "Position-aware spatio-temporal graph convolutional networks for skeleton-based action recognition," *IET Computer Vision*, t. 17, n° 7, p. 844-854, 2023. DOI : 10.1049/cvi2.12223.

[189] J. Ye, J. He, X. Peng, W. Wu et Y. Qiao, "Attention-driven dynamic graph convolutional network for multi-label image recognition," in *Computer Vision–ECCV 2020 : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, Springer, 2020, p. 649-665. DOI : 10.1007/978-3-030-58589-1_3.

[190] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang et K. Ramani, "Infogcn : Representation learning for human skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, p. 20 186-20 196. DOI : 10.1109/CVPR52688.2022.01955.

[191] B. Parsa, B. Dariush et al., "Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, p. 1080-1090. DOI : 10.1109/WACV45572.2020.9093368.

[192] H. Tian, Y. Zhang, H. Wu, X. Ma et Y. Li, "Multi-scale sampling attention graph convolutional networks for skeleton-based action recognition," *Neurocomputing*, t. 597, p. 128 086, 2024. DOI : 10.1016/j.neucom.2024.128086.

[193] Y. Dong, Q. Liu, B. Du et L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Transactions on Image Processing*, t. 31, p. 1559-1572, 2022. DOI : 10.1109/TIP.2022.3144017.

[194] A. Matin, M. R. Islam, X. Wang et H. Huo, "Robust Multimodal Approach for Assembly Action Recognition," *Procedia Computer Science*, t. 246, p. 4916-4925, 2024. DOI : 10.1016/j.procs.2024.09.448.

[195] L. Romeo, R. Marani, G. Cicirelli et T. D'Orazio, "Multimodal data extraction and analysis for the implementation of Temporal Action Segmentation models in Manufacturing," in *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, IEEE, 2024, p. 854-859. DOI : 10.1109/CoDIT62066.2024.10708238.

[196] T. Benmessabih, R. Slama, V. Havard et D. Baudry, "Graph-based framework for temporal human action recognition and segmentation in industrial context," *Engineering Applications of Artificial Intelligence*, t. 159, p. 111 710, 2025, ISSN : 0952-1976. DOI : 10.1016/j.engappai.2025.111710. adresse : https://www.sciencedirect.com/science/article/pii/S0952197625017129.

[197] P. Zhao, L. Xie, J. Wang, Y. Zhang et Q. Tian, "Progressive privileged knowledge distillation for online action detection," *Pattern Recognition*, t. 129, p. 108 741, 2022. DOI : 10.1016/j.patcog.2022.108741.

[198] G. Hinton, O. Vinyals et J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv :1503.02531*, 2015. DOI : 10.48550/arXiv.1503.02531.

[199] J. Gou, B. Yu, S. J. Maybank et D. Tao, "Knowledge distillation : A survey," *International journal of computer vision*, t. 129, n° 6, p. 1789-1819, 2021. DOI : 10.1007/s11263-021-01453-z.

[200] A. Moslemi, A. Briskina, Z. Dang et J. Li, "A survey on knowledge distillation : Recent advancements," *Machine Learning with Applications*, t. 18, p. 100 605, 2024. DOI : 10.1016/j.mlwa.2024.100605.

[201] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa et H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, t. 34, 2020, p. 5191-5198. DOI : 10.1609/AAAI.V34I04.5963.

[202] C. Yang, X. Yu, Z. An et Y. Xu, "Categories of response-based, feature-based, and relation-based knowledge distillation," in *Advancements in knowledge distillation : towards new horizons of intelligent systems*, Springer, 2023, p. 1-32. DOI : 10.1007/978-3-031-32095-8_1.

[203] J. Yim, D. Joo, J. Bae et J. Kim, "A gift from knowledge distillation : Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 4133-4141. DOI : 10.1109/CVPR.2017.754.

[204] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak et J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, p. 1921-1930. DOI : 10.1109/ICCV.2019.00201.

[205] W. Park, D. Kim, Y. Lu et M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, p. 3967-3976. DOI : 10.48550/arXiv.1904.05068Focustolearnmore.

[206] T. Akiba, S. Sano, T. Yanase, T. Ohta et M. Koyama, "Optuna : A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, p. 2623-2631. DOI : 10.1145/3292500.333070.

[207] M. Maselli, R. Marani, G. Cicirelli et T. D'Orazio, "Continuous Action Recognition in Manufacturing Contexts by Deep Graph Convolutional Networks," in *Intelligent Systems Conference*, Springer, 2023, p. 156-173. DOI : 10.1007/978-3-031-47718-8_11.

[208] P. Huang, Y. Peng, S. Hou et al., "Occluded gait recognition with mixture of experts : an action detection perspective," in *European Conference on Computer Vision*, Springer, 2024, p. 380-397. DOI : 10.1007/978-3-031-72658-3_22.

Toufik BENMESSABIH

Reconnaissance d'actions humaines basée sur l'apprentissage profond et la représentation squelettique : Propositions d'approches hors ligne et en ligne dans un contexte d'Industrie 5.0

Résumé : Dans le contexte de l'industrie 5.0, qui place l'humain au cœur du processus de production, notamment pour favoriser la collaboration humain-robot, améliorer l'ergonomie des postes de travail et mettre à jour les jumeaux numériques, la reconnaissance d'actions humaines (RAH) devient une composante importante. Cependant, malgré les avancées récentes de l'apprentissage profond, la RAH demeure un problème ouvert dans ce contexte industriel. Les approches existantes restent encore peu efficaces pour détecter avec précision et en temps réel des actions hétérogènes, présentant une dynamique temporelle importante et réalisées dans des environnements complexes. Les verrous scientifiques majeurs concernent donc la modélisation conjointe des dynamiques spatiales et temporelles des séquences d'actions à partir de jeux de données déséquilibrés, l'interprétabilité des modèles et leur adaptation aux contraintes du temps réel.

Plusieurs contributions ont permis de relever ces défis. Tout d'abord, la modalité squelettique a été exploitée, car elle est à la fois robuste et précise pour représenter le mouvement humain, et la première proposition a permis de l'intégrer dans une stratégie à multi-résolutions temporelles (coarse-to-fine). La méthodologie proposée améliore la distinction entre les actions de courte durée et dépasse les approches existantes en termes de performance. Ensuite, cette méthodologie a été enrichie d'une représentation en graphe, qui conserve la topologie du squelette tout au long de l'architecture du modèle, et intègre un mécanisme d'attention, permettant une représentation des mouvements plus expressive et une interprétabilité des décisions du modèle de segmentation d'actions. Enfin, comme troisième contribution, nous avons proposé une approche en ligne de détection d'actions, basée sur la distillation progressive des connaissances sous forme de leçons (temporelle, d'occultation et d'objet) transférées d'un modèle hors ligne enseignant vers un modèle en ligne étudiant, et adaptée aux contraintes industrielles. Cette stratégie renforce la stabilité temporelle, la robustesse face aux pertes de données dues aux occultations et la distinction des actions liées aux interactions avec les outils manipulés lors des tâches d'assemblage. Nos résultats montrent que notre méthode surpasse l'état de l'art, tant quantitativement que qualitativement, sur le jeu de données InHARD.

Ce travail ouvre des perspectives sur des applications industrielles concrètes. Il pourra notamment contribuer à l'évaluation ergonomique des postes de travail, à la collaboration humain-robot par l'assistance en temps réel aux opérateurs, ainsi qu'à la mise à jour de jumeaux numériques. Ces apports s'inscrivent dans une démarche plus large d'intégration de systèmes de perception intelligents, en cohérence avec la transition vers l'industrie 5.0.

Mots clés : Industrie 5.0, Reconnaissance d'actions, Apprentissage profond, Représentation basée graphe, Distillation des connaissances, Vision par ordinateur.

Abstract : In the context of Industry 5.0, which places humans at the center of production processes, particularly to foster human–robot collaboration, improve workplace ergonomics, and update digital twins, human action recognition (HAR) is becoming a key component. However, despite recent advances in deep learning, HAR remains an open problem in industrial environments. Existing approaches are still not sufficiently effective at accurately detecting heterogeneous actions in real time, especially when performed in constrained environments. The major scientific challenges lie in jointly modeling the spatial and temporal dynamics of action sequences from imbalanced datasets, ensuring model interpretability, and adapting models to real-time constraints.

Several contributions in this thesis addressed these challenges. First, skeletal data were exploited for their robustness and precision in representing human movement, and the initial contribution introduced a multi-resolution temporal strategy (coarse-to-fine) to process this modality. The proposed methodology improves the discrimination of short-duration actions and surpasses existing approaches in performance. Second, this methodology was extended using a graph-based representation that preserves the skeletal topology throughout the model architecture and incorporates an attention mechanism, enabling more expressive motion representations and improving the interpretability of action segmentation decisions. Finally, as a third contribution, we proposed an online action detection approach based on progressive knowledge distillation. This strategy transfers several types of "lessons" (temporal, occlusion-based, and object-related) from an offline teacher model to an online student model, making it suitable for industrial constraints. This approach enhances temporal stability, robustness to occlusions and missing data, and improves the distinction of actions involving tool manipulation during assembly. Our results show that the proposed method outperforms the state of the art, both quantitatively and qualitatively, on the InHARD dataset.

This work opens several perspectives toward concrete industrial applications. It can support ergonomic workstation assessment, human–robot collaboration through real-time operator assistance, and continuous updating of digital twins. These contributions align with a broader vision of integrating intelligent perception systems, in line with the transition toward Industry 5.0.

Keywords : Industry 5.0, Action Recognition, Deep learning, Graph-based representation, Knowledge distillation, Computer vision.