Méthodes d'apprentissage appliquées à l'analyse du comportement humain par vision

Astrid Orcesi

université PARIS-SACLAY

Méthodes d'apprentissage appliquées à l'analyse du comportement humain par vision

Learning methods applied to vision-based human behaviour analysis

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580,
Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Sciences du traitement du signal et des images
Graduate School : Sciences de l'ingénierie et des systèmes
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'**Institut LIST** (Université Paris-Saclay, CEA) sous la direction de **Quoc Cuong PHAM**, Directeur de recherche

Thèse soutenue à Paris-Saclay, le 14 Juin 2023, par

Astrid ORCESI

Composition du jury

Membres du jury avec voix délibérative

David PICARD Professeur, Ecole des Ponts ParisTech	Président
Alice CAPLIER Professeur, INP Grenoble Alpes	Rapporteur & Examinatrice
Thierry CHATEAU Professeur, Université Clermont Auvergne	Rapporteur & Examinateur
Adrien CHAN HON TONG Chercheur, ONERA	Examinateur

THESE DE DOCTORAT
NNT : 2023UPAST082

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Méthodes d'apprentissage appliquées à l'analyse du comportement humain par vision
Mots clés : Vision, Apprentissage, Comportement humain, Reconnaissance d'activités, Détection d'interactions

Résumé : L'analyse du comportement humain par vision est une thématique de recherche très étudiée car malgré les progrès apportés par l'apprentissage profond en vision par ordinateur, comprendre finement ce qui est en train de se passer dans une scène est une tâche loin d'être résolue car elle présente un très haut niveau sémantique. Dans cette thèse nous nous intéressons à deux applications : la reconnaissance d'activités longues temporellement dans des vidéos et la détection d'interaction dans des images. La première contribution de ces travaux est l'élaboration de la première base de données d'activités quotidiennes présentant de fortes variabilités intra-classe. La deuxième contribution est la proposition d'une nouvelle méthode de détection d'interaction en une seule passe sur l'image ce qui lui permet d'être beaucoup plus rapide que les méthodes de l'état de l'art en deux étapes et appliquant un raisonnement par paire d'instances. Enfin, la troisième contribution de cette thèse est la constitution d'un nouveau jeu de données d'interactions composé d'interactions à la fois entre des personnes et des objets mais également entre des personnes ce qui n'existait pas jusqu'à maintenant et qui permet pourtant une analyse des interactions humaines exhaustive. De manière à proposer des résultats de référence sur ce nouveau jeu de données, la précédente méthode de détection d'interactions a été améliorée en proposant un apprentissage multi-tâches ce qui permet d'obtenir les meilleurs résultats sur la base de données publique largement utilisée par la communauté.

Title : Learning methods applied to vision-based human behaviour analysis
Keywords : Vision, Learning, Human behavior, Activity recognition, Interaction detection

Abstract : The analysis of human behavior by vision is a strong studied research topic. Indeed despite the progress brought by deep learning in computer vision, understanding finely what is happening in a scene is a task far from being solved because it presents a very high semantic level. In this thesis we focus on two applications : the recognition of temporally long activities in videos and the detection of interaction in images. The first contribution of this work is the development of the first database of daily activities with high intra-class variability. The second contribution is the proposal of a new method for interaction detection in a single shot on the image which allows it to be much faster than the state of the art two-step methods which apply a reasoning by pair of instances. Finally, the third contribution of this thesis is the constitution of a new interaction dataset composed of interactions both between people and objects and between people which did not exist until now and which allows an exhaustive analysis of human interactions. In order to propose baseline results on this new dataset, the previous interaction detection method has been improved by proposing a multi-task learning which reaches the best results on the public dataset widely used by the community.

Remerciements

Je remercie Alice Caplier et Thierry Chateau d'avoir accepté de rapporter ce manuscrit de thèse. Je remercie également Adrien Chan Hon Tong et David Picard d'avoir accepté d'être membre du jury.

Un grand Merci à Quoc Cuong Pham d'avoir cru en moi depuis le début de ma carrière et de m'avoir offert l'opportunité de travailler sur des thématiques en lien avec mon projet personnel de recherche sur l'analyse du comportement humain. Merci à lui de m'inciter à voir toujours plus loin dans mes ambitions. C'est grâce à son soutien et à sa confiance que j'ai osé me lancer dans la démarche d'une thèse par VAE. Merci à lui également pour le temps qu'il a dédié à mon projet. J'associe Mohamed Chaouch et Bertrand Luvison à ces remerciements pour leur soutien et leurs nombreux conseils lors de la rédaction de ce manuscrit.

Merci à Patrick Sayd d'avoir également cru en moi en soutenant ma démarche de thèse par VAE.

Merci à Bertrand Luvison, mon binôme sur la thématique de l'analyse du comportement au laboratoire. Merci pour tout ce qu'il m'apprend au quotidien et ses idées toujours plus pertinentes les unes que les autres !

Je tiens également à remercier mes collègues avec lesquels j'ai collaboré sur les différents projets et sans lesquels ces résultats n'auraient pu voir le jour : Bertrand Luvison, Romaric Audigier, Sanaa Chafik, Geoffrey Vaquette, Laurent Lucas, Jaonary Rabarisoa, Guillaume Lorre, Julien Denize et Adrien Maglo.

Merci à Angélique Loesch, devenue ma cheffe et mon amie ! Je la remercie pour toutes ces belles années passées ensemble au VisionLab, pour nos échanges quotidiens, et son soutien sans faille.

Enfin Merci à tout le service d'intelligence artificielle pour le langage et la vision du CEA LIST, en particulier à Hervé Le Borgne qui m'a recrutée dans son équipe pour un stage en 2015. Il fait bon travailler au SIALV, je crois que c'est pour cette raison que je lui suis fidèle depuis le début !

Table des matières

1 - Introduction

Sommaire

1.1 . Contexte des travaux de recherche

Définition de l'analyse du comportement humain

L'analyse du comportement humain par l'interprétation de scènes visuelles est une thématique de recherche très active avec un nombre important d'applications dans des domaines variés. Tout d'abord pour la sécurité où il est important de pouvoir détecter des évènements anormaux et dangereux tels que des violences, des faits de vandalisme ou encore des bagages abandonnés. Pour les applications de santé, que ce soit pour l'assistance des personnes à domicile ou leur suivi à l'hôpital, la compréhension de l'activité des personnes est essentielle. En effet, comprendre les habitudes d'une personne permet de mieux détecter un début de perte d'autonomie si une activité n'est plus réalisée. A l'hôpital, l'intérêt sera de vérifier par exemple si une personne a bien reçu les différents soins nécessaires. Le sport est également un domaine où l'analyse automatique des gestes mais aussi des interactions sociales et la communication entre les joueurs se met au service de la performance. Par exemple, pour les sports collectifs de haut niveau, les performances techniques des joueurs ne sont plus à démontrer et l'issue d'un match se base bien souvent sur leur état psychologique. L'analyse du langage corporel et la détection d'attitudes révélatrices de la dynamique du groupe sont alors des nouvelles données très importantes pour les psychologues du sport. La compréhension des gestes est également en jeu pour les thématiques industrielles. Que ce soit pour vérifier qu'un technicien a correctement effectué un geste sur une chaîne de production ou pour le corriger pendant son apprentissage, une analyse précise de ses gestes techniques est impérative. Enfin dans l'industrie, certaines tâches ne

peuvent être complètement automatisées mais les robots peuvent aider les techniciens dans leur travail, c'est la robotique collaborative. Dans ce cadre, il est indispensable que le robot comprenne les actions du technicien pour l'assister correctement dans sa tâche et rendre les interactions entre robot et humain fluides.

Figure 1.1 – Différents domaines d'application de l'analyse du comportement humain.

La problématique de l'analyse du comportement humain peut être abordée par différentes questions. La première question est "Que se passe-t-il dans la scène ?", l'objectif est de reconnaître les actions en cours.

Ensuite, la deuxième question est "Qui réalise cette action ?" ou "Qui prend part à cette activité ?", là nous cherchons à associer les actions reconnues aux personnes dans la scène. L'analyse peut être faite pour un individu, un groupe de personnes lorsqu'il s'agit d'interactions sociales par exemple ou bien plus globalement sur toute une foule. Plus la densité de personne est importante, plus il est difficile d'analyser les individus indépendamment les uns des autres. L'analyse des comportements de foule est un axe d'étude à part entière [LCL+11, SRV+20, FLP16] car elle est très liée à la détection des évènements anormaux dont la définition elle-même est ambiguë car basée sur la question "qu'est-ce que la normalité ?".

Par le pronom "Qui", on peut également chercher à identifier la personne qui réalise l'action. En effet, pour les technologies d'assistance, la personnalisation du service est importante : il faut être capable d'analyser le bon sujet. L'axe de recherche lié à cette problématique est la ré-identification de personnes qui est également une thématique de recherche à elle seule [DAL+22, WZT+22]. La ré-identification regorge de défis scientifiques : Comment reconnaître une même personne avec des points de vue et des résolutions différentes ? Comment distinguer deux personnes habillées de façon très similaire, avec des uniformes par exemple ? Comment s'affranchir du fond et extraire une caractéristique propre à la personne ?

On peut vouloir caractériser la manière dont est réalisée une action en se posant la question "Comment ?". La réponse peut être déclinée selon différents aspects. Au niveau des émotions, la personne est-elle heureuse de réaliser cette activité ? Ou au contraire, est-elle en colère au moment de réaliser l'action ? S'il s'agit d'une activité de groupe, les personnes impliquées sont-elles complices ? Ou au contraire, expriment-t-elles de l'énervement l'une envers l'autre ? L'analyse des émotions est également un large champ de recherche à part entière en vision par ordinateur [ACC18, XWWL22]. Les méthodes actuelles fonctionnent bien pour des émotions "jouées" où l'on distingue très facilement les 6 émotions primaires qui sont la joie, la tristesse, le dégoût, la colère, la peur et la surprise, à condition d'avoir une résolution suffisante sur le visage de la personne. Mais dans la plupart des cas, les émotions sont beaucoup plus subtiles et dans ces cas, la problématique est encore loin d'être résolue.

Un autre aspect est l'intensité mise par l'individu pour réaliser l'action. Par exemple, cette personne mange-t-elle calmement ou est-elle pressée ? Porte-t-elle un carton facilement ou avec difficulté ? Tire-t-elle une autre personne doucement ou fortement ? On peut également chercher à mesurer la précision dans la réalisation d'un geste. Tel geste a été effectué mais était-il précis ? Correctement réalisé ?

Enfin, la dernière question à se poser pour obtenir une analyse du comportement complète est "Avec quoi ?". En effet, l'action peut impliquer des éléments de la scène comme certains objets : la personne tire cette valise, la personne lance cette balle, etc.

Dans ces travaux, nous abordons l'analyse du comportement de manière centrée sur chaque personne de la scène en cherchant à reconnaître leur activité et nous nous intéressons également aux objets impliqués.

La figure 1.2 illustre une description détaillée d'une scène dans le cadre d'une application de vidéo protection.

Définition de la vision par ordinateur

La vision par ordinateur quant à elle est la science qui cherche à comprendre et à automatiser les tâches que le système visuel humain est capable d'effectuer. Dans ce cadre, les données utilisées sont donc des images, des vidéos ou encore des données 3D. La vision est un moyen particulièrement adapté d'accéder à l'interprétation des comportements humains car on peut viser la même interprétation que le cerveau humain. Les images contiennent une information riche sur le contexte de réalisation de l'action mais les vidéos permettent une analyse temporelle des mouvements qui est indispensable pour certaines applications. Cependant le traitement vidéo est beaucoup plus gourmand en ressources de calcul en raison de l'ajout de la dimension

Figure 1.2 – Exemple d'une description complète d'une scène répondant aux questions : "Quoi ?", "Qui ?" par la localisation des actions uniquement, dans le contexte de vidéo protection, on ne connaît pas forcément l'identité de toutes les personnes présentes, "Comment ?" et "Avec quoi ?". Étant donné la résolution de l'image, il n'est pas possible de détecter les émotions sur les visages. Image issue du jeu de données MALL [CLGX12]

temporelle. Les données 3D peuvent également apporter une meilleure compréhension de la scène et notamment de la profondeur dans l'image pour modéliser la position des individus les uns par rapport aux autres ou estimer dans l'espace avec précision les poses des personnes.

Apprentissage machine pour la vision

Il existe différents paradigmes d'apprentissage pour aborder l'analyse du comportement humain. Le plus largement utilisé est l'apprentissage supervisé qui a pour but d'optimiser un modèle pour une tâche précise en exploitant des données pour lesquelles la réponse à la tâche est connue (ce sont les labels). Une fois entraîné, le modèle est capable de prédire la réponse à une nouvelle donnée d'entrée. L'apprentissage non supervisé quant à lui exploite des données non annotées. Une variante est l'apprentissage auto-supervisé qui se base sur une tâche prétexte pour créer automatiquement

des labels. Un intermédiaire est l'apprentissage semi-supervisé qui a pour caractéristique d'utiliser à la fois des données annotées et des données non annotées. Plus récemment, des techniques d'apprentissage par renforcement ont également été utilisées pour l'analyse du comportement [VRCHTA21]. Le problème est posé de façon à pouvoir être résolu par une prise de décision séquentielle. La figure 1.3 illustre ces différents paradigmes d'apprentissage automatique.

L'apprentissage supervisé est bien maîtrisé aujourd'hui et donne de bons résultats à condition d'avoir les données annotées nécessaires en nombre suffisant et souvent très conséquent. L'apprentissage auto-supervisé présente donc un gros avantage, celui de ne pas avoir besoin de données annotées mais la tâche prétexte n'est pas évidente à définir pour obtenir l'espace de représentation propice à la bonne résolution de la tâche finale visée. L'apprentissage semi-supervisé présente des résultats intéressants à condition que la distribution des données non annotées ne soient pas trop éloignée de celle des données annotées.

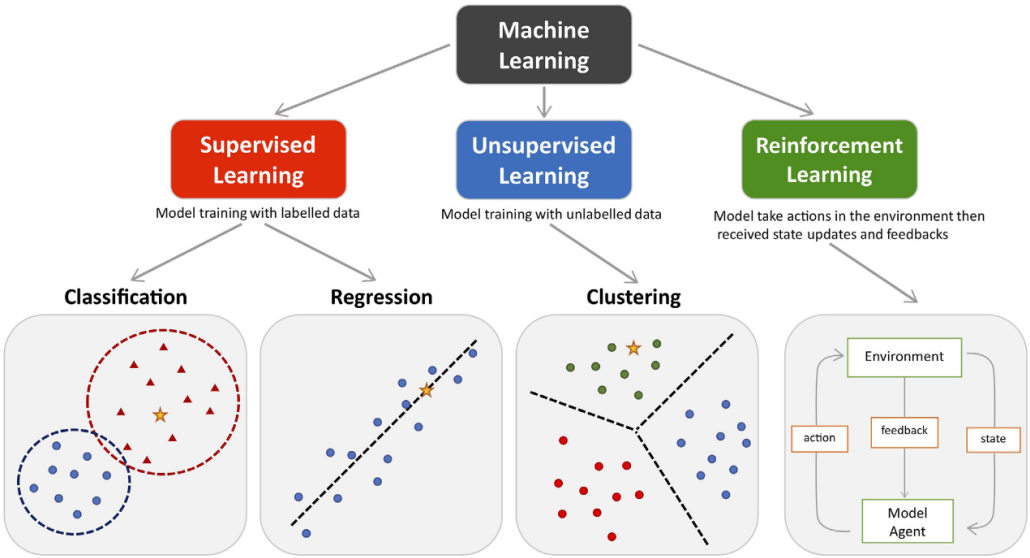


Figure 1.3 – Illustration des différents paradigmes d'apprentissage automatique. Image issue de l'article [PJDC21]

Dans les travaux présentés dans ce manuscrit, nous nous intéressons donc à l'automatisation de l'analyse du comportement centré sur une personne donnée en appliquant des méthodes d'apprentissage supervisé à des données de type image ou vidéo. Dans le chapitre 5, je présente d'autres travaux portant cette fois-ci sur un autre paradigme d'apprentissage : l'apprentissage auto-supervisé.

1.2 . Verrous scientifiques de l'analyse du comportement par vision

Généralisation du modèle aux paramètres d'acquisition

Contrairement à la segmentation d'objet qui demande une analyse à l'échelle pixellique et qui est une tâche en partie résolue aujourd'hui [WDC+22], l'analyse du comportement humain requiert une compréhension globale de l'image pour modéliser une situation, une scène, un évènement avec un niveau sémantique élevé.

De plus, l'analyse du comportement humain demande une généralisation à de nombreux paramètres tels que : le changement de sujet, la manière de réaliser une action, la vitesse d'exécution, le contexte et l'environnement de réalisation ou encore le point de vue de la caméra. Ces différents changements entraînent des problèmes de variabilités auxquels la méthode d'analyse doit être robuste. Le changement de point de vue est l'un des paramètres qui peut apporter le plus de variabilité dans la visualisation d'une même scène. Par exemple, les applications de vidéo protection utilisent un point de vue haut et plongeant pour pouvoir observer le plus grand champ possible et limiter les occultations. Au contraire, il existe de plus en plus d'applications comme les lunettes connectées qui utilisent une vue égo-centrique où la caméra est positionnée au niveau des yeux du sujet, le point de vue sur la scène est alors bas et rasant [DDF+22]. La figure 1.4 illustre ces différents paramètres d'acquisition.

Données d'apprentissage

Les données utilisées pour l'entraînement des modèles sont la clé pour tendre le plus possible vers cette généralisation. Sans données, l'apprentissage supervisé n'est pas réalisable or la création d'une base de données est une étape non négligeable car elle demande beaucoup de temps depuis la collecte des données brutes jusqu'à leur annotation.

Aujourd'hui, acquérir des données image ou vidéo est une étape relativement simple lorsque la tâche visée n'est pas spécifique à un certain cas d'usage, le Web regorge de données brutes. Cependant, si l'objectif est de reconnaître une tâche précise dans un contexte industriel par exemple, il faut nécessairement passer par une étape de récolte de ces données. L'acquisition peut alors être freinée par l'autorisation d'accès aux experts réalisant l'action et même une fois l'autorisation obtenue, la mise en oeuvre peut être longue pour obtenir un jeu de données conséquent.

Ensuite, obtenir l'annotation adéquate s'avère beaucoup plus complexe et coûteux car extrêmement chronophage surtout lorsque les données sont des vidéos. Effectivement, annoter un clip vidéo revient à annoter toutes les

Figure 1.4 – Illustration des différents paramètres auxquels les méthodes d'analyse du comportement humain doivent être invariantes : le changement de sujet, le changement de contexte, le changement de point de vue et le changement de vitesse d'exécution.

images qui le composent. De plus, segmenter précisément le début et la fin d'une action dans une vidéo n'est pas trivial. Par exemple pour une activité longue, choisir l'image précise du clip vidéo à partir de laquelle l'activité démarre est une tâche ambiguë tout comme l'annotation d'un évènement court dans une vidéo par une seule image. Si les scènes sont denses, il est néces-

saire d'avoir une annotation exhaustive sur l'ensemble des personnes présentes dans la scène. Le travail de localisation d'un individu et l'annotation de son activité est donc multiplié par le nombre de personnes.

De nombreux projets et travaux de recherche s'intéressent aujourd'hui au développement d'outils et de briques technologiques visant à faciliter l'annotation. Par exemple, mon laboratoire développe l'outil d'annotation Pixano qui intègre des modèles d'analyse de scène permettant une pré-annotation automatique des données (https ://pixano.cea.fr).

La communauté scientifique met de plus en plus à disposition des jeux de données annotés mais ces derniers sont trop souvent hétérogènes au niveau des annotations, non équilibrés et surtout trop peu représentatifs des nombreux cas réels.

Temps de calcul

La plupart des applications requièrent un traitement rapide, voire une réponse en temps réel tout en tournant sur un matériel embarqué. Aujourd'hui, pour obtenir des analyses performantes des données vidéo, la tendance est aux réseaux de neurones de plus en plus lourds [WLL+22]. Des technologies telles que la reconnaissance faciale peuvent être facilement embarquables en temps réel mais c'est encore loin d'être le cas pour des applications d'analyse de l'activité.

Fenêtre temporelle d'analyse

La fenêtre temporelle d'analyse est également un verrou de l'analyse du comportement humain. En effet, une même application peut vouloir analyser des comportements allant de la seconde à plusieurs minutes. Comment choisir la fenêtre d'analyse pour à la fois détecter des évènements très brefs et des situations demandant une agrégation d'informations temporelles plus longue ? Le risque de choisir une fenêtre d'analyse trop grande est d'effacer l'information utile à la détection des évènements de courte durée. A l'inverse, une fenêtre de temps trop petite limiterait la compréhension d'une activité longue.

Modalités d'entrée

Comment choisir la bonne modalité d'entrée ? Dans le cas de l'analyse du comportement humain, l'image seule permet d'obtenir le contexte de réalisation d'une action ou d'une activité ce qui est un fort indice pour reconnaître sa classe cependant dans la majorité des applications, certains comportements

nécessitent une analyse temporelle dans des séquences vidéo. En effet, la modalité image est souvent insuffisante pour distinguer deux gestes tels que "prendre" ou "poser" par exemple. Cependant, les approches exploitant des données d'entrées de type vidéo sont logiquement beaucoup plus coûteuses en ressources de calcul. La dimension supplémentaire nécessite une modélisation spatio-temporelle plus complexe, ainsi que de grands corpus de vidéos annotées. Se posent également la problématique de la détermination de l'échelle temporelle d'observation et le problème de modélisation des évènements de longue durée. Toutes ces difficultés font que les méthodes d'analyse vidéo ont une maturité moindre que les approches fondées sur l'analyse d'une seule image.

Un des défis des méthodes utilisant la vidéo est de correctement modéliser le mouvement en faisant abstraction du contexte. Pour contraindre la méthode à se focaliser sur le mouvement, l'utilisation du flot optique est intéressante car il permet de s'affranchir du contexte mais il est souvent coûteux en calcul. La différence d'image est alors une alternative pertinente. Cependant, l'utilisation du mouvement seul en s'affranchissant du contexte présente la faiblesse de n'utiliser qu'une information de très bas niveau sémantique.

1.3 . Différents niveaux de description sémantique du comportement humain

La compréhension du comportement humain peut s'étudier à différents niveaux sémantiques :

- **Le geste** : est un mouvement unitaire, court dans le temps, de l'ordre d'une à deux secondes. La variabilité de l'exécution d'un geste est limitée ce qui facilite sa reconnaissance. En effet, les gestes ne sont pas étudiés pour être reconnus mais plutôt pour analyser la précision d'un geste technique comme dans les domaines sportifs ou industriels. Le principal défi réside en la variabilité de point de vue et de contexte de réalisation du geste.
- **L'action** : est composée de plusieurs gestes, sa réalisation est de l'ordre de plusieurs secondes. La variabilité d'exécution d'une action est plus grande que celle du geste, la tâche de reconnaissance d'action n'est donc pas triviale.
- **L'activité** : est composée d'actions et est beaucoup plus longue dans le temps. En effet, une activité a un temps de réalisation de l'ordre de plusieurs minutes voire plusieurs heures. Son niveau sémantique est élevé tout comme la variabilité de son exécution. La reconnaissance d'activités est donc une tâche complexe à modéliser.
- **L'interaction** : met en relation deux instances via un geste, une action ou une activité. Une interaction peut avoir lieu entre un humain et un

objet ou entre plusieurs humains. L'analyse des interactions implique un niveau de description sémantique très élevé qui demande une compréhension plus fine du contexte de la scène. Le principal défi concerne l'association entre l'humain et la cible de l'interaction.

Pour illustrer ces quatre niveaux de description sémantique du comportement humain, prenons l'exemple de la préparation d'un repas : l'activité est "cuisiner" qui est composée de plusieurs actions telles que "couper", "verser" ou "mélanger" qui sont elles-mêmes composées de gestes tels que "tendre le bras" ou "porter à la bouche". L'interaction reprend ces différents verbes et y ajoute un objet cible : "couper avec ce couteau", "verser dans ce saladier", "porter à la bouche avec cette fourchette". La figure 1.5 illustre ces 4 niveaux sémantiques.

Figure 1.5 – Les différents niveaux sémantiques de l'analyse du comportement humain

Dans la pratique, ces différents niveaux sémantiques sont généralement traités individuellement compte tenu de la longueur de la fenêtre d'observation qui est très relative au niveau sémantique choisi. Cependant, malgré ces définitions, la frontière entre ces quatre niveaux sémantiques n'est pas toujours facile à placer ce qui constitue une complexité supplémentaire en soi. En effet, "ouvrir un tiroir", "fermer une porte", sont souvent interprétés comme des actions, tellement courtes qu'elles s'apparenteraient plutôt à des gestes.

Contre intuitivement, plus le niveau sémantique est élevé, plus il est facile de comprendre la situation grâce à une seule image. En effet, un humain est tout à fait capable de dire si une personne dort, mange ou lit, qui sont des activités, avec une seule image à sa disposition. Au contraire, distinguer les gestes "prendre" ou "poser" un objet nécessite quelques images consécutives pour comprendre le sens du mouvement. Certaines activités sont effectivement très liées au contexte alors que d'autres sont définies temporellement à partir des mouvements.

1.4 . Tâches de vision pour l'analyse des actions et Méthodes d'apprentissage supervisé

Tâches de vision pour l'analyse des actions

En vision par ordinateur, l'analyse des actions se déclinent en plusieurs tâches étudiées dans la littérature de façon disjointe les unes des autres. Pour chacune de ces tâches, des jeux de données sont disponibles et sont pour la plupart, spécifiquement annotés pour le type de tâche en question. La liste ci-dessous a pour but d'être la plus exhaustive possible quant aux différentes tâches étudiées par la communauté scientifique.

- **La classification** (ou reconnaissance) : Dans ce cas d'usage, l'image ou le clip vidéo contient un ou plusieurs comportements humains. L'objectif est de reconnaître ces comportements à partir de l'image ou de la vidéo prise dans sa globalité, autrement dit d'estimer la ou les bonnes étiquettes, en cas de classification multi-labels, parmi une liste prédéfinie. C'est par exemple le cas pour les jeux de données UCF-101 [SZS12] ou Kinetics [KCS+17].
- **La détection** : Dans ce cas d'usage, l'image contient plusieurs personnes où la vidéo enchaîne plusieurs comportements. L'objectif est ici de reconnaître ces comportements mais également de les localiser spatialement et/ou temporellement. C'est par exemple le cas pour les jeux de données AVA [GSR+18] et Epic-Kitchens [DDF+22]. La figure 1.6 illustre la différence entre la classification et la détection.
- **Le repérage des actions** (ou *action spotting* en anglais) : L'objectif est de détecter des évènements dans une vidéo en estimant précisément l'instant où ils sont apparus. Les applications sont principalement dans le domaine sportif où l'on souhaite détecter automatiquement une liste d'actions en vue d'établir le résumé automatique d'un match par exemple. SoccerNet [GADG18] est un exemple de jeu de données d'*action spotting* pour l'analyse de matches de football.
- **La description d'images ou de vidéos** (ou *image or video captioning*

en anglais) : L'objectif est de générer une phrase syntaxiquement correcte qui décrit ce qui se passe dans une image ou dans un clip vidéo. Plusieurs jeux de données existent également pour cette tâche comme ActivityNet captioning [KHR+17] ou YouCook2 [ZXC18].

- **La segmentation de vidéos non-supervisée** : Les labels des classes ne sont pas déterminés à l'avance, l'objectif est de découper la vidéo en fonction de la cohérence visuelle et temporelle de chaque segment. Les travaux [SY18] utilisent par exemple le jeu de données Breakfast [KAS14] pour évaluer cette tâche.

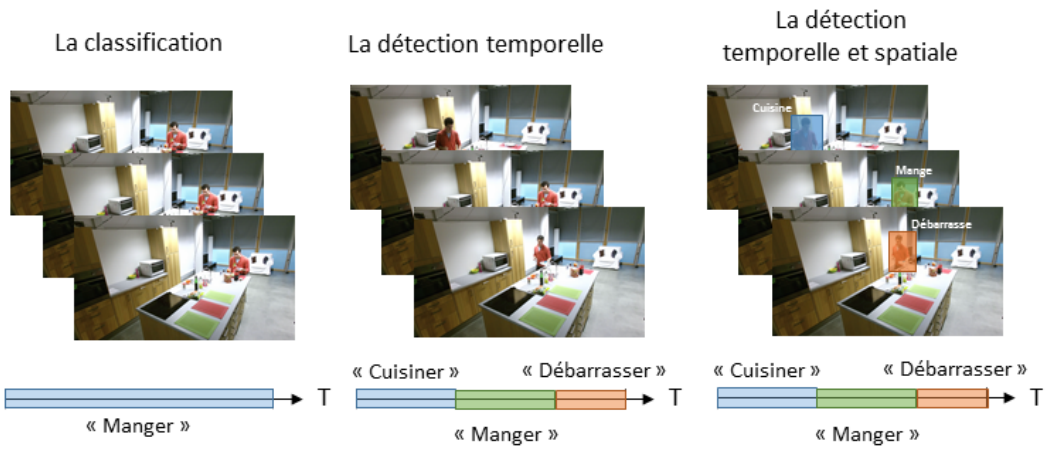
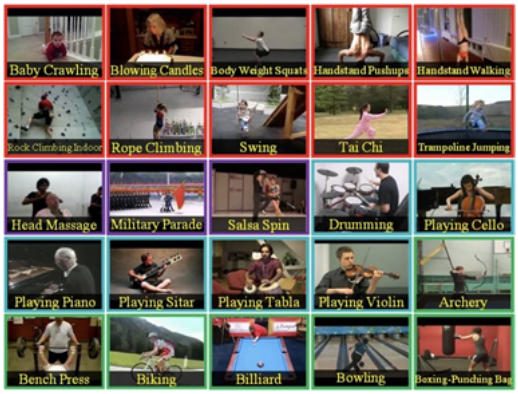


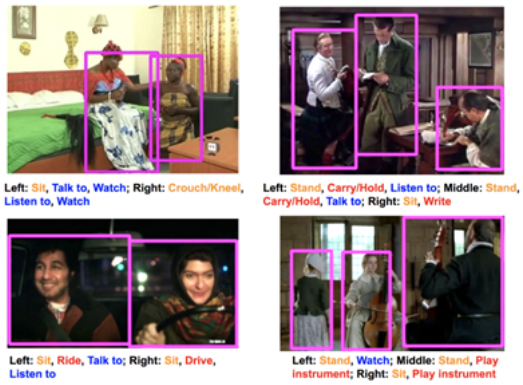
Figure 1.6 – Différences entre classification et détection d'actions

La figure 1.7 illustre les différentes tâches listées pour analyser des actions. La classification est la plus simple à la fois à annoter et à modéliser (bien sûr dans la limite des verrous à la généralisation cités précédemment et illustrés figure 1.4). Il n'y a pas d'ambiguïté possible puisque l'ensemble de l'image ou de la vidéo doit correspondre aux classes annotées. La détection a un plus grand intérêt dans l'objectif de déployer le système final dans le monde réel. En effet, dans la plupart des cas les scènes contiennent plusieurs personnes et les flux vidéo ne sont pas pré-segmentés. Cependant la tâche de détection est forcément plus complexe car elle nécessite de comprendre où et quand se passe l'action en plus de la reconnaître. L'*action spotting* est une tâche très difficile car elle demande une grande précision sur la localisation temporelle de l'action, les performances actuelles des modèles sur le jeu de données SoccerNet sont de l'ordre de seulement 60% de bonne localisation des actions. Enfin l'*image captioning*, dépend très fortement du jeu de données utilisé. Actuellement les phrases descriptives générées sont limitées au contexte de la scène mais ne rentrent pas dans les détails spatio-temporels. De plus, l'*image captioning* rajoute la difficulté de générer une phrase syntaxiquement correcte en plus de la compréhension de la

scène.

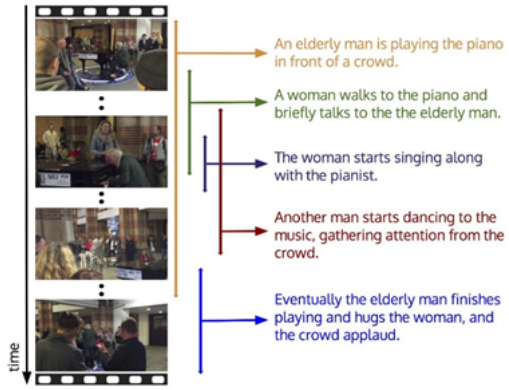


Classification d'action dans des vidéos
Jeu de données : UCF-101



Détection spatio-temporelle d'action dans des vidéos
Jeu de données : Dataset AVA

Action Spotting
Jeu de données : Dataset SoccerNet action spotting



Video captioning
Jeu de données : Dataset ActivityNet Captions

Figure 1.7 – Illustration des différentes tâches de vision pour l'analyse des actions. Jeu de données exemple : UCF-101 [SZS12], AVA [GSR+18], SoccerNet action spotting [GADG18], ActivityNet captioning [KHR+17]

Méthodes d'apprentissage supervisé

Le principal paradigme utilisé pour modéliser ces tâches est l'apprentissage supervisé qui a beaucoup évolué ces dernières années. On distingue aujourd'hui les méthodes classiques d'apprentissage automatique des méthodes d'apprentissage profond basé sur les réseaux de neurones.

La différence majeure entre les méthodes classiques d'apprentissage automatique et l'apprentissage profond est le type de données en entrée. En effet, les méthodes classiques d'apprentissage automatique nécessitent une étape de pré-traitement des images ou des vidéos avec des descripteurs ad-

hoc pour en extraire des caractéristiques visuelles qui seront les données d'entrées du modèle. Les méthodes d'apprentissage profond quant à elles, apprennent un réseau de neurones avec de nombreuses couches cachées. L'entrée de ces réseaux peut être les mêmes descripteurs "faits à la main" utilisés par les méthodes classiques d'apprentissage automatique mais, et c'est là la grande différence, également directement être les images ou les vidéos. On parle dans ce cas d'apprentissage de bout-en-bout. Ces méthodes utilisent directement l'image ou la vidéo et apprennent à extraire des caractéristiques pertinentes en même temps qu'elles apprennent le modèle.

L'apprentissage profond a littéralement révolutionné le monde de la vision par ordinateur. En effet, les performances des réseaux de neurones sur des tâches telles que la détection d'objet a nettement surpassé celles des méthodes classiques d'apprentissage à condition d'avoir une base de données d'exemples conséquente. Effectivement, un point limitant de l'utilisation des réseaux de neurone est la quantité de données nécessaires pour obtenir un modèle qui soit le plus généralisable possible. De plus, étant donné que l'apprentissage profond permet d'entraîner un modèle à extraire les caractéristiques, les résultats de ces méthodes sont beaucoup plus difficiles à expliquer que les estimations des méthodes classiques d'apprentissage automatique basées sur des caractéristiques extraites "à la main". En effet, même si des techniques telles que GradCam [SCD+17] facilitent quelque peu l'interprétation des résultats, l'explicabilité des réseaux de neurones est aujourd'hui un domaine de recherche à part entière. Enfin, si détecter des personnes dans une image est une tâche en grande partie résolue par l'apprentissage profond car analyse l'image à une échelle pixellique, c'est encore loin d'être le cas quand la tâche demande de préciser l'action ou l'activité réalisée par cette personne. En effet, ce niveau de description sémantique plus élevé demande une compréhension plus fine du contexte de la scène, ce qui est beaucoup plus complexe à modéliser.

1.5 . Positionnement des travaux et contributions

Parmi toutes les façons d'aborder l'analyse du comportement humain et tous les défis qui y sont liés, j'ai choisi d'étudier la détection temporelle des activités du quotidien dans des vidéos. Au moment de mes travaux, la reconnaissance d'action était surtout ciblée sur des actions courtes voire des gestes et les jeux de données disponibles présentaient des exemples filmés dans des environnements de laboratoire. La thématique de recherche autour des activités longues n'était pas encore développée notamment par manque de données publiques exploitables. La tâche est difficile car les activités présentent beaucoup de variabilité au sein d'une même classe mais les applications sont nombreuses et peuvent contribuer à répondre à de vrais enjeux de société.

Dans une société vieillissante, le sujet du suivi et du maintien des personnes à domicile est plus que jamais d'actualité, ce qui en fait une vraie source de motivation pour mes recherches.

J'ai donc constitué la première base de données publique d'activités longues, nommée DAHLIA (*DAily Home LIfe Activity Dataset*) dans laquelle une cinquantaine de personnes effectuent les activités du quotidien réalisables dans une cuisine : faire le ménage, cuisiner, manger, faire la vaisselle, etc. La création d'un jeu de données comme DAHLIA comporte ses propres difficultés car elle demande beaucoup de préparation (réflexion méthodologique sur les scénarios, la représentativité des activités, etc.), de temps d'enregistrement et de post-traitement. Le rendre public s'avère encore plus compliqué, après avoir récolté le consentement des acteurs, il est indispensable de respecter le processus RGPD de respect des données personnelles. La méthode d'apprentissage DOHT (*Deeply Optimized Hough Transform*) m'a permis d'élaborer des résultats de référence et aujourd'hui, DAHLIA est activement utilisé par la communauté scientifique [NGA+18, DMG+19, NB19, AAC+20].

Les performances de reconnaissance d'activité longue se dégradent rapidement lorsque le point de vue ou le contexte n'est pas le même que celui des données d'apprentissage. Le lieu unique est effectivement la principale limite du jeu de données DAHLIA. De manière à généraliser la tâche et à tenter de s'affranchir de la grande variabilité de réalisation des activités, je me suis intéressée à la détection des interactions entre les personnes et les objets. Cette tâche permet d'analyser les activités plus finement en les découpant en sous-actions. Le but de la détection des interactions dans des images est d'associer une personne et un objet par l'intermédiaire d'un verbe d'action afin de former un triplet $< sujet, verbe, objet >$. Au début de ces travaux, les approches existantes étaient toutes fondées sur l'analyse de l'ensemble des paires possibles de l'image composées d'un humain et d'un objet. La complexité de ces méthodes est de fait quadratique avec le nombre d'objets présents dans la scène. Visant des applications temps réel, j'ai cherché à proposer des méthodes dont le temps d'inférence n'est pas une contrainte pour le déploiement de l'algorithme. Ces travaux ont conduit à la conception de la méthode CALIPSO (*Classifying ALl Interacting Pairs in a Single shOt*) qui est la première méthode *single shot* de classification d'interactions, c'est à dire qu'elle estime les interactions en un seul passage dans le réseau de neurones en plus d'être compétitive sur les jeux de données de référence en détection d'interactions.

CALIPSO a la particularité d'être agnostique au détecteur d'objets. C'est-à-dire qu'au moment de l'inférence, n'importe quel détecteur d'objets peut être utilisé. Cependant, certaines applications, notamment les applications em-

barquées, ne peuvent se permettre de faire tourner deux réseaux distincts en raison des contraintes de puissance de calcul et de mémoire. J'ai donc étendu l'approche CALIPSO à l'architecture multi-tâches DIABOLO (*Detecting InterActions By Only Looking Once*) qui est capable d'apprendre la détection d'interactions et la détection d'objets au sein d'un même réseau de neurones. L'étude de l'apprentissage conjoint des deux tâches ont permis d'améliorer nettement les résultats puisque DIABOLO est la méthode donnant les meilleures performances sur le jeu de données le plus utilisé en détection d'interactions.

Enfin, les interactions entre personnes sont très peu représentées dans les jeux de données d'interactions. L'analyse du comportement humain ne peut être complet si la détection des interactions sociales ou des interactions violentes ne sont pas représentées. J'ai également pu noter que la taxonomie des jeux de données publiques pour la détection d'interactions est souvent trop liée au contexte de réalisation des interactions plutôt qu'à l'attitude des personnes et utilise souvent des verbes synonymes. Le contexte biaise fortement le modèle qui prédit cette interaction à chaque fois qu'il le détecte mais n'est plus capable d'estimer cette interaction dans un environnement différent. De manière à s'affranchir de ses limitations, j'ai créé le jeu de données H²O (*Human-to-Human-or-Object interaction dataset*) qui est aujourd'hui le seul à proposer à la fois des interactions entre personnes et des interactions entre des personnes et des objets. La taxonomie liée à ce jeu de données a également été revue pour être moins liée au contexte de réalisation mais plus à la gestuelle de l'interaction ce qui permet de décrire les scènes plus finement. H²O est aujourd'hui disponible publiquement pour la communauté scientifique.

Mes recherches sont structurées autour de deux axes : la détection d'activités et la détection d'interactions. Le chapitre 2 présente mes contributions sur la détection d'activités longues de la vie quotidienne. La première section de ce chapitre dresse un état de l'art sur les méthodes d'apprentissage supervisées appliquées à la détection d'activités et établit un aperçu des jeux de données disponibles. La suite de ce chapitre décrit la proposition du nouveau jeu de données DAHLIA et des résultats de référence de détection d'activités sur ce jeu.

Le chapitre 3 expose ma première contribution en détection d'interactions avec la proposition de CALIPSO, une nouvelle méthode *single shot*. Dans ce chapitre, nous donnons d'abord un état de l'art sur les méthodes d'apprentissage supervisées appliquées à la détection d'interactions et un panorama des jeux de données disponibles, puis nous décrivons en détail la méthode CALIPSO et les résultats obtenus.

Le chapitre 4 présente ma deuxième contribution en détection d'interactions : la création d'un nouveau jeu de données d'images d'interactions entre personnes et entre une personne et un objet, H^2O et la proposition d'une nouvelle approche, DIABOLO, une extension de CALIPSO capable de détecter simultanément les objets et leurs interactions. Cette méthode fournit des résultats de référence compétitifs pour H^2O.

Je présente dans le chapitre 5 mes autres travaux de recherche autour de la thématique de l'analyse du comportement humain qui ne sont pas l'objet principal du mémoire, mais permettent d'ouvrir des axes de travail complémentaires.

Le chapitre 6 résume mes contributions et propose des perspectives aux travaux présentés.

Enfin, le chapitre 7 liste l'ensemble de mes publications.

2 - Détection d'activités : DAHLIA, le premier jeu de données d'activités longues

Nous vivons aujourd'hui dans une société de plus en plus vieillissante où les besoins en technologies de maintien à domicile et de surveillance de l'activité des personnes âgées sont grands. Ce constat a motivé nos travaux de recherche sur la détection des activités longues du quotidien. L'objectif est de détecter les moments de la journée où une personne a cuisiné, pris son repas, travaillé, etc, afin de dresser un portrait de sa journée type et ainsi détecter une potentielle perte d'autonomie si elle ne réalise plus l'une de ces activités ou la réalise différemment.

Dans ce chapitre, nous commençons par dresser un état de l'art sur les méthodes de reconnaissance des actions et détaillons les jeux de données disponibles. Le manque de bases de données pour la détection des activités longues nous a poussé à créer le premier jeu de données d'activités du quotidien, DAHLIA (*DAily Home LIfe Activity Dataset*) : non segmenté, enregistré depuis 3 points de vue dans un environnement réaliste et faisant apparaître 44 sujets différents, ce qui est le plus grand nombre de sujet présent dans une base de données d'actions. DAHLIA est présenté dans la deuxième partie de ce chapitre.

Sommaire

2.1 . État de l'art de la détection d'activités dans des vidéos

2.1.1 . Principe

Il est important de distinguer la reconnaissance d'actions de la détection. L'objectif de la reconnaissance d'actions est de classifier un clip vidéo présegmenté contenant une seule ou plusieurs étiquettes d'action dans le cas multi-labels. Le modèle analyse le clip-vidéo dans sa globalité. Au contraire, la détection d'actions s'intéresse aux vidéos plus longues enchaînant plusieurs actions ou activités, l'objectif est de délimiter temporellement voir spatialement chaque action. Ces deux notions sont détaillées et illustrées dans le premier chapitre, section 1.4.

Les travaux de l'état de l'art s'intéressent en grande majorité à la reconnaissance d'actions et appliquent ensuite le modèle sur une fenêtre temporelle glissante pour résoudre la tâche de détection.

2.1.2 . Méthodes classiques d'apprentissage automatique pour la reconnaissance d'actions

Au moment de mes travaux, les méthodes d'apprentissages profond commencent tout juste à connaître l'essor qui les ont porté jusqu'à leur succès aujourd'hui. C'est pourquoi je réserve une partie de l'état de l'art aux méthodes classiques d'apprentissage automatique qui donnaient les meilleurs résultats de classification d'actions en 2016.

Les méthodes classiques d'apprentissage automatique s'appliquent en deux étapes : la première consiste à extraire des caractéristiques des données d'entrée tandis que la deuxième étape apprend de manière supervisée à classifier ces caractéristiques.

Extraction des caractéristiques

Les caractéristiques ont pour but de représenter le contenu du clip vidéo. Pour que le classifieur puisse ensuite généraliser les classes d'actions, les caractéristiques doivent être invariantes aux rotations, translations et changement d'échelles, invariantes face aux transformations affines et au bruit ainsi que robustes aux changements de luminosité.

La première étape pour extraire des caractéristiques est la détection de points d'intérêt. Différentes méthodes permettent de les localiser dans l'image : Le détecteur de coin Harris [Lap05] et sa variante 3D [Lap05], les SIFT [Low04], les points à partir de filtres de Gabor [DRCB05] ou encore les points Hessian 3D [WTVG08].

Pour une analyse temporelle, il est ensuite pertinent de suivre ces points d'intérêt temporellement. En les suivant au cours du temps, on obtient des trajectoires de points saillants. Une des premières méthodes de suivi de points d'intérêt est le *KLT tracker* [LK+81] qui se base sur une correspondance des points entre des images successives. En 2013, [WS13] introduisent les iDT (*improved Dense Trajectories*). Basés sur les trajectoires de points extraits sur une grille dense, les iDT estiment le mouvement global de l'image de manière à rendre l'extraction de trajectoires robustes aux mouvements de la caméra.

Une fois les points d'intérêt détectés et éventuellement suivis, il est nécessaire de les caractériser pour qu'ils puissent décrire le contenu spatiale et temporel local. Parmis les descripteurs les plus répandus, nous trouvons les Histogrammes de gradients orientés (HOG) [DT05] et les Histogrammes de flux optiques (HOF) [DTS06]. Les HOG décrivent l'apparence spatiale locale par le calcul d'un histogramme qui modélise les différentes directions du gradient sur la région étudiée. [KMS08] proposent une extension des HOG à la modalité spatio-temporelle en introduisant les HOG3D. Les histogrammes sont calculés en utilisant des gradient avec une dimension temporelle supplémentaire. Les HOF sont basés sur le même principe que les HOG mais utilisent les directions principales du flux optique à la place du gradient.

A partir des descripteurs HOF, [DTS06] proposent les descripteurs MBH (*Motion Boundary Histograms*). Ils sont fabriqués à partir des dérivées spatiales (horizontale et verticale) du flux optique. Seules les variations du flux optiques sont prises en compte ce qui correspond aux contours des mouvements d'où le nom de frontière.

Depuis 2010, des capteurs permettant de récupérer les cartes de profondeur de la scène à bas coût ont fait leur apparition tels que la *Kinect*, la *Asus Xtion* ou encore la *Camine PrimeSense*. Des descripteurs issus de cartes de profondeur ont donc été proposés. Nous pouvons par exemple citer : les DSTIPS [XA13], le BSC *Body Surface Context)* [STLY14] ou encore les DMM (*Depth Motion Maps*) [YZT12].

De manière à avoir une analyse plus centrée sur la gestuelle des sujets réalisant les actions, des travaux autour de l'analyse du squelette des

personnes ont vu le jour. L'objectif est de décrire la pose des personnes impliquées dans la scène à chaque instant de la vidéo. Ces squelettes peuvent être extraits à partir de dispositifs spécialisés comme la *motion capture* ou bien à partir des cartes de profondeur [GSK+11, SFC+11].

Classifieurs

Les SVM [HDO+98] ont eu beaucoup de succès pour la reconnaissance d'actions, particulièrement avec les méthodes utilisant des points d'intérêt [Lap05, DRCB05] mais la classification est globale et ignore la cohérence temporelle des descripteurs sur l'intégralité de la vidéo. De nouvelles méthodes ont donc proposé des approches séquentielles basées sur les modèles de Markov cachés (HMM) [AQMM07, LN06, FST98]. Les HMM ont l'avantage de pouvoir être appliqué à des tâches des détection cependant au moment de l'inférence, ils nécessitent d'analyser la vidéo dans son intégralité afin de réaliser l'optimisation et d'obtenir la segmentation idéale. En 2013, [CHTAL13] propose la méthode DOHT (*Deeply Optimized Hough Transform*) de détection en ligne basée sur les votes par transformée de Hough. Cette approche s'appuie sur une accumulation progressive de descripteurs et affine ses hypothèses au cours du temps. Elle a l'avantage de pouvoir utiliser en entrée n'importe quel type de descripteur. [CHTAL14] a montré que le DOHT est une formulation équivalente à un classifieur SVM. La figure 2.1 illustre le processus d'apprentissage de l'algorithme DOHT et la figure 2.2 le processus de test. En 2019, [VAL19] étudie différentes méthodes de fusion appliquées à différents niveaux de l'algorithme DOHT : au niveau des descripteurs, au niveau des votes ou au niveau des scores.

2.1.3 . Méthodes par apprentissage profond

Depuis mes travaux sur la reconnaissance d'actions, les méthodes par apprentissage profond sont devenues les grandes incontournables de l'état de l'art. Pour analyser un flux vidéo avec des réseaux convolutionnels, plusieurs stratégies sont possibles :

- Appliquer des convolutions 2D à chaque image du clip et fusionner les cartes de caractéristiques [KTS+14, WGGH18].
- Utiliser des convolutions 3D prenant également en compte la dimension temporelle [TBF+15, QYM17, CZ17, TWT+18, Fei20].
- Utiliser des réseaux récurrents comme les RNN ou les LSTM permettant de garder en mémoire l'information issue des images précédentes [DAHG+15, MS20]
- Utiliser une réseau à deux branches pour analyser en parallèle différentes modalités d'entrée qui sont généralement les données RGB et le flot optique [FPZ16, CZ17, FFMH19].

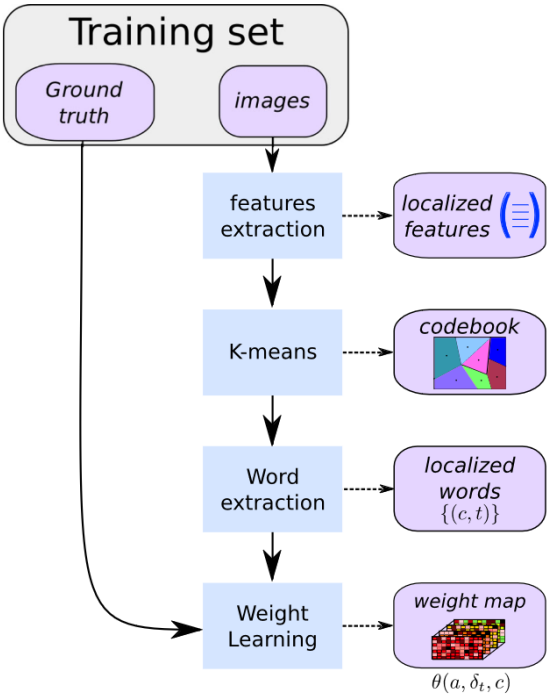


Figure 2.1 – Procédure d'apprentissage de l'algorithme DOHT [CHTAL13]. Image issue de l'article [VAL19]

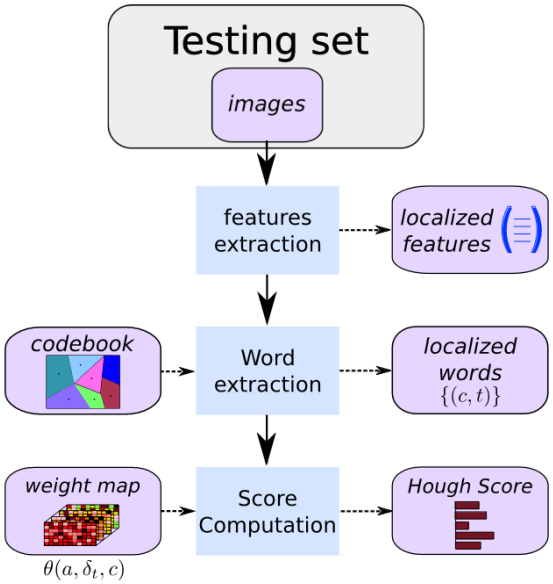


Figure 2.2 – Procédure de test de l'algorithme DOHT [CHTAL13]. Image issue de l'article [VAL19]

Depuis leur succès en traitement du langage, les méthodes à base de *transformer* font également leur apparition dans l'analyse vidéo et c'est aujourd'hui elles qui donnent les meilleurs résultats sur le jeu de données Kinetics 400 [KCS+17], devenu le jeu de données de référence pour la reconnaissance d'actions.

Le site de Pytorch Video (https ://pytorchvideo.org), propose une implémentation des modèles les plus connus pour l'analyse vidéo et dresse un tableau des méthodes de référence et leur résultat avec un apprentissage entièrement supervisé sur Kinetics [KCS+17]. La figure 2.3 est tirée du site web. Les deux dernières lignes correspondent à une architecture *transformer*, les autres sont des réseaux convolutifs. MViT [FXM+21], basé sur une architecture *transformer*, se classe donc effectivement à la première place sur cette sélection de méthode. Cependant, elle est loin d'être la plus légère et la plus rapide. En effet, c'est X3D [Fei20] qui présente le meilleur compromis entre performance et temps de calcul.

2.1.4 . Jeux de données

Ensembles de données pour la reconnaissance d'actions

Dans un premier temps, les ensembles de données proposés étaient assez simples, ils étaient composés de clips vidéo très courts, dans des environnements très contrôlés et avec des gestes brefs et peu diversifiés [GBS+07]. Par exemple, le jeu de données KTH [SLC04], largement utilisé pour la reconnaissance d'actions, contient de très courtes vidéos en noir et blanc capturées sur un fond homogène comme l'illustre la figure 2.4.

Par la suite, les vidéos ont été extraites de flux réels tels que des émissions de télévision, des films ou des vidéos provenant de sites Web [KJG+11]. Ces ensembles de données sont plus réalistes que les précédents, car les arrière-plans sont remplis et les caméras sont en mouvement. Cependant, les actions qui se produisent dans ces ensembles de données sont encore courtes, et contiennent principalement un ou quelques gestes.

Avec l'émergence de capteurs de profondeur à faible coût comme la Kinect, plusieurs ensembles de données RGB-D ont été publiés dont voici les plus populaires. L'un des premiers ensembles de données d'actions RGB-D publié est MSR-Action3D [ZG10] qui contient 20 actions exécutées 3 fois par 10 sujets. Les actions concernées sont courtes, comme "coup de poing avant", "coup de pied latéral", "jogging", "service de tennis", etc. Les

Kinetics-400

Année	arch	depth	pretrain	frame length x sample rate	top 1	top 5	Flops (G) x views	Params (M)
2018	C2D	R50	-	8x8	71.46	89.68	25.89 x 3 x 10	24.33
2017	I3D	R50	-	8x8	73.27	90.70	37.53 x 3 x 10	28.04
2019	Slow	R50	-	4x16	72.40	90.18	27.55 x 3 x 10	32.45
2019	Slow	R50	-	8x8	74.58	91.63	54.52 x 3 x 10	32.45
2019	SlowFast	R50	-	4x16	75.34	91.89	36.69 x 3 x 10	34.48
2019	SlowFast	R50	-	8x8	76.94	92.69	65.71 x 3 x 10	34.57
2019	SlowFast	R101	-	8x8	77.90	93.27	127.20 x 3 x 10	62.83
2019	SlowFast	R101	-	16x8	78.70	93.61	215.61 x 3 x 10	53.77
2019	CSN	R101	-	32x2	77.00	92.90	75.62 x 3 x 10	22.21
2018	R(2+1)D	R50	-	16x4	76.01	92.23	76.45 x 3 x 10	28.11
2020	X3D	XS	-	4x12	69.12	88.63	0.91 x 3 x 10	3.79
2020	X3D	S	-	13x6	73.33	91.27	2.96 x 3 x 10	3.79
2020	X3D	M	-	16x5	75.94	92.72	6.72 x 3 x 10	3.79
2020	X3D	L	-	16x5	77.44	93.31	26.64 x 3 x 10	6.15
2021	MViT	B	-	16x4	78.85	93.85	70.80 x 1 x 5	36.61
2021	MViT	B	-	32x3	80.30	94.69	170.37 x 1 x 5	36.61

Figure 2.3 – Tableau des résultats des méthodes de références sur le jeu de données Kinetics 400 [KCS+17]. Figure issue du site https ://pytorchvideo.org. Références des méthodes : C2D [WGGH18], I3D [CZ17], Slowfast et ses variantes [FFMH19], CSN [TWTF19], R(2+1)D [TWT+18], X3D [Fei20], MVit [FXM+21]

Figure 2.4 – Contenu du jeu de données KTH [SLC04]

auteurs ont fourni des séquences de profondeur et, plus tard, les données du squelette ont également été publiées. Wang et al. [WLWY12] ont présenté le jeu de données MSR Daily Activity, capturé avec une Kinect v1 et ont fourni les images RGB, les cartes de profondeur et les squelettes. Les 16 actions réalisées étaient orientées vers un usage quotidien comme "boire", "manger", "lire un livre", "jouer à un jeu", "s'asseoir", etc. La figure 2.5 illustre quelques exemples du jeu de données MSR Daily Activity.

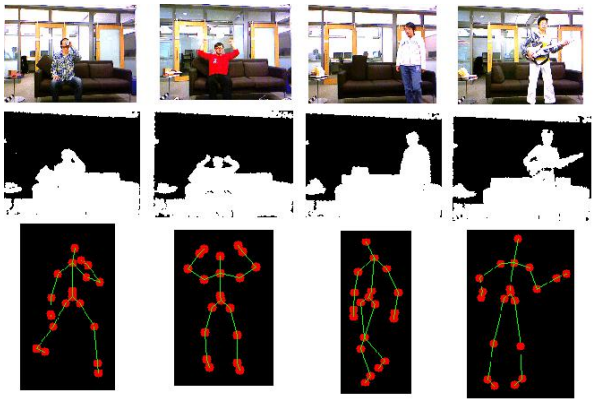


Figure 2.5 – Contenu du jeu de données MSR Daily Activity [WLWY12]

Sung et al. ont publié dans [SPSS11] un jeu de données appelé CAD-60 dans lequel 12 actions ont été capturées dans 5 environnements différents, à savoir la salle de bain, la chambre à coucher, la cuisine, le salon et le bureau, afin de diversifier le fond des actions.

Ces ensembles de données ne prennent en compte qu'un seul point de

vue dans chaque exemple. Afin d'augmenter la variabilité intra-classe, mais aussi de bénéficier de différents points de vue lors de l'étape de test, des jeux de données de reconnaissance d'action multi-vues ont été introduits [CQY+12, APNL13, OCK+13, NPMY13, KGS13]. Pour les jeux de données UWA3D Multiview [RMDHM14] et NJUST [STLY14] les sujets effectuent chaque action plusieurs fois, sous différentes vues latérales. La plupart de ces ensembles ont été capturés simultanément avec deux ou trois capteurs Kinect [NPMY13, KGS13].

ATC4² [CQY+12], l'un des premiers jeux de données multi-vues, a été collecté en 2012. Il contient 14 classes correspondant à des actions quotidiennes telles que "Boire", "Téléphoner", "Lire", "Jeter", etc. mais aussi deux actions en relation avec des applications de santé à savoir "Tomber" et "Trébucher". Les données de couleur, de profondeur et de squelette extraites de 4 capteurs Kinect sont enregistrées.

Le jeu de données Berkeley MHAD [OCK+13] a été capturé avec différents types de capteurs : un système de motion capture, 4 caméras stéréo, 2 Kinect, 6 accéléromètres et 4 microphones, pour explorer la complémentarité de plusieurs modalités.

Afin de pallier le manque d'exemples dans les bases de reconnaissance d'action, Shahroudy et al. [SLNW16] ont créé le jeu de données NTU RGB+D contenant 56 880 exemples, capturés avec une Kinect v2. Ce nouveau jeu de données, illustré figure 2.6, vise à faciliter l'utilisation des réseaux de neurones dans le domaine de l'analyse d'actions qui commencent à être utilisés au moment de mes travaux.

Depuis mes travaux, le jeu de données devenu référence en termes de reconnaissance d'actions est le jeu Kinetics. Il est comparable à ImageNet pour la modalité vidéo et est largement utilisé dans les recherches menées sur le pré-apprentissage non-supervisé. En effet, Kinetics400 [KCS+17] est composé de 216 000 vidéos pour l'apprentissage et de 18 000 pour la validation, le tout annoté selon 400 classes d'actions. Nous pouvons citer YouTube-8M [AEHKL+16] qui est également un des plus gros jeux de données de reconnaissance d'action avec 8 millions de vidéos soit presque 500 000 heures.

Jeux de données pour la détection en ligne des actions

Dans la plupart des applications, la détection en ligne des actions est beaucoup plus réaliste que la classification de clips vidéo. En effet, les flux

Figure 2.6 – Contenu du jeu de données NTU RGB+D [SLNW16]

extraits d'environnements non contrôlés ne sont pas segmentés. En suivant cette idée, les jeux de données composés de séquences continues contenant plusieurs actions ont été créés [YLW09, RAD17].

Tout d'abord, des actions simples ("se lever", "s'asseoir", "porter", "saluer", "boire", etc.) ont été effectuées dans un flux non segmenté [HYWT14]. Dans [WZZZ13], Wei et al. ont fourni un ensemble de données dans lequel les sujets effectuent plusieurs actions simultanément (parmi 12 actions). De plus, ces actions peuvent interagir les unes avec les autres. Dans [RAD17], plusieurs actions peuvent également se produire simultanément et les algorithmes doivent alors détecter les actions à la fois spatialement et temporellement.

Wu et al. ont publié dans [WZSS15] un jeu de données relativement important acquis avec le capteur Kinect v2. Il était demandé aux acteurs d'effectuer plusieurs actions soit dans une cuisine, soit dans un bureau en séquences continues. Ainsi, les vidéos ne sont pas coupées et conviennent aux algorithmes de segmentation. Cet ensemble de données a été capturé à 13 endroits différents, avec un seul point de vue.

Rohrbach et al. ont présenté dans [RAAS12] et son extension [RRA+12] un jeu de données contenant des séquences continues d'activités culinaires :

Figure 2.7 – Contenu du jeu de données MPII Cooking Activities Dataset [RAAS12]

MPII Cooking Activities Dataset. Il est composé de 65 activités différentes telles que "couper en tranches", "couper en dés", "éplucher", etc. réalisées par 12 acteurs. Ils ont augmenté la variabilité en donnant des instructions verbales aux acteurs pour qu'ils préparent un des 14 plats comme un sandwich ou une salade. La durée des vidéos varie de 3 à 41 minutes. La figure 2.7 illustre des exemples du jeu de données.

Dans le jeu de données DML Smart Actions [APNL13], les acteurs ont également effectué une série d'actions sans interruption et sans instructions précises. Il a été enregistré à l'aide de 2 caméras HD et un capteur Kinect v1, fournissant 3 points de vue différents et une carte de profondeur de la scène.

2.1.5 . Positionnement des travaux

Si de nombreuses bases de données de gestes [NWM11, SPSS11, OCK+13] ou d'actions [WLWY12, RAD17, WZSS15] existent dans la littérature, aucune base de données d'activités longues n'a été publiée au moment de mes travaux.

Nous souhaitons proposer une nouvelle base de données contenant des activités longues, durant plusieurs minutes, de la vie quotidienne effectuées dans un environnement domestique réaliste et qui sont exécutées sans contraintes pour obtenir le plus de variabilité intra-classe possible. Cette base de données ouvre un nouveau champ de recherche et de nouvelles problématiques liées aux activités de la vie courante de longue durée.

Aujourd'hui, depuis mes travaux, notre principal concurrent est la base de données Toyota Smart Home [DDK+19] où des personnes âgées effectuent des actions de la vie quotidienne dans un appartement. Les actions sont filmées depuis 7 points de vue avec un faible taux de recouvrement. D'abord publié sous forme de clip vidéos segmentés par actions, le jeu de données a ensuite été étendu [DDS+22] et propose à présent les vidéos en entier avec les actions qui s'enchaînent. Cependant les actions restent de durée courte et leur variabilité intra-classe est faible. La figure 2.8 illustre des exemples du jeu de données.

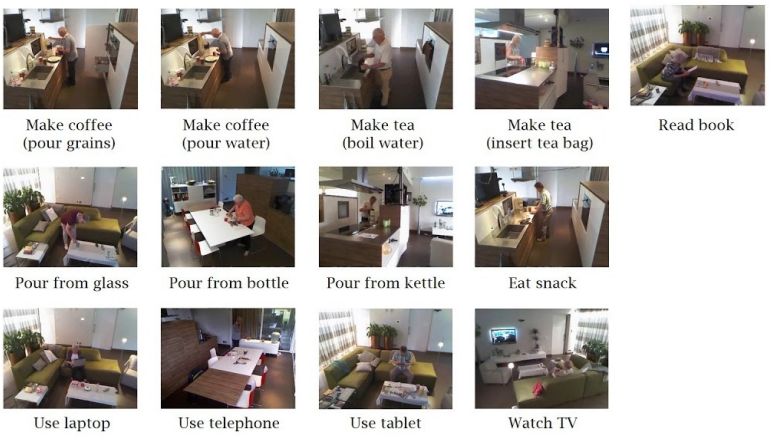


Figure 2.8 – Contenu du jeu de données Toyota Smart Home [DDK+19]

2.2 . Jeu de données proposé

Ce travail se concentre sur la surveillance de l'activité des personnes âgées à domicile. Il vise à détecter et à reconnaître des activités longues de la vie quotidienne telles que "prendre son repas". Ces activités sont composées de nombreuses actions, ce qui rend la tâche de reconnaissance difficile. Par exemple, la classe "laver la vaisselle" contient de nombreuses itérations de "prendre l'objet", "laver", "rincer" ou "mettre l'objet sur l'égouttoir". Ces sous-actions pourraient également être divisées en plusieurs gestes comme "avancer la main". Les différentes configurations de la cuisine, la vaisselle utilisée, les personnes effectuant les activités et l'ordre dans lequel elles effectuent ces sous-actions, conduisent à une très grande variabilité intra-classe.

Pour pallier le manque de jeu de données présentant des activités longues de la vie quotidienne, nous proposons un nouvelle base de données, appelée DAHLIA pour *DAily Home LIfe Activity Dataset*.

2.2.1 . Protocole expérimental

Nous avons enregistré 51 longues séquences exécutées par 44 personnes différentes à l'heure du déjeuner dans une cuisine réaliste. Les locaux du CEA LIST sont équipés d'une plateforme appelée MobileMii (https ://www-mobilemii.cea.fr/) qui est un appartement. Les pièces sont aménagées pour être les plus réalistes possible. L'objectif de cette plateforme est de créer et d'évaluer des services d'intelligence ambiante dans des espaces intelligents.

Avant l'enregistrement, nous avons donné aux participants des instructions très simples pour qu'ils réalisent les activités de manière très naturelle. En particulier, après une rapide présentation de la cuisine, nous leur avons demandé d'effectuer 7 activités de la vie quotidienne dans différents ordres (certaines sont naturellement ordonnées comme, par exemple, mettre la table avant de manger). L'ensemble de données ainsi obtenu présente donc une grande variété dans la manière de réaliser les activités.

Nous avons retenu 7 activités de haut niveau sémantique réalisables dans le contexte d'une cuisine. Elles sont représentatives des activités de la vie quotidienne et sont les suivantes :

- **Cuisiner** : Le sujet prépare son déjeuner. Il s'agit de prendre les aliments, les récipients, de couper les tomates et le fromage, de préparer une vinaigrette à base d'huile et de vinaigre.
- **Dresser la table** : Le sujet prépare la table pour son déjeuner : les assiettes, les couverts, la nourriture, l'eau, etc.
- **Déjeuner** : Cette activité correspond à l'intégralité de la prise du repas, sans instructions spécifiques, on demande simplement aux acteurs d'apprécier son repas.
- **Débarrasser la table** : Le participant retire les objets de la table et les pose près de l'évier ou à leur emplacement initial.
- **Faire la vaisselle** : Le sujet lave la vaisselle sale et l'essuie.
- **Travailler** : Les participants doivent répondre à un test en recherchant des informations dans des documents. Le test était régulièrement modifié pour ajouter des variations.
- **Faire le ménage** : Les participants devaient balayer autour de la table, nettoyer la table et changer le sac poubelle si nécessaire.

Afin d'être aussi réalistes que possible, les participants ont été invités à réaliser cette expérience à l'heure du déjeuner et à prendre leur repas pendant l'acquisition des données. La durée moyenne de l'ensemble des séquences est de 39 minutes, allant de 24 à 64 minutes, soit un total de 33,4 heures. En outre, certaines activités comme "déjeuner" prennent beaucoup plus de temps que "mettre la table". Il existe une grande variabilité dans la

durée de réalisation entre les classes. La figure 2.9 représente la proportion des 8 classes sur l'ensemble de la base (les sept classes précédemment décrites plus une classe neutre, introduite lorsqu'aucune des classes définies n'est présente). Les activités "Travailler" et "Déjeuner" représentent 25% de l'ensemble de données, tandis que l'activité "Débarrasser" ne représente que 5% de la base de données DAHLIA.

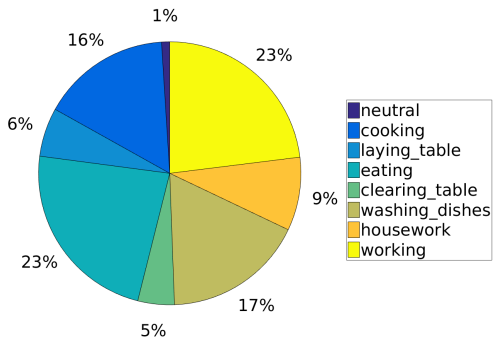


Figure 2.9 – Proportion de chaque classe d'activité sur l'ensemble du jeu de données DAHLIA.

29 hommes et 15 femmes âgés de 23 à 61 ans ont participé aux acquisitions ce qui augmente également la variabilité intra-classe.

2.2.2 . Environnement et paramètres d'acquisitions

Dans le contexte de maison intelligente, les équipements doivent être aussi génériques que possible et faciles à mettre en place. Il faut donc éviter de fixer des capteurs sur différents objets, ainsi que d'utiliser des capteurs intrusifs sur les personnes ou des capteurs nécessitant un calibrage compliqué. Les capteurs de profondeur abordables comme la Kinect conviennent bien à ce type d'application puisqu'ils peuvent être installés comme le serait une caméra classique. L'ensemble des données a donc été acquis dans une cuisine entièrement contrôlée, équipée de 3 Kinect v2 comme indiqué sur la figure 2.10. Plusieurs occultations peuvent apparaître selon la position de la personne et des objets présents sur la table.

Chaque acquisition est composée des quatre flux suivants enregistrés à une fréquence d'images de 15 fps :

- **Les vidéos RGB** enregistrées en haute résolution (1920×1080 pixels) et compressées avec le format H.264 à un débit de 2 Mbits/s.
- **Les cartes de profondeur** avec une résolution de 512×424. Pour chaque pixel, la valeur codée sur 16 bits représente la distance entre le capteur et le point correspondant. Nous avons appliqué un filtre passe-bas consistant en trois filtres médians (sur les dimension x, y et temporelle).

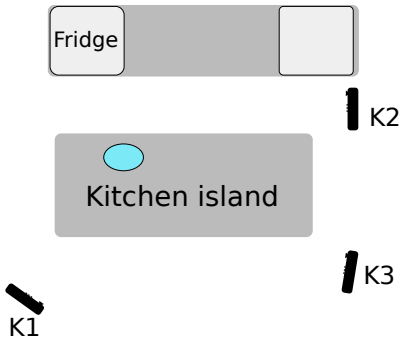


Figure 2.10 – Position des caméras pour l'enregistrement du jeu de données DAHLIA

La valeur filtrée est conservée si elle diffère de la valeur originale pour plus de d=10cm.

- **Les données du squelette** extraites à l'aide du SDK associé au capteur Kinect v2. Le squelette est constitué de la position 3D de 25 articulations du corps humain qui sont listées figure 2.11. Le capteur renvoie les coordonnées des articulations dans deux espaces différents (l'espace de la carte de profondeur et l'espace ponctuel 3D) et des informations sur l'état du suivi. Cet état de suivi vaut 0 si l'articulation n'est pas suivie, 1 si elle est déduite et 2 si elle est suivie. Enfin, toutes les fausses détections qui auraient pu se produire ont été supprimées manuellement pour assurer un unique squelette à chaque instant de la vidéo.
- **Une carte de segmentation des personnes** dans le point de vue du capteur de profondeur est également enregistrée.

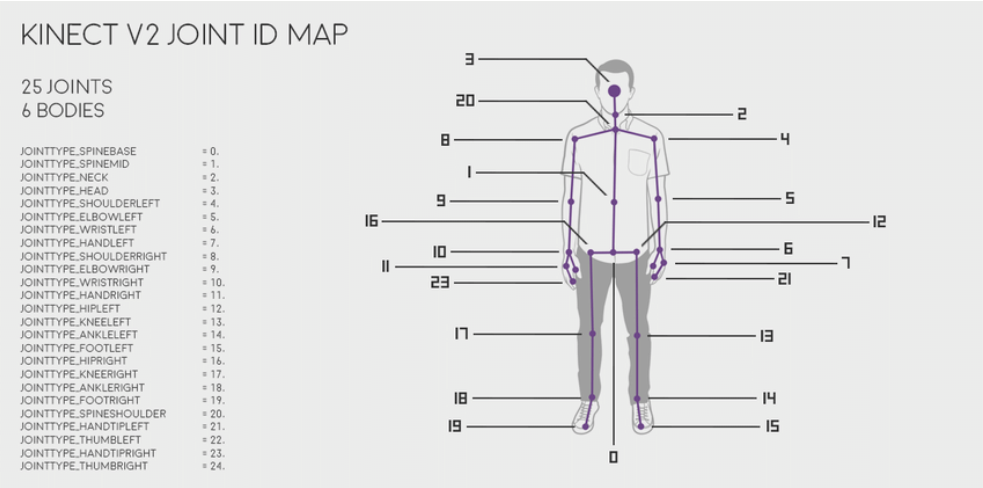
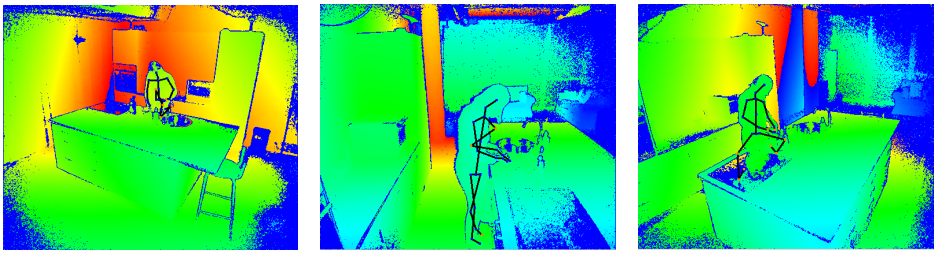


Figure 2.11 – Listes des articulations extraites par la Kinect V2

Comme le squelette est extrait du capteur Kinect, la qualité de l'estimation de la localisation des articulations du corps est irrégulière. En outre, en raison de la configuration de la cuisine, la partie inférieure du corps n'est généralement pas suivie, car occultée par l'îlot central de la cuisine.

Comme tous les exemples ont été capturés dans la même cuisine, nous avons ajouté de la variance en demandant aux acteurs de réaliser les actions à différents endroits de la scène. Un exemple extrait de l'ensemble de données est présenté sur la figure 2.12, avec les vues RGB et les cartes de profondeur provenant des trois caméras. Le squelette est superposé aux cartes de profondeur. L'estimation de la pose du squelette est de qualité médiocre en cas d'occultations importantes, comme sur les figures 3d et 3f. La figure 2.13 propose d'autres exemples issus de DAHLIA pour la modalité RGB et les activités "manger" et "faire la vaisselle".

(a) Camera 1 (b) Camera 2 (c) Camera 3



(d) Profondeur cam 1 (e) Profondeur cam 2 (f) Profondeur cam 3

Figure 2.12 – Exemple du jeu de données DAHLIA pour l'activité "cuisiner"

Figure 2.13 – Autres exemples du jeu de données DAHLIA pour la modalité couleur et les activités "manger" et "faire la vaisselle".

2.2.3 . Annotations et publication du jeu de données

Les 51 longues séquences de la base de données DAHLIA sont composées de 7 activités de la vie quotidienne. Après un travail de synchronisation des vues, chaque séquence a été étiquetée manuellement avec le début et la fin des 7 activités ou le label neutre. La figure 2.14 visualise l'annotation d'une séquence sous forme de chronogramme.

Dans l'objectif de rendre DAHLIA disponible à la communauté scientifique, chaque participant a signé une autorisation de réalisation, de reproduction et de représentation de son image. DAHLIA a ensuite été déclaré à la CNIL (Commission Nationale Informatique et Liberté) pour s'assurer que les droits à l'image sont bien respectés. DAHLIA est aujourd'hui hébergé sur un serveur

Figure 2.14 – Exemple de l'annotation d'une séquence de DAHLIA avec les temps de début et de fin de chaque activité.

du CEA en accès restreint. Si un chercheur souhaite obtenir l'accès, il doit préalablement avoir complété et signé un contrat de licence qui l'engage à n'utiliser DAHLIA que pour des travaux de recherche et non à des fins commerciales.

2.2.4 . Protocole d'évaluation

Afin de permettre des comparaisons de performance équitables, nous définissons deux protocoles d'évaluation précis :

a) L'évaluation inter-sujets

Nous avons défini deux groupes de participants : le groupe A et le groupe B, mis à disposition avec le jeu de données. Les deux représentent environ la moitié du jeu de données et doivent être utilisés alternativement comme ensemble d'entraînement et ensemble de test. Le résultat de l'évaluation inter-sujets est alors la performance moyenne des deux configurations.

b) L'évaluation inter-sujets et inter-vues

Pour contrebalancer le manque de variations de l'environnement, nous avons également défini un protocole d'évaluation en vue croisée. Dans ce protocole, un ensemble d'entraînement est défini pour chaque combinaison de

vue et de groupe de sujets (A et B) tandis que l'ensemble de test doit être effectué sur l'ensemble restant. Par exemple, l'un des ensembles d'entraînement serait formé sur la vue 1, groupe A et testé sur la vue 2, groupe B et la vue 3, groupe B indépendamment. Ensuite, la moyenne des 12 ensembles ainsi définis est retenue.

2.2.5 . Comparaison avec les bases de données existantes

Zhang et al. [ZLO+16] proposent une étude sur les bases de données de reconnaissance d'action basées sur la technologie RGB-D en introduisant certains critères tels que :

Les caractéristiques d'acquisition du jeu de données :

- Mode 1 : chaque action est stockée dans une séquence.
- Mode 2 : chaque séquence contient un ensemble continu d'actions.
- Mode 3 : chaque séquence contient un ensemble continu d'actions réalisées avec le même ordre.
- Mode 4 : chaque séquence contient un ensemble continu d'actions réalisées dans un ordre aléatoire.

Le type d'arrière plan et les occultations

- Faible : le fond est fixé et propre. Il n'y a pas d'occultations.
- Moyen : l'arrière-plan est fixé mais est encombré. Certaines occultations peuvent apparaître.
- Élevé : l'arrière-plan n'est pas fixé et/ou est encombré. Des occlusions sont présentes et peuvent affecter l'action.

La complexité cinématique

- Faible : les mouvements sont simples et de courte durée.
- Moyen : les mouvements sont de complexité moyenne et la durée est plus longue.
- Élevé : les mouvements sont complexes, avec une longue durée.
- Très élevé : les mouvements sont très complexes et composés de plusieurs sous-actions.

La variabilité entre les actions

- Faible : la variation des niveaux de complexité entre les actions dans un ensemble de données est faible.
- Moyen : la variation des niveaux de complexité entre les actions dans un ensemble de données est moyenne.
- Élevé : la variation des niveaux de complexité entre les actions dans un ensemble de données est élevée.

Le tableau 2.1 résume ces différentes caractéristiques évaluées sur plusieurs jeux de données déjà publiés au moment des travaux. Il est inspiré des tableaux présentés dans [ZLO+16] avec des critères adaptés. DAHLIA a donc été évalué selon ces critères : le fond est stable et encombré pour toutes les actions réalisées. L'emplacement des caméras et la configuration de la cuisine conduisent à des occultations partielles des sujets. DAHLIA contient des séquences très longues composées de plusieurs activités qui peuvent se dérouler dans un ordre variable en fonction du sujet. Ces activités impliquent de nombreuses interactions avec des objets de la scène et les mouvements sont très complexes. Par conséquent, la complexité cinématique est évaluée comme étant "très élevée" et la variabilité entre les actions comme étant "élevée".

DAHLIA a été réalisé par 44 sujets, ce qui est le plus grand nombre de sujets présent dans une base de données d'actions, avec 7 activités de haut niveau sémantique, contrairement aux jeux de données précédents qui traitaient d'actions de bas niveau. Il est important de noter que si 7 classes est l'un des nombres de classes les plus bas, les activités pourraient être décomposées en de nombreuses sous-actions, conduisant à une complexité cinématique très élevée. Ce haut niveau sémantique donne lieu à des activités d'une durée d'environ 6 minutes qui est la plus longue durée moyenne des bases de données analysées, et ce dans des vidéos non découpées (Mode 4).

Ainsi, DAHLIA est composé d'activités plus longues, avec un niveau sémantique plus élevé et une variabilité à la fois inter et intra-classe élevée.

2.2.6 . Métriques d'évaluation

Afin d'évaluer et de classer les algorithmes de détection d'activités, des métriques précises sont utilisées et dépendent fortement de l'application finale. Plusieurs métriques ont été définies dans des travaux antérieurs [KGS+05, Mun57, CZT05, YF05, RAAS12, WLT06, WLM+14] pour évaluer et/ou comprendre les performances des algorithmes.

Nous présentons ici les métriques sur lesquelles le jeu de données DAHLIA a été évalué. Pour chaque classe c du jeu de données, sont définis TP^c, FP^c, TN^c et FN^c comme le nombre d'images vraie positive, fausse positive, vraie négative et fausse négative.

Jeu de données	Année	#Sujet	#Action	Durée moyenne	Modalités	#Vue	Caractéristiques d'acquisition	Arrière plan	Cinématique	Variabilité	Niveau sémantique
TUM [TBB09]	2009	4	9	2	C,S	4	Mode 4	Faible	Moyen	Faible	Gestes
MSR-Action 3D [ZG10]	2010	10	20	2.8	D,S	1	Mode 1	Faible	Faible	Faible	Actions
CAD-60 [SPSS11]	2011	4	12	45	C,D,S	-	Mode 1	Moyen	Moyen	Moyen	Actions
RGBD-HuDaAct [NWM11]	2011	30	12	-	C,D	1	Mode 1	Moyen	Moyen	Moyen	Actions
MSRDaily-Activity3D [WLWY12]	2012	10	16	-	C,D,S	1	Mode 1	Moyen	Élevé	Faible	Actions
UTKinect [XCA12]	2012	10	10	1	C,D,S	1	Mode 3	Moyen	Moyen	Faible	Actions
ACT4 Dataset [CQY+12]	2012	24	14	4	C,D,S	4	Mode 1	Faible	Faible	Moyen	Actions
MPII Cooking Activities [RAAS12]	2012	12	65	6	C,S	1	Mode 4	Moyen	Moyen	Moyen	Actions
CAD-120 [KGS13]	2013	4	10	17	C,D,S	1	Mode 2	Élevé	Élevé	Élevé	Actions
UCFKinect [EMT+13]	2013	16	16	2	S	1	Mode 1	Aucun	Faible	Faible	Actions
3D Online [YLY14]	2014	36	7	3	C,D,S	1	Mode 1/4	Moyen	Moyen	Faible	Actions
Northwestern UCLA [WNX+14]	2014	10	10	-	C,D,S	3	Mode 1	Faible	Moyen	Moyen	Actions
RGB-D activity [WZSS15]	2015	7	21	5	C,D,S	1	Mode 4	Élevé	Élevé	Élevé	Actions
UTD-MHAD [CJK15]	2015	8	27	2	C,D,S	1	Mode 1	Faible	Faible	Faible	Actions
UWA 3D Multiview Dataset [RMHM16]	2016	10	30	-	C,D,S	4	Mode 1	-	-	-	Gestes
NTU RGB+D [SLNW16]	2016	40	60	5s	C,D,S	80	Mode 1	Faible	Moyen	Moyen	Actions
DAHLIA	2016	**44**	7	**6 min**	C,D,S	3	**Mode 4**	**Élevé**	**Très Élevé**	**Élevé**	**Activités**

Table 2.1 – Jeu de données existants avec leurs nom et référence | L'année de publication du jeu | Le nombre de sujet réalisant les actions | Le nombre d'action | La durée moyenne de chaque action | Les modalités fournies dans le jeu : Couleur, Profondeur et/ou squelette | Le nombre de vue | Les caractéristiques d'acquisition | Le type de fond et les occultations | La complexité cinématique | La variabilité entre chaque action | Le niveau sémantique des classes d'action.

a) La précision par image

Une métrique classique est la précision par image qui représente le rapport entre les images correctement classifiées et toutes les images de l'ensemble de données (2.1). Cette métrique est sensible à la distribution des classes mais fournit une mesure intuitive de la capacité de l'algorithme à reconnaître les actions.

$$FA_1 = \frac{\sum_{c \in C} TP^c}{\sum_{c \in C} N_c} \tag{2.1}$$

où N_c est le nombre d'images labellisées avec l'activité c dans la vérité terrain.

b) Le F-Score

Cette métrique combine la précision P^c et le rappel R^c pour chaque classe c et est défini par la moyenne harmonique de ces deux valeurs :

$$\mathcal{P}^c = \frac{TP^c}{TP^c + FP^c} \qquad \mathcal{R}^c = \frac{TP^c}{TP^c + FN^c} \tag{2.2}$$

$$\textit{F-Score} = \frac{2}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \times \frac{\mathcal{P}^c \times \mathcal{R}^c}{\mathcal{P}^c + \mathcal{R}^c} \tag{2.3}$$

Elle présente l'avantage de donner une importance égale à la précision et au rappel, deux grandeurs significatives de la capacité d'un algorithme à différencier des classes.

c) L'intersection sur l'union (IoU)

Cette métrique classique a été utilisée pour évaluer la segmentation dans le défi PVOC [EEVG+15]. Elle est définie comme le ratio de l'intersection sur l'union des prédictions avec les cibles de la vérité terrain :

$$IoU = \frac{1}{|C|} \sum_{c \in C} \frac{TP^c}{TP^c + FP^c + FN^c} \tag{2.4}$$

2.3 . Proposition de résultats de référence

Afin de fournir des résultats de référence sur le jeu de données DAHLIA, nous avons évalué trois algorithmes conçus pour la détection et/ou la reconnaissance d'actions en ligne, à savoir l'algorithme DOHT (*Deeply Optimized Hough Transform*) [CHTAL13], l'algorithme ELS (*Efficient Linear Search*) [16] et l'algorithme *Max-Subgraph search* [6].

2.3.1 . Algorithme DOHT (*Deeply Optimized Hough Transform*)

L'algorithme DOHT a été développé au laboratoire. Il utilise un paradigme de vote par transformée de Hough appliqué à des descripteurs précédemment extraits selon des intervalles de temps définis sur l'ensemble du clip vidéo. Ces descripteurs sont également encodés selon un dictionnaire généré par un algorithme K-means. Le DOHT a l'avantage d'être une méthode en ligne donc adaptée pour la tâche de détection, en plus d'être une méthode générique car agnostique aux descripteurs d'entrée.

Nous présentons des résultats calculés avec l'algorithme DOHT avec différents descripteurs : les trajectoires des articulations du squelette et des trajectoires denses comme le fait [VAL16]. En ce qui concerne les paramètres de vote, nous avons gardé les même valeurs pour les deux descripteurs. Étant donné que les activités contenues dans DAHLIA sont beaucoup plus longues que celles utilisées dans la proposition initiale de l'algorithme DOHT [CHTAL13], nous avons fixé la taille de la fenêtre temporelle pour les votes de Hough M à 1000 et le paramètre C (attachement des données) à 4, ce qui a donné les meilleurs résultats. L'entraînement multi-vues est réalisé en considérant les descripteurs issus des 3 vues.

a) Squelette 3D

Pour exploiter les articulations du squelette 3D extraites par la Kinect, les données brutes sont d'abord normalisées en appliquant une normalisation similaire à celle présentée par Raptis et al. dans [RKH11]. Cette normalisation est robuste à la variation du point de vue et ignore volontairement l'emplacement de la personne dans la pièce puisque nous ne voulons pas que l'algorithme bénéficie de cette information pour reconnaître les activités.

En raison de la configuration de la pièce et de l'emplacement des caméras (indiqués dans la figure 2.10), les occultations se produisent différemment sur chaque caméra. Pour éviter ces problèmes d'occultation, nous avons combiné les informations de toutes les vues en suivant le processus de [VAL16]. Les articulations utilisées sont les suivantes : la tête, les deux mains, les coudes, le bout de la main, le poignet, les épaules, les hanches, les genoux et la base de la colonne vertébrale. Dans chaque image, nous ne considérons que les articulations associées à l'état "suivi" (fourni par le capteur). Les images où les épaules ont un faible taux de confiance ne participent pas aux votes car la stratégie de normalisation est basée sur les épaules.

b) Trajectoires denses

Selon [VAL16], l'algorithme DOHT a également été appliqué sur les données RGB de DAHLIA. Les descripteurs issus de la forme des trajectoires denses (Traj) et les descripteurs HOG (*Histogram of Gradient*) [WKSL11] ont été utilisés, ainsi que la concaténation Traj+HOG comme présentée dans [VAL16].

c) Résultats de détection d'activités avec l'algorithme DOHT

Le tableau 2.2 résume les résultats obtenus avec l'algorithme DOHT. Dans chaque configuration, l'algorithme donne les meilleures prédictions

lorsqu'il est utilisé avec les caractéristiques HOG. Cela souligne l'importance du contexte spatial dans le processus de détection d'activité puisque le descripteur HOG capture la forme locale de l'image. L'utilisation de vues multiples pendant l'apprentissage et le test permet d'obtenir des résultats plus élevés, car les données extraites de différentes vues se complètent. En effet, une occultation dans une vue peut être compensée par les observations d'une autre vue.

	Squelette			**Trajectoires**			**HOG**			**Traj+HOG**		
	$\mathcal{FA}_1$	F-Score	IoU	$\mathcal{FA}_1$	F-Score	IoU	$\mathcal{FA}_1$	F-Score	IoU	$\mathcal{FA}_1$	F-Score	IoU
Vue 1	0.60	0.58	0.42	0.74	0.73	0.58	0.80	0.77	0.64	0.73	0.73	0.59
Vue 2	0.63	0.60	0.44	0.78	0.76	0.62	0.81	0.79	0.66	0.79	0.78	0.64
Vue 3	0.73	0.71	0.56	0.76	0.74	0.59	0.80	0.77	0.65	0.77	0.76	0.62
Multi-vues	0.77	0.75	0.60	0.81	0.80	0.67	0.85	0.82	0.71	0.82	0.80	0.68
Vues croisées	0.34	0.31	0.19	descripteur dépendant de la vue								

Table 2.2 – Résultat de l'algorithme DOHT sur le jeu de données DAHLIA

Les résultats sont plus faibles dans le protocole à vue croisée, ce qui est cohérent avec le fait que les caractéristiques extraites sur une vue peuvent être occultées dans une autre. Pour l'algorithme DOHT, ce protocole est le plus difficile malgré une normalisation des squelettes. Le descripteur issu des trajectoires denses étant dépendant de la vue, nous n'avons pas exécuté le protocole inter-vues pour ces caractéristiques.

Enfin, le tableau 2.3 présente les résultats par classe avec le descripteur HOG qui est celui donnant les meilleurs résultats. Les classes les mieux reconnues sont "Manger" et "Travailler" car ce sont les classes avec le moins de variabilité intra-classe. En effet, les personnes sont assises au même endroit et effectuent les mêmes gestes. Au contraire, "mettre la table" et "débarrasser" sont les classes les moins bien reconnues car elles sont confondues entre elles. Les gestes sont les mêmes et uniquement l'ordre dans lequel ils sont réalisés permet de distinguer ces deux activités. Ce sont également les deux classes les moins bien représentées du jeu de données.

2.3.2 . Algorithme ELS (*Online Efficient Linear Search*)

Meshry et al. [MHT16] ont proposé une méthode de détection d'action basée sur des séquences de squelettes 3D. Un dictionnaire est généré à partir des caractéristiques du squelette et un SVM linéaire permet d'apprendre des poids pour chaque mot du dictionnaire. Ensuite, la reconnaissance en ligne se fait par le maximum du score par classe basé sur ces poids.

	Multicam HOG	
	F-Score	IoU
Cuisiner	0.75	0.60
Mettre la table	0.69	0.53
Manger	0.91	0.84
Débarrasser	0.75	0.59
Faire la vaisselle	0.87	0.77
Faire le ménage	0.86	0.75
Travailler	0.92	0.86

Table 2.3 – Résultats par classe obtenus avec l'algorithme DOHT suivant le protocole inter-sujets en utilisant le descripteur HOG avec une approche multi-vues.

L'algorithme ELS en ligne a été calculé avec le descripteur local décrit dans l'article original, soit, la concaténation pondérée de l'angle θ [NS12], sa vitesse $\delta\theta$ et une adaptation du descripteur *Moving Pose* [ZLS13] P, de ses dérivées première et seconde, respectivement δP et $\delta^2 P$. La forme finale de leur descripteur est [P, $\alpha\delta P$, $\beta\delta^2 P\psi\theta$, $\delta\theta$] avec α, β et ψ trois paramètres de pondération. Nous avons évalué plusieurs ensembles de paramètres et présentons les meilleurs résultats que nous avons obtenus sur le jeu de données DAHLIA : $\alpha = 0, 1, \beta = 0, 1, \psi = 0, 1$. Nous avons fixé les paramètres de latence à 2 images et gardé les autres paramètres identique à ceux proposés par l'article original. Les résultats obtenus avec cet algorithme sont présentés dans le tableau 2.4 dans le protocole inter-sujet et mono-vue ou vues-croisées. Nous pouvons observer que ces résultats dépendent fortement de la vue considérée. Plus précisément, des résultats plus élevés ont été obtenus lorsque la caméra 3 est utilisée. Cette caméra est celle qui présente le moins d'auto-occultation en raison de son emplacement et de son angle par rapport à la scène.

Les performances sont présentés tableau 2.4. Comme avec la méthode DOHT, les résultats les plus faibles sont obtenus avec le protocole inter-vues et inter-sujets puisqu'un changement de point de vue affecte fortement les descripteurs extraits. Remarquons que les résultats varient fortement d'une vue à l'autre. La vue sur laquelle les prédictions sont les moins erronées est la vue 3, vue la moins propice aux auto-occultations par sa position dans la scène.

2.3.3 . Algorithme *Max-Subgraph Search*

Chen et Grauman [CG12] ont proposé une méthode de détection d'action basée sur une recherche de max-sous-graphe nommée *T-Jump-Subgraph* où

	Squelette		
	$\mathcal{FA}_1$	F-Score	IoU
Vue 1	0.18	0.18	0.11
Vue 2	0.27	0.26	0.16
Vue 3	0.52	0.55	0.39
Vues croisées	0.31	0.32	0.21

Table 2.4 – Résultats de l'algorithme ELS sur le jeu de données DAHLIA

un graphe est construit pour chaque action à détecter. Chaque nœud de ce graphe est associé à un score calculé à partir de descripteurs extraits sur une fenêtre temporelle. Les poids des descripteurs sont estimés par un SVM de manière similaire à [MHT16]. Puisque cet algorithme a été conçu pour faire de la détection plutôt que de la reconnaissance en ligne, plusieurs activités peuvent être détectées dans chaque image, la précision FA_1 par image ne peut donc pas être calculée. Les résultats sont présentés dans le tableau 2.5, en utilisant le même descripteur local que ELS et une taille de nœud fixée à 100.

	Squelette	
	F-Score	IoU
Vue 1	0.25	0.15
Vue 2	0.18	0.10
Vue 3	0.44	0.31

Table 2.5 – Résultats de l'algorithme Max-subgraph Search sur le jeu de données DAHLIA

2.4 . Conclusion et perspectives

Dans ce chapitre, nous avons présenté DAHLIA, le premier jeu de données de détection d'activités longues enregistré dans un environnement réel. Les vidéos de DAHLIA sont non segmentées et synchronisées depuis 3 points de vue différents. 44 personnes ont participé aux acquisitions de DAHLIA ce qui en fait le jeu de données présentant le plus grand nombre de sujets différents. La durée moyenne des activités est de 6 minutes environ, ce qui est significativement plus long que celles des ensembles de données existants non découpés. Les séquences vidéos ont été capturées de manière réaliste, avec des instructions très simples pour augmenter la variabilité intra-classe ainsi que la cinématique, ce qui fait de DAHLIA un jeu de données très intéressant pour la communauté scientifique qui travaille sur la problématique de détection d'activités longues.

Les statistiques actuelles sur l'utilisation de DAHLIA confirment l'intérêt porté par la communauté scientifique et le manque qui existait avant sa publication. En effet, depuis sa mise en ligne, DAHLIA a été téléchargé 23 fois par des chercheurs internationaux et 5 publications proposent des résultats évalués sur DAHLIA [NGA+18, DMG+19, NB19, AAC+20]. Ces récents travaux utilisent bien sûr des réseaux de neurones tel que [DMG+19] qui propose un nouveau module d'attention temporelle intégré dans une architecture encodeur décodeur.

D'une part, bien que nous ayons proposé le plus de variation possible dans la façon de réaliser les activités, la limite la plus importante de DAHLIA est le lieu unique. Pour pouvoir conserver les performances du modèle en le déployant dans un autre environnement, des caractéristiques indépendantes du lieu et du point de vue sont nécessaires.

Le squelette 3D retourné par la Kinect répond à ce problème mais aujourd'hui Microsoft ne maintient plus ce capteur, son utilisation est donc obsolète. Cependant les recherches autour de l'estimation du squelette 3D à partir d'une seule image RGB ont beaucoup progressé et des méthodes telles que [BCL+20] pourront bientôt obtenir des performances équivalentes au capteur Kinect. De plus, leurs performances peuvent être améliorées si on considère une estimation en fusionnant différents points de vue.

Les méthodes d'adaptation de domaine peuvent également permettre d'être indépendant au point de vue. Nous avons exploré l'adaptation de la méthode [PR21] au cas de DAHLIA et avons pu montrer un gain lors de l'évaluation inter-vues en utilisant le réseau de classification d'actions X3D [Fei20].

D'autre part, le travail d'annotation est une tâche fastidieuse et chronophage. Les activités annotées dans DAHLIA sont très intéressantes mais restent de haut niveau sémantique. Avec l'évolution des modèles qui pourront être utilisés pour l'annotation automatique ou interactive, on pourrait plus facilement compléter DAHLIA avec des annotations plus riches et plus précises comme les boîtes d'objet, la segmentation d'instances, les sous-actions et interactions, un squelette 3D plus précis, etc.

Mais pourrons-nous un jour, dans l'idéal, automatiser intégralement la création d'un jeu de données ? Le développement des approches de génération de données pourrait répondre à cette question. La création d'une base de données serait entièrement contrôlée et automatisée depuis la définition des scénarios jusqu'à l'annotation des séquences. De plus, la génération des données permettrait de produire facilement de nouveaux points de vue, de changer l'apparence des personnes et leur façon de réaliser une activité pour obtenir un jeu de données avec une grande variabilité.

3 - Détection d'interactions : CALIPSO, un réseau rapide en une étape

Dans l'objectif de mieux gérer la grande variabilité de réalisation des activités longues, qui ont fait l'objet des travaux du chapitre 2, il est intéressant de changer l'approche holistique analysant la vidéo dans sa globalité pour adopter une méthode qui modélise de manière explicite les sous-actions composant l'activité. L'intérêt est de reconnaître une activité par le sous-ensemble d'actions qui la compose permettant ainsi de s'affranchir de l'ordre d'exécution de ces sous-actions. C'est ce qui m'a motivé à étudier la détection des interactions entre les personnes et les objets dans une image pour analyser plus finement les activités.

Dans ce chapitre, nous dressons un état de l'art des méthodes d'apprentissage supervisées appliquées à la détection d'interactions et un panorama des jeux de données disponibles, puis nous présentons en détail la méthode CALIPSO (*Classifying ALl Interacting Pairs in a Single shOt*) et les résultats obtenus.

Sommaire

3.1 . État de l'art de la détection d'interactions dans des images

3.1.1 . Principe

Plusieurs tâches de vision par ordinateur abordent le problème de la compréhension du contenu sémantique des images, comme la reconnaissance de relations visuelles qui a pour but de comprendre le positionnement des objets les uns par rapport aux autres dans une scène. Plus spécifique que la relation visuelle, la détection de l'interaction humain-objet (HOI) vise à détecter ce qui se passe et où cela se passe dans l'image en accordant une attention exclusive aux interactions centrées sur les personnes. La détection des HOI est un problème difficile, essentiel pour diverses applications telles que la compréhension des activités, la surveillance, l'assistance à l'autonomie à domicile, la cobotique, etc. Dans le cas des systèmes de surveillance, il est particulièrement intéressant de comprendre rapidement les interactions centrées sur l'humain. Comme les images peuvent contenir un grand nombre de personnes et d'interactions, il est essentiel qu'une méthode de détection des HOI soit scalable en fonction du nombre d'objet et d'interactions visibles. Ce problème de passage à l'échelle a motivé notre travail. Dans ce qui suit, les "objets" affectés de la classe humaine sont appelés *sujets* tandis que ceux de la classe non-humaine sont appelés *cibles*.

La détection des HOI consiste à déterminer et à localiser la liste de triplets $< sujet, verbe, cible >$ décrivant toutes les interactions visibles dans l'image. Bien que la détection des HOI ait été classiquement basée sur la vidéo (en général, avec un focus sur une seule action), des approches récentes basées sur une seule image ont montré des résultats impressionnants sur la détection d'interactions simultanées. D'une manière générale, la tâche de détection des HOI basée sur l'image est réalisée en résolvant les sous-tâches suivantes : détection des objets en interaction (le *problème de détection des objets*), association correcte de ces objets (le *problème d'association*) et classification des interactions (le *problème de classification des verbes*). La figure 3.1 illustre le principe de la détection des HOI.

Dans la suite, les interactions entre une personne et un objet sont nommées HOI alors que les interactions entre deux personnes sont appelées HHI.

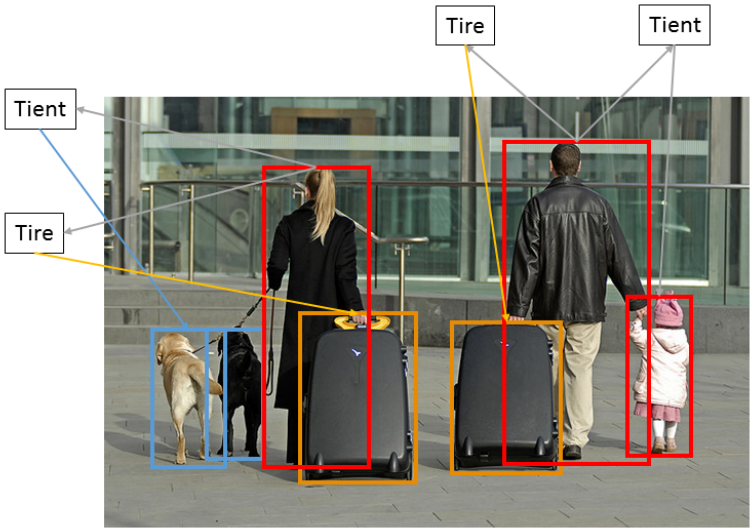


Figure 3.1 – Principe de la détection des interactions : détecter les objets de la scène, association de ces objets et classification des interactions. Image issue du jeu de données V-COCO [GM15]

3.1.2 . Méthodes de détection d'interactions

Méthodes à deux étapes

La plupart des approches de détection d'interaction s'appuient sur un détecteur d'objets qui identifie les paires de candidats sujet-cible, ces boîtes sont ensuite traitées dans une deuxième étape pour évaluer la présence et le type d'interaction.

Gupta et al. [GM15] détectent successivement un sujet, classifient l'action et associent la cible en fonction d'un score d'interaction. Plusieurs approches [CLL+18, GZH18, GGDH18, LZH+19, UIM20] construisent sur un modèle de détecteur d'objet, à savoir Faster R-CNN [RHGS15], avec des branches supplémentaires soit pour prédire les actions et l'estimation d'une densité de probabilité sur l'emplacement de l'objet cible pour chaque action [GGDH18], soit pour prédire les relations spatiales des paires humain-objet [CLL+18], soit une mesure d'attention centrée sur l'instance [GZH18], ou bien le filtrage des paires humain-objet non-interactives en croisant les données d'apprentissage de différents jeux de données [LZH+19]. Qi et al. [QWJ+18] présentent un cadre générique combinant des graphes et un réseau neuronal profond, capturant les interactions humain-objet de manière itérative. Li et al. [LOW17] introduisent une communication inter-branches avec un message guidé par une phrase pour assurer une modélisation conjointe de la classification des

actions et de l'association des cibles.

Plus récemment, des améliorations de la deuxième étape ont été proposées en utilisant des informations supplémentaires dans l'image. Par exemple, [WZL+19, LYC20] utilisent la pose humaine pour avoir une analyse plus fine de la posture du sujet de l'interaction. D'autres méthodes ajoutent l'encodage de mots [SHRW20, LKBFF16] ou la segmentation [ZWQ+20]. Enfin, [SHRW20] combine tous ces types d'informations supplémentaires. Certaines techniques [BRSC20, XWL+19, YLMD17] intègrent des *connaissances linguistiques* pour résoudre le problème de la distribution à longue queue des classes d'interaction humain-objet. En effet certaines classes sont sur-représentées par rapport à beaucoup d'autres ne présentant que peu d'exemples. Ils exploitent les informations contextuelles présentes dans les à priori linguistiques appris avec un réseau "word2vec", pour généraliser les interactions entre des objets ayant des fonctions similaires. Alternativement, Peyre et al. [PLSS18] apprennent une représentation visuelle des relations combinant le sujet, la cible et le prédicat avec une représentation visuelle des phrases pour la détection des HOI.

Finalement, toutes ces méthodes ont un traitement en deux temps avec une deuxième étape basée sur les paires précédemment détectées, leur temps de calcul est donc quadratique avec le nombre d'instances dans l'image, ce qui peut poser un problème de scalabilité lorsque l'image présente un grand nombre d'instances d'objets et d'interactions.

Certaines méthodes offrent une alternative à l'étude de toutes les paires possibles et accélèrent ainsi le temps d'inférence. Par exemple, [LLW+20, WYD+20] modélisent le problème de détection des interactions comme un problème de détection de points clés.

D'autres méthodes utilisent l'apprentissage par métrique pour estimer un espace de représentation où les instances en interaction seront proches les unes des autres sans étudier toutes les paires possibles. L'apprentissage par métrique a été appliqué à de nombreuses tâches différentes, de la recherche d'images [FSSM07] à la reconnaissance des visages [SKP15]. En plus de fournir une mesure de similarité pour comparer des images, il peut également être utilisé pour transformer des caractéristiques visuelles et textuelles dans un espace commun [FCS+13, KFF17], les vecteurs de représentation, ou rassembler des caractéristiques visuels pour reconnaître un groupe d'éléments. Par exemple, Newell et Deng [NHD17] ont proposé un espace de représentation associatif pour regrouper les articulations du corps afin d'estimer la pose humaine. L'apprentissage par métrique est également

appliqué à la détection de relations visuelles [ND17, PLSS18, ZSE+19]. En particulier, Pixel2Graphs [ND17] produit en une seule passe un ensemble d'objets et de liens d'interaction représentés par un graphe qui est déduit de deux cartes de chaleur. Puis, dans un deuxième temps, chacune de ces caractéristiques d'objets ou de liens est passée dans un réseau entièrement connecté pour prédire les propriétés d'interaction (verbe, association sujet-cible, classe d'objet et boîte englobante). La figure 3.2 détaille le principe de la méthode. La deuxième étape est une fois de plus dépendante du nombre d'interactions. De plus, le fait qu'un nombre fixe d'interactions est pré-supposé à chaque emplacement de l'image peut être limitant pour les images densément peuplées car plusieurs relations peuvent être localisées au même endroit.

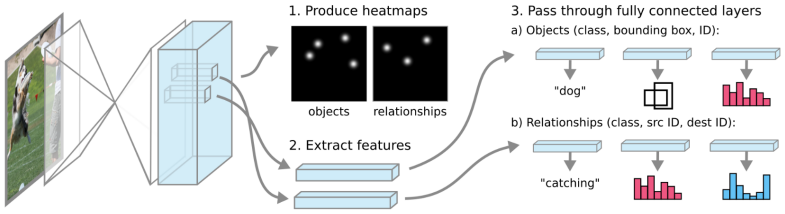


Figure 3.2 – Principe de la méthode Pixel2Graph, figure extraite de l'article [ND17]. Un premier réseau est appris pour estimer deux cartes de chaleur qui s'activent pour l'une, à l'emplacement des objets et pour l'autre à la position centrale entre deux objets, elle représente donc les interactions. Ensuite les vecteurs de caractéristiques sont extraits des points chauds et donnés en entrée d'un second réseau qui prédit la nature de l'objet ou de la relation ainsi qu'un vecteur de représentation qui permet de relier les objets lors de l'inférence.

Méthodes à une étape

Certains travaux récents proposent des détecteurs de HOI à une étape. Les méthodes UnionDet [KCKK20] et DIRV [FXSL21] s'appuient uniquement sur la régression et la classification pour prédire les interactions. Ils intègrent tous deux une branche de détection d'instance similaire aux détecteurs d'objets classiques. La détection des interactions est ensuite basée sur l'union des boîtes englobantes régressées [KCKK20]. [FXSL21] note qu'il est préférable de se concentrer sur les régions d'interaction plutôt que sur la boîte d'union entière qui présente trop d'informations inutiles. Par conséquent, [FXSL21] propose une branche centrée sur les régions d'interaction pour détecter l'interaction. Ces deux méthodes initialisent leur détecteur d'objets interne avec un modèle pré-entraîné, puis figent ces poids pour

apprendre la branche d'interaction.

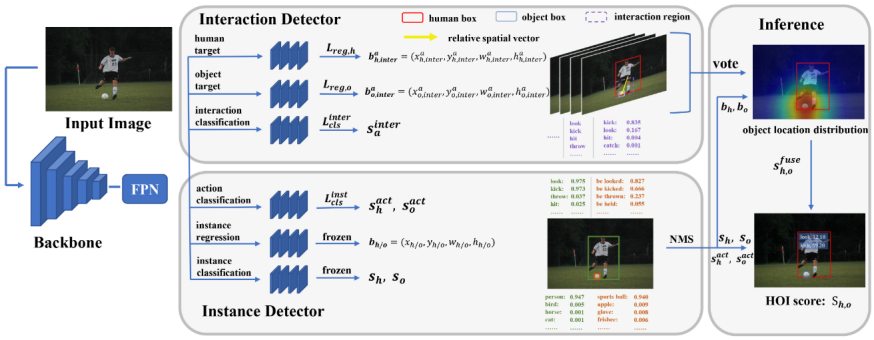


Figure 3.3 – Principe de la méthode DIRV, figure extraite de l'article [FXSL21]. L'architecture de la méthode DIRV est composée d'un détecteur d'interaction et d'un détecteur d'instances. La branche de détection d'instances est basée sur un détecteur d'objet classique auquel a été ajouté une sortie qui estime la probabilité de réalisation de chaque verbe d'interaction. La branche d'estimation des interactions introduit le concept de région d'interaction qui sont des parties des boîtes union de paire d'instances. 3 sorties sont alors estimées, la régression des régions d'interactions pour obtenir la boîte du sujet, la régression des régions d'interactions pour obtenir la boîte de l'objet et enfin la probabilité des verbes d'interaction pour chaque région d'interaction.

Ces méthodes sont plus efficaces en terme de temps de calcul puisque leur complexité est indépendante du nombre d'objet dans l'image.

3.1.3 . Jeux de données

Malgré les progrès rapides de la recherche dans l'analyse des humains et de leurs activités par vision par ordinateur, la reconnaissance des interactions humaines à partir d'une seule image reste un défi. Alors que les vidéos contiennent de riches indices temporels, les images présentent beaucoup d'informations contextuelles utiles pour déduire les relations entre les objets. L'un des principaux problèmes de la détection des relations visuelles est la nécessité de disposer d'énormes quantités d'exemples variés, car les apparences et les classes du sujet et de la cible doivent varier pour assurer la généralisation de chaque classe d'interaction. La publication de grands ensembles de données [CLL+18, GM15, KZG+17, ZWS+17] a permis le développement de plusieurs détecteurs de relations visuelles ces dernières années [DZL17, KLF18, LOW17, LKBFF16, ND17, YLMD17, ZYTC18, ZKCC17, ZSE+19] ainsi que des détecteurs HOI [BRSC20, CLL+18, GZH18, GGDH18, GM15, PLSS18, QWJ+18, XWL+19]. Dans cette partie, nous présentons les différents types de jeux de données.

La compréhension du comportement humain a souvent donné lieu à la reconnaissance d'actions dans les clips vidéo. Un nombre important de jeux de données [CZ17, APGS14, SZS12, SVW+16] consistent en des clips vidéo de quelques secondes qui doivent être classés parmi un ensemble d'actions possibles, ils sont présentés partie 2.1.4. Dernièrement le grand jeu de données de clips vidéo AVA [GSR+18] introduit partiellement la localisation à la fois spatiale et temporelle des actions. Cependant, seules les boîtes des personnes réalisant des actions sont annotées dans l'image centrale de chaque clip. Aucune information n'est fournie sur les cibles des actions. La figure 3.4 illustre des exemples du jeu de données AVA.

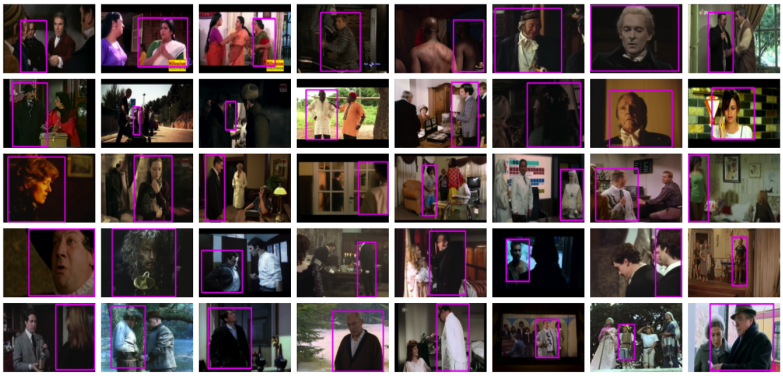


Figure 3.4 – Contenu du jeu de données AVA [GSR+18]

Plus récemment, certains ensembles de données d'images se concentrent sur les relations entre les objets dans une image. Par exemple, Visual Genome [KZG+17] propose 108 000 images annotées avec 18 relations visuelles. Ces relations peuvent concerner deux objets quelconques de l'image. Mais les prédicats ont des niveaux sémantiques assez bas et concernent surtout des relations positionnelles (comme "derrière" ou "à côté de"). Ce jeu de données est généralement utilisé pour étudier les relations visuelles, mais pas les interactions humaines.

HCVRD [ZWS+18] est un sous-ensemble de Visual Genome qui sélectionne les relations centrées sur l'humain selon 927 catégories, y compris les actions, les relations (pré)positionnelles et comparatives. Ce grand nombre d'interactions implique que la plupart d'entre elles apparaissent moins de 10 fois.

HICO [CWH+15] est un jeu de données pour la classification des HOI : les images sont centrées sur le sujet et l'image entière correspond à une liste

d'interactions non localisées. HICO-DET [CLL+18] utilise 47 051 images de HICO (37 536 images dans l'ensemble d'entraînement et 9 515 images dans l'ensemble de test) et ajoute des annotations de boîtes englobantes pour créer un jeu de données de détection des HOI. HICO-DET contient 117 verbes sur 80 catégories d'objets. Cependant certains verbes sont très spécifiques à une catégorie d'objets cibles, comme "ajuster" qui est toujours associé à une cravate et "dribbler" associé à un ballon de basket, et d'autres sont des synonymes tels que "tenir" et "porter". La figure 3.5 illustre des exemples du jeu de données HICO-DET.

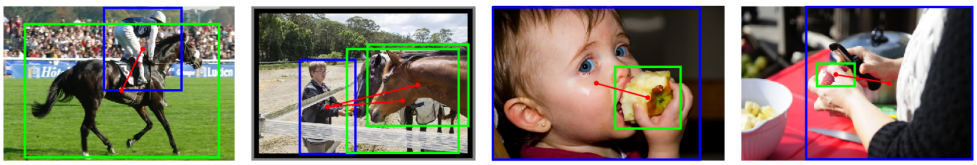


Figure 3.5 – Contenu du jeu de données HICO-DET [CLL+18]

V-COCO dataset [GM15] est un sous-ensemble du jeu de données COCO [LMB+14] pour la détection des interactions humain-objet. Il comprend 10 346 images (2 533 images dans l'ensemble d'entraînement, 2 867 images dans l'ensemble de validation et 4 946 images dans l'ensemble de test). V-COCO contient 16 199 instances humaines, où chaque personne est annotée avec 29 catégories d'action sur 80 catégories d'objet. Les objets cibles de l'ensemble de données sont classés en deux types : "objet" ou "instrument". La dénomination "objet" concerne les cibles qui subissent l'action (par exemple, "découper un gâteau"), tandis que "instrument" est employé pour les objets permettant l'interaction (par exemple, "découper avec un couteau"). Quatre verbes n'ont pas de cible ("debout", "sourir", "courir" et "marcher"). Cependant la cible d'une interaction donnée est limitée à un seul objet par verbe. La figure 3.6 illustre quelques exemples du jeu de données V-COCO.

En ce qui concerne les ensembles de données HHI, TVHI [HZ14] ne propose que quelques interactions sociales. [CS12] se concentre uniquement sur le mouvement relatif entre les personnes. Aucun de ces ensembles de données ne propose de fusionner HOI et HHI.

Dans un soucis d'exhaustivité, nous pouvons également citer les jeux de données d'interactions égocentriques tels que [BNW+18, FBGF19, SEL18]. Ce sont des bases de données vidéo dans lesquels la caméra est placée sur la personne et filme le mouvement des mains. L'objectif est différent de celui

Figure 3.6 – Contenu du jeu de données V-COCO [GM15]

des HOI classique car il ne s'agit plus d'associer les paires en interaction mais de classifier l'action des mains ou de chaque main indépendemment d'une de l'autre.

Depuis mes travaux, un jeu de données d'interaction vidéo est sorti, il s'agit de VidHOI [CLW+21]. Contrairement à AVA, les objets en interaction sont explicitement annotés. Ce jeu de données est composé de 7 122 vidéos qui représentent 7,3 millions d'images annotées selon 50 prédicats. Ce jeu de données est très intéressant pour les futurs travaux sur la détection des HOI dans des vidéos.

3.1.4 . Positionnement des travaux

Aucune méthode à une étape n'a encore été proposée pour résoudre le problème des interactions. Notre proposition a été motivée par le problème de passage à l'échelle et l'intérêt de pouvoir gérer des images contenant de nombreux objets et interactions tout en ayant un temps de traitement constant. Nous proposons donc l'approche CALIPSO (Classifying ALl Interacting Pairs in a Single shOt), une architecture multi-tâches qui estime les interactions en un seul passage sur l'image de façon dense sur une grille d'ancres. CALIPSO n'intègre pas de détecteur d'objet, c'est pourquoi nous l'appelons classifieur et non détecteur. Au moment de l'inférence, n'importe quel détecteur d'objet peut être utilisé pour pointer les objets de la scène. Pour lier les instances en interactions, CALIPSO calcule un vecteur de représentation pour toutes les ancres de la scène, il est donc également basé sur le paradigme de l'apprentissage par métrique. Mais, contrairement à Pixel2Graphs, il n'utilise pas de graphe pour modéliser explicitement chaque objet et chaque relation. Au contraire, il fournit simultanément des caractéristiques associatives et des types d'interaction pour tous les emplacements des sujets et des cibles potentiels en une seule prise. Une autre différence fondamentale est que Pixel2Graphs vise à définir une caractéristique unique pour chaque objet indépendamment du verbe, et une caractéristique unique pour chaque relation. À l'inverse, CALIPSO vise à définir, pour chaque verbe d'interaction, un

vecteur de représentation où tous les objets impliqués dans une instance d'interaction devraient avoir des caractéristiques similaires. Cela permet à une paire sujet-cible d'avoir plusieurs interactions tout en résolvant le problème des interactions superposées de Pixel2Graphs. De plus, le fait d'avoir un espace de représentation différent pour chaque verbe devrait intuitivement laisser plus de flexibilité pour modéliser des types d'interactions très différents (interaction de contact, interaction à distance, etc.).

3.2 . Description de la méthode proposée

Ce travail propose une nouvelle approche de détection des interactions, appelée CALIPSO (*Classifying ALl Interacting Pairs in a Single shOt*) dont la complexité est indépendante du nombre d'interactions. Le réseau de neurones proposé estime simultanément toutes les interactions entre tous les objets en un seul passage sur l'image. Il gère les problèmes d'association et de classification des verbes tandis que n'importe quel détecteur d'objets externe peut être utilisé pour traiter le problème de la détection des objets. À cette fin, l'approche CALIPSO exploite un schéma d'apprentissage multi-tâches, réalisant trois tâches complémentaires : une tâche de classification prédit le verbe de l'interaction, une tâche d'estimation de la présence de la cible évalue la présence de l'objet cible de l'interaction et une tâche de calcul de vecteurs de représentation fait correspondre une paire sujet-cible en interaction à une représentation similaire. Enfin, au moment de l'inférence, n'importe quel détecteur d'objets peut être utilisé pour pointer les objets d'intérêt et produire les interactions correspondantes. Notez que l'approche proposée n'utilise aucune information ontologique telle qu'une liste préalable d'interactions d'intérêt, afin de promouvoir la généralisation sur les classes cibles. La figure 3.7 présente le principe de la méthode CALIPSO.

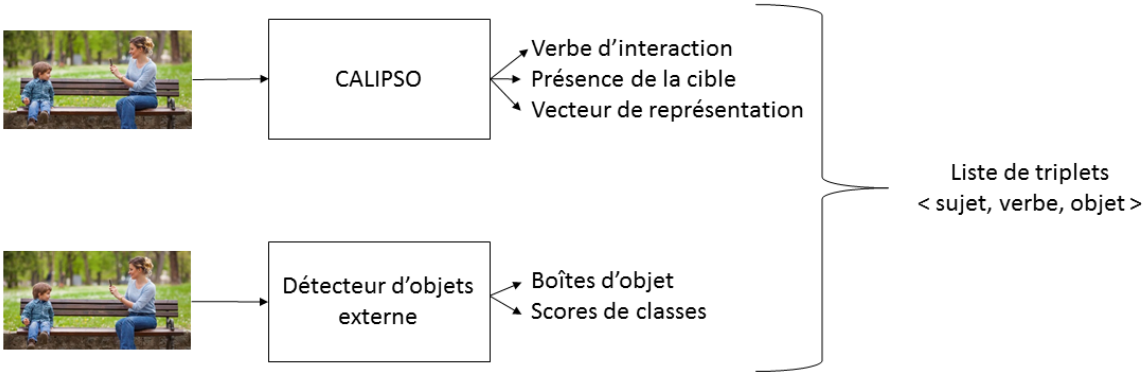


Figure 3.7 – Principe de la méthode CALIPSO

La tâche de détection des interactions humain-objet consiste à localiser et à reconnaître les humains et les objets dans une image donnée et à identifier les actions (c'est-à-dire les verbes) qui les relient. Formellement, il s'agit de localiser et de reconnaître l'ensemble $\mathcal{T}$ de triplets d'interaction $< sujet, verbe, cible >$ avec $verbe$, un verbe d'interaction parmi V verbes. L'approche proposée traite de l'association et de la classification des paires sujet-cible en interaction avec une complexité indépendante du nombre d'interactions. À cette fin, CALIPSO décorrèle la tâche de détection d'objet des tâches d'association et de classification d'interaction. C'est une méthode agnostique au détecteur d'objets qui requiert ce dernier seulement au moment de l'inférence, afin d'indiquer et de classer les objets à considérer réellement pour l'interaction.

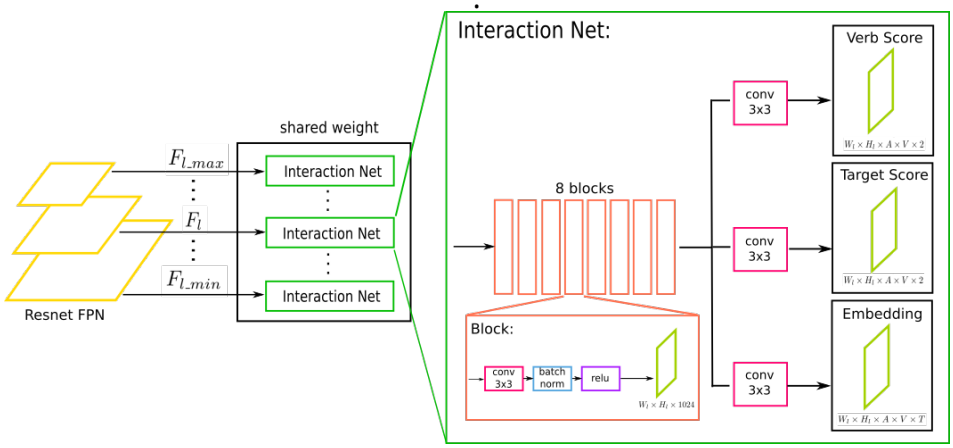


Figure 3.8 – L'architecture de CALIPSO commence par une *backbone* Resnet-50 FPN avec une pyramide de carte de caractéristiques ($F_{l_{min}}$ à $F_{l_{max}}$). La carte de caractéristique d'un niveau donné l a une taille $W_l \times H_l$. Le réseau d'interaction est appliqué à chaque niveau. Il est composé d'une succession de 8 blocs convolutifs. Le réseau se divise ensuite en 3 branches calculant 3 tâches complémentaires. A est le nombre d'ancres, V est le nombre de verbes et T est la taille du vecteur de représentation.

L'architecture du modèle proposé est un réseau neuronal multi-tâches, illustré figure 3.8. Il se compose d'une *backbone* suivie d'un réseau d'interaction. À partir d'une image I, une pyramide de caractéristiques est construite à l'aide d'une *backbone* FPN (*Feature Pyramid Network*) [LDG+17], capturant une sémantique multi-échelle de haut niveau. La *backbone* FPN prend en entrée une image I de taille $W \times H$, et produit des cartes de caractéristiques pour chaque niveau : F_l de taille $W_l \times H_l$, où $W_l = \frac{W}{2^l}$, $H_l = \frac{H}{2^l}$ et l est le niveau de la pyramide, $l \in [l_{min}, l_{max}]$. Le FPN est construit au-dessus du réseau résiduel selon l'architecture RetinaNet [LGG+17].

Ensuite, la pyramide des cartes de caractéristiques alimente un réseau d'interaction entièrement convolutif qui se termine par trois têtes spécialisées pour chacune des trois tâches. La première tâche est une classification d'action qui prédit le verbe décrivant le type d'interaction entre le sujet et la cible. La deuxième tâche estime la présence de la cible en fournissant la probabilité que l'objet avec lequel un humain interagit soit visible ou non, ou soit en dehors des classes connues du détecteur d'objets, pour un verbe donné. La troisième tâche associe le sujet et la cible en interaction, en les faisant correspondre à la même représentation. Le réseau global est entraîné de bout en bout : les trois tâches sont entraînées simultanément, partageant une *backbone* commune ce qui aide à la généralisation et régularise l'entraînement.

L'approche CALIPSO estime simultanément toutes les interactions possibles entre tous les humains et les objets de l'image, en un seul passage dans l'architecture. Ainsi, CALIPSO est indépendant du nombre de sujets, de cibles et d'instances d'interaction. De plus, en estimant de manière dense les vecteurs de représentation pour chaque verbe, l'extraction d'exemples négatifs est exhaustive sur l'image. Par exemple, toutes les personnes qui ne font pas une action spécifique sur l'image seront fournies au réseau comme échantillons négatifs pour apprendre l'espace de représentation de cette action.

Pour effectuer l'estimation des trois tâches sur toute l'image, à chaque emplacement de la carte des caractéristiques, un ensemble de boîtes de référence appelées ancres est utilisé. Ces ancres sont définies à plusieurs échelles et plusieurs aspect ratio alignés sur les objets. Nous utilisons des boîtes d'ancrage similaires à celles du réseau de proposition de région de [GDDM14]. A chaque niveau de la pyramide de carte de caractéristiques, les ancres sont définies selon 3 ratios d'aspect, $\{1{:}2, 1{:}1, 2{:}1\}$ et 3 coefficients $\{2^0, 2^{1/3}, 2^{2/3}\}$ appliqués aux 3 ratios d'aspect ce qui donne un total de $A = 9$ ancres à chaque emplacement de la carte de caractéristiques. Plus la carte de caractéristique est petite, plus les ancres définies à ce niveau de la pyramide sont grandes.

Enfin, lors de l'inférence, après avoir généré des cartes denses, un détecteur d'objets externe est utilisé pour pointer les sujets et les cibles candidats. Ainsi, les triplets d'interaction finaux sont déterminés grâce à la classe d'objet des cibles fournies par le détecteur ainsi qu'aux informations d'association et au verbe d'interaction donnés par CALIPSO.

3.2.1 . Module d'interaction

Tout d'abord, la détection des interactions nécessite d'identifier les objets (humains et non humains) en interaction. Pour chaque niveau l de la pyramide des caractéristiques, nous définissons un ensemble d'ancres

$\mathcal{A}_l$, contenant $W_l \times H_l \times A$ ancres, où $A = 9$ est le nombre d'ancres à chaque emplacement de la carte des caractéristiques. Par souci de clarté, nous définissons $\mathcal{A} = \{a_i | \, i \in [1, A_{all}]\}$ comme l'ensemble de toutes les ancres de la pyramide, où A_{all} est le nombre total d'ancres. Chaque ancre dans $\mathcal{A}$ est étiquetée comme avant-plan ou arrière-plan. Nous désignons $\mathcal{G} = \{g_j | \, j \in [1, B]\}$ comme l'ensemble des boîtes englobantes de la vérité terrain, où B est le nombre d'objets dans l'image. Comme cela se fait classiquement [LGG+17], une ancre est attribuée à une boîte de vérité terrain si son intersection-sur-union (IoU) est supérieure à $0,5$. Nous définissons $\mathcal{A}_{g_j}$ comme l'ensemble des ancres assignées aux boîtes de vérité terrain g_j et $\mathcal{A}_\mathcal{G}$ comme l'union de toutes les ancres assignées à une boîte de vérité terrain.

Le sous-réseau d'interaction est responsable de trois tâches apprises simultanément. Ce sous-réseau est appliqué avec les mêmes poids à chaque niveau de la pyramide des cartes de caractéristiques de la backbone, capturant ainsi les relations entre les instances de différentes tailles pourtant localisées à différents niveaux du FPN. De plus, les poids partagés du réseau appliqués à chaque niveau de la pyramide améliorent l'apprentissage des tâches corrélées. Ces tâches partagent une succession de huit blocs de couches de convolution, batchnorm et ReLU. Le nombre de blocs a été trouvé empiriquement. La taille spatiale de la sortie de chaque tâche pour un niveau donné de la pyramide l est égale à la taille de la carte de caractéristiques à ce niveau : $W_l \times H_l$.

3.2.2 . Apprentissage du modèle multi-tâches

La fonction de perte globale du modèle L_{total} est la somme des fonctions de perte des trois tâches estimées par le réseau : la fonction de perte de classification du verbe L^{verb}, la fonction de perte de l'estimation de la présence de la cible L^{target}, et la moyenne des fonction de perte d'estimation des représentations L_v^{emb}.

$$L_{total} = \frac{1}{|V|} L^{verb} + \frac{1}{|V|} L^{target} + \frac{1}{|V|} \sum_{v \in V} L_v^{emb} \tag{3.1}$$

Prédiction du verbe

En considérant que les sujets de l'interaction peuvent entreprendre simultanément plusieurs actions, la tâche de prédiction des verbes minimise une perte d'entropie croisée binaire multi-label L^{verb} entre les verbes prédits et les verbes de la vérité terrain :

$$L^{verb} = \sum_{v \in V} \sum_{a \in \mathcal{A}_\mathcal{G}^p} -q_a^v \log(p_a^v) \tag{3.2}$$

où v est un verbe de l'ensemble V, $\mathcal{A}_{\mathcal{G}}^{p}$ désigne l'ensemble des ancres actives associées aux personnes exécutant les actions, p_a^v est la probabilité estimée du modèle pour la classe v et l'ancre a, q_a^v est égal à 1 si la classe de vérité terrain de l'ancre a est v, 0 sinon.

Contrairement aux autres méthodes, nous introduisons une estimation supplémentaire des verbes passifs centrée sur l'objet afin d'améliorer réciproquement la détection des relations. Par exemple la forme passive du verbe "tenir" associé à une ancre de personne est "être tenu" pour l'ancre de l'objet qui subit l'action. La tâche de prédiction des verbes est effectuée sur la base de l'apparence contextuelle qui est très informative pour distinguer les actions que les personnes effectuent et celles que les objets subissent. Parmi l'ensemble des ancres $\mathcal{A}_{\mathcal{G}}$, nous trouvons les ancres actives, $\mathcal{A}_{\mathcal{G}}^{p}$, représentant les ancres associées aux personnes exécutant les actions, et les ancres passives, $\mathcal{A}_{\mathcal{G}}^{o}$, associées aux objets subissant l'action.

La classification de la forme passive est une tâche optionnelle qui améliore les performances. La tâche de prédiction des verbes produit ensuite, pour chaque ancre, une sortie de classification sur les verbes dans les formes active et passive, ce qui donne une sortie de taille $2V$ avec V le nombre total de verbe. La perte de l'estimaton des verbes passifs est la même que L^{verb} mais elle est calculée sur $\mathcal{A}_{\mathcal{G}}^{o}$, l'ensemble des ancres passives associées aux objets subissant l'action. Cette estimation de l'interaction réciproque est une manière d'obtenir une supervision renforcée en utilisant à la fois les annotations initiales centrées sur l'humain et en introduisant la notion de réciproque centrée sur l'objet.

Estimation de la présence de la cible

L'estimation de la présence de la cible une tâche complémentaire à la tâche de prédiction du verbe. Elle vise à estimer la probabilité que l'objet, avec lequel une personne interagit, soit visible ou non ou qu'il appartienne ou non aux classes d'objets de la base étudiée (par exemple, les 80 classes du jeu de données COCO). Comme pour la tâche de prédiction de verbe, l'estimation de l'objet cible est effectuée sur l'apparence contextuelle de chaque ancre de personne, capturant la position spatiale et l'environnement de la personne dans l'image. Pour chaque ancre, la sortie de taille $2V$ consiste en des classifieurs sigmoïdes binaires. L'objectif de l'apprentissage est de minimiser la perte d'entropie croisée binaire, L^{target}, entre les étiquettes d'objet cible de vérité terrain et l'estimation de la cible prédite :

$$L^{target} = \sum_{v \in V} \sum_{a \in \mathcal{A}_{\mathcal{G}}^{p}} -q_a^v \log(p_a^{1v}) - (1 - q_a^v) \log(p_a^{2v}) \tag{3.3}$$

où v est un verbe de l'ensemble V, $\mathcal{A}_{\mathcal{G}}^{p}$ désigne l'ensemble des ancres associées aux personnes exécutant les actions, p_a^{1v} est la probabilité de présence de la cible, estimée par le modèle pour le verbe v et l'ancre a, p_a^{2v} est la probabilité d'absence de la cible, estimée par le modèle pour le verbe v et l'ancre a, q_a^v est égal à 1 si la cible de l'ancre a pour le verbe v est présente, 0 sinon.

Apprentissage de la représentation de l'interaction

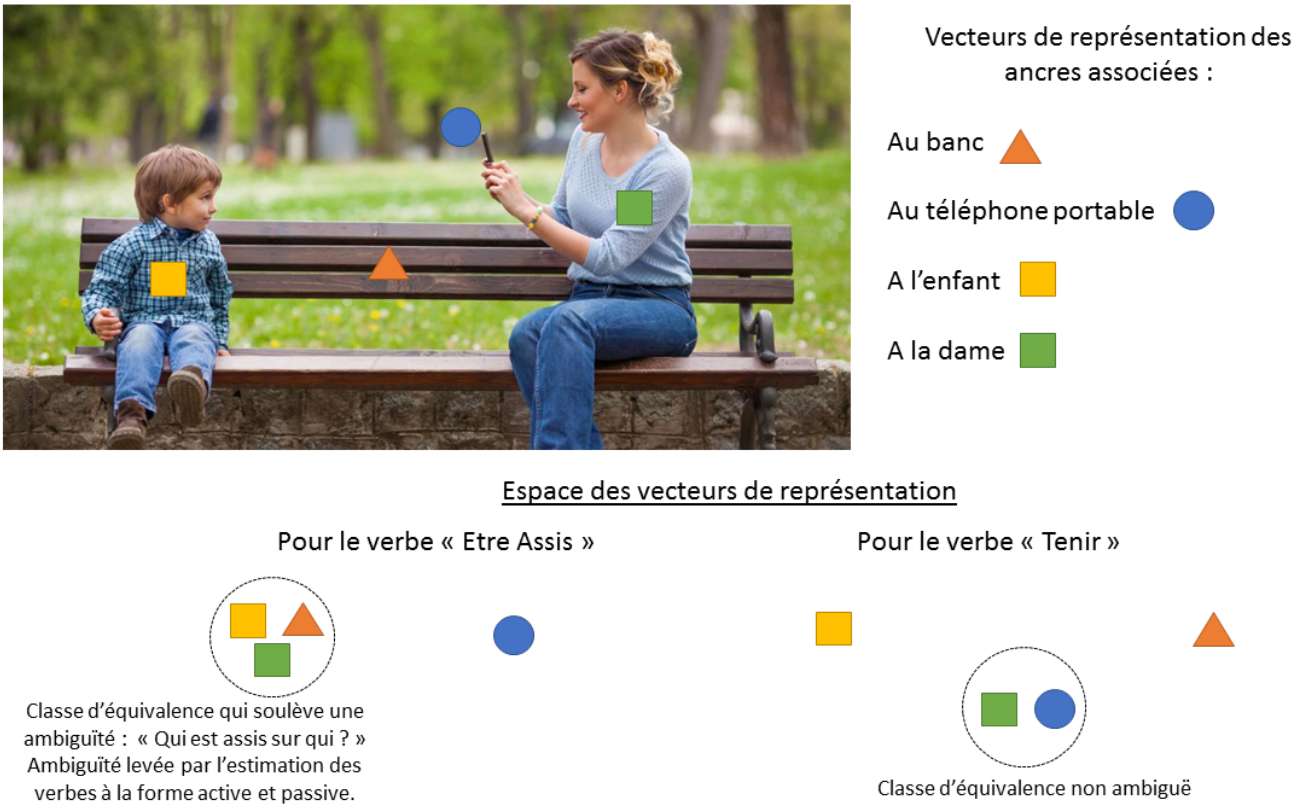


Figure 3.9 – Exemple de la répartition des vecteurs de représentation pour une image donnée.

Cette tâche vise à associer plusieurs ancres correspondant à des sujets et des cibles en interaction à la même représentation pour un verbe donné. Le sous-réseau de calcul des vecteurs de représentation (en anglais, *embeddings)* est une fonction faisant correspondre l'espace des ancres $\mathcal{A}$ à un

nouvel espace tel que : emb: $\mathcal{A} \rightarrow \mathbb{R}^{V \times T}$ où T est la dimension de l'espace des vecteurs de représentation de l'interaction spécifique à un verbe. Pour un verbe donné, la tâche d'estimation du vecteur de représentation vise à s'assurer d'attribuer la même représentation aux ancres liées à la même instance d'objet et également aux ancres appartenant à la même interaction.

La figure 3.9 illustre l'espace des vecteurs de représentation pour une image où deux personnes sont assises sur un banc et l'une des personnes tient un téléphone portable. Pour le verbe "être assis", les vecteurs de représentation des ancres associées à l'enfant, à la dame et au banc seront proches car à la fois l'enfant et la dame sont assis sur le banc. Dans le cadre du jeu de données V-COCO, les cibles sont les objets, il n'y a donc pas d'ambiguïté possible dans la constitution des triplets d'interaction. Cependant si on souhaite élargir la méthode à un cas plus général où on autorise les personnes à faire parti des cibles des interactions, alors nous sommes face à une ambiguité où l'enfant pourrait être assis sur la dame au lieu d'être assis sur le banc. C'est l'estimation du verbe à la forme active et passive qui va permettre de lever l'ambiguïté et savoir "qui est assis sur qui ou quoi ?". En effet, l'enfant et la dame auront un score élevé pour le verbe "être assis" à la forme active alors que le banc aura un score élevé pour ce même verbe à la forme passive, c'est donc lui qui subit l'action. Cette ambiguité est levée seulement au premier ordre, s'il y a une succession de personnes assises les unes sur les autres, CALIPSO ne permet pas de déterminer "qui est assis sur qui ?".

Formellement, étant donné les ancres $a_i, a_j \in \mathcal{A}^2_{\mathcal{G}}$ où pour rappel, G est défini comme l'ensemble des boîtes englobantes de la vérité terrain, a_i et a_j sont en interaction selon le verbe v, c'est-à-dire $a_i \sim_v a_j$, si :

$$\exists\, g_n \in \mathcal{G} \mid (a_i, a_j) \in \mathcal{A}^2_{g_n} \tag{3.4}$$

ou

$$\left.\begin{array}{c} \exists\, (g_n, g_m) \in \mathcal{G}^2,\ n \neq m, \\ < g_n, v, g_m > \text{ or } < g_m, v, g_n > \in \mathcal{T} \end{array}\right| (a_i, a_j) \in \mathcal{A}_{g_n} \times \mathcal{A}_{g_m} \tag{3.5}$$

Ainsi, à chaque verbe v, correspond un ensemble de classes d'équivalence associées à une relation d'équivalence $\sim_v$, notée $\mathcal{C}_v = \{\mathcal{C}^i_v | i \in [1, E_v]\}$, avec E_v le nombre de classes d'équivalence pour le verbe v. Soit $|\mathcal{C}^i_v|$ le nombre d'ancres appartenant à la classe d'équivalence $\mathcal{C}^i_v$. La référence de la classe d'équivalence est définie par la moyenne des représentations estimées pour la même classe d'équivalence comme suit :

$$\overline{e_{\mathcal{C}^i_v}} = \frac{1}{|\mathcal{C}^i_v|} \sum_{j \in \mathcal{C}^i_v} e^v_j \tag{3.6}$$

où e_j^v est le vecteur de représentation prédit pour l'ancre a_j et le verbe v.

Le réseau d'estimation des vecteurs de représentation vise à apprendre l'espace des classes d'équivalence $\mathcal{C}_v$, en minimisant la perte d'équivalence L_v^{emb}, définie sous une forme d'apprentissage par métrique. Pour un verbe donné v, la fonction de perte est définie comme :

$$L_v^{emb} = L_v^{pull} + L_v^{push} \tag{3.7}$$

La fonction de perte d'attraction (en anglais *pulling loss*) qui vise à rassembler les éléments correspondants, est définie comme suit :

$$L_v^{pull} = \frac{1}{E_v} \sum_{\mathcal{C}_v^i \in \mathcal{C}_v} \frac{\lambda_{\mathcal{C}_v^i}}{|\mathcal{C}_v^i|} \sum_{j \in \mathcal{C}_v^i} \left(e_j^v - \overline{e_{C_v^i}}\right)^2 \tag{3.8}$$

Sur la base des annotations de la vérité terrain définissant les instances en interaction, le premier terme de l'équation vise à fusionner les instances en interaction dans la même classe d'équivalence en calculant la distance quadratique moyenne entre les références d'équivalence $\overline{e_{C_v^i}}$ et le vecteur de représentation prédit e_j^v pour chaque ancre j dans la classe d'équivalence $\mathcal{C}_v^i$. Le poids $\lambda_{\mathcal{C}_v^i}$ vise à se concentrer davantage sur les classes d'équivalence représentant des sujets et des cibles en interaction réelle plutôt que sur la classe d'équivalence associée à un objet unique n'appartenant à aucune interaction (cf. équation 3.4). Il est défini comme suit :

$$\lambda_{\mathcal{C}_v^i} = \begin{cases} \lambda_{pull} & \text{si } \exists\, a_j, a_k \in \mathcal{C}_v^i \text{ tel que} \\ & (a_j, a_k) \in \mathcal{A}_{g_n} \times \mathcal{A}_{g_m}, n \neq m, \\ & < g_n, v, g_m > \text{ or } < g_m, v, g_n > \in \mathcal{T}; \\ 1 & \text{sinon.} \end{cases} \tag{3.9}$$

La fonction de perte de répulsion (en anglais *pushing loss*) permet de séparer des ancres d'instances qui n'interagissent pas, dans différents clusters en utilisant une fonction exponentielle décroissante avec un paramètre fixe σ. Elle est définie comme suit :

$$L_v^{push} = \frac{1}{E_v^2} \sum_{\substack{\mathcal{C}_v^i, \mathcal{C}_v^j \in C_v^2 \\ i \neq j}} \gamma_{\mathcal{C}_v^i, \mathcal{C}_v^j}\, exp\left(\frac{-1}{2\sigma^2}\left(\overline{e_{\mathcal{C}_v^i}} - \overline{e_{\mathcal{C}_v^j}}\right)^2\right) \tag{3.10}$$

Le poids $\gamma_{\mathcal{C}_v^i, \mathcal{C}_v^j}$ introduit une pénalité douce à la fonction de perte pour forcer le réseau à associer la cible correcte parmi plusieurs objets présents dans l'image qui sont des cibles habituelles pour ce verbe. Par exemple, le vecteur de représentation d'une personne assise sur une chaise donnée ne doit pas être regroupé avec celui d'autres chaises ou objets sur lesquels on peut s'asseoir (par exemple, un canapé, un lit, une table, ...), présents dans

l'image. Ce poids est un moyen d'imposer la sélection de la bonne cible parmi plusieurs candidats, même s'ils conviennent à cette interaction. Plus formellement, lab_i est l'étiquette de classe de l'ancre a_i et $\mathcal{L}_v$ l'ensemble des classes d'objets pouvant être impliquées dans le type d'interaction donné par le verbe v selon les statistiques sur le jeu de données (par exemple chaise, canapé, lit, table... pour le verbe "s'assoir"). Le poids $\gamma_{\mathcal{C}_v^i,\mathcal{C}_v^j}$ est défini comme suit :

$$\gamma_{\mathcal{C}_v^i,\mathcal{C}_v^j} = \begin{cases} \gamma_{push} & \text{si } \exists (a_k, a_l) \in \mathcal{C}_v^i \times \mathcal{C}_v^j \text{ tel que} \\ & (a_k, a_l) \in \mathcal{A}_{g_n} \times \mathcal{A}_{g_m},\ n \neq m, \\ & (lab_k, lab_l) \in \mathcal{L}_v^2; \\ 1 & \text{sinon.} \end{cases} \tag{3.11}$$

Ce processus d'estimation du vecteur de représentation est effectué pour chaque verbe, ce qui permet au réseau d'apprendre les différentes façon d'interagir en fonction du verbe. De plus, les prédictions des représentations sont effectuées simultanément sur toutes les ancres, quel que soit le nombre d'instances. Cela permet également une meilleure gestion des interactions négatives lors de l'apprentissage en traitant tous les cas d'absence d'interaction dans l'image. De plus, cela permet à l'inférence de connecter les instances de façon précise et rapide. La tâche d'estimation des vecteurs de représentation ne fait pas d'hypothèses spécifiques entre les positions du sujet et de la cible, elle peut donc modéliser des interactions aussi bien distantes que proches. De plus, la tâche d'estimation des vecteurs de représentation apprend à associer des objets de tailles éventuellement différentes, c'est-à-dire localisés sur différents niveaux de la pyramide.

3.2.3 . Inférence

De la même manière que les approches existantes, nous prédisons les triplets $HOI < sujet, verbe, cible >$, ce qui implique de prédire les paires de boîtes englobantes humain-objet, d'identifier le verbe et calculer le score du triplet. Les trois tâches du modèle proposé fournissent trois cartes de caractéristiques. La carte d'ancres de la première tâche définit le score d'action de chaque emplacement dans l'image. La deuxième tâche fournit une carte de caractéristiques estimant pour chaque verbe, la présence d'une cible en interaction pour chaque ancre relative à une personne. La troisième carte de caractéristiques fournit une représentation pour chaque ancre dans l'image, afin de déterminer les ancres en interaction. La méthode extrait toutes les cartes de caractéristiques simultanément et indépendamment du nombre d'instances d'objets qui se trouvent à des emplacements et des échelles d'image arbitraires, contrairement à la plupart des approches existantes où chaque paire humain-objet sélectionnée est traitée individuellement.

La prédiction des triplets HOI nécessite la détection préalable de toutes

les personnes et objets de la scène. Ainsi, lors de l'inférence, CALIPSO nécessite un détecteur d'objet externe pour pointer les ancres d'intérêt sur les sorties du réseau. Le détecteur externe peut être n'importe quel détecteur d'objets fournissant les positions des boîtes englobantes et les scores de classe, notés s_h^{det} pour une personne et s_o^{det} pour un objet.

Le détecteur fournit un ensemble de boîtes de délimitation d'objets candidats qui sont ensuite mises en correspondance avec la grille d'ancre. L'ancre sélectionnée est celle optimisant un graphe biparti basé sur les IoU entre les ancres et les boîtes fournies par le détecteur. Ainsi, à partir de cette mise en correspondance, pour chaque verbe v et pour chaque boîte candidate, différents scores peuvent être lus : les scores des verbes (plus précisément, le score actif $s_{v,h}^{active}$ pour une personne et le score passif $s_{v,o}^{passive}$ pour un objet), les scores de présence de la cible $s_{v,h}^{target}$ pour une personne, et les représentations e_i^v de chaque instance détectée. Ces vecteurs sont comparés entre eux en définissant un score de connexion $s_{v,h,o}^{emb}$ calculé comme suit :

$$s_{v,h,o}^{emb} = exp^{(-|e_h^v - e_o^v|)} \tag{3.12}$$

Nous faisons l'hypothèse grossière mais simplificatrice que tous les scores sont indépendants, cela nous permet d'utiliser la moyenne géométrique pour définir le score du triplet comme suit :

$$s_{v,h,o}^{triplet} = \sqrt[6]{s_h^{det}\ s_{v,h}^{active}\ s_o^{det}\ s_{v,o}^{passive}\ s_{v,h}^{target}\ s_{v,h,o}^{emb}} \tag{3.13}$$

Tous les triplets possibles sont calculés pour chaque personne détectée et chaque verbe. Un score de paire $< subject, verb >$ est également calculé pour le cas d'absence d'objet cible :

$$s_{v,h}^{pair} = \sqrt[3]{s_h^{det}\ s_{v,h}^{active}\ (1 - s_{v,h}^{target})} \tag{3.14}$$

La moyenne géométrique nous permet de comparer les scores $s_{v,h,o}^{triplet}$ et $s_{v,h}^{pair}$ qui n'ont pas le même nombre de terme. Pour un verbe et une personne donnés, tous les triplets et la paire sont triés en fonction de leurs scores et celui qui a le score le plus élevé est conservé après seuillage.

3.3 . Expériences

Les expériences sont menées sur deux ensembles de données largement utilisés pour la détection des interactions, avec une comparaison entre l'approche proposée et l'état de l'art.

3.3.1 . Jeux de données

V-COCO dataset [GM15] est un sous-ensemble du jeu de données COCO [LMB+14] pour la détection des interactions humain-objet. Il comprend 10 346 images (2 533 images dans l'ensemble d'entraînement, 2 867 images dans l'ensemble de validation et 4 946 images dans l'ensemble de test). V-COCO contient 16 199 instances humaines, où chaque personne est annotée avec 29 catégories d'action sur 80 catégories d'objet. Les objets cibles de l'ensemble de données sont classés en deux types : "objet" ou "instrument". La dénomination "objet" concerne les cibles qui subissent l'action (par exemple, "découper un gâteau"), tandis que "instrument" est employé pour les objets permettant l'interaction (par exemple, "découper avec un couteau"). Quatre verbes n'ont pas de cible ("debout", "sourir", "courir" et "marcher").

HICO-DET dataset [CLL+18] est un sous-ensemble du jeu de données HICO pour la détection des interactions humain-objet. Il est plus grand et plus diversifié que le jeu de données V-COCO. HICO-DET comprend 47 051 images (37 536 images dans l'ensemble d'entraînement et 9 515 images dans l'ensemble de test). HICO-DET contient 117 catégories d'actions sur 80 catégories d'objets comme le jeu de données COCO. Toutes les combinaisons d'actions et d'objets ne sont pas pertinentes, selon une ontologie définie. Par conséquent, seules 600 catégories spécifiques d'interaction humain-objet sont annotées et évaluées.

3.3.2 . Métriques d'évaluation

En suivant les paramètres d'évaluation standard des jeux de données V-COCO [GM15] et HICO-DET [CLL+18], nous évaluons les performances de détection des HOI à l'aide de la métrique de précision moyenne : AP_{role} . Le triplet $< sujet, verbe, cible >$ prédit est considéré comme un vrai positif, lorsque toutes les composantes prédites du triplet sont correctes. Les boîtes de délimitation prédites de la personne et de l'objet sont supposées être correctes si elles ont un IoU supérieur à 0,5 avec les boîtes de vérité terrain.

Suivant les travaux précédents [CLL+18, GZH18, GGDH18, QWJ+18], l'évaluation sur le jeu de données V-COCO est basée sur la précision moyenne du rôle appelé AP^1_{role} sur 24 catégories de verbes. En effet, pour une comparaison équitable avec les approches de l'état de l'art, 5 actions ("debout", "sourir", "courir", "marcher" et "pointer") sont ignorées dans l'évaluation, comme dans les approches précédentes. Les quatre premières car ce sont des verbes sans objet cible et "pointer" car V-COCO contient trop peu d'exemple pour qu'un entraînement sur ce verbe soit pertinent.

En ce qui concerne le jeu de données HICO-DET [CLL+18], nous présentons le AP moyen sur trois ensembles différents de catégories de HOI : (a) l'ensemble des 600 catégories de HOI dans HICO-DET (*Full*), (b) 138 catégories de HOI avec moins de 10 instances d'apprentissage (*Rare*), et (c) 462 catégories de HOI avec 10 ou plus d'instances d'apprentissage (*Non-Rare*).

3.3.3 . Détails d'implémentation

Nous initialisons la backbone du ResNet-50 FPN avec les poids correspondants de RetinaNet [LGG+17] entraîné sur l'ensemble de données de COCO dont les images V-COCO ont été préalablement retirées.

CALIPSO est entraîné par descente de gradient stochastique (SGD), avec un taux d'apprentissage initial de 0,016, qui est ensuite divisé par 10 à 25 000 itérations sur un lot de taille 10. Un *flip* horizontal de l'image est appliqué pour l'augmentation des données. Le *weight decay* est fixé à 10^{-4} et le *momentum* à 0,9. σ, λ_v et γ_v sont fixés expérimentalement à 2, 10 et 100.

Lors de l'inférence, CALIPSO a besoin d'un détecteur externe pour filtrer les boîtes en interaction sur les sorties des trois sous-tâches du réseau. Comme dans la plupart des méthodes de l'état de l'art, le Faster RCNN [RHGS15] issu de Detectron [1] est utilisé comme détecteur externe. Il s'appuie sur une *backbone* ResNet-50-FPN pour générer toutes les boîtes d'objets. D'autres détecteurs d'objets sont également testés pour montrer leur influence sur la détection des HOI.

3.3.4 . Résultats Qualitatifs

Les figures 3.10 et 3.11 illustrent les résultats des interactions détectées par le modèle proposé. Elles montrent tous les triplets présents dans l'image. Chaque triplet est représenté par une boîte en trait plein pour le sujet et une boîte en pointillé pour l'objet cible. En haut à gauche de la boîte du sujet, l'action effectuée est indiquée sur un fond de la même couleur que la boîte de la cible correspondante.

La figure 3.10 illustre les interactions détectées par notre approche. Comme on peut le voir, CALIPSO peut déduire les HOI dans diverses situations telles que :

- Une personne effectuant différentes actions sur un même objet (par exemple, "une personne conduit, s'assoit et tient un vélo", dans la Figure 3.10-a-b-c-f).
- Une personne interagissant avec différents objets (par exemple, dans les figures 3.10-b et 3.10-f, "une personne travaille sur un ordinateur tout

1. https ://github.com/facebookresearch/Detectron

en étant assise sur une chaise/un canapé").

- Plusieurs personnes interagissant avec un même objet (par exemple, dans la figure 3.10-e, "deux personnes tiennent le même couteau"). Remarquez que CALIPSO assigne correctement l'objet cible à l'action correspondante, et peut détecter avec succès les interactions sans contact (dans la Figure 3.10-d, "regarder et lancer un frisbee").

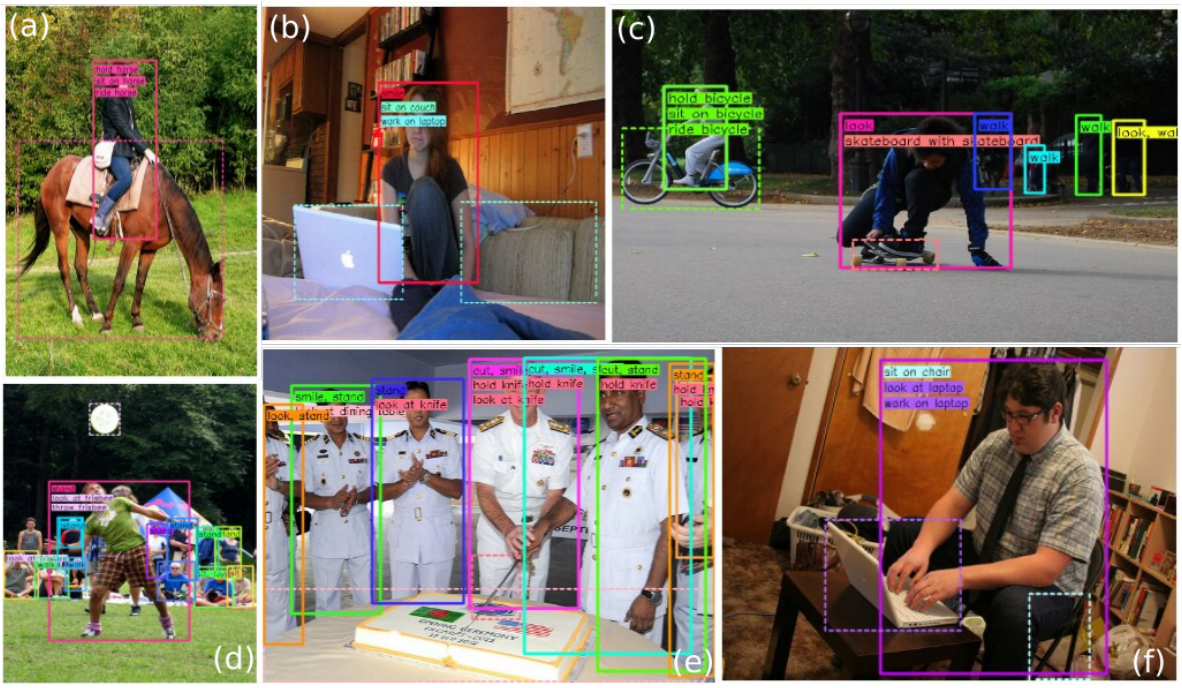


Figure 3.10 – Exemples d'interactions humain-objet détectées par CALIPSO sur certaines images du test de V-COCO. Un triplet d'interaction est composé d'un sujet humain représenté par une boîte en trait plein, d'un objet cible représenté par une boîte en pointillés et, en haut à gauche de la boîte du sujet, l'action effectuée est inscrite sur un fond de la même couleur que la boîte de l'objet cible. Les différentes couleur de traits sont utilisées pour distinguer chaque objet détecté. On peut voir que CALIPSO arrive à reconnaître quand une personne effectue différentes actions sur un même objet, c'est le cas de l'image (a) où la personne tient, est assise sur et monte le cheval. CALIPSO est également capable de détecter si une personne interagit avec différents objets, c'est le cas pour l'image (f) où la personne est à la fois assise sur la chaise et travaille sur l'ordinateur. Enfin, CALIPSO est également capable de détecter lorsque différentes personnes utilisent le même objet. C'est le cas pour l'image (e) où les deux personnes à droite tiennent le même couteau.

La figure 3.11 illustre un autre échantillon d'images de test de V-COCO, où

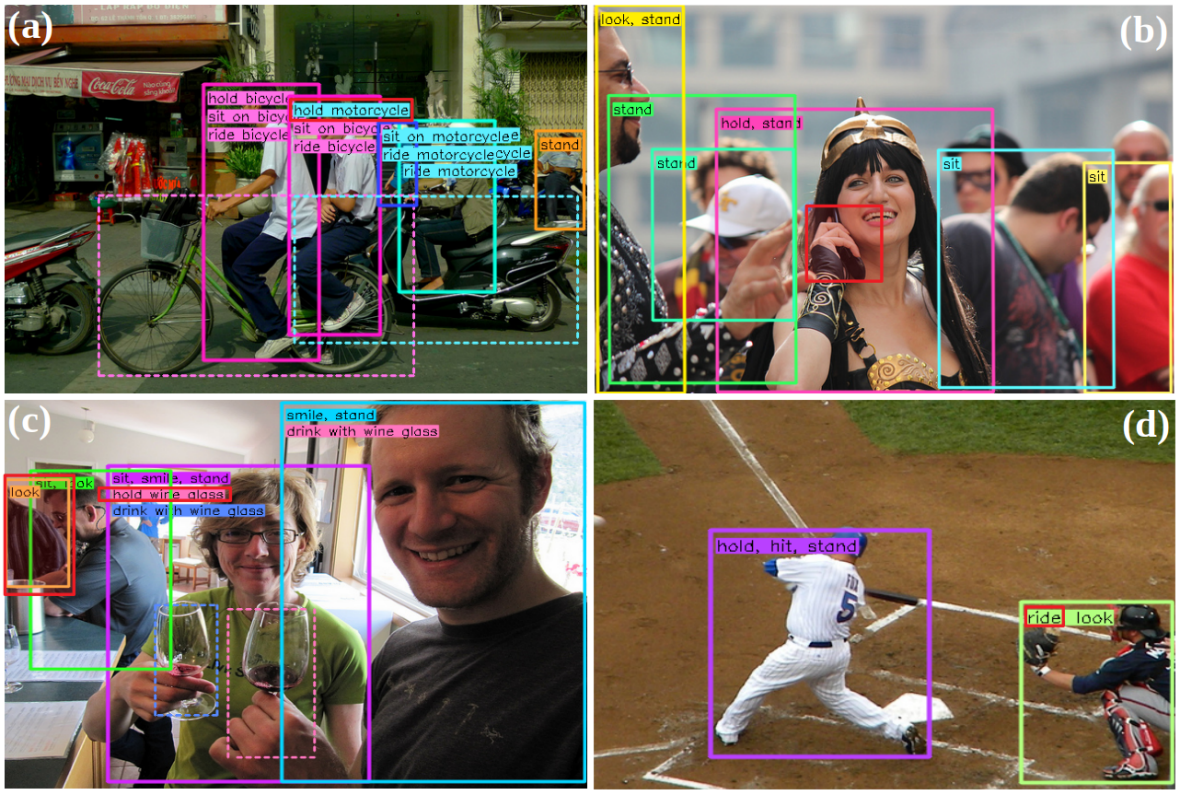


Figure 3.11 – Illustration de détections incorrectes d'interaction humain-objet sur quelques images de test de V-COCO. Dans l'image (a), CALIPSO a faussement estimé que la personne au premier plan tient la moto au second plan, cela peut s'interpréter par une mauvaise compréhension de la profondeur des objets dans la scène. Dans l'image (b), le téléphone portable n'a pas été détecté donc l'association ne peut être réalisée. Dans l'image (c), le sac à dos a été classifié comme un humain et le verre à vin est tenu par la mauvaise personne. Dans l'image (d), la position ambiguë de la personne implique une mauvaise estimation du verbe "monter sur".

CALIPSO détecte quelques triplets incorrects. Ceci est principalement dû à :

- Une détection d'objet erronée, avec soit l'absence d'objet détecté (comme le montre la figure 3.11-b où le téléphone portable n'est pas détecté), soit un objet mal classé (illustré dans la figure 3.11-c où le sac à dos est classé comme humain).
- Une mauvaise estimation du verbe, illustrée dans la figure 3.11-d où la personne a une posture ambiguë.
- Une mauvaise association de cibles, illustrée sur la figure 3.11-c où le verre à vin est tenu par la mauvaise personne.

La figure 3.11-a montre un exemple où toutes ces difficultés apparaissent simultanément. En effet, la forte densité d'objets entraîne davantage d'occul-

tations, une mauvaise compréhension de la profondeur des objets dans la scène et, par conséquent, des confusions dans les associations sujet-cible.

3.3.5 . Résultats Quantitatifs

Étude d'ablation

Dans le Tableau 3.1 nous évaluons sur le jeu de données V-COCO les contributions des différentes composantes de la méthode.

Méthode	$\mathbf{AP^1_{role}}$ (%)
CALIPSO	**46.36**
CALIPSO sans le partage des poids	43.86
CALIPSO sans le mode passif du verbe	36.86
CALIPSO sans l'estimation de la présence de la cible	25.51
CALIPSO 5 blocs convolutifs	44.35
CALIPSO 8 blocs convolutifs	**46.36**
CALIPSO 11 blocs convolutifs	45.05

Table 3.1 – Étude d'ablation pour CALIPSO sur le jeu de test de V-COCO.

Les poids partagés : Le partage des poids entre les niveaux du réseau de la pyramide des caractéristiques montre une amélioration de $2,25$ p.p. (points de pourcentage) des performances de détection des interactions. Intuitivement, cela permet de mieux capturer les relations entre les instances appartenant à différents niveaux du FPN correspondant à différentes tailles d'objets.

Le mode passif : Alors que le mode actif est centré sur le sujet, le mode passif est un moyen d'introduire un point de vue complémentaire centré sur la cible et, ainsi, d'introduire une redondance pour améliorer la robustesse. Sans tâche en mode passif, notre modèle atteint un AP^1_{role} de 36.86%. Il augmente d'environ 10 p.p. et atteint un AP^1_{role} de 46.36% lorsque le mode passif est utilisé.

La présence de la cible : La présence de la cible a un impact énorme sur les performances de CALIPSO, augmentant les résultats d'environ 20 p.p. Une telle variation des performances est due à la difficulté de fixer une distance maximale (dans l'espace des représentations) en dessous de laquelle un sujet peut être considéré en interaction avec la cible. Il est bien connu que le seuillage direct d'une métrique apprise n'est pas trivial. En effet,

l'apprentissage par métrique ne contraint pas la distance absolue entre les échantillons mais impose seulement un classement entre eux. La tâche de présence de la cible est un moyen de contourner ce problème.

Profondeur du réseau d'interaction : Le nombre de blocs utilisés dans le réseau d'interaction a été choisi de manière empirique. Une succession de 8 blocs convolutionnel a donné le meilleur résultat.

Résultats sur le jeu de données V-COCO

Comme la méthode proposée se concentre sur la classification des HOI indépendamment de la tâche de détection d'objets, elle peut avantageusement utiliser n'importe quel détecteur d'objets externe au moment de l'inférence. En effet, le changement de détecteur ne nécessite pas de réentraîner ou d'adapter le réseau, ce qui est une propriété très intéressante lorsque de meilleurs détecteurs d'objets apparaissent dans l'état de l'art.

Par conséquent, nous évaluons notre modèle avec deux détecteurs d'objets externes différents : Faster RCNN [RHGS15] avec une backbone ResNet50 (*Faster R50*) qui est généralement utilisée par les méthodes de l'état de l'art comme base pour apprendre les interactions, et Faster RCNN avec une backbone ResNext101 [XGD+17] (*Faster RNext101*). Pour une comparaison équitable, nous présentons les résultats RP_DC_D de l'approche Interactiveness [LZH+19] qui correspond au modèle entraîné sans jeux de données supplémentaires.

Le tableau 3.2 présente les résultats d'évaluation des variantes de CALIPSO par rapport aux méthodes de l'état de l'art sur le jeu de données V-COCO. CALIPSO atteint des performances similaires à la meilleure méthode Interactiveness [LZH+19] avec une *backbone* plus complexe, Faster RNext101, mais il est beaucoup plus efficace en termes de temps de calcul comme illustré et expliqué figure 3.12 et section 3.3.6.

De plus, afin de décorréler la tâche de détection d'objet de celle de détection d'interaction, nous utilisons à l'inférence le détecteur d'objet parfait (c'est-à-dire les vérités terrain de détection du jeu de données) et reportons les résultats dans le tableau 3.2. Le détecteur parfait permet d'augmenter les performances ce qui montre que la qualité de la détection a un impact significatif sur les résultats de la détection d'interaction. Néanmoins, l'amélioration n'est que de 7 p.p., la performance finale de 54.48% montre qu'il y a encore une grande marge de progrès et que le problème d'association sujet-cible et celui de classification des verbes est loin d'être résolu.

Méthode	Détecteur / Backbone	$\mathbf{AP}^{\mathbf{1}}_{\mathbf{role}}$ (%)
V-COCO [GM15]	Faster R50	31.8
InteractNet [GGDH18]	Faster R50	40.0
GPNN [QWJ+18]	Deform. CNN	44.0
iCAN late(early) [GZH18]	Faster R50	44.7 (45.3)
Xu [XWL+19]	Faster R50	45.9
Interactiveness [LZH+19]	**Faster R50**	**47.8**
CALIPSO	Faster R50	46.36
CALIPSO	Faster RNext101	47.65
CALIPSO	**Vérité terrain**	**54.48**

Table 3.2 – Comparaison des résultats de l'évaluation de CALIPSO sur l'ensemble de test de V-COCO avec les résultats des méthodes de l'état de l'art. Les détecteurs d'objets ou les backbones utilisés sont mentionnés dans la colonne du milieu.

Résultats sur le jeu de données HICO-DET

Les objets du jeu de données HICO-DET étant annotés grossièrement (plusieurs boîtes peuvent être attribuées au même objet), nous adoptons le même protocole que [GGDH18] pour nettoyer les annotations. Nous utilisons un détecteur d'objets ResNext101 entraîné sur COCO pour détecter les objets et attribuer les étiquettes de vérité terrain des annotations HICO-DET aux objets détectés qui chevauchent fortement les boîtes HICO-DET.

Selon les paramètres de [CLL+18], nous présentons l'évaluation quantitative des interactions *Full*, *Rare*, et *Non-Rare* sur le paramètre d'évaluation "*default*". Le tableau 3.3 présente les résultats de précision moyenne de notre méthode sur le jeu de données HICO-DET, par rapport aux approches de détection des HOI les plus récentes. Une fois de plus, pour une comparaison équitable, nous avons présenté des méthodes qui utilisent uniquement le jeu de données sans l'aide de données supplémentaires, telles que les connaissances linguistiques, provenant de jeux de données externes par exemple.

L'approche proposée montre des résultats compétitifs, atteignant la deuxième place avec le détecteur Faster RNext101.

	Précision Moyenne (*Default*)		
Méthode	**Full**	**Rare**	**Non-Rare**
HO-RCNN [CLL+18]	7.81	5.37	8.54
InteractNet [GGDH18]	9.94	7.16	10.77
GPNN [QWJ+18]	13.11	9.34	14.29
Xu [XWL+19]	14.70	13.26	15.13
iCAN [GZH18]	14.84	10.45	16.15
Interactiveness [LZH+19]	**17.03**	**13.42**	**18.11**
CALIPSO (Faster R50)	14.31	10.43	15.46
CALIPSO (Faster RNext101)	14.89	11.12	16.01

Table 3.3 – Résultats d'évaluation sur l'ensemble de test HICO-DET, comparés aux méthodes les plus récentes.

3.3.6 . Complexité algorithmique

En ce qui concerne la complexité par rapport au nombre de personnes (N) et d'objets (M) dans l'image, CALIPSO n'effectue qu'un seul passage dans l'image avec une complexité $O(1)$, alors que toutes les autres approches de l'état de l'art au moment des travaux ont une complexité $O(P)$ avec P le nombre de paires traitées, $T \leq P \leq N \times M$ avec $T = |\mathcal{T}|$ le nombre de triplets de vérité terrain. L'impact sur le temps de calcul est illustré dans la figure 3.12 : CALIPSO s'exécute en temps constant (460 ms sur NVIDIA Titan X Pascal) indépendamment du nombre de personnes et d'objets dans l'image. En revanche, les méthodes de l'état de l'art qui fournissent leurs codes, Interactiveness [LZH+19] et iCAN [GZH18], ont un temps de calcul très variable selon le nombre d'objets dans l'image. Ce temps est de moins d'une seconde pour des images avec peu d'objets à plus de 40 secondes pour des scènes denses.

3.4 . Conclusion et perspectives

Nous avons proposé un nouveau modèle de détection d'interactions, nommé CALIPSO. Il estime toutes les interactions de manière efficace et simultanée entre toutes les personnes et les objets en effectuant un seul passage sur l'image et quel que soit le nombre d'objets et d'interactions dans l'image. Cette complexité constante est obtenue grâce à une stratégie d'apprentissage par métrique qui regroupe le sujet et la cible en interaction, et écarte tous les objets qui n'interagissent pas. En outre, l'ajout d'une tâche d'estimation de la présence de la cible ainsi que d'une estimation du verbe passif redondant centré sur l'objet permet d'améliorer les performances. CALIPSO présente des résultats compétitifs par rapport à l'état de l'art sur

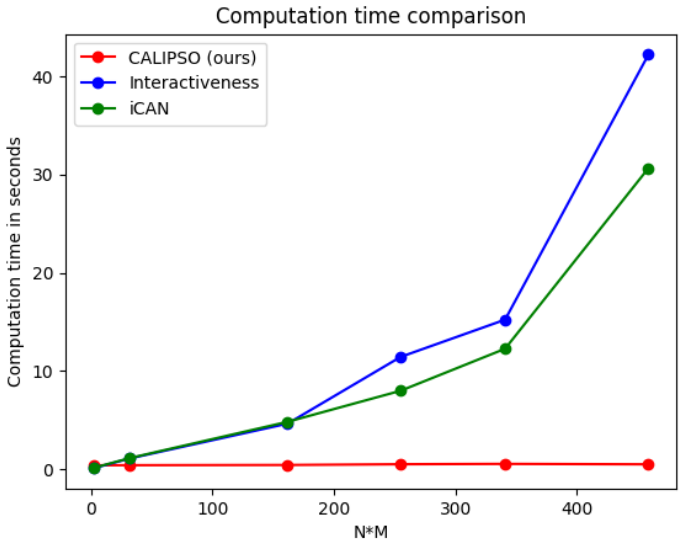


Figure 3.12 – Temps de calcul en secondes pour CALIPSO, Interactiveness [LZH+19] et iCAN [GZH18] pour des nombres croissants de paires potentielles présentes dans l'image.

deux ensembles de données largement utilisés, tout en étant beaucoup plus scalable avec le nombre d'interactions dans l'image.

L'un des intérêts de CALIPSO est d'être agnostique au détecteur d'objet, si un nouveau détecteur d'objet plus performant est disponible, il peut être facilement utilisé par CALIPSO sans ré-apprendre le modèle de détection d'interactions. Cependant le fait d'utiliser un détecteur d'objet externe peut être limitant dans le cas où le temps de calcul ou les contraintes d'exécution (en terme de mémoire par exemple) ne permettent pas l'usage de deux modèles séparés.

Une nette amélioration de CALIPSO est étudiée dans le chapitre suivant. Notre idée consiste à apprendre à reconnaître les objets en même temps que l'entraînement des interactions pour un partage efficace des représentations. Notre approche permet une meilleure compréhension des interactions et ainsi augmente les résultats de détection des HOI.

4 - Détection simultanée d'instances et d'interactions

Les interactions humaines, telles que les interactions sociales et violentes, ne sont généralement pas prises en compte dans les jeux de données publics pour la détection des interactions humain-objet (HOI). Comme nous pensons que ces types d'interactions ne peuvent être ignorés et décorrélés des HOI lors de l'analyse du comportement humain, nous proposons un nouveau jeu de données d'interaction pour traiter les deux types d'interactions humaines : Human-to-Human-or-Object (H^2O). De plus, nous introduisons une nouvelle taxonomie de verbes, destinée à être plus proche d'une description de l'attitude du corps humain en relation avec les cibles environnantes de l'interaction, et plus indépendante de l'environnement. Contrairement à certains ensembles de données existants, nous nous efforçons d'éviter de définir des verbes synonymes lorsque leur utilisation dépend fortement du type de cible ou nécessite un niveau élevé d'interprétation sémantique.

De manière à proposer des résultats de référence compétitifs sur le nouveau jeu de données H^2O, nous proposons une extension de CALIPSO, appelée DIABOLO (*Detecting InterActions By Only Looking Once*), qui a l'avantage d'intégrer un détecteur d'objets. En effet, CALIPSO est une architecture modulaire séquentielle dont le déploiement peut être freiné par le coût mémoire qu'implique de lancer deux réseaux distincts. DIABOLO est une architecture multi-tâches dont l'optimisation est réalisée de bout en bout. De plus, nous montrons que le partage des représentations entre les deux tâches améliore la détection des interactions et permet à DIABOLO d'afficher les meilleurs résultats de l'état de l'art sur le jeu de données V-COCO. La figure 4.1 présente les différences d'architecture entre CALIPSO et DIABOLO.

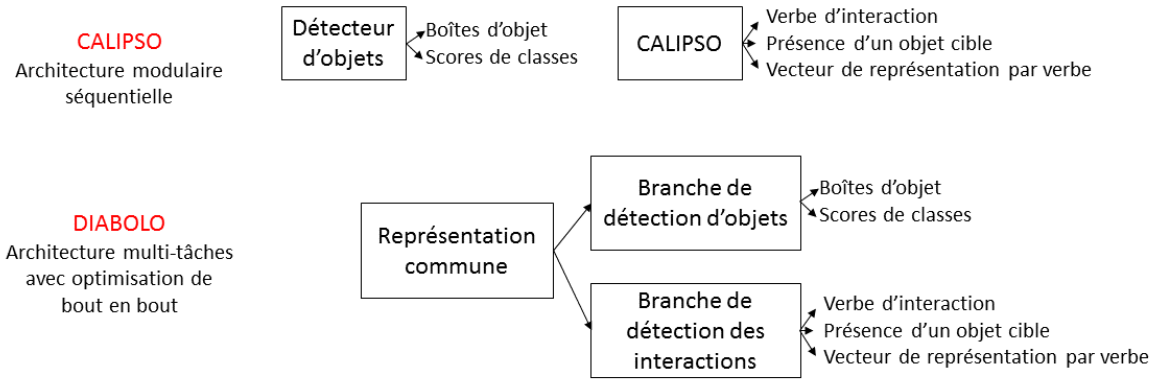


Figure 4.1 – Différences d'architecture entre CALIPSO et DIABOLO

L'état de l'art des travaux présentés dans ce chapitre est le même que celui du chapitre précédent, établit section 3.1.

Sommaire

4.1 . Jeu de données proposé

Plusieurs jeux de données d'images pour HOI ont été mis à disposition [GM15, CLL+18]. Cependant, ils se concentrent sur des cibles non humaines, appelées "objets" ci-après. Par conséquent, de nombreuses interactions humaines, telles que les interactions sociales ou violentes, ne sont pas prises en compte. Pour analyser le comportement humain, les interactions Humain-à-Humain (HHI), c'est-à-dire les interactions entre personnes, ne peuvent être ignorées et décorrélées des HOI. Par exemple, dans les applications de vidéo protection, il est intéressant de reconnaître les personnes qui se battent, de les distinguer des personnes qui s'enlassent, mais aussi de détecter les personnes qui donnent des coups de pied dans du matériel urbain. Le manque de jeux de données traitant des deux types d'interactions est la première motivation pour proposer un nouveau jeu de données appelé H^2O (*Human-to-Human-or-Object*).

De plus, les taxonomies utilisées par les ensembles de données existants [GM15, CLL+18] sont parfois ambiguës. Par exemple, certains types d'interactions correspondent à des verbes synonymes (par exemple, inspecter ou regarder un objet, tenir ou porter un téléphone portable, lire ou regarder un livre...) qui parfois ne diffèrent que par le type de cible (par exemple, surfer, faire du snowboard, ou faire du skateboard) ou nécessitent un haut niveau d'interprétation sémantique du contexte ou de l'intention (par exemple, tenir, cueillir ou acheter une pomme, monter ou s'asseoir sur un cheval). Inversement, certains verbes anglais peuvent fusionner différents

types de postures ou d'attitudes humaines qui pourraient être distinguées dans une autre langue (par exemple, "to ride a horse or a bus" ne correspondent pas exactement à la même attitude corporelle par rapport à l'objet cible). Ceci nous motive à introduire une nouvelle taxonomie de verbes, destinée à être plus proche d'une description de l'attitude du corps humain en relation avec les cibles d'interaction environnantes, et moins dépendante de l'environnement, du type de cible ou de l'arbitraire linguistique. Pour constituer le jeu de données H^2O, nous avons ré-annoté les images du jeu de données V-COCO [GM15] avec cette nouvelle taxonomie incluant à la fois les cibles objet et humaine, et ajouté de nouvelles images pour enrichir le jeu de données avec les verbes HHI. La figure 4.2 présente des images du jeu de données H^2O contenant à la fois des HOI et des HHI.

Figure 4.2 – Exemples d'images issues du jeu de données H^2O contenant à la fois des HOI et de HHI.

Nous présentons d'abord les modalités qui constituent le jeu de données, la taxonomie choisie pour annoter les interactions et nous la comparons aux jeux de données actuellement disponibles. Ensuite, nous présentons les métriques permettant d'évaluer les performances des méthodes de détection d'interactions humaines.

4.1.1 . Composition de H^2O

H^2O est composé des 10 301 images de V-COCO [GM15] auxquelles sont ajoutées 3 666 images sélectionnées sur le web comme pour le jeu de données COCO [LMB+14] et qui contiennent principalement des interactions entre personnes. Ainsi, contrairement aux jeux de données actuellement disponibles, H^2O présente des interactions entre personne et objet mais

aussi entre personne et personne. En ce qui concerne les annotations d'objets, toutes les instances en interaction sont annotées avec des boîtes englobantes, même si elles n'appartiennent pas aux 80 classes de COCO [LMB+14]. Au total, les instances sont réparties dans 214 classes. Cependant, sur un total de 128 969 instances annotées (58 225 personnes et 70 744 objets), 96% font partie des 80 classes de COCO. Les interactions sont annotées de manière exhaustive pour chaque personne, qu'elles soient réalisées avec un objet ou une autre personne. Ces annotations ont été réalisées avec l'outil d'annotation développé par le laboratoire, Pixano (https ://pixano.cea.fr).

Les annotations des images sont disponibles publiquement sous la licence "Creative Commons Attribution Non Commercial 4.0". Nous fournissons également les liens de téléchargement des images supplémentaires au jeu de données V-COCO sans avoir à les redistribuer directement.

4.1.2 . Taxonomie de H^2O

Pour annoter H^2O, nous avons défini une nouvelle taxonomie de verbes comprenant à la fois HOI et HHI. Nous avons créé cette taxonomie à la main en nous efforçant de lever les limites rencontrées dans les jeux de données déjà existants. En effet, cette taxonomie, présentée figure 4.3, se veut plus proche de l'attitude du corps humain par rapport aux cibles d'interaction environnantes, et moins dépendante de l'environnement dans lequel se déroulent les interactions, du type de cible ou du biais linguistique. Nous nous efforçons donc d'éviter les verbes synonymes lorsque leur utilisation dépend fortement du type de cible ou les verbes qui nécessitent un niveau élevé d'interprétation sémantique.

Le jeu de données H^2O est annoté avec 51 verbes répartis en cinq catégories :

- (i) les verbes décrivant la posture générale du sujet
- (ii) les verbes liés à la façon dont le sujet se déplace
- (iii) les verbes utilisés pour les interactions avec des objets
- (iv) les verbes décrivant les interactions entre humains
- (v) les verbes d'interactions impliquant la force ou la violence qui peuvent toucher soit des objets soit des personnes.

Les catégories de posture et de mouvement ont des verbes qui sont exclusifs et obligatoires. Cela signifie que chaque personne dans l'image doit être annotée avec un seul verbe de posture et un seul verbe de mouvement. H^2O contient 58 225 verbes de posture et autant de verbes de mouvement. Les verbes des trois autres catégories ne sont ni exclusifs, ni obligatoires. Cela signifie que les sujets peuvent être annotés avec aucun, un ou plusieurs de ces verbes d'interaction.

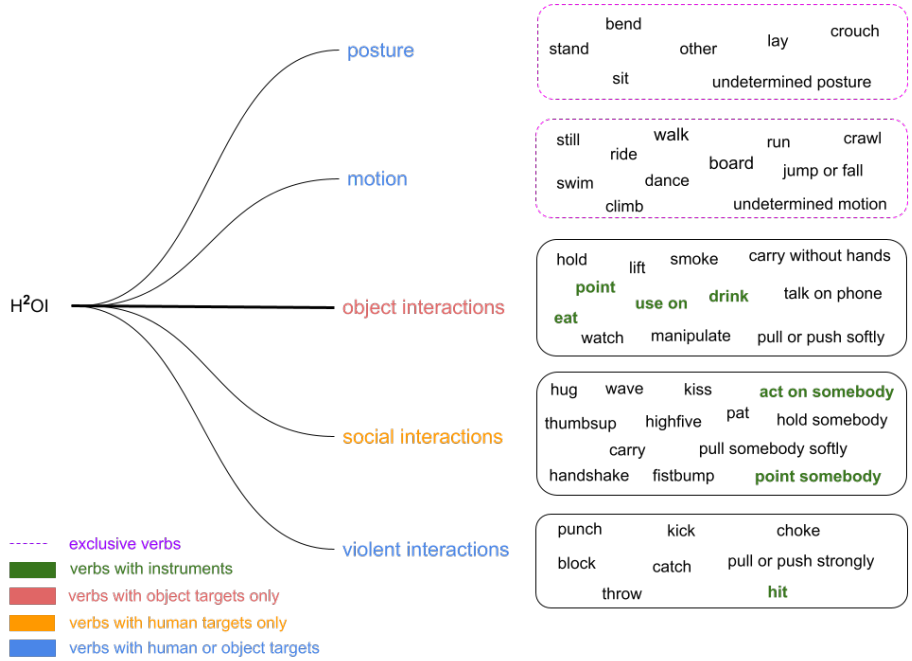


Figure 4.3 – Hiérarchie de la taxonomie proposée dans H^2O

Les verbes de posture – Les verbes de posture sont : "debout", "penché", "assis", "accroupi", "couché", "autre" et "posture indéterminée". La "posture indéterminée" est dédiée aux personnes tronquées dont la posture ne peut être déterminée avec certitude. Au contraire, "autre" signifie que le sujet est entièrement vu mais que sa posture n'est pas habituelle ou ne peut être décrite simplement. C'est par exemple le cas de certaines positions acrobatiques dans les images sportives.

Les verbes de mouvement – Les verbes de mouvement sont : "immobile", "marcher", "courir", "chevaucher", "sur une planche", "ramper", "sauter ou tomber", "danser", "nager", "grimper" et "mouvement indéterminé". L'expression "immobile" est dédiée aux personnes qui ne bougent pas. Le "mouvement indéterminé" est annoté pour les personnes qui sont tronquées et dont le mouvement ne peut être décrit avec certitude. Nous avons choisi d'annoter avec "sur une planche" tous les types de mouvement sur une planche (par exemple, skateboard, surf, snowboard, ski) pour ne pas être lié au contexte des interactions. Tous les verbes de posture et de mouvement peuvent avoir une cible et celle-ci est nécessairement la même pour les deux verbes. Par exemple, une personne peut être "debout" et "immobile" sur un tabouret. Dans la troisième ligne de la Figure 4.4, la distinction entre posture et mouvement avec la possibilité d'annoter une cible permet une description

précise d'une "personne accroupie se déplaçant sur un skateboard".

Les verbes d'interactions avec des objets – Ces verbes ne peuvent être réalisés qu'avec des objets. On sépare finement le fait de tenir un objet en quatre verbes : "tenir", "soulever", "porter sans les mains" et "tirer ou pousser doucement" car ils sont visuellement différents. L'expression "soulever" est réservée aux objets lourds qui doivent être soulevés à deux mains (par exemple, un canapé). "Porter sans les mains" est utilisé pour les objets qui sont portés sans les mains (sac à main ou sac à dos sur le dos ou sur l'épaule). "Tirer ou pousser doucement" est réservé aux objets roulants comme la valise, le caddie ou la poussette. Nous ne distinguons pas "tirer" de "pousser" car cette expression est ambiguë sur une seule image. "Manipuler" est annoté pour les personnes qui utilisent un objet pour sa fonction spécifique. Par exemple, ce verbe regroupe des verbes comme "couper", "brosser" ou "coller". Quatre verbes de cette catégorie acceptent jusqu'à deux types d'objets en interaction : la cible finale de l'interaction et l'outil ou l'instrument utilisé pour exécuter l'interaction sur l'objet cible. Ces verbes sont : "pointer", "utiliser sur", "manger" et "boire". "Manger" et "boire" sont annotés comme tels, uniquement lorsque le sujet fait le geste de porter quelque chose à la bouche (par exemple, s'asseoir autour d'une table avec un plat ne signifie pas que la personne est en train de le manger). Enfin, "regarder", "parler au téléphone" et "fumer" font également partie de cette catégorie.

Les verbes d'interactions sociales – Les verbes de cette catégorie sont exclusivement liés aux interactions entre les personnes. Les verbes d'interaction sont : "faire un câlin", "embrasser", "serrer la main", "faire un signe de la main", "taper dans la main", "tape du poing", "lever le pouce", "tapoter", "tenir quelqu'un", "tirer ou pousser doucement quelqu'un", "porter quelqu'un", "pointer quelqu'un" et "agir sur quelqu'un". Les expressions "pointer" et "agir sur" peuvent être utilisées avec un instrument (le cas échéant) qui permet de réaliser l'interaction, ainsi qu'avec la cible humaine finale (par exemple, un médecin "agit sur" un patient "avec" un stéthoscope).

Les verbes d'interactions violentes – Les cibles des verbes d'interaction de cette catégorie peuvent être soit une personne, soit un objet. Les interactions sont réalisées avec force ou violence. Les verbes sélectionnés dépendent fortement des parties du corps impliquées : "coup de poing", "coup de pied", "étranglement", "bloquer", "tirer ou pousser fortement", "lancer", "attraper" et "frapper". Si cela est approprié, "frapper" peut être annoté avec un instrument (par exemple, "frapper" la balle "avec" une batte de baseball).

L'annotation de toutes ces interactions est une tâche très lourde. Pour éviter le plus possible les erreurs, nous avons développés un module d'annotation pour chacune des 5 catégories de verbe et nous avons donné comme consignes aux annotateurs de traiter toutes les images avec une catégorie de verbe avant de passer au module suivant. Cela permet à l'annotateur de se concentrer sur une sous-partie des verbes. De plus, une pré-annotation automatique était réalisée en utilisant les annotations de détection des images issues de V-COCO ainsi que les annotations d'interactions, converties dans la nouvelle taxonomie. Enfin, pour la catégorie des verbes de posture, toutes les personnes étaient pré-annotées avec le verbe "debout" et avec le verbe "immobile" pour la catégorie des verbes de mouvement. En effet, statistiquement, une grande majorité des images présentent des personnes debout et immobiles. Les différentes règles expliquées dans chaque catégorie sur l'exclusivité des verbes et le type d'objet cible ont également été intégrées dans l'outil pour éviter les erreurs et les oublis.

4.1.3 . Comparaison avec les bases de données existantes

Le jeu de données V-COCO [GM15] est un sous-ensemble du jeu de données COCO [LMB+14] où chaque personne est annotée avec 26 verbes d'interaction sur 80 catégories d'objets. Contrairement à H^2O, quatre verbes ("debout", "sourire", "courir" et "marcher") ne mettent pas en jeu de cible. De plus, chaque interaction est limitée à un seul objet cible pour un sujet donné. Ces deux points ne permettent pas de décrire la scène de manière exhaustive. Par exemple, si une personne est debout sur un tabouret et tient deux objets différents en même temps, ces limitations forcent la description partielle d'une personne debout tenant un seul objet. Dans l'image de la deuxième ligne de la figure 4.4, V-COCO annote seulement une des deux valises tirées par la personne. De plus, contrairement à H^2O, toutes les interactions qui font partie des 26 verbes ne sont pas annotées de manière exhaustive dans l'image (par exemple, la femme immobile de la première ligne de la Figure 4.4 n'est pas annotée). La figure 4.4 illustre les différences entre les annotations H^2O et V-COCO.

Le jeu de données HICO-DET [CLL+18] est plus grand et plus diversifié que le jeu de données V-COCO car il contient 117 catégories d'actions sur les mêmes 80 catégories d'objets que le jeu de données COCO. Cependant, ces prédicats sont très spécifiques et trop liés au contexte de la scène. Par exemple, "dribbler" est toujours fait avec un ballon de sport ou "glisser" est toujours fait avec une planche.

Contrairement à H^2O, dans ces deux jeux de données, les objets en interaction en dehors des 80 classes de COCO ne sont pas annotés. Les tableaux 4.1

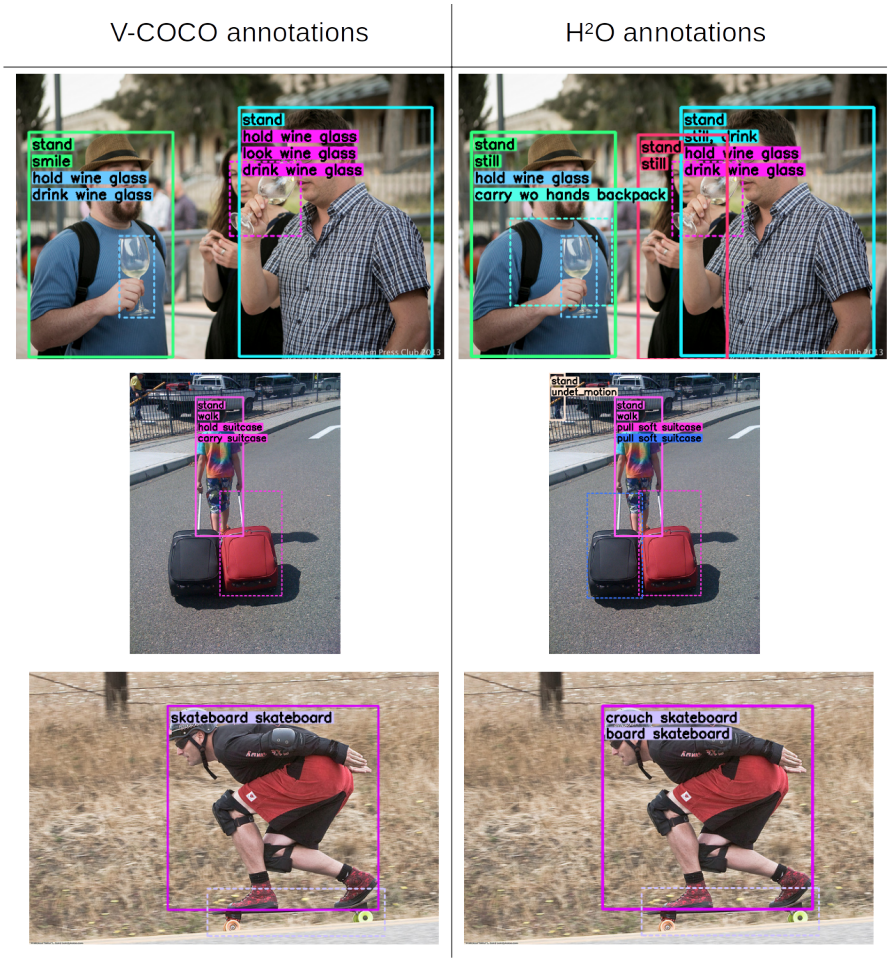


Figure 4.4 – Comparaison des annotations entre H^2O et V-COCO

et 4.2 présentent des statistiques sur H^2O par rapport à V-COCO et HICO-DET. Le tableau 4.1 montre la quantité d'images, le nombre de verbes et leur type de cible. Le tableau 4.2 compare le nombre global d'interactions entre H^2O et V-COCO, et détaille le nombre d'interactions par catégories. Comme les catégories de posture et de mouvement sont obligatoires, donc exhaustivement annotées dans H^2O, leur quantité est beaucoup plus importante que pour V-COCO. Le tableau compare également le nombre moyen de personnes par image, et le nombre moyen d'objets par image.

4.1.4 . Protocole d'évaluation

V-COCO [GM15] propose une évaluation des détections de HOI basée sur deux métriques. La première est la précision moyenne de l'agent, AP_{agent}, qui mesure la précision de la paire $< sujet, verbe >$. La seconde, la précision moyenne du rôle, appelée AP_{role}, analyse l'ensemble du triplet $< sujet, verbe, cible >$ en le considérant comme un vrai positif lorsque toutes

Jeu de données	#images	#verbes	Type de cible
HICO-DET [CLL+18]	47 774	117	objet
V-COCO [GM15]	10 346	26	objet
H^2O	13 967	51	objet, personne

Table 4.1 – Comparaison des jeux de données selon le nombre d'images et de verbes

Jeu de données	#interactions		#personne par image	#objet par image
V-COCO [GM15]	**Total :**	**49 019**	3.8	4.9
	Posture + Mouvement :	22 480		
	HOI :	26 539		
H^2O	**Total :**	**151 816**	4.2	5.1
	Posture + Mouvement :	116 450		
	HOI :	25 984		
	Social :	5 413		
	Violence :	3 969		

Table 4.2 – Comparaison des jeux de données selon le nombre d'annotations

les composantes sont correctes. Les boîtes prédites de l'humain et de l'objet sont correctes si elles ont un IoU supérieur à 0,5 avec les boîtes de vérité terrain. Deux métriques AP_{role} différentes sont proposées. Elles diffèrent dans l'évaluation du triplet $< sujet, verbe, \emptyset >$ qui apparait lorsque l'objet cible n'est pas vu, n'existe pas ou ne fait pas partie des 80 classes de COCO. Dans le premier cas (AP_{role1}), le fait de ne pas prédire le $\emptyset$ comme cible est pénalisé alors que dans le second cas (AP_{role2}), il ne l'est pas. De plus, le jeu de données V-COCO n'attend qu'un seul objet cible pour un verbe et une personne donnés. Par conséquent, le calcul de l'AP_{role} est limité à cette hypothèse.

Le jeu de données H^2O fournit une annotation des cibles même si l'objet ne fait pas partie des 80 classes de COCO. Le triplet spécifique $< sujet, verbe, \emptyset >$ dans H^2O signifie que la cible n'est pas vue ou n'existe pas. Par conséquent, nous proposons un nouveau scénario appelé "Objectness" où un objet cible en dehors des 80 classes de COCO doit être détecté et correctement associé à l'interaction. L'étiquette de classe d'une telle cible

est "autre".

H^2O fournit également plusieurs objets en interaction pour un verbe donné et une personne donnée s'ils existent. Par conséquent, la métrique AP_{role} a été adaptée pour prendre en compte cette nouvelle caractéristique. Pour plus de clarté, le scénario original de V-COCO sera appelé "original" par opposition au nouveau scénario "Objectness".

4.2 . DIABOLO : Détection et classification simultanées des interactions par une approche multi-tâches

Nous proposons une extension de CALIPSO intégrant un détecteur d'objets : DIABOLO (*Detecting InterActions By Only Looking Once*). Comme CALIPSO, DIABOLO est une méthode efficace de détection de toutes les interactions en un seul passage sur l'image, avec un temps d'inférence constant et indépendant du contenu de l'image. Contrairement à CALIPSO, ce réseau multi-tâches détecte simultanément toutes les personnes et tous les objets de la scène en même temps que l'estimation de toutes les interactions. Nous montrons comment le partage d'un réseau pour ces tâches permet non seulement d'améliorer les performances de manière collaborative mais aussi d'économiser des ressources de calcul.

4.2.1 . Méthode proposée

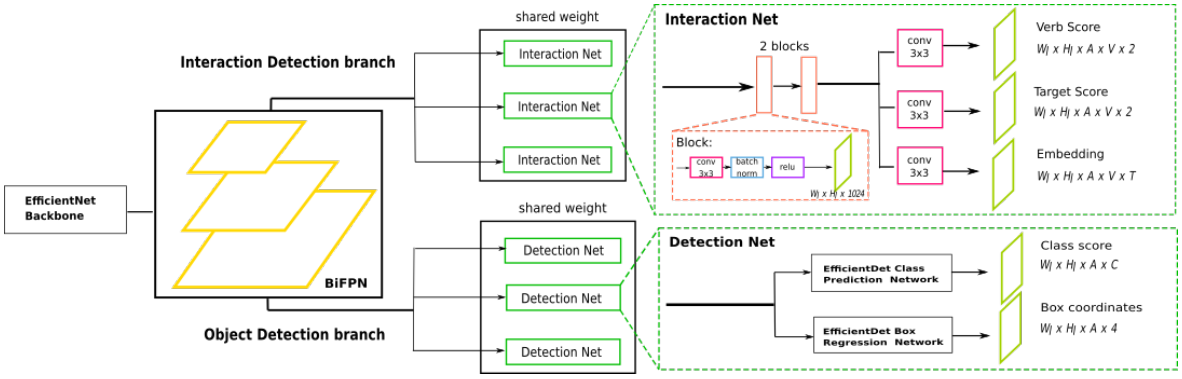


Figure 4.5 – Architecture du réseau DIABOLO avec deux branches pour la détection des interactions (en haut) et la détection des objets (en bas). W_l et H_l désignent respectivement la largeur et la hauteur de la carte de caractéristiques de la pyramide au niveau l. A est le nombre d'ancres, V est le nombre de verbes et T est la taille du vecteur de représentation.

DIABOLO est basée sur la méthode CALIPSO, présentée dans le chapitre 3, qui estime de manière dense les interactions sur une grille d'ancres de détection classique grâce à trois tâches. La première tâche définit le score

d'action de chaque ancre. La deuxième tâche estime pour chaque verbe la présence d'une cible en interaction pour chaque ancre humaine. La troisième tâche donne un vecteur de représentation pour chaque ancre dans l'image afin d'associer les paires en interaction. Ainsi, au moment de l'inférence, CALIPSO a besoin d'un détecteur d'objets externe pour indiquer les ancres qui correspondent réellement à des personnes ou des objets. Contrairement à CALIPSO, DIABOLO intègre un détecteur d'objets, basé sur la méthode EfficientDet [TPL20]. Ainsi, DIABOLO détecte simultanément les personnes, les objets et leurs interactions dans l'image en une seule passe.

La figure 4.5 illustre l'architecture du réseau de neurone multi-tâches utilisée par DIABOLO. Les branches de détection d'objet et d'estimation d'interaction partagent la même backbone EfficientNet suivie d'un BiFPN introduit par [TPL20].

Branche de détection d'objet

Notre architecture a été conçue pour pouvoir intégrer facilement tout modèle de détection d'objets en une passe sur l'image et basé sur une grille d'ancres. C'est par exemple le cas des détecteurs SSD [LAE+16] ou YOLO [WBL22]. Pour notre étude, nous avons choisi EfficientDet [TPL20] car il est le détecteur d'objet le plus performant au moment des travaux. Les sorties de la branche de détection d'objets sont la classification des instances et la régression de leurs boîtes englobantes. Pendant l'apprentissage, DIABOLO supervise ces deux tâches de la même manière que [TPL20].

Branche d'estimation des interactions

La branche d'estimation d'interaction de DIABOLO est proche de celle de CALIPSO. En effet, cette branche estime de manière dense trois tâches sur la grille d'ancre : la tâche de prédiction du verbe à la forme active et à la forme passive, la tâche d'estimation de la présence de la cible et la tâche de calcul du vecteur de représentation de l'interaction. L'utilisation d'une backbone EfficientNet et d'un BiFPN au lieu d'un FPN ResNet permet de réduire le nombre de convolutions contenues dans le réseau d'interaction. Contrairement à CALIPSO, le réseau d'interaction de DIABOLO est composé de seulement deux blocs convolutifs après le BiFPN. La supervision de la tâche d'estimation de la présence de la cible et de la tâche du calcul du vecteur de représentation de l'interaction sont les mêmes que dans CALIPSO : respectivement une perte d'entropie croisée binaire et une perte d'apprentissage par métrique "pull-push".

Cependant, DIABOLO utilise une *focal loss* [LGG+17] pour prédire de manière dense le verbe d'interaction. La *focal loss* FL vise à mieux gérer le déséquilibre d'occurrence des classes et est calculée comme suit :

$$FL = \sum_{v \in V} \sum_{a \in A^+} -\alpha (1 - p_a^v)^\gamma \log(p_a^v) \tag{4.1}$$

où v est un verbe de l'ensemble V, α et γ sont des paramètres de la *focal loss* et A^+ désigne l'ensemble des ancres associées aux personnes en interaction. Dans l'équation de la *focal loss*, p est la probabilité estimée du modèle pour la classe v et l'ancre a, p_a^v est égal à p si la classe de vérité terrain de l'ancre a est v, $1 - p$ sinon.

Inférence

À partir de la branche de détection, la localisation des personnes et des objets dans l'image est calculée avec un algorithme traditionnel de suppression de non-maxima (NMS), sauf que les indices des ancres sélectionnées sont utilisés pour lire directement les informations relatives aux différentes interactions dans la branche d'interaction.

Pour chaque personne détectée, un score d'interaction pour chaque verbe et chaque cible possible pour ce verbe, selon sa catégorie (voir figure 4.3), est calculé comme expliqué section 3.2.3. Les catégories d'interactions exclusives sont traitées indépendamment en ne fournissant que les triplets du verbe ayant la plus forte probabilité dans sa catégorie.

4.2.2 . Expériences

Les expériences visant à évaluer les performances de DIABOLO par rapport à l'état de l'art sont présentées ici et une première base de référence sur le jeu de données H^2O est proposée.

Jeux de données et métriques utilisés

Les jeux de données – Nous avons choisi d'évaluer DIABOLO sur V-COCO [GM15] et sur notre nouveau jeu de données H^2O mais pas sur HICO-DET [CLL+18]. Comme mentionné dans la section 4.1.3, V-COCO présente toujours l'inconvénient de définir des prédicats trop étroitement liés au contexte, et HICO-DET présente le même problème dans une plus large mesure.

Les métriques d'évaluation – Pour évaluer les résultats sur V-COCO [GM15], nous utilisons le protocole d'évaluation standard, tel que présenté dans la section 4.1.4, en utilisant le scénario AP_{role1} qui est le plus difficile.

Remarquez que, comme dans les travaux précédents [GGDH18, GZH18] sur V-COCO, le verbe "pointer" n'est pas pris en compte car il a trop peu d'échantillons. Sur H^2O, les scénarios "Original" et "Objectness" sont tous deux utilisés pour l'évaluation.

Détails d'implémentation

La *backbone* d'EfficientNet, le BiFPN et la branche de détection sont initialisés avec des poids préalablement appris sur le jeu de données COCO de détection d'objets [LMB+14]. Nous utilisons EfficientDet-D3 pour une comparaison équitable avec l'état de l'art et EfficientDet-D1 pour l'étude d'ablation car il est plus léger que D3.

Pour l'apprentissage multi-tâches, nous comparons les stratégies consistant à apprendre la branche de détection d'instance (sur COCO) et à la geler ou non pendant l'apprentissage de la détection d'interaction, avec la stratégie consistant à poursuivre l'apprentissage de la détection d'instance en même temps que la détection d'interaction. Dans cette dernière stratégie, nous utilisons V-COCO ou H^2O pour entraîner l'ensemble du réseau de DIABOLO et, en même temps, COCO (images V-COCO exclues) pour poursuivre l'entraînement de la branche de détection. Nous verrons que les données V-COCO ne sont pas assez variées pour apprendre correctement la détection d'instance. En effet, V-COCO ne dispose que de 5 400 images d'entraînement alors que EfficientDet est généralement appris sur les 120 000 images d'entraînement de COCO. C'est pourquoi nous utilisons des lots mixtes avec des images de V-COCO et des images de COCO pour améliorer la branche de détection d'objets. Les proportions d'images de chaque jeu de données dans le lot sont constantes pendant l'entraînement (lot de 24 images COCO + 24 images H^2O sur le jeu de données H^2O et le tableau 4.4 précise la composition des lots sur la base de données V-COCO).

Deux entraînements différents sont effectués selon les scénarios d'évaluation mentionnés dans la section 4.1.4. Pour le scénario "Original", seuls les objets cibles des 80 classes de COCO sont pris en compte et pour le scénario "Objectness", tous les objets inhabituels exclus par les 80 classes sont ajoutés dans une nouvelle classe appelée "autre" que la branche de détection doit apprendre.

L'apprentissage est effectué sur des GPU NVIDIA A100-SXM4. DIABOLO est entraîné par descente de gradient stochastique (SGD), avec un taux d'apprentissage initial de 0,016, qui est ensuite réduit de 10 à 10 000 itérations. Les paramètres de la *focal loss* α et γ sont respectivement fixés à $0,25$ et 2.

Le retournement horizontal de l'image et le "color jittering" sont appliqués pour l'augmentation des données.

Résultats de DIABOLO sur V-COCO et comparaison avec l'état de l'art HOI

Le tableau 4.3 présente les résultats de DIABOLO sur le jeu de données V-COCO comparé aux meilleures méthodes de l'état de l'art et aux méthodes en une étape. DIABOLO est la méthode la plus performante, dépassant la meilleure méthode de l'état de l'art au moment des travaux, DIRV [FXSL21] de 12.1 p.p. avec 76.7% pour l'AP_{agent} et de 1.2 p.p. avec 57.3% pour l'AP_{role}. Nous pensons qu'une telle augmentation de l'AP_{agent} peut être expliquée par l'aspect centré sur le sujet de DIABOLO qui est plus adapté à la reconnaissance d'actions multiples. Remarquez que le résultat AP_{role} pour DIABOLO est obtenu sans utiliser aucune information ontologique (c'est-à-dire, sans utiliser de règles sur les types d'objet cible possible pour un verbe donné), contrairement à DIRV [FXSL21] qui perd 1.3 p.p. si aucune information ontologique n'est fournie.

Méthode	**Une étape**	$AP_{agent}(\%)$	$AP_{role}(\%)$
InteractNet [GGDH18]	✗	69.2	40.0
CALIPSO [COAL20]	✗	-	46.4
PMFNet [WZL+19]	✗	-	52.0
ConsNet [LYC20]	✗	-	53.2
MLCNet [SHRW20]	✗	-	55.2
UnionDet [KCKK20]	✓	-	47.5
DIRV [FXSL21] sans ontologie	✓	64.7	54.8
DIRV [FXSL21] sans flip	✓	64.1	55.2
DIRV [FXSL21]	✓	64.6	56.1
DIABOLO	✓	**76.7**	**57.3**

Table 4.3 – Méthodes de l'état de l'art pour la détection des HOI : Performances évaluées sur le jeu de test de V-COCO

Étude d'ablation pour les stratégies d'apprentissage de DIABOLO

Les méthodes UnionDet [KCKK20] et DIRV [FXSL21] entraînent d'abord le détecteur d'objets sur l'ensemble de données COCO, puis gèlent à la fois la backbone et la branche de détection des instances pour maintenir les performances de détection. Dans ce travail, nous choisissons d'apprendre conjointement la détection et l'interaction sur les jeux de données V-COCO et COCO. Pour mesurer l'impact de cette stratégie, nous utilisons la backbone

EfficientDet-D1 car le temps d'apprentissage est beaucoup plus rapide que D3.

Pour une analyse plus fine des résultats, les mesures sont également calculées avec des détections parfaites de la vérité terrain (GT), afin de dé-corréler les performances de détection des instances et des interactions. Comme le montre le tableau 4.4, l'architecture de la branche d'interaction de DIABOLO fonctionne mal si la backbone et la branche de détection sont figées puisque l'AP_{role} n'est que de 46.2% avec des détections parfaites. Cela s'explique par les liens entre les objets de différentes tailles qui ne peuvent pas s'apprendre correctement si le BiFPN est figé.

DIABOLO entraîné sans geler la backbone ni la branche de détection, atteint 61.3% avec une détection parfaite mais seulement 43% avec les détections du modèle. Ce résultat s'explique par les mauvais résultats de DIABOLO en matière de détection d'instance (18% de mAP sur l'ensemble de test de V-COCO) lorsqu'il est appris sur V-COCO uniquement pour les deux tâches. En continuant à alimenter DIABOLO avec COCO et V-COCO, les performances de détection d'instances reviennent à la normale (35% de mAP contre 42% qui est la performance de EfficientDet-D1 appris uniquement sur la détection d'objets) et l'AP_{role} atteint 51.1% avec les détections du modèle, ce qui montre l'effet positif de l'apprentissage conjoint.

Pour mieux évaluer la part d'erreur dans la détection d'interaction induite par les mauvaises détections d'instance, nous entraînons DIABOLO avec une meilleure backbone, EfficientDet-D3, et comparons une fois encore les résultats donnés par les détections du modèle avec ceux des détections parfaites (voir tableau 4.4). La détection des instances par DIABOLO atteint 47% de mAP sur le jeu de test de V-COCO. La tâche d'association sujet-cible est plus fortement impactée par la détection des instances que par l'estimation des verbes. En effet, l'utilisation de la vérité terrain pour la détection n'augmente que de 3.5 p.p. l'AP_{agent} alors que l'amélioration est de 8.7 p.p. pour l'AP_{role}. De plus, la branche d'interaction avec une détection parfaite atteint 80.2% pour l'AP_{agent} mais seulement 66.0% pour l'AP_{role}, ce qui laisse une marge d'amélioration pour la tâche d'association sujet-cible.

Résultats de DIABOLO sur le nouveau jeu de données H^2O

Résultats quantitatifs – Nous évaluons DIABOLO sur le nouveau jeu de données H^2O, les résultats sont présentés dans le tableau 4.5. DIABOLO obtient des scores d'AP_{agent} de 41% et 40.6% respectivement pour les scénarios "Original" et "Objectness". Pour la métrique AP_{role}, les scores sont respectivement de 25.26% et 23.68% pour les scénarios "Original" et

Backbone	**Branche de détection**	**Taille et composition du lot**	**Détection d'objet mAP (%)**	AP_{agent} **(%)**	AP_{agent} **avec détections GT (%)**	AP_{role} **(%)**	AP_{role} **avec détections GT (%)**
EfficientDet-D1	figée	12 V-COCO	**42**	59.8	61.9	37.3	46.2
EfficientDet-D1	apprise	12 V-COCO	18	71.1	77.5	43.0	61.3
EfficientDet-D1	apprise	12 V-COCO + 16 COCO	35	**75.7**	**79.5**	**51.1**	**64.0**
EfficientDet-D3	apprise	24 V-COCO + 32 COCO	**47**	**76.7**	**80.2**	**57.3**	**66.0**

Table 4.4 – Étude d'ablation pour les stratégies d'apprentissage multi-tâches DIABOLO. Les métriques sont évaluées sur l'ensemble de test de V-COCO. De gauche à droite, les colonnes correspondent respectivement à : le réseau utilisé pour la backbone, si la branche de détection est figée ou apprise, la taille et la composition du lot d'images en entrée, la mAP de détection d'objets, l'AP_{agent} calculé avec les détection du modèle, l'AP_{agent} calculé avec des détections parfaites, l'AP_{role} calculé avec les détections du modèle, et enfin l'AP_{role} calculé avec des détections parfaites.

"Objectness". Les résultats sur le scénario "Objectness" sont moins bons que sur le scénario "Original". En effet, dans le premier cas, l'évaluation demande que le détecteur ait correctement détecté un objet cible ne faisant pas parti des 80 classes de COCO ce qui est plus compliqué que de limiter l'estimation à un triplet $< sujet, verbe, \emptyset >$ dans le cas du scénario "Original". L'AP_{role} global est inférieur à celui de V-COCO, ce qui suggère que H^2O est plus difficile que V-COCO. Nous pensons que cela est dû à l'annotation de H^2O qui est plus exhaustive (le nombre d'interactions par image a plus que doublé), et à la taxonomie qui est plus liée à l'attitude corporelle du sujet, plutôt qu'au contexte d'interaction qui n'est plus un indice supplémentaire. En effet, malgré des performances quantitatives plus basses que pour V-COCO, H^2O et DIABOLO permettent d'accéder à des informations plus précises comme détaillé dans l'étude des résultats qualitatifs.

Scénario	AP_{agent}	AP_{role}
Original	41.0	25.3
Objectness	40.6	23.7

Table 4.5 – Baseline de DIABOLO sur H^2O

Résultats qualitatifs – La figure 4.6 montre les résultats qualitatifs de DIABOLO appris sur H^2O. Dans les exemples (b), (e) et (g), on peut voir que DIABOLO détecte bien les HHI et leur réciprocité comme "faire un calin", "donner un coup de poing", "taper dans la main" et "serrer la main". Dans l'illustration (a), DIABOLO est capable de détecter que deux objets sont tenus. De même, dans l'exemple (d), le livre et le parapluie sont correctement détectés comme étant tenus. Dans l'exemple (c), la taxonomie H^2O permet à DIABOLO de détecter correctement une personne debout sur ses skis. Enfin, dans l'exemple (f), DIABOLO prédit correctement que les valises sont tirées et non tenues. Cela illustre le fait que la taxonomie H^2O est plus proche de l'attitude corporelle et de la manière d'interagir avec une cible que le contexte d'interaction.

Temps de calcul

Le tableau 4.6 montre les temps de calcul de notre méthode par rapport à DIRV [FXSL21]. Les deux méthodes sont exécutées sur un GPU NVIDIA RTX2080Ti. Le temps d'inférence de DIABOLO avec EfficientDet-D3 est compétitif avec DIRV. Pour obtenir les meilleures performances, [FXSL21] appliquent leur méthode sur l'image et sa version retournée (voir tableau 4.3). Avec une telle stratégie, le temps d'inférence mesuré pour la méthode DIRV est de 132 ms, et de 129 ms pour DIABOLO.

Méthode	Backbone	Temps d'inférence (ms)
DIRV [FXSL21] sans flip	EfficientDet-D3	108
DIRV [FXSL21]	EfficientDet-D3	132
DIABOLO	EfficientDet-D1	89
DIABOLO	EfficientDet-D3	129

Table 4.6 – Temps de calcul

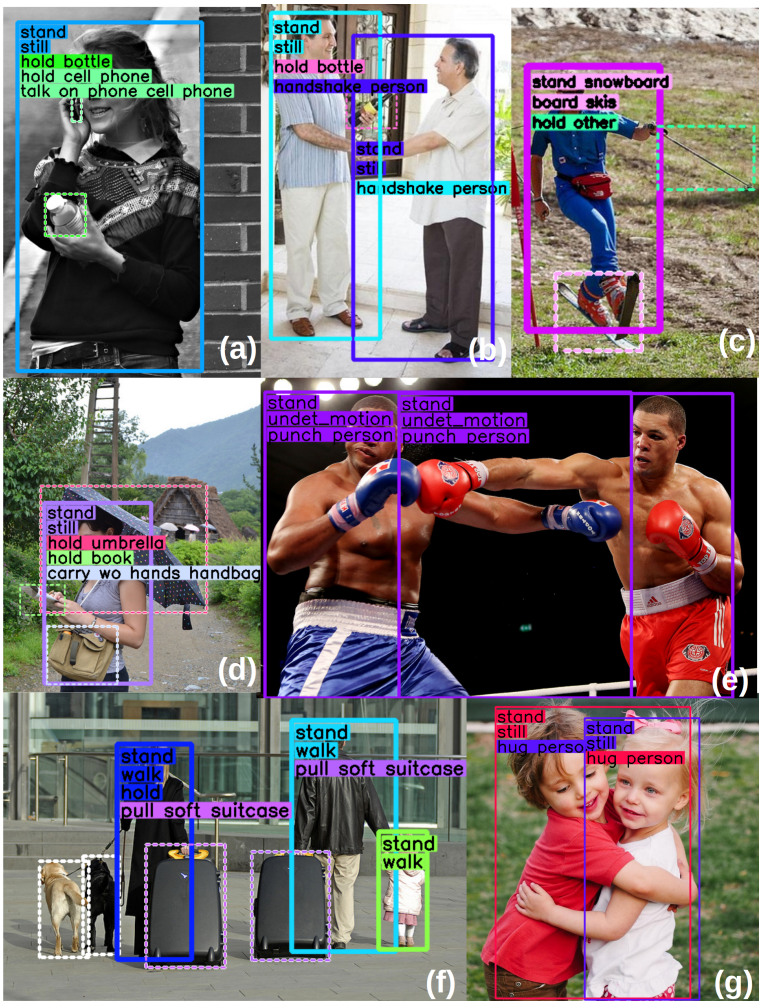


Figure 4.6 – Résultats qualitatifs de DIABOLO sur H^2O. Le verbe de l'interaction a la couleur de la boîte de la cible s'il y en a une, sinon le verbe est de la couleur de la boîte du sujet. Les boîtes en pointillé sont pour les objets et les boîtes en trait plein pour les personnes. Les boîtes blanches correspondent aux objets détectés qui n'interagissent pas.

4.3 . Conclusion et perspectives

Dans ce chapitre, un nouveau jeu de données complexe appelé H^2O a été présenté, il traite à la fois des interactions humaines avec d'autres personnes et avec des objets. Toutes ces interactions suivent une nouvelle taxonomie qui se concentre sur l'attitude corporelle du sujet plutôt que sur le type d'objet impliqué ou l'environnement.

Comme première base de référence pour H^2O, nous proposons DIABOLO, basé sur CALIPSO, une nouvelle méthode multi-tâches pour détecter

à la fois les instances et les interactions sans avoir besoin d'un détecteur externe. Il s'agit d'une méthode centrée sur le sujet et en une seule étape qui s'exécute en un temps rapide et constant, indépendamment du nombre d'instances dans l'image. Enfin, nous montrons expérimentalement que l'entraînement conjoint des détections d'interactions et d'instances améliore grandement la détection des interactions, ce qui fait de DIABOLO une approche solide qui surpasse l'état de l'art sur le jeu de données V-COCO.

L'AP_{role} de 57.3% sur le jeu de test de V-COCO nous montre bien que la tâche est encore loin d'être résolue, c'est principalement la tâche d'association entre sujet et cible qui reste complexe. En effet, la profondeur de l'image n'est pas explicitement modélisée dans l'architecture de DIABOLO, ce qui implique des erreurs d'association souvent liées au positionnement des objets dans les différents plans de l'image. Par exemple sur l'image figure 4.7, la personne à droite avec le sac à dos noir, détectée par une boîte verte, se trouve derrière la moto et non dessus mais DIABOLO l'a classifié comme conducteur de la moto à droite de l'image. Donner en entrée de DIABOLO une modélisation 3D de la scène permettrait d'améliorer les mauvaises associations liées à la profondeur de l'image. Différentes méthodes permettent aujourd'hui d'estimer la 3D à partir d'une simple image 2D [RLH+20, AW18, LRB+16, RBK21].

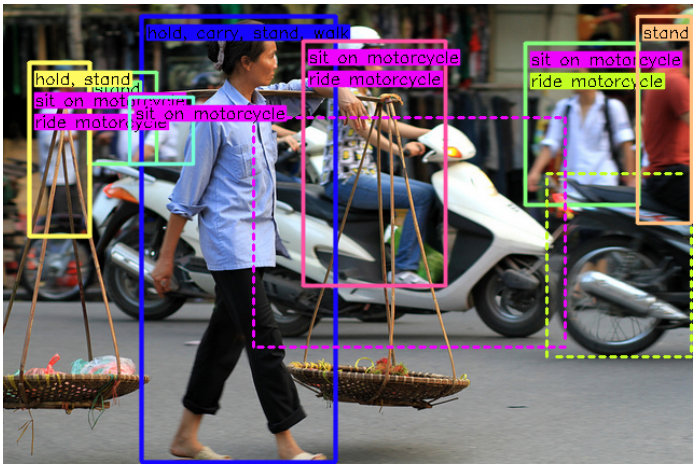


Figure 4.7 – Exemple de mauvaise classification d'interaction liée à la profondeur de l'image.

La nouvelle taxonomie proposée dans H^2O prend mieux en compte l'attitude corporelle des personnes. Intégrer une information sur la pose des personnes dans l'image en entrée du réseau permettrait au réseau de se concentrer sur les zones de contact souvent impliquées dans les interactions comme les bras et le bout des mains. De nombreux travaux cherchent à estimer la pose 3D des personnes à partir d'une seule image [BCL+20].

Bien que l'AP_{agent} lui, soit de 76.7%, ce qui signifie que la classification seule des interactions sans l'association de la cible atteint déjà des performances correctes sur V-COCO, ces performances se dégradent rapidement lorsque nous testons DIABOLO sur d'autres types d'image. Effectivement, les images de V-COCO sont plutôt centrées sur les personnes et le point de vue de la caméra est rasant. Or ce n'est pas le cas pour des caméras de vidéo protection où le point de vue est plongeant. DIABOLO n'est donc pas robuste aux changements de domaine. Une piste d'amélioration est d'utiliser des méthodes d'adaptation de domaine non supervisée qui sont des variantes de l'apprentissage semi-supervisé mais où les données non annotées proviennent d'une distribution différente de celle de l'ensemble de données annotées. L'avantage est de pouvoir améliorer les performances sur une base cible différente du jeu de données source sans avoir besoin d'annotation supplémentaire. Des méthodes telles que [LMH+21, KVRM19, LKL+20, ZY19] ont prouvé leur efficacité sur la détection d'objet mais aucune n'a été testé sur la tâche de détection d'interactions.

Récemment des travaux utilisant des *transformers* ont été publiés et surpassent les résultats de l'état de l'art. Les *transformers* permettent grâce aux mécanismes d'attention globale de faire apparaître des relations spatiales entre différentes régions de l'image, ils sont donc effectivement idéaux pour représenter des liens entre des instances. Par exemple, la méthode proposée dans [ZLW+22] atteint un AP^1_{role} de 66.2% sur le jeu de données V-COCO.

5 - Autres travaux de recherche sur l'analyse du comportement humain

Dans ce chapitre sont évoqués mes travaux dans le prolongement de la thématique de l'analyse du comportement humain auxquels j'ai contribué directement ou par un co-encadrement.

5.1 . Apprentissage auto-supervisé de représentations pour la classification d'actions

La création de jeux de données tels que DAHLIA, présenté chapitre 2, ou H^2O présenté chapitre 4, est une tâche coûteuse et fastidieuse. Or acquérir des données brutes images ou vidéos est relativement simple car le Web en accumule de plus en plus aujourd'hui. C'est pourquoi j'ai également axé mes recherches sur l'apprentissage de représentations sans données annotées de manière à exploiter l'ensemble des données brutes à notre disposition sans coût d'annotation.

Initialement, les travaux sur l'apprentissage non supervisé sont menés pour la classification d'images qui demande une analyse globale, ils ont ensuite été transférés à des tâches denses et également à la modalité vidéo. Cependant, cette dernière demande une compréhension spatio-temporelle largement plus complexe. Notre objectif est de proposer des méthodes efficaces de pré-entraînement d'un réseau de neurones par l'intermédiaire d'une tâche prétexte ou constrastive pour obtenir des représentations pertinentes pour l'analyse vidéo. C'est-à-dire des représentations encodant des caractéristiques visuelles à la fois spatiales et temporelles. En intégrant des connaissances sur la structure générique des données vidéo, il est possible d'apprendre un modèle robuste pour l'utiliser comme une initialisation efficace pour résoudre différentes tâches de reconnaissance à partir de ces données vidéo. Cette approche permet ainsi d'optimiser les résultats et limiter les coûts liés à l'annotation manuelle de données.

Au début de nos travaux en 2018, l'apprentissage non-supervisé est dominé par les méthodes génératives tels que les GAN (*Generative Adversarial Network*) ou les VAE (*Variational Auto Encoder*). Les méthodes de prédiction des images futures ont été les premières à être étudiées pour le pré-entraînement de modèles de représentation vidéo et ont suscité un grand intérêt [BFE+17, MCL15, TLYK18, D+17]. Cependant l'accent a été mis davantage sur la qualité des prédictions des données vidéo que sur l'apprentissage de représentations pertinentes pour des tâches d'interprétation sémantique. L'amélioration des performances en reconnaissance d'actions restait donc assez limitée. Les mé-

thodes auto-supervisée ont fait leur apparition. Elles sont basées sur l'apprentissage d'une tâche prétexte qui extrait une vérité terrain à partir de l'exemple et tente de la prédire [MZH16, WLZF18, AME19]. Ces méthodes sont les premières à avoir vraiment donné des résultats marquants pour l'apprentissage de représentations.

Dans le cadre de nos travaux, nous nous interessons à l'apprentissage auto-supervisé et la tâche cible choisie est la reconnaissance d'actions dans des vidéos, un modèle de classification est donc appris de façon supervisé lors d'une étape de spécialisation ou *fine-tuning* en anglais à partir des représentations extraites du réseau pré-entraîné de manière auto-supervisé et d'un faible nombre de clips vidéo annotés.

La méthode CPC pour l'auto-apprentissage vidéo

J'ai co-encadré la thèse de Guillaume Lorre, soutenue en 2021 au cours de laquelle nous avons proposé une nouvelle méthode auto-supervisée pour prédire le futur d'une vidéo. Elle se base sur la méthode *Contrastive Predictive Coding* (CPC), beaucoup étudié dans le cadre des méthodes génératives. La figure 5.1 présente le principe de la méthode CPC pour l'auto-apprentissage vidéo. Nous avons testé cette nouvelle approche sur 3 modalités issues de la vidéo : le flot optique, la différence d'images et les séquences d'images. Nous avons également appliqué différents protocoles d'évaluation pour l'apprentissage de la couche linéaire de classification : avec *fine-tuning* ou non de l'encodeur pré-appris, avec ou sans transfert entre base de données vidéo, avec une quantité de données annotées variable, en évaluant sur la tâche de classification d'actions ou sur la tâche de recherche de vidéos. Les résultats montrent que la méthode permet d'apprendre des représentations efficaces pour une tâche cible de reconnaissance d'actions, en particulier lorsque les données annotées sont rares. Cette méthode surpasse l'état de l'art pour la classification linéaire en utilisant les flots optiques sur UCF-101 [SZS12]. Les résultats prouvent également que notre méthode peut être utilisée avec des entrées diverses. Même si les performances sont plus faibles en utilisant les différences d'images et les images, les résultats sont grandement améliorés sur ces modalités par rapport à l'état de l'art, ce qui est intéressant car ces modalités sont beaucoup plus faciles à obtenir que le flot optique qui est coûteux à calculer. Les expériences prouvent également que nos représentations apprises de façon non supervisée peuvent être transférées à des ensembles de données similaires (de UCF-101 à HMDB51 [KJG+11] par exemple).

Cependant la méthode CPC fondée sur une approche contrastive a besoin d'un grand nombre d'exemple négatifs pour être performante,

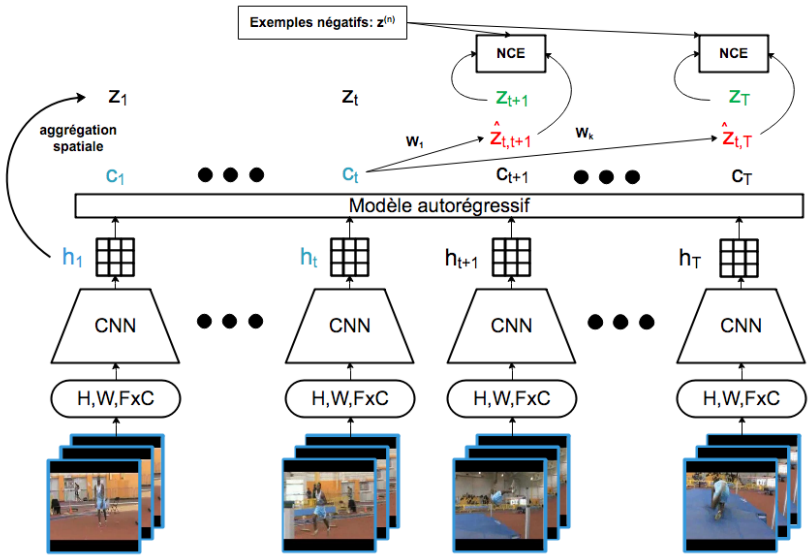


Figure 5.1 – Schema illustrant la méthode *Contrastive Predictive Coding* pour les vidéos. Les différents segments vidéos sont encodés par un CNN. Les cartes de caractéristiques passées en sortie sont agrégées par un modèle autorégressif. Le contexte c_t permet de prédire les représentations des segments futurs. Image tirée du manuscrit de thèse de Guillaume Lorre [Lor21]

ce qui pose vite un problème de capacité en mémoire pendant la phase d'apprentissage. Une autre limitation réside dans la définition d'un exemple négatif. En effet, deux images peuvent appartenir à la même classe et donc être proches visuellement. Cependant les méthodes contrastives ont pour objectif de repousser leur représentations au même titre que deux images complètement différentes.

La méthode SCE pour un apprentissage contrastif doux

Je co-encadre actuellement la thèse de Julien Denize, débutée en décembre 2020. Nous avons proposé une nouvelle approche pour répondre à la limitation liée à la définition d'un exemple négatif. L'originalité a été de proposer une nouvelle formulation de la fonction de perte, appelée SCE (*Similarity Contrastive Estimation loss*). Elle met en contraste des paires de vue augmentées de la même image avec d'autres instances tout en maintenant les relations entre les instances. SCE tire parti de l'apprentissage contrastif [HFW+20] et de l'apprentissage relationnel [ZYW+21] et améliore les performances par rapport à l'optimisation d'un seul des deux aspects. Les différences entre les espaces de représentation formés par l'apprentissage contrastif, l'apprentissage relationnel et notre apprentissage contrastif

doux, SCE, sont illustrées figure 5.2. SCE donne des résultats compétitifs par rapport aux méthodes de l'état de l'art, surtout en termes de temps d'apprentissage. En effet, pour obtenir des résultats similaires, SCE a une convergence beaucoup plus rapide.

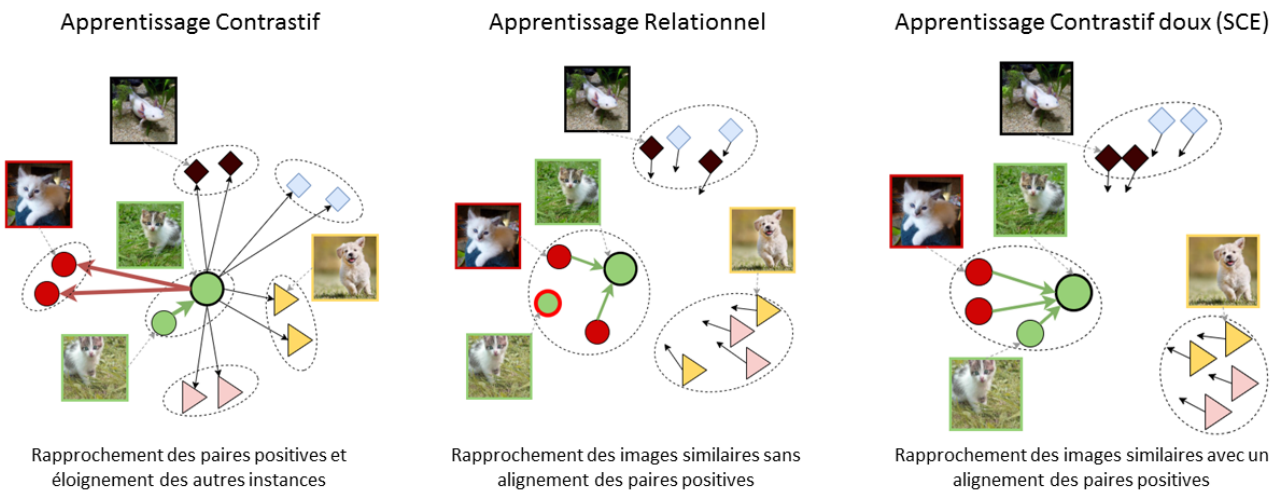


Figure 5.2 – Différences entre les espaces de représentation formés par l'apprentissage contrastif, l'apprentissage relationnel et notre apprentissage contrastif doux, SCE. Illustration issue du poster de la présentation de l'article [DRO+23] à la conférence WACV 2023.

Les publications associées à l'ensemble de ces travaux de recherche sont listées chapitre 7.

5.2 . Analyse du comportement dans le domaine sportif

Mon implication récente dans le projet TeamSports (appel à projet ANR JO 2024 sur la performance sportive) dont je suis cheffe de projet et référente technique, m'a permis d'être confrontée à d'autres limitations liées à l'analyse du comportement humain telles qu'une dynamique de mouvement différente, une densité de personnes potentiellement importante à gérer, et également un manque important de données annotées spécifiques au sport.

L'objectif du projet est de développer des outils d'analyse de la dynamique de groupe à destination des entraîneurs d'équipes de sports collectifs. Nous nous sommes vite rendu compte que pour analyser la qualité des interactions entre les joueurs tels que des regroupements, la technologie de base incontournable est le suivi des joueurs. Si le suivi de personnes est largement étudié dans le contexte de la video-protection grâce à un nombre important de jeux de données de référence à disposition tels que la base de données

MOT [MLTR+16], les méthodes de l'état de l'art sur ces jeux de données ne permettent pas de gérer les défis propres au contexte sportif : joueurs portant les mêmes maillots, entrant et sortant du champ de la caméra, faible résolution, postures inhabituelles, déplacements rapides, etc. La figure 5.3 illustre les différents défis liés au contexte sportif.

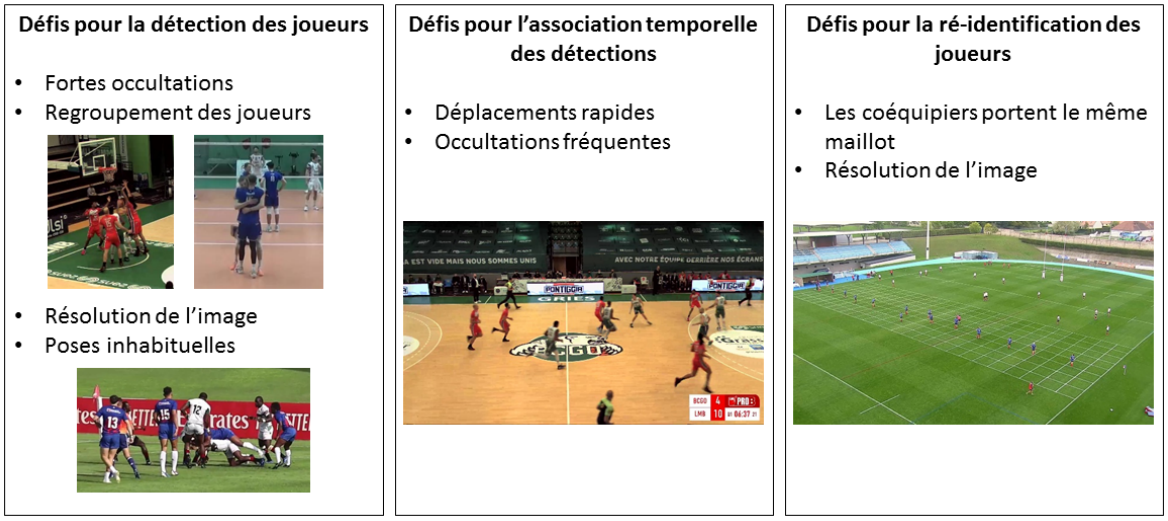


Figure 5.3 – Illustration des différents défis liés au contexte sportif

Notre première contribution porte sur la ré-identification des joueurs de sport collectif. Partant du constat qu'il était difficile pour les algorithmes classiques de ré-identification de personne de distinguer les joueurs d'une même équipe à cause de leurs maillots similaires, nous avons proposé une méthode incrémentale dont l'apprentissage se met à jour en fonction des corrections d'identité apportées par l'utilisateur. La figure 5.4 illustre ce mécanisme. Par manque de jeux de données annotés pour la ré-identification dans le domaine sportif, nous avons proposé et rendu publique une base de données de ré-identification composée de 3 extraits de matches de rugby à 7 de 40 secondes. Les performances de l'algorithme classique de ré-identification ont été fortement améliorées sur l'intégralité d'un match grâce à seulement 6 annotations par joueur.

De manière à automatiser la spécification du modèle de ré-identification à un nouveau match sans annotations supplémentaires, notre deuxième contribution a été de proposer une méthode de fine-tuning basée sur des *tracklets* (séquence temporelle de détection d'une même identité) courtes mais sans ambiguïté sur l'identité suivie, c'est-à-dire sans croisement ni occultation. Les triplets utilisés pour optimiser le modèle sont ensuite créés selon l'hypothèse qu'un joueur ne peut apparaître deux fois au même instant.

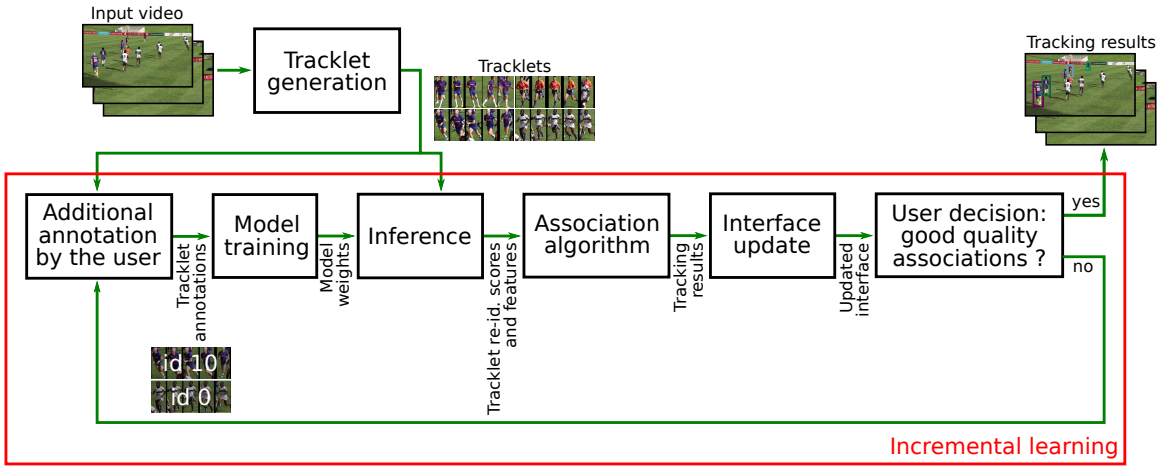


Figure 5.4 – Apprentissage incrémental de la classification des *tracklets*. L'utilisateur fournit des annotations pour entraîner le modèle à classer correctement les *tracklets* en fonction de l'identité du joueur.

La figure 5.5 illustre la sélection des exemples positifs et négatifs pour la création des lots d'apprentissage. Cette méthode a été intrégrée à un algorithme de suivi des joueurs qui a obtenu la première place au challenge SoccerNet Tracking 2022 [GCD+22]. Nous avons également proposé une méthode de calibration automatique des terrains de sports collectifs permettant suite à notre algorithme de suivi des joueurs de remonter à leur position 3D sur le terrain. La figure 5.6 illustre les résultats finaux de l'ensemble de la chaîne de traitement.

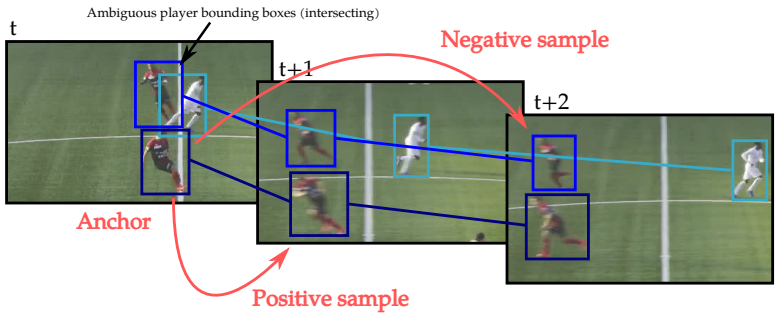


Figure 5.5 – Exemple de création des triplets pour l'optimisation du réseau de ré-identification sur un match donné. Les liens entre les boîtes de détection représentent les tracklets générées simultanément.

Les publications associées à ces travaux de recherche sont listées chapitre 7.

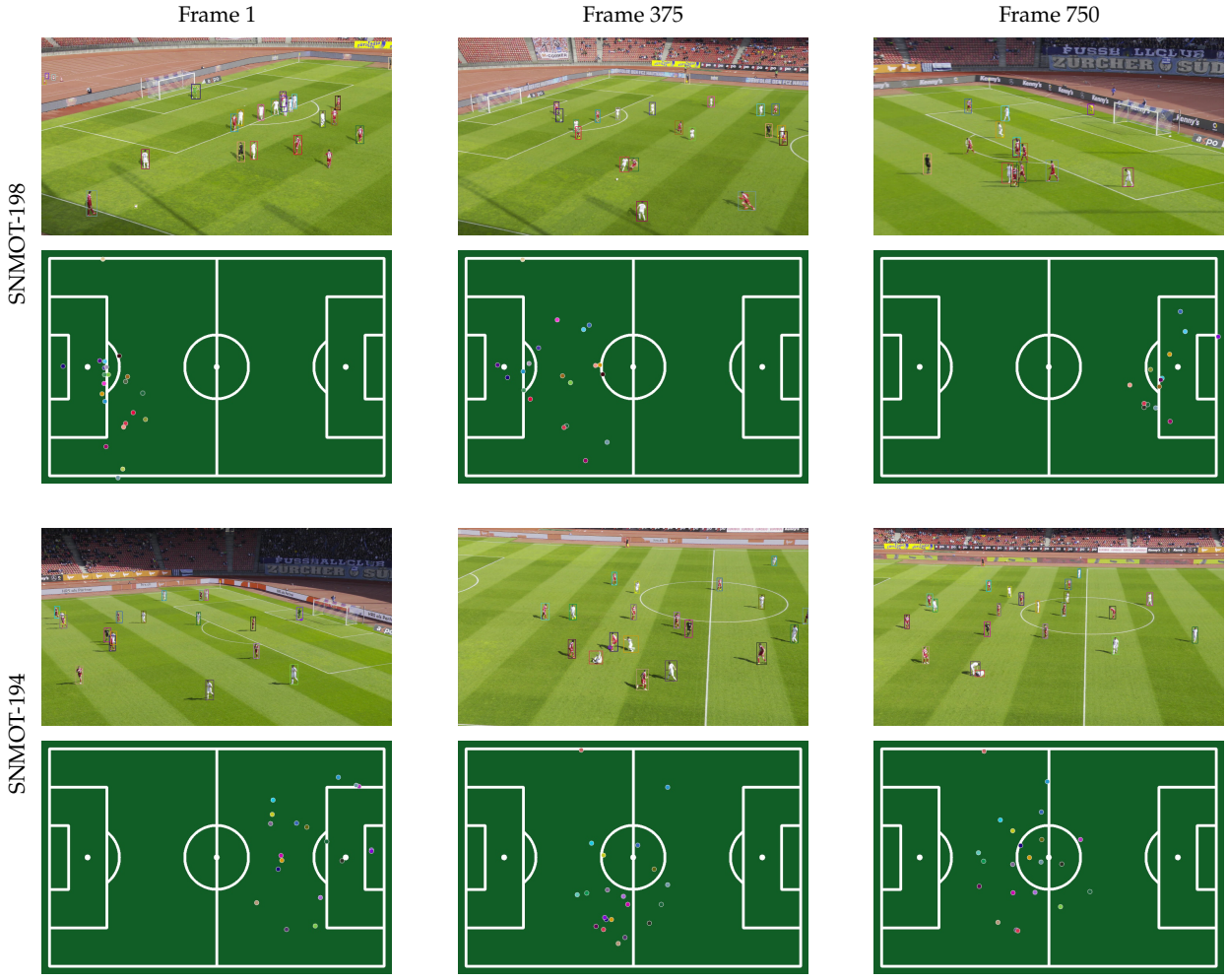


Figure 5.6 – Résultats de localisation individuelle des joueurs sur le terrain issus de la chaîne de traitement composée de la calibration automatique du terrain, de la spécialisation de la ré-identification à un match donné et du suivi des joueurs. Images issues du jeu de données SoccerNet Tracking [GCD+22]

6 - Conclusion et Perspectives

6.1 . Conclusion

L'analyse du comportement humain par vision permet d'aborder un large champ d'applications. Que ce soit pour le sport, la vidéo-protection, la santé ou l'industrie manufacturière, les algorithmes d'analyse de données images ou vidéos visent un niveau égal aux performances d'un humain qui observerait la même scène. Cependant les recherches actuelles se heurtent à de nombreux verrous scientifiques (détaillés chapitre 1). D'une part, une analyse à l'échelle pixellique ne permettant pas d'obtenir une compréhension proche de celle d'un humain, les modèles doivent être en capacité d'analyser les images dans leur globalité. D'autre part, les modèles d'apprentissage doivent également être robustes face à de nombreux critères tels que : le changement de point de vue, la vitesse d'exécution ou encore le contexte de réalisation d'une action. Cela implique des données d'apprentissage suffisamment représentatives de tous les cas réels. Or le coût de collecte et d'annotation de tels jeux de données sont un frein pour le développement des méthodes d'analyse du comportement. Enfin, beaucoup d'applications nécessitent des algorithmes embarqués, alors que l'analyse vidéo demande au contraire des ressources de calcul importantes ce qui limite aussi la généralisation et le déploiement de ces technologies.

Parmi toutes les façons d'aborder l'analyse du comportement humain, j'ai choisi d'orienter mes travaux selon deux problématiques de recherche : la reconnaissance d'activités dans des vidéos et la détection d'interactions dans des images. Mes contributions scientifiques sont partagées selon deux axes : les contributions algorithmiques et les contributions sur les données d'apprentissage.

Contributions algorithmiques

Lorsque j'ai commencé mes travaux sur la détection des interactions dans des images, toutes les méthodes de l'état de l'art proposaient des approches en deux étapes : une détection d'objets suivie par une estimation de l'interaction entre toutes les paires humain-objet possibles. Si la scène à analyser est dense, comme c'est souvent le cas en vidéo-protection par exemple, leur complexité quadratique avec le nombre d'objets dans la scène entraîne un temps de calcul qui explose. Ce constat m'a poussé à proposer la méthode **CALIPSO**, présentée chapitre 3, et son extension **DIABOLO**, présentée chapitre 4.

CALIPSO est la première méthode de classification d'interactions en un

seul passage sur l'image. CALIPSO a l'avantage d'être agnostique au détecteur d'objet puisque n'importe quel détecteur externe peut être utilisé sans réapprendre le modèle d'estimation des interactions. CALIPSO donne des résultats compétitifs avec l'état de l'art sur le jeu de données V-COCO tout en affichant un temps de calcul constant en plus d'être le plus efficace.

Étant donné que faire tourner deux modèles distincts pour chacune des tâches, détection d'objet et estimation des interactions, peut être un frein pour certaines applications dont les ressources en mémoire sont limitées, j'ai proposé une extension à CALIPSO, appelée DIABOLO, qui intègre un détecteur d'objet. DIABOLO est une méthode multi-tâches qui se révèle bénéfique pour l'estimation des interactions. En effet, l'étude de l'apprentissage conjoint des deux tâches a permis d'améliorer nettement les résultats puisque DIABOLO est la méthode donnant les meilleures performances sur le jeu de données V-COCO.

Aujourd'hui, la méthode [ZLW+22] enregistre les plus hauts résultats sur le jeu de données V-COCO avec un AP^1_{role} de 66.2%. Ce chiffre nous montre bien que la tâche est encore loin d'être résolue. Même si les architectures *Transformers* augmentent les performances en analysant les images de façon naturelle dans leur ensemble permettant ainsi d'accéder à un plus haut niveau sémantique, l'association humain-objet cible reste complexe à modéliser à partir d'une seule image. Des informations supplémentaires pourraient être données en entrée du réseau pour l'aider à mieux appréhender la scène comme la profondeur [RLH+22] ou l'estimation de la pose des personnes [BCL+20]. Ces données permettraient par exemple de mieux distinguer les interactions de contact des interactions à distance.

Même si la problématique de détection des interactions dans des images est résolue dans le futur, est-elle vraiment exploitable telle quelle? Le traitement d'une vidéo image par image donne très souvent des réponses non stables d'une image à l'autre. Il faudrait donc envisager une intégration temporelle, soit par l'utilisation d'un modèle traitant directement en entrée des clips vidéo soit en utilisant un filtre en post-traitement qui permettrait aux détections d'être plus stables dans le temps. Un algorithme de suivi d'objets par détection serait également nécessaire pour avoir une analyse temporelle complète. Mais comment passer de la détection brève d'interactions à la reconnaissance d'une activité longue ou d'un cas d'usage spécifique? Par exemple, le cas du bagage abandonné pourrait être traité avec des règles : si une personne portait un sac, que le sac en question n'est plus porté et que son propriétaire n'est plus à proximité, l'alerte du bagage abandonné peut être levée. Dans le cas d'une activité longue composée de multiple sous-actions, les règles ne sont pas aussi triviales à décrire même si intuitivement cela semble possible. Par exemple, l'activité "prendre un repas"

est a priori composée des actions "manger", "boire", "tenir un couteau, une fourchette", etc. L'utilisation d'une intelligence artificielle (IA) symbolique pourrait faciliter la modélisation de ces règles. Les IA symboliques ont l'intérêt d'être simple à utiliser en plus d'être facilement explicables et de proposer un raisonnement haut niveau. La figure 6.1 illustre une proposition pour passer de la détection des interactions par image à une détection d'activités longues.

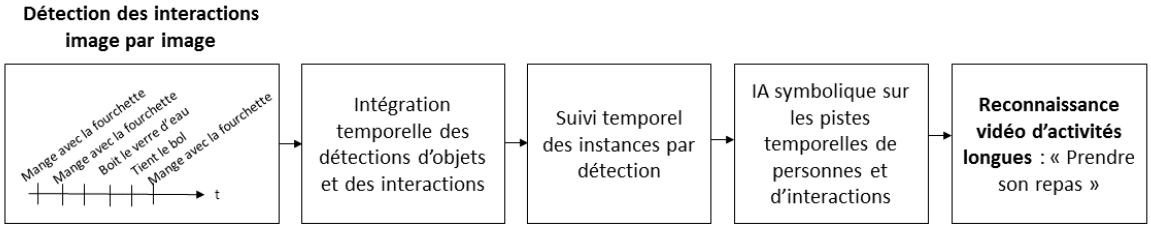


Figure 6.1 – Proposition d'une approche possible pour obtenir une détection vidéo d'activités longues à partir de détection d'interactions image par image.

Contributions sur les données d'apprentissage

Concernant les données d'apprentissage, mon intérêt pour l'analyse des activités longues de la vie quotidienne a vite été freiné par le manque de jeux de données publics. C'est pourquoi j'ai constitué la première base de données d'activités longues mise à la disposition de la communauté scientifique, **DAHLIA**, présentée chapitre 2. Cette dernière est composée de 51 séquences de 40 minutes en moyenne jouées par 44 sujets différents, non segmentées, enregistrées par 3 points de vue différents et annotés selon 7 activités. La réalisation de cette base de données a demandé beaucoup d'investissements en temps : de la réflexion sur les scénarios, au post-traitement des vidéos en passant par la mise en place de la captation. DAHLIA a été rendue publique après que nous nous soyons assurés qu'elle respectait les normes RGPD. Aujourd'hui, DAHLIA est activement utilisée par la communauté scientifique et permet d'inciter les chercheurs à proposer de nouvelles méthodes et à les évaluer de façon comparative.

Concernant la détection d'interactions, la problématique du manque de données n'était pas la même car il existait déjà des jeux sur lesquels la communauté scientifique s'évaluait. Cependant ces jeux de données ne permettaient pas de couvrir l'ensemble des interactions. Le premier point limitant était le manque de verbes d'interaction entre personnes notamment pour détecter les relations sociales ou les comportements agressifs. Le deuxième point était le choix optimal des taxonomies pour annoter les

bases de données. Les verbes d'interactions sont sémantiquement trop liés au contexte de réalisation qui devient un très fort indice pour le modèle et ne permet pas une généralisation de l'interaction à un autre contexte. Ces différents constats m'ont amenée à constituer le jeu de données d'images **H²O**, présenté chapitre 4, qui est aujourd'hui le seul à proposer à la fois des interactions entre des personnes et entre des personnes et des objets. Un travail de réflexion a également été mené pour proposer une nouvelle taxonomie constitué de verbes permettant de décrire la gestuelle de l'interaction plutôt que le contexte d'exécution. La baisse de performance de DIABOLO entre l'évaluation sur V-COCO et celle sur H²O montre que la nouvelle taxonomie présente de nouveaux défis qui n'étaient pas proposés jusqu'alors dans les anciennes bases de données.

Ce travail de constitution de base de données présente de nombreuses contraintes et ne peut être envisagé pour tous les cas d'usage de l'analyse du comportement. Des alternatives doivent être proposées dans le futur pour l'alléger voire s'en affranchir.

Une première solution possible est l'utilisation de données synthétiques. Elles ont l'avantage de pouvoir être générées très facilement et en grande quantité, de plus leur annotation est entièrement automatique. L'équilibre des exemples par classe est également maîtrisé. Les techniques récentes de génération d'images donnent aujourd'hui des résultats impressionnants et très réalistes, notamment lorsqu'il s'agit de visages synthétiques. Cependant la génération de vidéo montre encore des défauts surtout sur les mouvements humains. En effet, la cinématique n'est pas bonne et un oeil humain se rend très vite compte que la vidéo n'est pas réaliste. Des modèles appris sur ces données seront difficilement transférables au monde réel.

Une deuxième solution possible est l'apprentissage auto-supervisé qui a l'avantage d'exploiter des données non annotées. J'ai eu l'occasion de travailler sur cette thématique de recherche par le co-encadrement de deux thèses, ces travaux sont évoqués chapitre 5. Les dernières techniques "classiques" d'apprentissage auto-supervisé devancent aujourd'hui les méthodes entièrement supervisées dans le cadre de la classification d'images. Ce n'est pas encore le cas quand elles sont déclinées sur d'autres tâches telles que la détection et d'autant plus sur la modalité vidéo. Cependant, très récemment, l'apparition des modèles de fondation est en train de révolutionner le monde de l'apprentissage non-supervisé. Ce sont des modèles de très grande taille, de l'ordre du milliard de paramètres, entraînés sur une quantité de données non annotées gigantesque et dont les performances de transfert sur différentes tâches cibles sont très bonnes. C'est notamment le cas de la méthode InternVideo [WLL+22] qui affiche des performances dépassant l'apprentissage supervisé sur plusieurs tâches d'analyse vidéo. Cependant l'en-

traînement de ces giga-modèles n'est accessible qu'à une minorité disposant d'énormes ressources de calcul. De nombreuses questions sont donc soulevées, notamment : Comment avoir accès à ces pré-apprentissages ? Comment être sûr des données utilisées pour les générer ? Sans parler du fait que ces nouveaux modèles ne suivent pas du tout la tendance actuelle en terme d'économie d'énergie. Je reste donc convaincue que les recherches sur des architectures classiques voire plus légères ont toutes leur intérêt.

6.2 . Perspectives de recherche sur l'analyse du comportement humain par vision

Dans cette section je présente des perspectives pour de futurs travaux de recherche. Ces perspectives font l'objet de deux sujets de thèse que je propose d'encadrer.

Fédération de données visuelles hétérogènes pour l'apprentissage d'un réseau multi-tâches

La communauté scientifique aborde aujourd'hui l'analyse du comportement humain par des tâches très distinctes les unes des autres. Les différentes tâches de vision pour l'analyse du comportement humain sont détaillées dans le chapitre 1, section 1.4. Ces différents problèmes sont actuellement traités avec des architectures différentes, sur des bases de données annotées différemment, avec des contextes et des ontologies très variés alors que la compréhension visuelle de l'image est la même. Ce qui fait que les données à disposition sont divisées selon la tâche cible alors que le problème général à résoudre est le même et qu'il demande pourtant une quantité gigantesque de données.

L'objectif ultime pour résoudre le problème d'analyse du comportement humain tout en ayant la possibilité de répondre aussi bien à une tâche de classification, de détection ou de description textuelle, serait d'avoir :

- Une architecture multi-modale qui puisse prendre en entrée aussi bien une image qu'un clip vidéo.
- Une architecture multi-tâches aux performances de l'état de l'art sur chaque tâche de l'analyse du comportement.

Ce nouveau modèle utiliserait l'ensemble des jeux de données hétérogènes à notre disposition lors d'un seul apprentissage, nous pouvons donc espérer que le temps d'entraînement soit plus rapide que si les sous-tâches sont apprises indépendemment les unes des autres. De plus le modèle devrait avoir des grandes capacités de transfert vers de nouveaux contextes et de nouvelles ontologies sans avoir besoin de beaucoup de données an-

notées. Cette idée de fédération de données visuelles hétérogènes pour l'apprentissage d'un réseau multi-tâches dirige mes recherches vers des développements de plus en plus frugaux à la fois en terme de données et d'énergie nécessaire à leur réalisation.

Je propose et porte ce sujet avec mon collègue, Bertrand Luvison.

Fusion de caractéristiques pour la ré-identification multi-vues

Ce projet de recherche fait écho aux travaux sur l'analyse du comportement humain dans le domaine sportif évoqués chapitre 5 où la ré-identification des joueurs est essentielle pour obtenir une analyse spécifique par joueur. Cependant la ré-identification des joueurs de sports collectifs fait face à deux principales limites : les joueurs portent les même maillots donc le modèle de ré-identification doit pouvoir s'attacher à des détails. Cependant la faible résolution sur les joueurs ne permet pas de capter facilement ces détails.

Les matches sont pour la plupart filmés selon différents points de vue avec des résolutions différentes. Certaines caméras filment le match avec un plan large permettant de voir une grande partie du terrain et donc de localiser les individus facilement. Cependant la résolution des joueurs dans ces vues ne permet pas de les reconnaître aisément. Au contraire, d'autres caméras suivent l'action de façon beaucoup plus rapprochée. Il est plus facile d'identifier les joueurs dans cette vue. Une mise en correspondance entre ces deux points de vue permettrait de fusionner les informations d'identité et de localisation des joueurs. Cette association est loin d'être triviale du fait que les caméras sont en mouvement et qu'elles filment l'action avec des points de vue différents.

L'objectif est donc d'apprendre des caractéristiques de ré-identification robustes à l'incidence et au niveau de zoom sur les individus, ce qui permettrait de reconnaître un même individu quel que soit le point de vue sur celui-ci et de mettre en correspondance différentes caméras qui observent les mêmes individus de deux points de vue différents. Remplir ces deux objectifs permettrait de rendre plus robuste le suivi multi-caméras des individus.

7 - Publications

Les **contributions présentées dans ce mémoire** ont fait l'objet de trois publications listées ci-dessous :

La présentation du jeu de données DAHLIA et ses résultats de référence de reconnaissance d'activité obtenus avec la méthode DOHT ont été publiés à la conférence internationale avec comité de lecture IEEE FG2017 :

- Geoffrey Vaquette, Astrid Orcesi, Laurent Lucat, and Catherine Achard. The daily home life activity dataset : a high semantic activity dataset for online recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 497–504. IEEE, 2017

La méthode CALIPSO a fait l'objet d'un article accepté à la conférence internationale avec comité de lecture IEEE WACV2020 :

- Sanaa Chafik, Astrid Orcesi, Romaric Audigier, and Bertrand Luvison. Classifying all interacting pairs in a single shot. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2892–2901, 2020

Les contributions sont égales entre les deux premiers auteurs.

H^2O et ses résultats de référence issus de la méthode DIABOLO ont été publiés à la conférence internationale avec comité de lecture IEEE FG2021 :

- Astrid Orcesi, Romaric Audigier, Fritz Poka Toukam, and Bertrand Luvison. Detecting human-to-human-or-object (h^2o) interactions with diabolo. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021

Les travaux sur l'**apprentissage auto-supervisé de représentations vidéos** ont fait l'objet de 3 publications listées ci-dessous :

Les résultats de la méthode CPC adaptée et appliquée à la vidéo ont été publiés à deux conférences internationales avec comité de lecture, WACV 2020 et Workshop ICML 2019 :

- Guillaume Lorre, Jaonary Rabarisoa, Astrid Orcesi, Samia Ainouz, and Stephane Canu. Temporal contrastive pretraining for video action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 662–670, 2020
- Guillaume Lorre, Jaonary Rabarisoa, Astrid Orcesi, Samia Ainouz, and Stéphane Canu. Contrastive predictive coding for video representation learning. In *ICML 2019 36th International Conference on Machine Learning-*

Workshop on Self-Supervised Learning, 2019

La méthode SCE proposant une nouvelle fonction de perte pour le pré-apprentissage contrastif de représentations a été publiée à la conférence internationale avec comité de lecture, WACV 2023 :

- Julien Denize, Jaonary Rabarisoa, Astrid Orcesi, Romain Hérault, and Stéphane Canu. Similarity contrastive estimation for self-supervised soft contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2706–2716, 2023

Nos travaux sur l'**analyse du comportement dans le domaine sportif** ont fait l'objet d'une publication et d'un premier prix au challenge Soccer-Net Tracking 2022 :

La méthode incrémentale de ré-identification de joueurs a été publiée à la conférence internationale avec comité de lecture, Workshop CVPR CVsports 2022 :

- Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Efficient tracking of team sport players with few game-specific annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3461–3471, 2022

La méthode de ré-identification intégrée à un algorithme de suivi des joueurs a permis d'obtenir la première place au challenge SoccerNet Tracking 2022 :

- Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, et al. Soccernet 2022 challenges results. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, pages 75–86, 2022

Bibliographie

[AAC+20] Hazem Abdelkawy, Naouel Ayari, Abdelghani Chibani, Yacine Amirat, and Ferhat Attal. Spatio-temporal convolutional networks and n-ary ontologies for human activity-aware robotic system. *IEEE Robotics and Automation Letters*, 6(2) :620–627, 2020.

[ACC18] Dawood Al Chanti and Alice Caplier. Deep learning for spatio-temporal modeling of dynamic spontaneous emotions. *IEEE Transactions on Affective Computing*, 12(2) :363–376, 2018.

[AEHKL+16] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m : A large-scale video classification benchmark. *arXiv preprint arXiv :1609.08675*, 2016.

[AME19] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw : Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189. IEEE, 2019.

[APGS14] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation : New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014.

[APNL13] S Mohsen Amiri, Mahsa T Pourazad, Panos Nasiopoulos, and Victor CM Leung. Non-intrusive human activity monitoring in a smart home environment. In *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, pages 606–610. IEEE, 2013.

[AQMM07] Catherine Achard, Xingtai Qu, Arash Mokhber, and Maurice Milgram. Action recognition with semi-global characteristics and hidden markov models. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 274–284. Springer, 2007.

[AW18] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv :1812.11941*, 2018.

[BCL+20] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Pandanet : Anchor-based single-shot multi-person 3d pose estimation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6856–6865, 2020.

[BFE+17] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv :1710.11252*, 2017.

[BNW+18] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. *Proceedings of the IEEE European Conference on Computer Vision*, 2018.

[BRSC20] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10460–10469, 2020.

[CG12] Chao-Yeh Chen and Kristen Grauman. Efficient activity detection with max-subgraph search. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1274–1281. IEEE, 2012.

[CHTAL13] Adrien Chan-Hon-Tong, Catherine Achard, and Laurent Lucat. Deeply optimized hough transform : Application to action segmentation. In *Image Analysis and Processing–ICIAP 2013*, pages 51–60. Springer, 2013.

[CHTAL14] Adrien Chan-Hon-Tong, Catherine Achard, and Laurent Lucat. Simultaneous segmentation and classification of human actions in video streams using deeply optimized hough transform. *Pattern Recognition*, 47(12) :3807–3818, 2014.

[CJK15] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad : A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 168–172. IEEE, 2015.

[CLGX12] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *Bmvc*, volume 1, page 3, 2012.

[CLL+18] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018.

[CLW+21] Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. St-hoi : A spatial-temporal baseline for human-object interaction detection in videos. In *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 9–17, 2021.

[COAL20] Sanaa Chafik, Astrid Orcesi, Romaric Audigier, and Bertrand Luvison. Classifying all interacting pairs in a single shot. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2892–2901, 2020.

[CQY+12] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. Human daily action analysis with multi-view and color-depth data. In *European Conference on Computer Vision*, pages 52–61. Springer, 2012.

[CS12] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, pages 215–230, 2012.

[CWH+15] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO : A benchmark for recognizing human-object interactions in images. In *ICCV*, pages 1017–1025, 2015.

[CZ17] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[CZT05] Robert Collins, Xuhui Zhou, and Seng Keat Teh. An open source tracking testbed and evaluation web site. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, volume 35, 2005.

[D+17] Emily L Denton et al. Unsupervised learning of disentangled representations from video. *Advances in neural information processing systems*, 30, 2017.

[DAHG+15] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[DAL+22] Fabian Dubourvieux, Romaric Audigier, Angélique Loesch, Samia Ainouz, and Stéphane Canu. A formal approach to good

practices in pseudo-labeling for unsupervised domain adaptive re-identification. *Computer Vision and Image Understanding*, 223 :103527, 2022.

[DDF+22] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision : Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130 :33–55, 2022.

[DDK+19] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome : Real-world activities of daily living. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 833–842, 2019.

[DDS+22] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed : Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[DMG+19] Rui Dai, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and François Bremond. Self-attention temporal convolutional network for long-term daily living activity detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7. IEEE, 2019.

[DRCB05] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance*, pages 65–72. IEEE, 2005.

[DRO+23] Julien Denize, Jaonary Rabarisoa, Astrid Orcesi, Romain Hérault, and Stéphane Canu. Similarity contrastive estimation for self-supervised soft contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2706–2716, 2023.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[DTS06] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer, 2006.

[DZL17] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 3298–3308. IEEE, 2017.

[EEVG+15] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge : A retrospective. *International journal of computer vision*, 111(1) :98–136, 2015.

[EMT+13] Chris Ellis, Syed Zain Masood, Marshall F Tappen, Joseph J Laviola Jr, and Rahul Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 101(3) :420–436, 2013.

[FBGF19] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 2019.

[FCS+13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise : A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26*, pages 2121–2129, 2013.

[Fei20] Christoph Feichtenhofer. X3d : Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[FFMH19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[FLP16] Hajer Fradi, Bertrand Luvison, and Quoc Cuong Pham. Crowd behavior analysis using local mid-level visual descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3) :589–602, 2016.

[FPZ16] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[FSSM07] Andrea Frome, Yoram Singer, Fei Sha, and Jitendra Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[FST98] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model : Analysis and applications. *Machine learning*, 32(1) :41–62, 1998.

[FXM+21] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[FXSL21] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. DIRV : Dense interaction region voting for end-to-end human-object interaction detection. In *The AAAI Conf. on Artificial Intelligence (AAAI)*, 2021.

[GADG18] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet : A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721, 2018.

[GBS+07] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12) :2247–2253, 2007.

[GCD+22] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, et al. Soccernet 2022 challenges results. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, pages 75–86, 2022.

[GDDM14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[GGDH18] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367. IEEE, 2018.

[GM15] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv :1505.04474*, 2015.

[GSK+11] Ross Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *2011 International Conference on Computer Vision*, pages 415–422. IEEE, 2011.

[GSR+18] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA : A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018.

[GZH18] Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN : Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference*, 2018.

[HDO+98] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4) :18–28, 1998.

[HFW+20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[HYWT14] Dong Huang, Shitong Yao, Yi Wang, and Fernando De La Torre. Sequential max-margin event detectors. In *European conference on computer vision*, pages 410–424. Springer, 2014.

[HZ14] Minh Hoai and Andrew Zisserman. Talking heads : Detecting humans and recognizing their interactions. In *CVPR*, pages 875–882, 2014.

[KAS14] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions : Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[KCKK20] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. UnionDet : Union-level detector towards real-time human-object interaction detection. In *ECCV*, pages 498–514, 2020.

[KCS+17] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv*, abs/1705.06950, 2017.

[KFF17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4) :664–676, April 2017.

[KGS+05] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, V Manohar, Mathew Boonstra, and Valentina Korzhova. Performance evaluation protocol for text, face, hands, person and vehicle detection & tracking in video analysis and content extraction (vace-ii). *Protocol Document*, 2005.

[KGS13] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8) :951–970, 2013.

[KHR+17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

[KJG+11] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb : a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[KLF18] Alexander Kolesnikov, Christoph H Lampert, and Vittorio Ferrari. Detecting visual relationships using box attention. *arXiv preprint arXiv :1807.02136*, 2018.

[KMS08] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference*, pages 275–1. British Machine Vision Association, 2008.

[KTS+14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[KVRM19] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019.

[KZG+17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome : Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1) :32–73, 2017.

[LAE+16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd : Single shot multibox detector. In *Computer Vision–ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[Lap05] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2) :107–123, 2005.

[LCL+11] Bertrand Luvison, Thierry Chateau, Jean-Thierry Lapreste, Patrick Sayd, and Quoc Cuong Pham. Automatic detection of unexpected events in dense areas for videosurveillance applications. In *Video Surveillance*. InTech, 2011.

[LDG+17] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 4, 2017.

[LGG+17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[LK+81] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*, volume 81. Vancouver, 1981.

[LKBFF16] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[LKL+20] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *European conference on computer vision*, pages 440–456. Springer, 2020.

[LLW+20] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. PPDM : Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, pages 482–490, 2020.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO : Common objects in context. In *ECCV*, pages 740–755, 2014.

[LMH+21] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv :2102.09480*, 2021.

[LN06] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *European conference on computer vision*, pages 359–372. Springer, 2006.

[Lor21] Guillaume Lorre. *Apprentissage non-supervisé de représentations pour l'analyse de séquences vidéos*. PhD thesis, Normandie, 2021.

[Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2) :91–110, 2004.

[LOW17] Yikang Li, Wanli Ouyang, and Xiaogang Wang. ViP-CNN : A visual phrase reasoning convolutional neural network for visual relationship detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[LRB+16] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[LRO+19] Guillaume Lorre, Jaonary Rabarisoa, Astrid Orcesi, Samia Ainouz, and Stéphane Canu. Contrastive predictive coding for video representation learning. In *ICML 2019 36th International Conference on Machine Learning-Workshop on Self-Supervised Learning*, 2019.

[LRO+20] Guillaume Lorre, Jaonary Rabarisoa, Astrid Orcesi, Samia Ainouz, and Stephane Canu. Temporal contrastive pretraining for video action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 662–670, 2020.

[LYC20] Ye Liu, Junsong Yuan, and Chang Wen Chen. ConsNet : Learning consistency graph for zero-shot human-object interaction detection. In *Int. Conf. on Multimedia*, pages 4235–4243, 2020.

[LZH+19] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[MCL15] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv :1511.05440*, 2015.

[MHT16] Moustafa Meshry, Mohamed E Hussein, and Marwan Torki. Linear-time online action detection from 3d skeletal data using bags of gesturelets. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[MLTR+16] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16 : A benchmark for multi-object tracking. *arXiv preprint arXiv :1603.00831*, 2016.

[MOP22] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Efficient tracking of team sport players with few game-specific annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3461–3471, 2022.

[MS20] Mahshid Majd and Reza Safabakhsh. Correlational convolutional lstm for human action recognition. *Neurocomputing*, 396 :224–229, 2020.

[Mun57] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1) :32–38, 1957.

[MZH16] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Unsupervised learning using sequential verification for action recognition. corr abs/1603.08561 (2016). *arXiv preprint arXiv :1603.08561*, 2016.

[NB19] Farhood Negin and François Brémond. An unsupervised framework for online spatiotemporal detection of activities of daily living by hierarchical activity models. *Sensors*, 19(19) :4237, 2019.

[ND17] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*, pages 2171–2180, 2017.

[NGA+18] Farhood Negin, Abhishek Goel, Abdelrahman G Abubakr, Francois Bremond, and Gianpiero Francesca. Online detection of long-term daily living activities by weakly supervised recognition of sub-activities. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[NHD17] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding : End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems 30*, pages 2277–2287, 2017.

[NPMY13] Bingbing Ni, Yong Pei, Pierre Moulin, and Shuicheng Yan. Multilevel depth and image fusion for human activity detection. *IEEE transactions on cybernetics*, 43(5) :1383–1394, 2013.

[NS12] Sebastian Nowozin and Jamie Shotton. Action points : A representation for low-latency online human action recognition. *Microsoft Research Cambridge, Tech. Rep. MSR-TR-2012-68*, 2012.

[NWM11] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact : A color-depth video database for human daily activity recognition. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 1147–1153. IEEE, 2011.

[OATL21] Astrid Orcesi, Romaric Audigier, Fritz Poka Toukam, and Bertrand Luvison. Detecting human-to-human-or-object (h^2o) interactions with diabolo. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[OCK+13] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley mhad : A comprehensive multimodal human action database. In *2013 IEEE workshop on applications of computer vision (WACV)*, pages 53–60. IEEE, 2013.

[PJDC21] Junjie Peng, Elizabeth C Jury, Pierre Dönnes, and Coziana Ciurtin. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases : applications and challenges. *Frontiers in pharmacology*, 12 :720694, 2021.

[PLSS18] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting rare visual relations using analogies. *arXiv preprint arXiv :1812.05736*, 2018.

[PR21] AJ Piergiovanni and Michael S Ryoo. Recognizing actions in videos from unseen viewpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4124–4132, 2021.

[QWJ+18] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision*, pages 407–423. Springer, 2018.

[QYM17] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[RAAS12] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1194–1201. IEEE, 2012.

[RAD17] MS Ryoo, JK Aggarwal, and UT-Interaction Dataset. Icpr contest on semantic description of human activities (sdha), 2010, 2017.

[RBK21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN : Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[RKH11] Michalis Raptis, Darko Kirovski, and Hugues Hoppe. Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 147–156. ACM, 2011.

[RLH+20] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation : Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[RLH+22] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation : Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3) :1623–1637, 2022.

[RMDHM14] Hossein Rahmani, Arif Mahmood, Q Du Huynh, and Ajmal Mian. Hopc : Histogram of oriented principal components of 3d pointclouds for action recognition. In *European conference on computer vision*, pages 742–757. Springer, 2014.

[RMHM16] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. 2016.

[RRA+12] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *European conference on computer vision*, pages 144–157. Springer, 2012.

[SCD+17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam : Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[SEL18] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. LSTA : Long short-term attention for egocentric action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[SFC+11] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011.

[SHRW20] Xu Sun, Xinwen Hu, Tongwei Ren, and Gangshan Wu. Human object interaction detection via multi-level conditioned network. In *Int. Conf. on Multimedia Retrieval*, pages 26–34, 2020.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet : A unified embedding for face recognition and clustering. 2015.

[SLC04] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions : a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[SLNW16] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d : A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[SPSS11] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from rgbd images. In *Workshops at the twenty-fifth AAAI conference on artificial intelligence*, 2011.

[SRV+20] Kuldeep Singh, Shantanu Rajora, Dinesh Kumar Vishwakarma, Gaurav Tripathi, Sandeep Kumar, and Gurjit Singh Walia. Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. *Neurocomputing*, 371 :188–198, 2020.

[STLY14] Yan Song, Jinhui Tang, Fan Liu, and Shuicheng Yan. Body surface context : A new robust feature for action recognition from depth videos. *IEEE transactions on circuits and systems for video technology*, 24(6) :952–964, 2014.

[SVW+16] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes : Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016.

[SY18] Fadime Sener and Angela Yao. Unsupervised learning and segmentation of complex activities from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8368–8376, 2018.

[SZS12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101 : A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv :1212.0402*, 2012.

[TBB09] Moritz Tenorth, Jan Bandouch, and Michael Beetz. The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1089–1096. IEEE, 2009.

[TBF+15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[TLYK18] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan : Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[TPL20] Mingxing Tan, Ruoming Pang, and Quoc V Le. EfficientDet : Scalable and efficient object detection. In *CVPR*, pages 10781–10790, 2020.

[TWT+18] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[TWTF19] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

[UIM20] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet : Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, pages 13617–13626, 2020.

[VAL16] Geoffrey Vaquette, Catherine Achard, and Laurent Lucat. Information fusion for action recognition with deeply optimised hough transform paradigm. In *11th International Conference on Computer Vision and Applications (VISAPP)*, 2016.

[VAL19] Geoffrey Vaquette, Catherine Achard, and Laurent Lucat. Robust information fusion in the doht paradigm for real-time action detection. *Journal of Real-Time Image Processing*, 16(5) :1511–1524, 2019.

[VOLA17] Geoffrey Vaquette, Astrid Orcesi, Laurent Lucat, and Catherine Achard. The daily home life activity dataset : a high semantic activity dataset for online recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 497–504. IEEE, 2017.

[VRCHTA21] Guillaume Vaudaux-Ruth, Adrien Chan-Hon-Tong, and Catherine Achard. Actionspotter : Deep reinforcement learning framework for temporal action spotting in videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 631–638. IEEE, 2021.

[WBL22] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7 : Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv :2207.02696*, 2022.

[WDC+22] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage : Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv :2211.05778*, 2022.

[WGGH18] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[WKSL11] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[WLL+22] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo : General video foundation models via generative and discriminative learning. *arXiv preprint arXiv :2212.03191*, 2022.

[WLM+14] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127 :14–30, 2014.

[WLT06] Jamie A Ward, Paul Lukowicz, and Gerhard Tröster. Evaluating performance in continuous context recognition using event-driven error characterisation. In *International Symposium on Location-and Context-Awareness*, pages 239–255. Springer, 2006.

[WLWY12] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In

2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 1290–1297. IEEE, 2012.

[WLZF18] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[WNX+14] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2649–2656. IEEE, 2014.

[WS13] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[WTVG08] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008 : 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*, pages 650–663. Springer, 2008.

[WYD+20] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, pages 4116–4125, 2020.

[WZL+19] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, pages 9469–9478, 2019.

[WZSS15] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch : Unsupervised understanding of actions and relations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4362–4370, 2015.

[WZT+22] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4754–4763, June 2022.

[WZZZ13] Ping Wei, Nanning Zheng, Yibiao Zhao, and Song-Chun Zhu. Concurrent action detection with structural prediction. In *Pro-*

ceedings of the IEEE International Conference on Computer Vision, pages 3136–3143, 2013.

[XA13] Lu Xia and JK Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2834–2841, 2013.

[XCA12] Lu Xia, Chia-Chih Chen, and JK Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27. IEEE, 2012.

[XGD+17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[XWL+19] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[XWWL22] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. Mdan : Multi-level dependent attention network for visual emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9479–9488, June 2022.

[YF05] David P Young and James M Ferryman. Pets metrics : On-line performance evaluation service. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 317–324, 2005.

[YLMD17] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. *Proceedings of the IEEE international conference on computer vision*, 2017.

[YLW09] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative subvolume search for efficient action detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2442–2449. IEEE, 2009.

[YLY14] Gang Yu, Zicheng Liu, and Junsong Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*, pages 50–65. Springer, 2014.

[YZT12] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060, 2012.

[ZG10] Jianguo Zhang and Shaogang Gong. Action categorization with modified hidden conditional random field. *Pattern Recognition*, 43(1) :197–203, 2010.

[ZKCC17] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, volume 1, page 5, 2017.

[ZLO+16] Jing Zhang, Wanqing Li, Philip O Ogunbona, Pichao Wang, and Chang Tang. Rgb-d-based action recognition datasets : A survey. *Pattern Recognition*, 60 :86–105, 2016.

[ZLS13] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose : An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2752–2759, 2013.

[ZLW+22] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022.

[ZSE+19] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[ZWQ+20] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, pages 4263–4272, 2020.

[ZWS+17] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Care about you : towards large-scale human-centric visual relationship detection. *arXiv preprint arXiv :1705.09892*, 2017.

[ZWS+18] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. Hcvrd : a benchmark for large-scale human-centered visual relationship detection. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[ZXC18] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[ZY19] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *arXiv preprint arXiv :1912.11164*, 2019.

[ZYTC18] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs : Scene graph parsing with global context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[ZYW+21] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl : Relational self-supervised learning with weak augmentation. In *Advances in Neural Information Processing Systems 34 : Annual Conference on Neural Information Processing Systems*, pages 2543–2555, 2021.